Machine Learning in Forensic Evidence Examination

A New Era

Edited by
Niha Ansari

CRC Press
Taylor & Francis Group

Machine Learning in Forensic Evidence Examination

The availability of machine-learning algorithms, and the immense computational power required to develop robust models with high accuracy, has driven researchers to conduct extensive studies in forensic science, particularly in the identification and examination of evidence found at crime scenes. *Machine Learning in Forensic Evidence Examination* discusses methodologies for the application of machine learning to the field of forensic science.

Evidence analysis is the cornerstone of forensic investigations, examined for either classification or individualization based on distinct characteristics. Artificial intelligence offers a powerful advantage by efficiently processing large datasets with multiple features, enhancing accuracy and speed in forensic analysis to potentially mitigate human errors. Algorithms have the potential to identify patterns and features in evidence such as firearms, explosives, trace evidence, narcotics, body fluids, etc. and catalogue them in various databases. Additionally, they can be useful in the reconstruction and detection of complex events, such as accidents and crimes, both during and after the event. This book provides readers with consolidated research data on the potential applications and use of machine learning for analyzing various types of evidence. Chapters focus on different methodologies of machine learning applied in different domains of forensic sciences such as biology, serology, physical sciences, fingerprints, trace evidence, ballistics, anthropology, odontology, digital forensics, chemistry and toxicology, as well as the potential use of big data analytics in forensics. Exploring recent advancements in machine learning, coverage also addresses the challenges faced by experts during routine examinations and how machine learning can help overcome these challenges.

Machine Learning in Forensic Evidence Examination is a valuable resource for academics, forensic scientists, legal professionals and those working on investigations and analysis within law enforcement agencies.

Machine Learning in Forensic Evidence Examination

A New Era

Edited by
Niha Ansari

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Designed cover image: Getty Images

First edition published 2026

by CRC Press
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN
CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 978-1-032-58236-8 (hbk)
ISBN: 978-1-032-58233-7 (pbk)
ISBN: 978-1-003-44916-4 (ebk)

DOI: 10.4324/9781003449164

Typeset in MinionPro
by Deanta Global Publishing Services, Chennai, India

This book is dedicated to my parents

and

To all those who encouraged me to fly during my hard times

Contents

List of Figures

Acknowledgements

The editor would like to thank all the contributors of the book. The editor also like to thank Mark Listewnik and all the members of the Taylor & Francis Group for their support through the journey of book publication. Further, the editor would like to thank the National Forensic Sciences University for its support.

Editor

Niha Ansari is an Assistant Professor at the National Forensic Science University in Gandhinagar, India. She earned her PhD in Forensic Science from Gujarat University, Ahmedabad, India, where she conducted the pioneering research 'Study on Changes in Vitreous Humours Concerning Time Since Death', utilizing nano sensor smartphone applications and microfluidic devices. Dr Ansari has also held positions at Jain University in Bangalore, India. She has published a number of chapters in edited volumes and 17 articles in peer-reviewed international journal publications. Her research interests encompass forensic nanotechnology, microfluidics and smartphone-based sensors. She has participated in numerous conferences, workshops and training sessions, imparting knowledge and skills to professionals and students alike. Among her accolades, Dr Ansari has been awarded the Maulana Azad National Fellowship by the University Grant Commission and the Best PhD Thesis Award by CHARUSAT. She is a part of the Editorial Board of The Science publishing group.

Contributors

Sachin Sharma
Department of Computer Science
Gran Sasso Science Institute
L'Aquila, Italy

Sumit Kumar Choudhary
School of Forensics, Risk Management and National Security
Rashtriya Raksha University
Gujarat, India

Dharmesh Shah
Department of Computer Science and Engineering
Indrashil University
Mehsana, India

Parvesh Sharma
Department of Forensic Science
Narayan Shastri Institute of Technology
Ahmedabad, India

Sidheswar Routray
Department of Computer Science and Engineering
Pandit Deendayal Energy University
Gandhinagar, India

Yasasvikumar Vala
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Madhavi Dave
DRDO Industry Academia Sardar Vallabhbhai Patel Center of Excellence
Gujarat University
Ahmedabad, India

Jennifer Johnson
School of Forensic Science
National Forensic Sciences University
Gandhinagar, India

Digvijaysinh Rathod
School of Cyber Security and Digital Forensics
National Forensic Sciences University
Gandhinagar, India

Satish Kumar
School of Forensic Science
National Forensic Sciences University
Gandhinagar, India

Vaishali
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Akanksha Singh Kachhawaha
School of Medico-Legal Studies
National Forensic Sciences University
Gandhinagar, India

Niha Ansari
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Vijeta Khare
School of Cyber Security and Digital Forensics
National Forensic Sciences University
Gandhinagar, India

Jeet Dasgupta
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Ahlad Kumar
School of Cyber Security and Digital Forensics
National Forensic Sciences University
Gandhinagar, India

Athulya Rajan
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Santhiya Raghavan
School of Medico-Legal Studies
National Forensic Sciences University
Gandhinagar, India

Muskan Singal
Department of Forensic Science
Government Institute of Forensic Science
Nagpur, India

Arunima Dutta
Department of Forensic Science
SGT University
Gurgaon, India

Neeti Kapoor
Department of Forensic Science
Government Institute of Forensic Science
Nagpur, India.

Vineeta Saini
Department of Forensic Science
SGT University
Gurgaon, India

Ashish Badiye
Department of Forensic Science
Government Institute of Forensic Science
Nagpur, India.

Kanica Chugh
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Ujaala Jain
College of Traffic Management
Institute of Road Traffic Education
Faridabad, India

Pooja Ahuja
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Tanurup Das
School of Forensic Sciences
The West Bengal National University of Juridical Sciences
Kolkata, India

Anubhav Singh
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Ankit Srivastav
School of Forensic Sciences
The West Bengal National University of Juridical Sciences
Kolkata, India

Dhwani Patel
School of Medico-Legal Studies
National Forensic Sciences University
Gandhinagar, India

Surbhi Mathur
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Kritika Sood
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Vinay Aseri
School of Forensics, Risk Management and National Security
Rashtriya Raksha University
Lavad, India

Divyant Jain
School of Forensic Sciences
National Forensic Sciences University
Gandhinagar, India

Introduction

Over the years, scientific and legal scholars have consistently challenged the forensic science community by questioning the legitimacy and trustworthiness of many forensic examination methods, which often rely on subjective interpretations by forensic experts. The President's Council of Advisors on Science and Technology (PCAST) has recommended that forensic analysis should be as objective as possible so that it can be performed either by automated systems or human experts with minimal judgement. PCAST also emphasized that the use of algorithms and automated systems can help overcome the limitations of human judgement.

The availability of machine-learning algorithms and the immense computational power required to develop robust models with high accuracy has driven researchers to conduct extensive studies in forensic science, particularly in the identification and examination of evidence found at crime scenes. Artificial intelligence technologies offer the ability to mitigate human errors and act as expert systems. These algorithms have the potential to identify objects and weapons, match faces and analyze structured materials. Additionally, they can reconstruct and detect complex events, such as accidents and crimes, both during and after the event.

The vision for the proposed book is to provide readers with consolidated research data on the use of machine learning for analyzing evidence found at various crime scenes. The book discusses the applicability and various approaches of machine learning in relation to the identification and examination of evidence typically encountered in different types of criminal activities. It focuses on the methodologies of machine learning as applied to forensic biology, serology, physics, ballistics, anthropology, odontology, digital forensics, chemistry, toxicology and, importantly, big data analytics in the forensic field. Each chapter addresses the problems faced by experts during routine examinations and how machine learning can help overcome these challenges. Further chapters will explore recent advancements in machine learning as they pertain to the specific types of evidence under discussion.

DOI: 10.4324/9781003449164-1

1 Understanding the Fundamentals of Machine Learning and its Applications in Forensic Evidence Examination

SACHIN SHARMA, DHARMESH SHAH, SIDHESWAR ROUTRAY, MADHAVI DAVE, AND DIGVIJAYSINH RATHOD

Introduction to Machine Learning

Definition and Importance of Machine Learning in Forensic Science

Machine learning (ML) is a brand-new system which falls under artificial intelligence; it has been designed to enable systems to understand information through data without the need for human intervention or programming [1]. The need for machine learning in forensic science is greater than in any other scientific discipline. Machine learning adds a new dimension to investigative processes by changing the way that we analyze and interpret evidence given by forensic science experts.

In essence, machine learning equips forensic scientists with the capacity to utilize very large datasets, which helps the system detect patterns, trends and irregularities that may bypass traditional approaches. The extraction of important insights from various data sources through automation is one way that machine learning improves the efficiency and accuracy of forensic investigations.

The significance of machine learning in forensic investigation is found in its ability to address intricate and extensive datasets associated with investigations. Machine-learning algorithms excel at detecting subtle patterns whether it is fingerprints, DNA profiles, document analysis or cybercrimes, thus helping investigators make better choices.

DOI: 10.4324/9781003449164-2

Also, a dynamic approach towards examining evidence is fostered by machine learning. It adjusts and changes as new data are introduced so that investigative techniques are still flexible enough to cope with emerging problems. This adaptability becomes even more important in this age where forensic science must deal with an expanding range of data types and sources.

Machine-Learning Workflow: Data Collection, Preprocessing, Model Training and Evaluation

To carefully employ machine learning for forensic purposes requires implementing a systematic workflow involving data handling and preprocessing as well as model development and rigorous evaluation [2]. Here we describe the basic stages involved in the workflow for machine learning, delineating in what way every phase contributes to the success of forensic investigations.

- **Data Collection:** The basis of any machine-learning project is the quality and relevance of its training data. Forensic science can have various datasets which include fingerprints and genetic profiles among others. This entails careful collection from diverse sources including proper representation and inclusiveness. In order to ensure comprehensiveness and representativeness of collected data, subsequent analyses' integrity depends on it.
- **Preprocessing:** Raw forensic data often comes with noise, outliers or incomplete entries. Preprocessing involves cleaning up and transforming raw forensic data such that it is suitable for feeding into machine-learning models. Techniques used comprise normalization, missing values handling and outlier detection aimed at placing data in a state that will facilitate effective model training. Properly done, preprocessing can greatly increase the accuracy and robustness of the ML model.
- **Model Training:** With preprocessed data in hand, the next phase involves selecting an appropriate machine-learning algorithm and training the model. The algorithm learns patterns and relationships within the data during this training process. In forensic science, depending on the algorithm chosen or the nature of the task—be it classification, regression or clustering. The model is fine-tuned iteratively to optimize its performance, striking a balance between complexity and generalization.

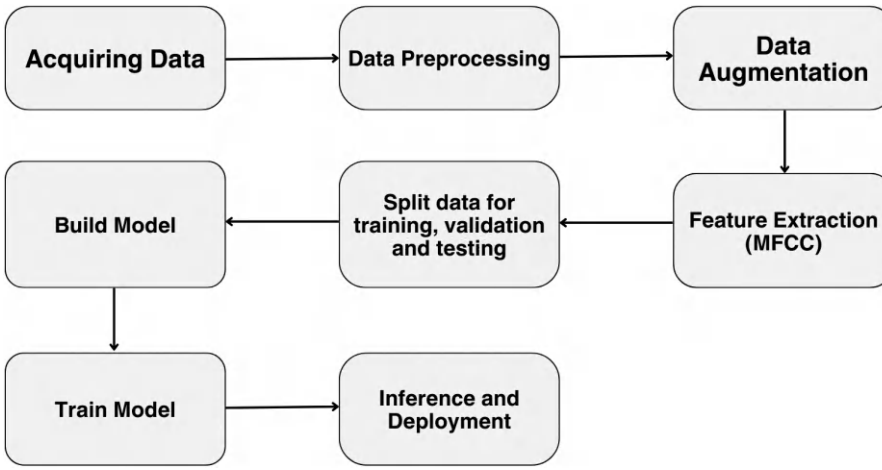

Figure 1.1 Block diagram for the workflow of a typical machine-learning process [9].

- **Evaluation:** The effectiveness of a machine-learning model is gauged through rigorous evaluation against independent datasets. Evaluation metrics such as accuracy, precision, recall and F1-score provide insights into the model's performance [3]. Cross-validation techniques are often employed to ensure the model's generalizability to new, unseen data. In forensic investigations, the reliability and accuracy of predictions directly influence the outcome of analyses. Therefore, robust evaluation methodologies are crucial to validate the model's efficacy in real-world forensic scenarios. Figure 1.1 shows a block diagram for the workflow of a typical machine-learning process.

Supervised Learning Algorithms

Decision Trees and Random Forests

- **Decision Trees:** Decision trees are a type of versatile, easy-to-understand machine-learning algorithm that can be used for both classification and regression tasks [4]. At its heart, the decision tree data is broken into subsets according to features resulting in a tree-like structure with every internal node representing a feature on which decisions are made while each leaf node symbolizes the outcome. In cases such as fingerprinting or DNA profiling, where decision boundaries are highly complex, the interpretability of a decision tree

makes it possible to peer into the model's decision-making process, which would otherwise remain unseen. Figure 1.2 represents a common decision tree algorithm.

- **Random Forests:** To avoid the overfitting problem, random forests build many decision trees and merge them together to obtain a more stable prediction. Each tree is trained on a different random subset of the data. Democratic voting across individual trees collectively produces a final prediction [5]. Such an ensemble technique improves the strength and generalization of these models; thus, random forests are a preferred option for more complex forensic and dementia prediction tasks, as they are able to predict outcomes with a high level of success.

Random forests are used in forensic evidence examination to enhance accuracy and reliability across several domains. One example is the use of random forests in fingerprint analysis because they can generalize much better than neural networks in arcane scenarios involving classifying and matching prints and as a result output more detailed predictions with higher precision.

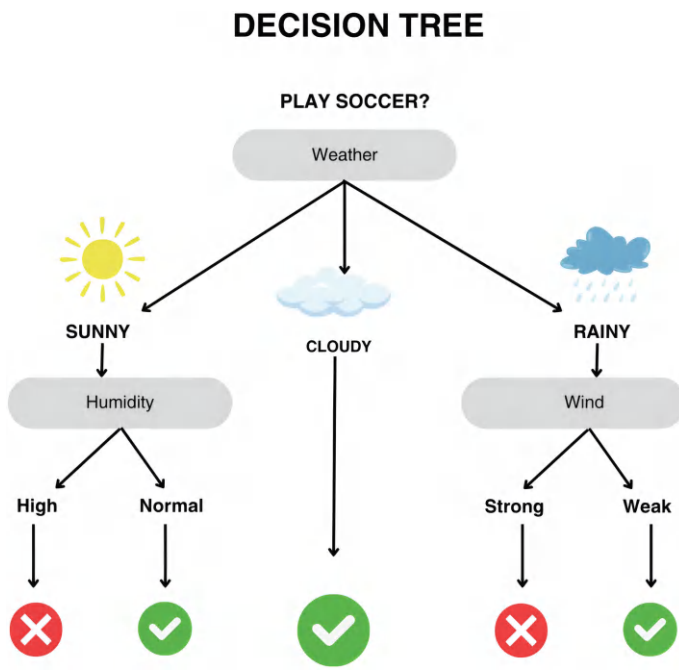

Figure 1.2 An example of a decision tree algorithm [4].

Support Vector Machines (SVMs)

Overview: Support vector machines (SVMs) represent a high-performing supervised learning algorithm used in many applications, including forensic science. SVMs are used in classification and regression tasks by building optimal decision boundaries, called hyperplanes, within a space of high dimensionality [6]. At the highest level, SVMs are ideal for forensic evidence analysis because this is a very clear classification problem: finding patterns in fingerprints or categorizing genetic profiles.

Operating Principle: Support vector machines (SVMs) seek to identify hyperplanes which segregate data points of different classes, while at the same time trying to maintain a margin between them. The efficacy of SVMs depends on the data points—known as support vectors—that are useful in determining an optimal hyperplane. SVMs are great in dealing with highly non-linear relationships as they use kernel functions to generate linear boundaries for different classes where a simple line cannot separate one class from another. Figure 1.3 shows a diagram for SVMs.

Applications in Forensic Science: For example, in fingerprint analysis, SVMs are used to classify and match prints based on dissimilarities, with the features collectively responsible for correct identifications. SVMs can perform genetic profiling analysis, where in a linear classifying manner the SVM is suited to pick out patterns within DNA data which would then provide vital information for identifying individuals.

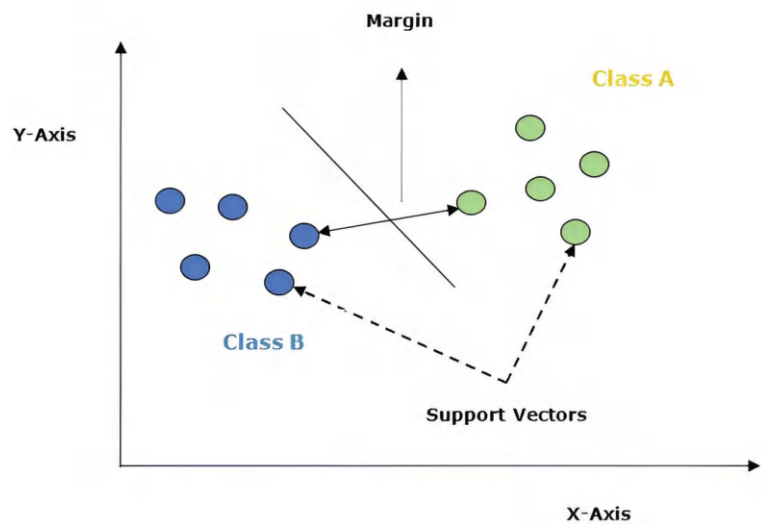

Figure 1.3 Support vector nachine diagram [21].

Naive Bayes Classifiers

Introduction: Naive Bayes classifiers are probabilistic models based on the principle of Bayes' theorem and work well for classification tasks, especially in forensic sciences Committed to their underlying simple assumptions, the naive Bayes classifiers are efficient and effective tools in some scenarios such as when computational resources or training examples is limited [7].

Operating Principle: The basic premise behind naive Bayes classifiers is their assumption of feature independence with respect to the class label—hence, 'naive'. Nevertheless, naive Bayes classifiers are likely to perform very well in practice. What these do is simply crunch numbers, such as calculating the likelihood (*a priori*) of obtaining a specific class given input features, and then selecting the class with the highest probability. In forensic evidence examination, naive Bayes classifiers find application in tasks such as document analysis or handwriting recognition.

Applications in Forensic Science: For example, in document analysis, they can be utilized to categorize and authenticate documents based on textual features.

Neural Networks and Deep Learning Models

Introduction: Neural networks and deep learning represent an exciting class of machine learning that works by modelling the way the human brain processes complex information. In the field of forensic science, these models have shown state-of-the-art performance at dealing with intricate patterns and relationships among a broad array of data which boosts accuracy levels as well as automation.

- **Neural Networks:** In neural networks, each of the networks is built to mimic the human brain and contain a bunch of nodes configured in layers—input, hidden and output. Each connection has an associated weight with it, and during training the network learns patterns from data by adjusting these weights [8]. Indeed, neural networks are powerful in capturing non-linear relationships and have shown promising results in a wide range of forensic fields including fingerprint analysis, DNA profiling and image forensics.
- **Deep Learning Models:** Deep learning expands the idea of neural networks to deeper architectures in terms of many hidden layers and gives rise to the name "deep" learning [9]. Convolutional neural networks (CNNs) perform well for tasks involving images and

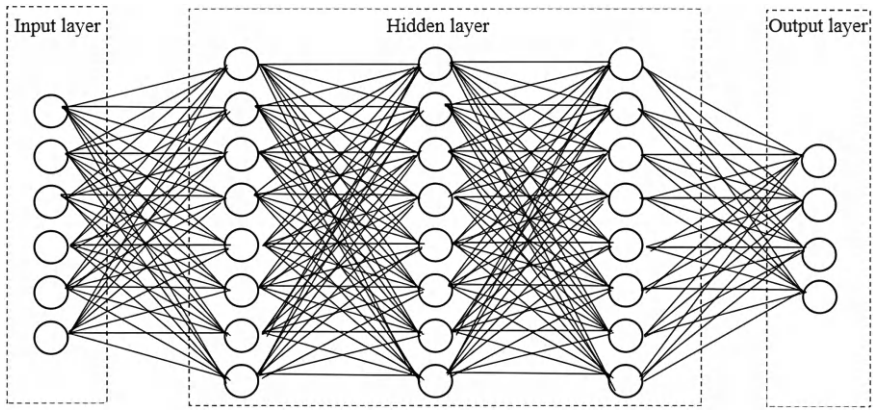

Figure 1.4 Typical deep learning neural network architecture. Source: https://insidebigdata.com/2020/10/16/whats-under-the-hood-of-neural-networks/

so are extensively used in forensic applications such as face recognition and image forensics. CNNs are also used for voice and speaker recognition which are sequential information so CNN can manage sentences in those examples. Therefore, a deep learning model can automatically learn the hierarchical features of objects with great depth and complexity over time, which provides vitality for its application in forensic investigation. Figure 1.4 outlines a deep learning neural network architecture.

Applications in Forensic Science: They have significantly driven advancements in the field of forensic science by automating laborious processes. Neural networks are useful in fingerprint analysis where the intricate patterns would help to improve identification accuracy. By using deep learning models, DNA profiling can be improved and it enables more detailed genetic information for accurate individual identification. The hierarchical feature extraction capabilities of deep learning are also ideal for image and video forensics.

Unsupervised Learning Algorithms

Clustering Algorithms: K-Means, Hierarchical Clustering

- **Clustering Algorithms:** These are a core part of unsupervised learning, which allows the discovery of patterns in datasets. Clustering algorithms, like k-means and hierarchical clustering, prove to be

useful in forensic science, in which patterns might not only exist but are also unknown; this method is able to provide an intuitive structure of relationships among data points generated by forensics and bring about noteworthy insight.

- **K-Means Clustering:** K-means structures data into k-clusters while also being partitional. It is designed to assign data points into clusters of similarity, iteratively reassigning data points until every single point has been appropriately grouped. K-means are both efficient and widely used in forensic scenarios where distinctive boundaries exist between classes [10]. For example, in a cybersecurity investigation, k-means can identify unique behaviours or patterns within network data, helping law enforcement quantify proper thresholds for specific tactics, techniques and procedures (TTPs).
- **Hierarchical Clustering:** Hierarchical clustering makes a tree-like ladder of clusters, known as a dendrogram, by iterative integration or splitting clusters based on their similarity [11]. This is especially useful not only for looking at how many clusters we need to make but also for identifying patterns of hierarchy within forensic data. Hierarchical clustering can be used in voice and speaker recognition to group similar audio patterns thereby helping identify individuals by their voices.

Applications in Forensic Science: In fingerprint analysis, k-means clustering can help to determine unique features within a set of prints so that these patterns appear different from each other. Applications of hierarchical clustering to DNA profiling can reveal hierarchical relationships between genetic markers, and detail familial information.

Dimensionality Reduction Techniques: Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE)

Introduction: Dimensionality reduction techniques are a necessity when working with high-dimensional data, which is typical in forensic science. Like principal component analysis (PCA), these techniques try and reduce high-dimensional data into a lower dimension while keeping important information. PCA and t-distributed stochastic neighbour embedding (t-SNE) are two powerful methods that have applications in forensic evidence examination by simplifying data for more effective analysis.

- **Principal Component Analysis (PCA):** PCA is a process that aims to reduce dimensions in some way by detecting components which are axes of dimension where data have the highest variances [12]. PCA preserves the information while reducing the dimensions in such a way as to project onto these numbers. In forensic science and other fields, PCA is useful to analyze markers in DNA profiles (genetics) coordinates, the most important correlating element of each fingerprint index/features dataset.
- **t-Distributed Stochastic Neighbour Embedding (t-SNE):** t-SNE is a nonlinear method for dimensionality reduction that aims to globally structure data by performing local similarities present in high-dimensional space [13]. It is excellent for discovering patterns (structures, clusters) in the data. A key finding about t-SNE is that when results are used wisely they can help us to understand relationships within data, which is critical during image and video forensics scenarios or for spotting cybercrime-related data.

Applications in Forensic Science: Both PCA and t-SNE are tools used frequently in science among datasets that are high dimensional (complex). These approaches go a long way towards speeding up the process of feature detection during fingerprint analysis. As a case in point, t-SNE is used to assist in the investigation of cybercrime as it unveils regional structures that aid interpretation around nuanced causation between any two data points.

Feature Selection and Extraction

Feature Selection Methods: Filter, Wrapper and Embedded Approaches

Introduction: Feature selection is an essential part of data preparation for a machine-learning model, especially in forensic science, where the datasets can be complicated and large. This is precisely the process by which different important features that enhance the predictive power of a model are selected, while irrelevant or redundant ones are removed. There are three main ways to select features: filter, wrapper and embedded—the appropriate way depends on the approach.

- **Filter Methods:** Filter methods train for multiple different machine-learning algorithms over the same dataset by evaluating features independently There are two general methods to rank the individual

significance of features, either by statistical tests or correlation analyses. Finally, features are chosen or ruled out based on actions [14]. Filter methods are computationally efficient and work well with high-dimensional datasets. Filter methods are useful in forensic applications because they can improve the performance of tasks such as DNA profiling by selecting the most discriminative genetic markers.

- **Wrapper Methods:** Wrapper methods work by putting the machine-learning algorithm to work against each subset. It examines various feature subsets and perceives the capability of functioning or assessing this model to detect its suitability function in each class [15]. This iterative process is more computationally expensive than using filter methods, but typically provides a finer-tuned feature set for specific models. Wrapper methods are important when configuring features for voice or speaker recognition tasks, that is, in forensic sciences.
- **Embedded Approaches:** Embedded approaches integrate feature selection into the model-training process. During the training, machine-learning algorithms select features and assign a weight or coefficient to them based on their importance. Regularization techniques, such as L1 regularization, can enhance the embedded feature selection process by promoting sparsity in the model [16]. On the other hand, embedded approaches are specially designed for a condition where feature selection is part of learning—one such being in neural networks because applied to image forensics.

Applications in Forensic Science: Feature selection methods help forensic models to develop better efficiency and accuracy. In matching fingerprint analysis, filter methods can help identify the best discriminative features for identifying fingerprints. The DNA profile features can be adjusted for specific identity verification tasks using the wrapper methods. By jointly targeting the most discriminative image features, embedded approaches can improve neural networks' performance in image forensics.

Feature Extraction Techniques: Autoencoders, Principal Component Analysis (PCA)

Introduction: Feature extraction techniques try to convert the original raw data into a smaller, and more useful format, facilitating efficient model training and enhancing interpretability. Two prominent techniques—autoencoders and principal component analysis (PCA)—offer distinct approaches to feature extraction in forensic applications.

- **Autoencoders**: Autoencoders are a type of neural network which is very useful for unsupervised learning. An autoencoder, which consists of an encoder and a decoder learns to compress the input data into some lower-dimensional representation (encoding), while also learning how to reconstruct the original data through this encoding [17]. Autoencoders can be quite useful in forensic science, especially when it comes to image forensics, as they can capture detailed patterns and even anomalies within visual data. These encodings, being learned, often unveil important structures involving forensic evidence.
- **Principal Component Analysis (PCA):** PCA is a classical technique in the linear realm that projects high-dimensional data to lower dimensions after identifying the principal components which capture maximum variance [12]. These are orthogonal axes which represent the biggest vectors of variability in your data. PCA is frequently used as a basic technique in many forensic applications, such as DNA profiling, where it enables the reduction of the dimensionality of genetic data while maintaining information regarding its underlying structure.

Applications in Forensic Science: Feature extraction techniques play a critical role in identifying informative patterns from forensic data. The ability of the autoencoder to work with voice data makes it possible for use in speech and speaker recognition. PCA is a linear dimensionality reduction method that helps improve the efficiency of fingerprint analysis by decomposing a large set of features into key components.

Model Evaluation and Performance Metrics

Evaluation Metrics: Accuracy, Precision, Recall and F1-Score

Introduction: Evaluation metrics are very important to determine how machine-learning models perform in our forensic science work. They supply numerical metrics designed to indicate how well a model generalizes with unseen data and help forensic practitioners make informed decisions. A model can be evaluated on four basic metrics—exactness/accuracy, preciseness/precision, recall and F1-score. These four metrics give a good overall indication of how a model is working in a forensic task [3].

- **Accuracy:** The exactness metric is the most common metric and the most basic. It tells how many of the predictions made by a model

were correct and how many were not. It is the ratio of correctly predicted instances to the total instances as given in Equation 1.1 where TN indicates true negative, TP represents true positive, FN represents false negative and FP represents false positive in the equation. For classification, true positives (TP) and true negatives (TN) are the most relevant and accurate parameters. However, it does not tell how many of the instances that a model was supposed to identify were in fact identified or how many of the things that a model was supposed to classify as negative were classified in fact as positive (which is not exactly what you want). So, to gain a more comprehensive insight into the effectiveness of a model across various forensic tasks, it is better to look at all four metrics instead of just relying on one.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.1}$$

- **Precision:** The focus of precision is on the accuracy of the positive predictions made by a model. It is the ratio of the number of truly positive predictions to the sum of truly pleasing and falsely pleasing predictions, as given in Equation 1.2. Forensic science particularly relates to precision when the cost of a false positive is very high, as it is in the case of fingerprint analysis. In that application, precision measures the accuracy of the model in positively identifying a match among the instances it predicts will match.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1.2}$$

- **Recall:** Recall, is also known as 'sensitivity' or the 'true positive rate'. Recall assesses how well the model captures all the positive instances in the dataset. Again, assuming our resolution to a forensic problem is a model that makes predictions, recall assesses the accuracy of the model in capturing all the truly pleasing predictions. An instance of this is found in DNA profiling, where the recall metric measures how well individuals can discern one from another based on their unique genetic markers. The profiling's correct identification rate averages well below 100%. Recall is given as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.3}$$

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives as given in Equation 1.4. Both metrics are vitally important because they directly assess the balance between two types of errors—false positives and false negatives—that are pivotal for scoring the success of forensic science outcomes.

$$\text{F1-Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{1.4}$$

Applications in Forensic Science: Evaluation metrics play a very important role in measuring the reliability of forensic models. In the context of applications, such as voice and speaker recognition, overall accuracy in identified speakers is what we generally seek. With precision, we can ensure that matched fingerprints in an analysis are correct and not false positives. Recall becomes critical when we want to ensure that all 'recognized' instances of nefarious activity in a system have been apprehended. With these three core metrics—accuracy, precision and recall—forensic scientists and investigators can see inside the black box of a model's performance and make better decisions based on what appearances that model gives when it is working well and when it is not.

Cross-Validation and Overfitting

Cross-Validation: In the forensic sciences, the applied use of machine learning requires a high level of assurance that a model can be trusted to operate on data it has not seen before. For instance, a model should not be expected to work on new DNA samples if it has only been trained on previously known samples. The primary technique for assessing the predictive power of a model is called cross-validation. The basic idea is to hold out some of the data from the model during training and then see how well it performs when the model is applied to that held-out data [18].

Overfitting: When a model becomes too finely attuned to the training data, it overfits. By overfitting, a model learns not only the essential features of the training data but also the unimportant, irrelevant details. In this context, a model may give us 'results we can trust' for the training data, but it is far less effective on fresh, new data—in our case, new forensic evidence [19].

ROC Curves and AUC (Area Under the Curve): Receiver operating characteristic (ROC) curves and area under the curve (AUC) are important tools for evaluating how well machine-learning models work in forensic science [3]. They help understand the balance between true positive rates and false positive rates, which is crucial in tasks such as fingerprint analysis or DNA profiling.

The true positive rate (sensitivity) and false positive rate (1—specificity) tell us how well a model performs only at fixed, predetermined thresholds. ROC curves provide much more information than that. They show how the true positive rate varies with the false positive rate at all possible thresholds and help us select the 'best' operating point. AUC collapses the ROC curve down to a single number. AUC is higher the better the voice recognition model is and the better a model can distinguish between normal and malicious activities in the cyber domain. Figure 1.5 shows an ideal ROC–AUC curve.

Applications in Forensic Science: In forensic work, we need to pay much more attention to ROC curves and AUC than we historically have—if we do not want to maximize false positives or false negatives.

Types of Machine Learning: Supervised, Unsupervised and Reinforcement Learning

In the field of forensic science, diverse machine-learning methodologies are applied, with the three fundamental types—supervised learning, unsupervised learning and reinforcement learning—being the most common. Each plays a distinctive and valuable role, rendering its own form of interpretation of the data. Figure 1.6 shows different types of machine learning.

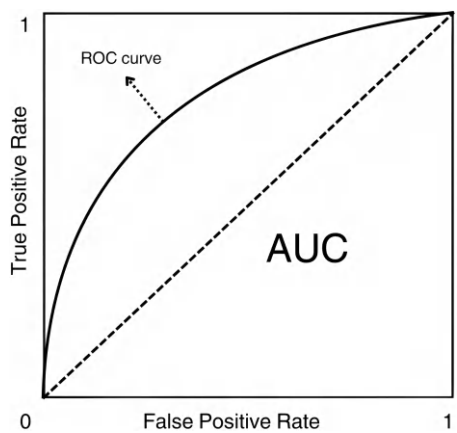

Figure 1.5 ROC–AUC curve.

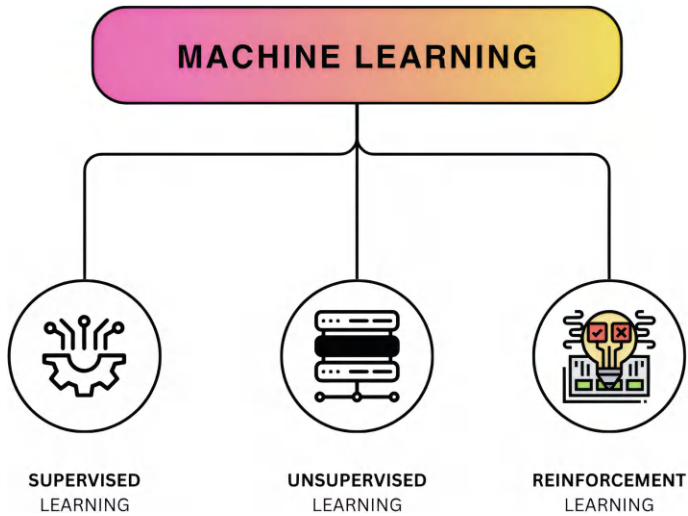

Figure 1.6 Types of machine learning. Source: https://www.dreamstime.com/machine-learning-types-supervised-vs-unsupervised-reinforcement-glance-vector-editable-stroke-colors-image297519539

Supervised Learning: Supervised learning is a method where algorithms are trained on labelled data to understand the relationship between inputs and outputs. These algorithms are then applied to projects that require a clear, known target for the algorithm to achieve.

Unsupervised Learning: Unsupervised learning is a method where the algorithm works on unlabelled data. When analyzing unlabelled data, the algorithm attempts to find inherent patterns or structures in the data. Forensics experts can better utilize machine learning when they understand the primary operating conditions and use cases of the two primary methods: supervised and unsupervised learning. Clustering algorithms, such as k-means and hierarchical clustering, are important to understand when forensic experts think about applying unsupervised learning to their projects.

Reinforcement Learning: Reinforcement learning involves an agent learning from feedback in a dynamic environment, which can help in cybercrime investigations and resource optimization.

Understanding these machine-learning types helps forensic experts uncover patterns, streamline analysis and make better decisions.

Challenges, Ethical Considerations, Interpretability and Explainability of Machine-Learning Models

When it comes to machine learning in forensic applications, two things are extremely important: data quality and bias. And for good reason: in the

contexts where learning algorithms are being taught to work, the correctness and fairness of their results simply have to be taken as a given if they are to be relied upon at all. Forensic science, after all, is not only about working with evidence; it is about working with evidence that has been found to be accurate, complete and consistent. If it is not, then the results being worked upon—and, in the case of machine learning, the models being worked with—aren't worth much in terms of both predicting future results and ensuring those results are fair.

Enhancing data quality involves identifying and repairing errors and removing outliers. We must involve data from all sorts of groups and conditions to reduce bias. We also use algorithms that train with a reduced bias and thus result in a much fairer set of answers. And, of course, we try to keep those models as understandable as possible. That is especially important in forensics, where we affect people's lives (and make decisions that can ultimately result in someone being imprisoned or not) and where the work we do needs to be understood by experts, investigators and lawyers (who also have to make decisions that can affect someone's being imprisoned or not).

To enhance clarity, we can employ more straightforward models, emphasize the key parts of our data, elucidate predictions and utilize rule-based systems. In a courtroom, the evidence provided by these models must be understood and trusted by judges and juries. When we apply machine learning to forensic science, we must ensure that we are protecting privacy and security, because the data we are working with—often biometric or genetic in nature—is very sensitive. The challenges are not small: the need to not lose data usefulness while adding privacy necessary for data to be useful; the need to have effective model protection while enabling necessary access to data for use by law enforcement; all of this break down into the three facets of privacy, security and ethics. Addressing these involves a number of strategies: using differential privacy, secure multiparty computation and encryption; obtaining informed consent; and ensuring that the forensic analysis of models not only is accurate but also respects the privacy rights of individuals.

Future Directions and Emerging Trends

Deep Learning and Neural Networks

Forensic science is now part of the machine-learning revolution, and the most powerful tools of this revolution are deep learning and neural networks. These have given us new architectural forms—such as CNNs, recurrent neural networks (RNNs), transformers and generative adversarial networks (GANs)—that handle the most basic types of forensic evidence, such as images, videos, audio and text. These networks, however, need a lot

of data to work—and a lot of human judgment to interpret their outputs—for forensic evidence to follow the rules of due process, which is a basic requirement for anything called 'forensics' to occur.

Explainable AI and Interpretable Machine-Learning Models

The forensic importance of XAI (explainable artificial intelligence) and the machine-learning models that produce interpretable outcomes lies in their ability to facilitate understanding and foster trust [20]. Forensic scientists must know the legal and ethical particulars of machine-generated evidence as they reason and make decisions about clear and present dangers to society. If an AI algorithm is deciding which facial recognition match is good for legal admissibility, we want to be sure the algorithm is producing understandable outcomes—not just for the algorithm's developers but for all of us. As this chapter demonstrates, some machine-learning models are better at this than others.

References

1. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2, 160. https://doi.org/10.1007/s42979-021-00592-x
2. Hickman, S. E., Woitek, R., Le, E. P. V., Im, Y. R., Mouritsen Luxhøj, C., Aviles-Rivero, A. I., Baxter, G. C., MacKay, J. W., and Gilbert, F. J. (2022). Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. *Radiology*, 302(1), pp. 88–104. https://doi.org/10.1148/radiol.2021210391
3. Sharma, S. U., and Shah, D. J. (2017). A Practical Animal Detection and Collision Avoidance System Using Computer Vision Technique. *IEEE Access*, 5, pp. 7358–7365. https://doi.org/10.1109/ACCESS.2016.2642981
4. Song, Y. Y., and Lu, Y. (2015). Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, 27(2), pp. 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
5. Breiman, L. (2001). Random Forests. *Machine Learning*, 45, pp. 5–32. https://doi.org/10.1023/A:1010933404324
6. Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. In C. Liu, L. Wang, and A. Yang (Eds.), *Information Computing and Applications* (Vol. 308, pp. 260–267). Springer. https://doi.org/10.1007/978-3-642-34041-3_27
7. Yang, F.-J. (2018). An Implementation of Naive Bayes Classifier. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 301–306). IEEE. https://doi.org/10.1109/CSCI46756.2018.00065

8. Goel, A., Goel, A. K., and Kumar, A. (2023). The Role of Artificial Neural Network and Machine Learning in Utilizing Spatial Information. *Spatial Information Research*, 31, pp. 275–285. https://doi.org/10.1007/s41324-022-00494-x
9. Sharma, S., Pandey, S., and Shah, D. (2023). Enhancing Medical Diagnosis with AI: A Focus on Respiratory Disease Detection. *Indian Journal of Community Medicine*, 48(5), pp. 709–714. https://doi.org/10.4103/ijcm.ijcm_976_22
10. Na, S., Xumin, L., and Yong, G. (2010). Research on K-Means Clustering Algorithm: An Improved K-Means Clustering Algorithm. In *2010 Third International Symposium on Intelligent Information Technology and Security Informatics* (pp. 63–67). IEEE. https://doi.org/10.1109/IITSI.2010.74
11. Patel, S., Sihmar, S., and Jatain, A. (2015). A Study of Hierarchical Clustering Algorithms. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 537–541). IEEE.
12. Jolliffe, I. T., and Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A*, 374(2065), p. 20150202. https://doi.org/10.1098/rsta.2015.0202
13. Belkina, A. C., Ciccolella, C. O., Anno, R., et al. (2019). Automated Optimized Parameters for T-Distributed Stochastic Neighbor Embedding Improve Visualization and Analysis of Large Datasets. *Nature Communications*, 10, p. 5415. https://doi.org/10.1038/s41467-019-13055-y
14. Sánchez-Maroño, N., Alonso-Betanzos, A., and Tombilla-Sanromán, M. (2007). Filter Methods for Feature Selection – A Comparative Study. In H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2007* (Vol. 4881, pp. 178–187). Springer. https://doi.org/10.1007/978-3-540-77226-2_19
15. El Aboudi, N., and Benhlima, L. (2016). Review on Wrapper Feature Selection Approaches. 2016 International Conference on Engineering & MIS (ICEMIS), Agadir, Morocco. IEEE, pp. 1–5. https://doi.org/10.1109/ICEMIS.2016.7745366
16. Lal, T. N., Chapelle, O., Weston, J., and Elisseeff, A. (2006). Embedded Methods. In I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh (Eds.), *Feature Extraction* (Vol. 207, pp. 137–165). Springer. https://doi.org/10.1007/978-3-540-35488-8_6
17. Zhai, J., Zhang, S., Chen, J., and He, Q. (2018). Autoencoder and Its Various Variants. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 415–419). IEEE. https://doi.org/10.1109/SMC.2018.00080
18. Refaeilzadeh, P., Tang, L., and Liu, H. (2009). Cross-Validation. In L. Liu and M. T. Özsu (Eds.), *Encyclopedia of Database Systems*. Springer. https://doi.org/10.1007/978-0-387-39940-9_565
19. Das, P. L., Manoj, A. S., Sharma, S., and Jayaraj, P. B. (2021). Early Detection of COVID-19 From CT Scans Using Deep Learning Techniques. In S. M. Thampi, E. Gelenbe, M. Atiquzzaman, V. Chaudhary, and K. C. Li (Eds.), *Advances in Computing and Network Communications* (Vol. 736, pp. 45–54). Springer. https://doi.org/10.1007/978-981-33-6987-0_5

20. Došilović, F. K., Brčić, M., and Hlupić, N. (2018). Explainable Artificial Intelligence: A Survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 0210–0215). IEEE.
21. Srivenkatesh, M. (2020). Prediction of Prostate Cancer Using Machine Learning Algorithms. *International Journal of Recent Technology and Engineering*, 8(5), pp. 5353–5362. https://doi.org/10.35940/ijrte.E6754.018520

Scope of Machine Learning in Forensic Trace Evidence Examination

2

VAISHALI, NIHA ANSARI
AND JEET DASGUPTA

Introduction

Artificial intelligence (AI) is an ever-evolving field that involves teaching machines to perform tasks that would typically require human intelligence [1]. One of the primary goals of AI is to develop machines that can think, reason and learn as people do. This involves creating algorithms that can analyse data, recognize patterns and make decisions based on that information [2]. It is a rapidly developing discipline that is also being used for enhancement in the field of forensics and the legal system as a whole. Considering the enormous quantity of data; the small size of the evidence in the incoherent, complicated environment; conventional laboratory frameworks, as well as occasionally inadequate expertise, professionals in forensic science and criminal investigation are currently facing many difficulties that could result in a failed investigation or injustice.

In the court of justice, evidence is defined as something which tends to either prove or disprove a fact under consideration related to the ongoing case. According to Section (3) of the Indian Evidence Act evidence can be of three types, oral or testimonial, documentary and electronic evidence, further, evidence can be classified into two types direct and indirect evidence. Direct evidence is described as proving a fact beyond doubt, whereas indirect evidence either supports a circumstance of the crime or points against this fact [3]. For example, a fingerprint found at the scene of a crime is circumstantial evidence that a specific person was present at the scene of the crime. Evidence can also be classified as real or testimonial. Real evidence is physical evidence that can be seen, touched or smelled. Testimonial evidence is the testimony of witnesses who saw or heard something about the case [4].

Physical evidence is any tangible object that can be used to establish proof or disprove a fact in a legal proceeding [5]. It can be anything from

DOI: 10.4324/9781003449164-3

fingerprints and DNA to weapons and clothing. Physical evidence is often used to corroborate witness testimony or to establish a timeline of events. There are six types of physical evidence, namely, trace evidence, transfer evidence, impression evidence, striated evidence, geometric evidence and chemical evidence [6].

Scientific trace evidence was first recognized by Edmund Locard in the year 1910 [7]. Trace evidence is a type of physical evidence which is defined as small, often minuscule items left at the scene of a crime or on someone who was involved in a crime [6]. Trace evidence examination is the assessment of physical trace evidence obtained at crime scenes, such as hair, fibres, dirt and glass fragments [8]. It can provide significant details with regard to the perpetrator's identity, the circumstances surrounding a crime and the location of individuals or goods. It may be possible to train machine-learning algorithms to assess traces of evidence and offer insights that could help with criminal investigations [9]. As machine learning enables machines to learn patterns and look for similarities and differences which humans may miss, it is gaining popularity in physical-evidence analysis. Machine learning is also the foundation that enables AI to create intuitive, human-friendly interfaces that can be easily used by humans [10].

Significance of Trace Evidence

1. Trace evidence is used to establish a link between the crime scene, the victim and the culprit [11]. For example, a paint chip found at a car accident and the same type of chip found on the cloth of the victim can establish a link between the two.
2. Trace evidence either strengthens or weakens the witness testimony, providing an objective ground to the truth value of their evidence. For example, evidence such as DNA and fingerprints can prove the factual presence or absence of a person at a scene of crime.
3. Trace evidence can be crucial in filling in the gaps while reconstructing the events of a crime providing a logical series of steps [12]. Elemental analysis of trace evidence provides a unique perspective which may be overlooked at an initial glance, which can change the narrative of the investigation significantly.
4. Trace evidence such as tool marks and glass fracture patterns help us to identify the type of force, direction of force and the tool used which provides us with an idea of events undertaken at a scene.
5. Trace evidence plays an important role in defining the timeline, validating a hypothesis and introducing new angles and leads in an

investigation. The presence or absence of trace evidence can be crucial for validating timelines and connecting one event with another or connecting two crime scenes.

Problems Encountered during Trace Evidence Examination

Trace evidence analysis comes with its fair share of difficulties; this can be due to the nature of the evidence or the limitation and availability of instruments for their analysis. One of the main issues with trace evidence, as the name suggests, is that it is found in very small amounts, making identification, handling, preservation and contamination prevention difficult. These issues can be easily overlooked until trained eyes look at the trace evidence and the evidence is at a greater risk of becoming lost or damaged. Trace evidence can be easily contaminated from multiple sources including environmental factors, such as rain and temperature, contamination while handling, such as from gloveless hands, contamination can also occur during collection, packaging, transportation or storage, which ultimately compromises the evidentiary value of the evidence in the court of law. Background noise during the analysis of trace evidence can interfere with the identification of the elements; to overcome we require expertise and sophisticated analytical techniques which can accurately identify the elements with great sensitivity. Forensic experts handling such evidence should be well-trained in skills and knowledge to accurately perform the required analysis with precision. There should be universal standardization in the steps of analysis to avoid inconsistencies.

Machine Learning in Gait Pattern Analysis

Gait is defined as the movement of the limbs of an organism when it moves from one place to another, most animals have a quadruple gait in which they use both pairs of limbs, the front and the rear, to move, whereas humans have a bipedal gait in which we only use our lower limbs to move. An animals gait has been traditionally used to track an organism, mark its movements and locate its habitual area, we use gait similarly in the case of forensic science, other than this, gait is used in medical science to diagnose neuromuscular problems and diseases such as Parkinson's and cerebral palsy. Gait is also used in animation to create human-like movements for graphically developed characters and in many other fields. In forensics we work with two aspects of gait pattern, one is gait impression and the other is using phases

of gait that are captured on video cameras and surveillance recordings. Traditional methods of analysis had limitations in accurately understanding gait patterns; however, with machine-learning approaches, many new potential avenues have been unlocked. These data-driven algorithms have proven to be more effective in recognizing patterns and relationships in gait data, leading to improved accuracy and efficiency in identifying individuals based on their walking styles. Zixuan Zhang et al. explored an effective and broadly applicable approach for increasing the durability of triboelectric sensors in motion detection as shown in Figure 2.1 [13].

Sharon Jemimah Peace et al. released research on a pose estimate strategy for gait evaluation using machine learning, highlighting the connection between variations in gait patterns and degradation in perceptual and motor abilities in the elderly. It emphasized the usefulness of gait analysis in diagnosing senility and frailty diseases [14].

The gait cycle occurs with the initial heel contact of either foot and concludes with the subsequent heel contact of the same foot [15]. Hence, a singular gait cycle encompasses the execution of two distinct steps, wherein each step corresponds to either the right or left foot or, conversely, the left or right foot. The gait cycle is essentially composed of two distinct phases, namely the swing phase and the stance phase [16]. The term 'stance phase' describes the

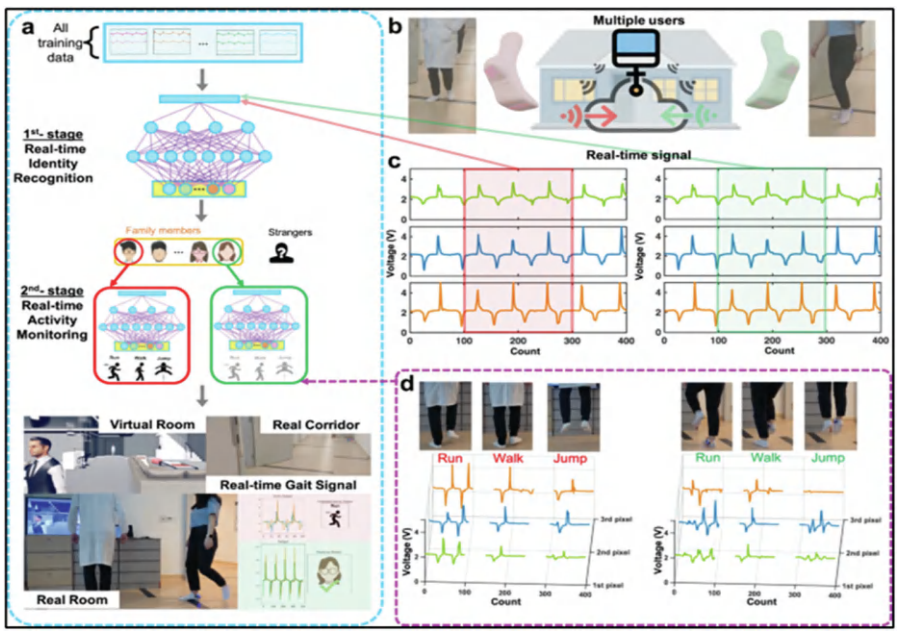

Figure 2.1 Deep learning enabled gait recognition by wearing sensing socks [13].

time that the foot is in contact with the earth. The swing phase refers to the time when the foot is not touching the ground. Examiners are better able to pinpoint deviant gait traits when the stance phase and swing phase are split into eight distinct stages, referred to as critical occurrences. Then gait cycle is divided into eight separate subperiods, cutting between the two phases. [16]. The stance phase has five subphases: initial contact, loading response, mid-stance, terminal stance and pre-swing. The second phase, that is, the swing phase has three subphases: initial swing, mid-swing and terminal swing represented in Figure 2.2 [16].

Gait features can be collected by various methods, which include photography of gait impressions, video surveillance tapes, ground reaction force plates, which record the kinematic features of the gait and wearable sensors. Abdul Saboor et al. provide an overview of current developments in gait analysis, encompassing aspects such as publication specifics, sample rates, machine-learning models (MLMs), wearable sensors and their respective placements [17].

The dataset commonly consists of video sequences or time-series data that depict the walking pattern of an individual. The preprocessing of information is a critical step for obtaining data ready for use in machine-learning algorithms. Some potential techniques that could be employed are noise reduction, normalization and feature extraction. The process of feature extraction entails the identification of pertinent attributes from unprocessed

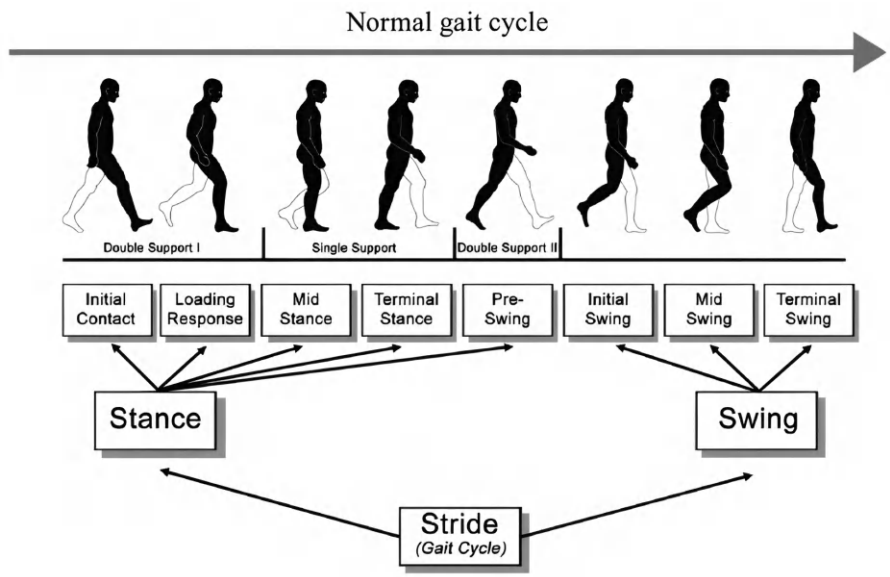

Figure 2.2 Classic gate model shows the stance and swing phase [48].

data, including but not limited to joint angles, step length or temporal factors. Feature selection is the process of identifying the most relevant variables for a specific task. Due to the large number of dimensions in gait data, various extracted features such as stride length, step duration, joint angles, foot pressure distribution and other gait-related metrics are employed in order to select the most pertinent and distinguishing aspects. This stage is crucial in directing the model's attention towards the fundamental elements of gait patterns, resulting in enhanced performance and efficiency. For the process of training a model, the dataset is partitioned into separate subsets for training and testing purposes [18]. The training data is utilized for the purpose of training the machine-learning model, whereas the testing data is employed to assess its performance. A range of machine-learning algorithms are employed in the construction of a gait pattern identification model. Algorithms such as support vector machines (SVMs), neural networks, random forests and deep learning architectures, such as convolutional neural networks (CNNs), are widely employed in various domains. Jayati Ghosh Dastidar et al. used three machine-learning models, support vector machines (SVMs), k-nearest neighbours (KNN) and random forest, for the classification of biometric identification based on gait parameters analysis from walking video sequences. Gait parameters such as heel strike angle, toe-off angle and stride angle were analysed in the system [19]. The algorithms acquire knowledge from the provided training data in order to discern patterns and build correlations between gait variables and the unique identities of individuals. During the training phase, the model acquires the ability to establish a correspondence between the extracted features and the distinct gait patterns of people, relying on annotated data. The evaluation of the trained model involves the utilization of distinct datasets to measure its performance, accuracy and ability to generalize. The aforementioned assessment procedure aids forensic professionals in refining the model and determining the most appropriate algorithm for the given forensic testing objective. Once the model has been validated for its accuracy, it can be utilized to discern individuals in practical situations, categorize patterns of walking or establish contact between the accused and the crime location through the examination of the gait marks.

The application of machine-learning techniques has enabled forensic professionals to automate and optimize the process of analysing gait patterns. Algorithms possess the capability to examine extensive datasets with more efficiency and accuracy in comparison to conventional methodologies. In addition, the impartiality of machine learning mitigates the inherent biases linked to human visual evaluations, hence enhancing the dependability and credibility of findings in the field of forensic investigations.

Machine Learning in Paint Evidence Analysis

The term 'paint evidence' encompasses paint residue or fragments discovered at crime scenes, on automobiles or on other things that may be associated with illegal actions [7]. The analysis of paint evidence can yield valuable insights pertaining to several characteristics, including but not limited to the type of paint, its composition, colours, texture and the manner in which different layers of paint are arranged [20]. When subjected to thorough analysis, paint evidence has the potential to facilitate the establishment of correlations among suspects, victims and crime scenes. Paint residue can be found at the location of various incidents, such as hit-and-run accidents, burglaries, assaults and so on. Paint residue can be found in the form of chips or streaks on clothing, vehicles or items, or it may be present in a loose state at the site. Furthermore, there is a possibility of paint transferring between two cars, a vehicle and an object or two objects [21]. Using machine learning in paint sample analysis helps us to negate the chances of damage to evidence, as paint samples are usually very brittle in nature and analysing them under scanning electron microscopy (SEM) and other microscopic examination carries a risk of breaking the sample itself, whereas in machine learning approaches we can analyse the paint chips and the hue marks via photographs taken and through the algorithm. Paint samples have one question in common and that is whether this questioned sample is the same as that of control or is the same as that used in a particular car or model of vehicle, for this we often need to match the sample results against the paint data query which can be a tedious task done manually but with machine-learning algorithms, the algorithms have already been trained on the dataset so they can give a similarity score automatically. Although current algorithms cannot replace analytical techniques such as Gas Chromatography-Mass Spectrometry (GC-MS) and Fourier transform infrared spectroscopy (FTIR), training machine learning models on their results can produce similarity scores and can allow for limited prediction of paint composition .

The aforementioned skill possesses the potential to make a substantial contribution to criminal investigations by perhaps facilitating the identification of suspects or establishing connections between crime scenes. Francis Kwofie et al. analyzed automobile paint samples using a machine-learning technique from 26 vehicles to identify the manufacturer and make and model of the vehicle, as shown in Figure 2.3 [22].

Paint analysis with machine-learning algorithms involves a few steps, which include data collection from the scene or object involved in the crime and the clothes of the person, if any, involved. Machine-learning algorithms have been used to develop a pattern recognition approach that combines

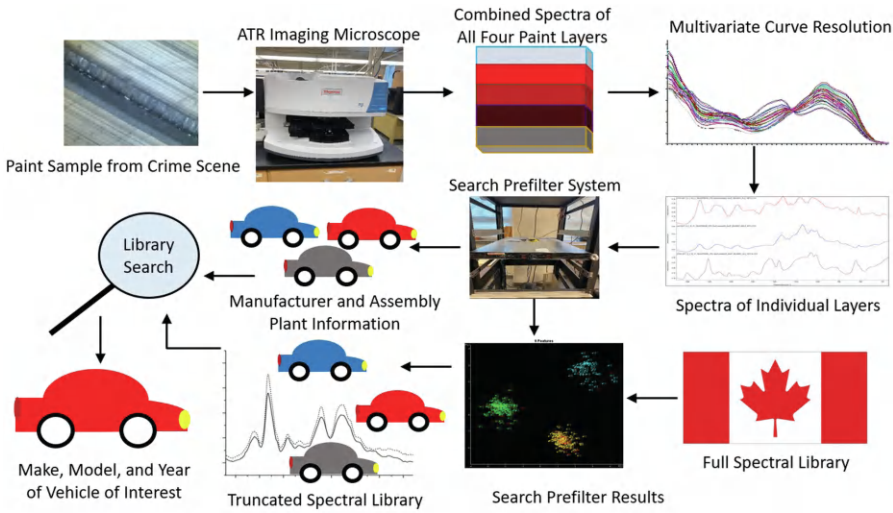

Figure 2.3 Forensic paint chip examination using a machine-learning model [22].

infrared spectral libraries with cross-correlation search algorithms. This improves the accuracy of the search and the comparison of the original equipment manufacturer (OEM) of automotive paint layers using IR spectra alone [23].

The initial steps for sample processing involve preprocessing of the sample, this helps us to remove noise, variance and other useless information. Paint samples vary in their composition, colour, texture and age. Data preprocessing helps form a better feature extraction process. The feature extraction process involves using numerical values to characterize the paint samples. Histograms, pigment composition, spectral reflectance and particle size distribution, are possible features that could be considered for paint samples and other pertinent factors. Next comes the data preparation process which involves arranging the dataset of already identified samples and categorizing on the basis of their source of origin to be divided into two sets for training and testing in an 80:20 ratio various machine-learning algorithms such as support vector machines (SVMs), k-nearest neighbours (KNN), random forests and neural networks can be used. This training dataset can be labelled for supervised learning methods and non-labelled for unsupervised learning methods.

In this case, the trained model is assessed by means of the testing dataset aimed at evaluating the model's accuracy, precision and recall. All predictions are composed of the entirety of positive (P) and negative (N) examples. P is composed of TP and false positives (FP), and N is composed of TN and

false negatives (FN). Thus, we can define accuracy as ACC = TP + TNTP + TN + FN + TP = TP + TNP + N. Thus, such techniques are aimed at evaluating the results of a model on different divisions of data and are used quite often. Once this model is proven to work, it can then be used towards the interpretation of paint samples collected from crime scenes or any other object.

The application of automated learning methods for the analysis of forensic paint evidence allows the handling of data originating from various sources. Spectroscopic analysis covers a broad spectrum of techniques such as FTIR (Fourier transform infrared spectroscopy) and Raman spectroscopy that are used to obtain spectra from paint samples [24]. Kwofie et al. have demonstrated that it is feasible to obtain infrared spectrum data of all four layers of the cross-sectioned unbroken multi-layered OEM paint chip in a single examination. This is carried out by restoring the whole structure in the infrared (IR) transmission mode employing an infrared transmission microscope [22]. Spectra can then be analyzed using machine-learning methods in an attempt to be able to get a much deeper insight into the microscope and its composition and characteristics of the paint layers are determined by analyzing paint samples. Microscope employing imaging in addition to microscopy techniques. For instance, in a situation where one has to compare various samples of paint, it is easy to have automated systems or machine-learning models to analyze images of tissues at the microscopic level and then determine certain attributes. For the identification of paint samples and to determine the elemental content then techniques such as SEM with EDS is a useful technique that integrates scanning electron microscopy (SEM) with energy dispersive X-ray spectroscopy (EDS). People assumed that with the application of machine-learning approaches, it would be beneficial to perform complex chemical data analysis to identify the sources of paints.

Machine Learning in Soil Evidence Analysis

Soil evidence refers to the analysis of soil samples collected from crime scenes or associated with suspects, victims or objects involved in a crime [25]. 'Soil evidence is often a silent witness to a crime', these words underscore the remarkable potential that soil evidence holds within the realm of forensic investigations. Like a quiet observer, soil can bear witness to the unfolding of a crime, capturing vital information that may otherwise remain hidden. One of the foremost strengths of soil evidence is to reveal information about the origin of a suspect. Soil is not uniform; it carries distinct characteristics depending on its location. Every region, and often sub-regions within regions,

possesses unique soil compositions. Thus, when soil samples are collected from a suspect's person, clothing or vehicle, they can provide a geographical location or geographic fingerprint that links the individual to a specific area or even a particular crime scene [26]. This association can be essential in verifying or challenging a suspect's alibi. The Lindbergh Kidnapping Case (1932) is a case in which the child of Charles Lindbergh was kidnapped from their home in New Jersey. Soil evidence was crucial evidence in this case which is proven effective in solving this case. Soil found on a ladder used in the kidnapping was analysed and traced back to a location near the home of Bruno Hauptmann, who was subsequently arrested and convicted. The case is a landmark example of how soil evidence can link a suspect to a crime scene.

However, despite its potential, the analysis of soil evidence traditionally relied on manual methods. Vung Pham et al. examined soil spectral information using portable X-ray fluorescence (pXRF) spectroscopy in conjunction with machine learning or deep learning algorithms for predicting soil attributes, as shown in Figure 2.4 [27].

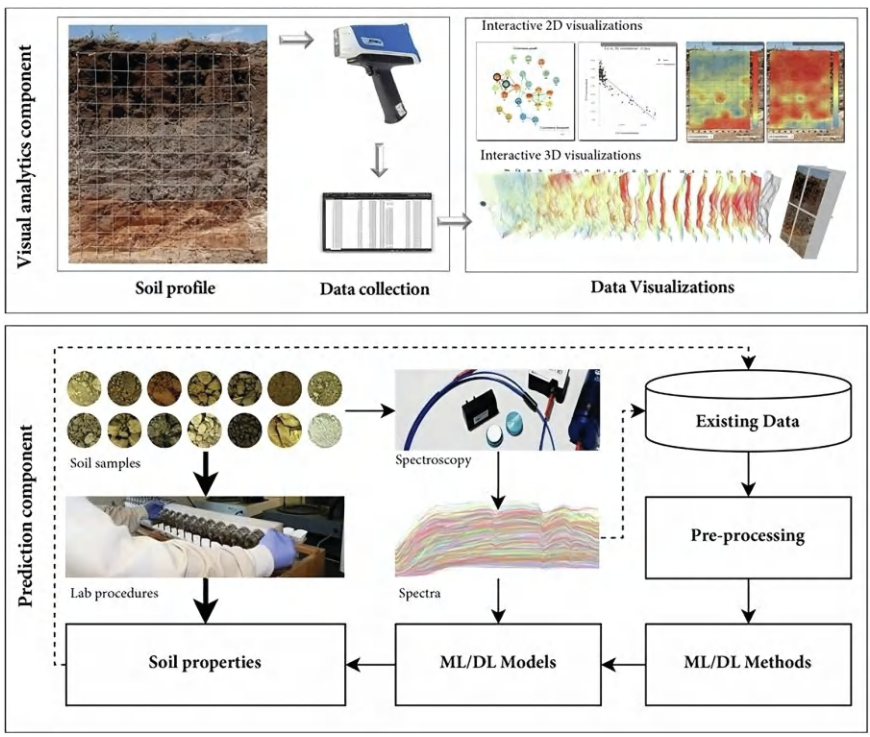

Figure 2.4 Soil profiling using machine/deep learning approaches [27].

Forensic scientists would painstakingly compare the physical and chemical properties of soil samples, a process that is both time-consuming and susceptible to human error. Recognizing these limitations, the integration of machine learning (ML) into soil evidence analysis has accompanied a new era of forensic science. The following are the contributions of machine learning to the field of forensic science:

Pattern Recognition: Due to their capability of identifying complex patterns in the data, ML algorithms are useful in the analysis of large datasets. This ability proves highly useful in the context of soil evidence analysis. Soils collected from different sites may appear to be similar when tested for physical properties but they are different. This generated data are then analysed by different ML models which on the basis of feature extraction can differentiate between particle size, distribution, mineral content and other organic matter. For example, the presence of certain minerals and ores can be helpful in identifying the source of a particular region.

Source Attribution: Establishing the origin of a soil sample is one of the primary goals of forensic soil analysis. ML greatly improves this process. Based on the database of soil profiles, the ML algorithms can carry out the search for compositions similar to the sample. Such a comparison neglects the time taken by a manual approach and is data-driven. The end-product therefore has a higher degree of accuracy in relating the collected soil sample to its probable source. When other approaches can be inconclusive, ML can give a more clear-cut answer, which can assist in linking evidence to places.

Geospatial Analysis: The geospatial analysis in soil evidence analysis involves the integration of both the physical properties of the soil samples and geographical data. This fusion is a unique strength of each machine-learning algorithm. Explaining these ideas, it is possible to state that the machine-learning algorithm can estimate the geographical origin of a sample with high accuracy if it regards not only the chemical and mineralogical composition of the soil samples but the geographic information related to the samples as well. This capability is especially important when the exact location of the crime scene is not well-known or when there is a dispute about its location. Machine learning on its part can correlate soil composition with geographical information to possibly cut down the list of likely crime scenes, thus saving time and effort in investigations. Chandan and Ritula Thakur have made a comparative analysis of the assessment of the classification of soil with the methods that are based on the use of machine learning for identifying many distinguishable

characteristics such as moisture content of the soil nutrients found in the soil, the physical structure of the soil, the state of the soil, the pH value of the soil and the consistency of the soil [28].

Data Integration: In forensic investigations, the analysis of different kinds of evidence is commonly conducted. The ability to compile types of evidence is a major advantage of ML. In the context of soil evidence, this implies integrating the soil data with other forms of evidence for instance tyre impressions, footprints, DNA samples or even eyewitness testimony. The combination of these various sources of information gives a much fuller and more rounded picture of the crime scene. For instance, the soil type may be matched with a tyre impression to make a stronger association between the suspect's car and the crime scene, making the evidentiary and investigative merit of the soil higher.

Some of the cases that have demonstrated the ability of machine learning in soil evidence are presented below. For instance, in a hit-and-run accident investigation, soil samples were analysed using machine-learning algorithms and the results were compared with the samples collected from the suspect's vehicle and the result led to the successful identification of the involved vehicle.

In another case, a missing person's investigation involved the application of machine learning in analyzing soil samples detected on the clothes of the missing person. When integrated with geospatial data, one can easily pinpoint the area of interest and in this specific case, the investigators were able to find the victim.

Machine Learning in Glass Evidence Analysis

Forensic experts bear a significant responsibility in the processing of forensic evidence to gather essential information aiding criminal investigations. Glass, as a tangible piece of evidence, is commonly encountered in a range of criminal activities, including burglary, road accidents, homicide, sexual assault, shooting events, arson and vandalism. The fragments of shattered glass resulting from a broken window have the potential to become embedded in the footwear or clothing of individuals involved in a burglary [29]. Similarly, the presence of minute particles of headlight glass discovered at the location of a hit-and-run incident can provide valuable evidence that verifies the identification of a suspected vehicle. Furthermore, glass traces may also be spotted on the clothing of the suspect, particularly in cases where a bottle

is utilized as a weapon. Moreover, in illustrations of violence, glass objects such as bottles, window pane glass, mirrors, eyeglasses and other similar items may unintentionally shatter, resulting in pieces that could potentially attach themselves to the perpetrator's attire or footwear [30]. Glass fragments possess the unique ability to connect shattered objects to crime scenes, victims and potential suspects, providing insights into the circumstances surrounding the 'how', 'when' and 'what' of an incident [31]. These minute glass fragments readily adhere to articles such as clothing, shoes, hair, skin and various objects, making them a valuable resource in forensic analysis.

Traditionally, the investigation of glass evidence in the discipline of forensic science relied on careful manual scrutiny. These encompass techniques such as refractive index (RI), density gradient analysis, physical attribute matching and advanced microscopy [32, 33]. Moreover, there is an increasing tendency among forensic scientists, researchers and investigators to utilize elemental analysis as a technique for examining glass particles. Various elemental analysis techniques are employed in scientific research, such as laser ablation inductively coupled plasma mass spectrometry (LA–ICP–MS), particle-induced X-ray emission (PIXE), scanning electron microscopy with energy-dispersive X-ray spectrometry (SEM–EDS) [34], instrumental neutron activation analysis (INAA) and prompt-gamma activation analysis (PGAA) [35]. Grzegorz Zadora discusses the classification of glass fragments based on their elemental composition obtained by SEM coupled with an energy-dispersive X-ray spectrometer instrument and refractive index values and compares the efficiency of likelihood ratio-based methods to other classification methods such as support vector machines and naive Bayes classifiers [36]. However, the introduction of machine learning (ML) has revolutionized this practice, resulting in improved precision and effectiveness in the analysis of glass evidence. Ruthmara Corzo et al. evaluated the forensic glass evidence collected from automotive windshields using refractive index, micro–X-ray Fluorescence Spectroscopy (μXRF) and laser-induced breakdown spectroscopy (LIBS) the data from this is shown in Figure 2.5 [37].

The Significance of Glass Evidence: Glass evidence incorporates various aspects of forensic investigations. It can originate from broken windows, shattered bottles or even gunshot damage, offering critical information in criminal cases [30].

Source Attribution: Because different types of glass have various qualities, it is possible to determine the origin of a certain glass shard. This can assist detectives in tracing the origin of a shattered window, a weapon or even the glass of a vehicle.

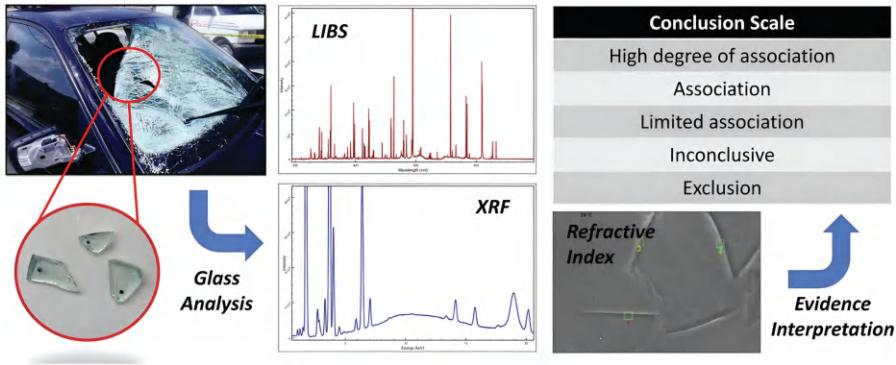

Figure 2.5 Forensic glass evidence examination using refractive index, micro–X-ray fluorescence spectroscopy (μXRF) and laser-induced breakdown spectroscopy (LIBS) data [37].

Trajectory Analysis: Glass fragment trajectory analysis is a method used to replicate the path taken by bullets or projectiles that caused the glass to shatter. Forensic professionals can uncover vital information about the direction and location of impact by meticulously examining the patterns and intricacies of glass fracture.

Connecting Suspects with Crime Scenes: Forensic glasses are a link between the suspects and scenes of the crimes. Whenever glass particles are traded on the clothes or the shoes of a suspect, then they turn into very persuasive incriminating materials, more so when they are of the same type as those found at the designated crime scene.

Identifying the Type of Glass: Common glasses such as float glass or tempered glass have different types of fracture properties. Based on this unique pattern of glass fracture a forensic expert can identify the type of glass involved in an occurrence.

Role of Machine Learning in Glass Evidence Analysis: Machine learning has brought a new dimension to the analysis of glass evidence in forensic science. The following are the contributions of machine learning in the field of forensic science:

Pattern Recognition: Machine-learning algorithms are useful in identifying complex patterns in a large database. They can study the fracture pattern of fragments of glass in the case of glass-related incidents. Machine learning can help in identifying the type of force or projectile that was used in the incident based on the characteristic of breakage that may be valuable in reconstructing the crime scene.

Source Attribution: Machine-learning algorithms specialize in comparing the distinguishing features of glass pieces to a large and detailed database of known glass samples. This capacity provides forensic scientists with a useful tool for determining the origin of a given glass fragment with exclusive precision. Machine learning provides a rigorous system to determine source attribution, whether it determines the brand of a window, identifies the manufacturer of glass items or even the specific make and model of a vehicle from which the glass originates. This not only improves the forensic investigator's ability to track evidence back to its source, but it also improves the forensic investigator's ability to distinguish among apparently identical fragments of glass.

Trajectory Analysis: Machine learning aids in the difficult task of reconstructing the trajectory of bullets or projectiles that caused glass shattering [38]. Machine learning assists investigators in determining the angle and direction of the shooting by examining the placements of glass fragments, impact sites and other pertinent factors. This procedure solves the unresolved issues surrounding shooting occurrences by providing investigators with the invaluable ability to determine the precise path taken by projectiles. This, in turn, offers important insights into shooting situations, leading to the formation of a thorough picture of the crime scene.

Linking Suspects: Machine learning is being used to compare glass evidence found on suspects' clothing, footwear or personal belongings with fragments carefully collected from crime scenes. Machine-learning algorithms systematically evaluate the data for matches, and when such links are made, they give a solid platform for legal processes to proceed with certainty. As a result, significant relationships between suspects and criminal activity have been discovered.

Forensic glass evidence has been widely used in several forensic investigations as a decisive factor in solving criminal cases and the delivery of justice. Machine learning has revolutionized the profession when integrated into the analysis of glass evidence and forensic specialists have been able to extract deeper information with accuracy and efficiency. Considering the advancement in technology, there is a combination of machine learning (ML) and forensic science, which offer the probability to enhance the accuracy and effectiveness of investigations. This, in turn, may help in the achievement of justice and provide answers to victims and their loved ones. Glass has not only clarity but also the capacity to show the truth in even the most complicated criminal cases and, thus, is the gateway to the past.

Advantages of the Approaches

Forensic investigators can benefit from machine-learning techniques as they are capable of handling huge datasets related to investigations much more efficiently than conventional methods. The rapid development of artificial intelligence technologies helps assist law enforcement agencies and forensic experts, police personnel in not only detecting crimes but also in preventing and predicting them [2]. Machine-learning algorithms are designed to identify crime patterns and discover suspicious anomalies, forecast future crime locations, evaluate criminal risk factors and reveal criminal networks [39]. Machine-learning models can be trained to automatically sort and classify different types of trace evidence. For example, in hit-and-run cases, it can automatically tell from which model of vehicle the paint chip may come from or what type of fibres are used. It is quick and gives more accurate results. Machine learning is able to function with greater efficiency since its performance is not affected by subjective judgement, it never becomes exhausted, and it is not influenced by sentiments [40].

Disadvantages of the Approaches

The implementation of machine learning carries inherent risks which gives unexpected negative outcomes [41]. It is complex, expensive and has the potential for showing bias similar to that of a human forensic expert, dependent upon the type of data utilized to train the machine learning model [42–44]. The use of AI in evidence analysis carries a probability of giving rise to either false positives or false negatives, which could have major implications for both suspects and victims [44]. Obermeyer et al. discovered indications of racial bias in a healthcare algorithm which has the data of 200 million people [45].

Future Aspects of Machine Learning

Machine learning has the capacity to improve the manner through which forensic experts examine evidence [38]. The future of machine learning in trace evidence examination holds the potential to transform forensic science. The relationship between technology developers and forensic professionals will play a crucial role in creating ethical, legitimate and effective applications of machine learning in the complex field of trace evidence analysis [41]. In contrast to human learning, machine learning is primarily advantageous due to its capacity to process vast quantities of data, identify patterns and operate more effectively in environments that are less predictable [46]

Furthermore, it is being employed to enhance the overall productivity and efficacy of forensic laboratories. AI-powered software can accelerate as well as improve evidence processing in laboratories by automating routine activities and advancing work processes which enables labs to effectively meet the growing requirements of contemporary criminal investigations [47].

References

1. Singh, L., 2023. Generative AI vs Machine Learning vs Deep Learning Differences. [Online] Available at: https://redblink.com/generative-ai-vs-machine-learning-vs-deep-learning
2. Jadhav E., Sankhla M., Kumar R., 2020. Artificial Intelligence: Advancing Automation in Forensic Science & Criminal Investigation. *Journal of Seybold Report*, 15(8), pp. 2064–2075.
3. Haldhar N, 2022. B&B Associates LLP. [Online] Available at: https://bnblegal.com/article/types-of-evidence-direct-evidence-vs-circumstantial-evidence/
4. CASD, 2015. Coatesville Area School District ORGANISATION. [Online] Available at:https://www.casdschools.org/site/handlers/filedownload.ashx?moduleinstanceid=7201&dataid=6177&FileName=02-TypesOfEvidence.pdf
5. Vikas, 2023. The EDU LAW. [Online] Available at: https://portal.theedulaw.com/singlearticle?uid=469#:~:text=Real%20Evidence%3A%20it%20refers%20to,circumstances%20related%20to%20the%20case.
6. Dahiya, M., 2015. *Principles and Practices in Contemporary Forensic Sciences*. Shanti Prakashan.
7. Saferstein, R., 2017. *Criminalistics: An Introduction to Forensic Science*. 12th ed. Pearson.
8. Global Forensic and Justice Center, 2013. Trace Evidence: Introduction. [Online] Available at: https://www.forensicsciencesimplified.org/trace/
9. Chinnikatti, S. K., 2018. Artificial Intelligence in Forensic Science. *Forensic Science & Addiction Research*, 2(5), pp. 182–183.
10. Tucci L., 2023. Tech Target. [Online] Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML
11. Dillon, A., 2021. [Online] Available at: https://study.com/academy/lesson/trace-evidence-definition-analysis-examples.html#:~:text=Interaction%20Between%20Perpetrator%20and%20Victim&text=Bodily%20fluids%2C%20gunshot%20residue%2C%20hair,other%20parts%20of%20the%20body.
12. New Jersey State Police, 2023. NJSP. [Online] Available at: https://www.nj.gov/njsp/division/investigations/trace-evidence.shtml
13. Zhang, Z., He, T., Zhu, M., et al., 2020. Deep Learning-Enabled Triboelectric Smart Socks for IoT-Based Gait Analysis and VR Applications. *NPJ Flexible Electronics*, 4, p. 29. https://doi.org/10.1038/s41528-020-00092-7
14. Sharon, C., Ebenezer, V., Bijolin, E., Abinayaa, R., Sharan, D., & Roshni, T., 2023. Pose Estimation Approach for Gait Analysis using Machine Learning. 2nd International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, pp. 1071–1075, doi: 10.1109/ICEARS56392.2023.10085311.

15. Prakash C., Kumar R., Mittal N., 2018. Recent Developments in Human Gait Research: Parameters, Approaches, Applications, Machine Learning Techniques, Datasets, and Challenges. *Artificial Intelligence Review*, 49, pp. 1–40.
16. Perry J., and Burnfield M. J., 2010. Gait Analysis: Normal and Pathological Function. *Journal of Sports Science & Medicine*, 9(2), p. 353.
17. Saboor, A., Kask, T., Kuusik, A., Alam, M., Moullec, Y., Niazi, I., Zoha, A., Ahmad, R.., 2020. Latest Research Trends in Gait Analysis Using Wearable Sensors and Machine Learning: A Systematic Review. *IEEE Access*, 8, pp. 167830–167864.
18. Kesavulu, D., & Kannadasan, R., 2024. A Systematic Review and Applications of How AI Evolved in Healthcare. *Optical and Quantum Electronics*, 56(301), pp. 1–16.
19. Dastidar J., Samanta, S., Basu, A., & Purkait, S., 2023. Identification of Humans by Using Machine Learning Models on Gait Features. In: *Frontiers of ICT in Healthcare*. Edited by Mandal J. and Debashis De, Springer, pp. 137–149.
20. Sigman, D. M., 2016. Forensic Paint Analysis and Comparison Guidelines. [Online] Available at: https://ncfs.ucf.edu/research/chemical-evidence/paint/
21. California Department of Justice, n.d. Physical Evidence Bulletin. [Online] Available at: https://oag.ca.gov/sites/all/files/agweb/pdfs/cci/reference/peb_5.pdf
22. Kwofie, F., Perera, U. D. N., Allen, M. D., & Lavine, B. K., 2018. Transmission Infrared Imaging Microscopy and Multivariate Curve Resolution Applied to the Forensic Examination of Automotive Paints. *Talanta*, 186, pp. 662–669. https://doi.org/10.1016/j.talanta.2018.02.025
23. Lavine, B. K., White, C. G., Allen, M. D., & Weakley, A. 2017. Pattern Recognition-Assisted Infrared Library Searching of the Paint Data Query Database to Enhance Lead Information from Automotive Paint Trace Evidence. *Applied Spectroscopy*, 71(3), pp. 480–495. https://doi.org/10.1177/0003702816666287.
24. Ryland, S. G., Jergovich, T. A., & Kirkbride, K. P. (2006). Current Trends in Forensic Paint Examination. *Forensic Science Review*, 18(2), pp. 97–117.
25. Fitzapatrick, R. W., 2013. Soil: Forensic Analysis. In: *Wiley Encyclopedia of Forensic Science*. Edited by Jamieson A. and Moenssens AA. Wiley.
26. Fitzpatrick, R. W., 2008. Nature, Distribution and Origin of Soil Materials in the Forensic Comparison of Soils. In: *Soil Analysis in Forensic Taphonomy: Chemical and Biological Effects of Buried Human Remains*. Edited By Mark Tibbett, David O. Carter, CRC Press, pp. 1–28.
27. Vung, P. V., Weindorf, D. C., & Dang, T., 2021. Soil Profile Analysis Using Interactive Visualizations, Machine Learning, and Deep Learning. *Computers and Electronics in Agriculture*, 191(106539), pp. 1–9.
28. Taluja, C., & Thakur, R., 2018. Recent Trends of Machine Learning In Soil Classification: A Review. *International Journal of Computational Engineering Research (IJCER)*, 8(9), pp. 25–32.
29. Caddy, B., 2001. *Forensic Examination of Glass and Paint*. CRC Press.

30. Maxwell, V. M., 2001. Forensic Interpretation of Glass Evidence. *Journal of Forensic Identification*, 51(597), pp. 597.
31. Aitken, C., & Lucy, D., 2004. Evaluation of Trace Evidence in the Form of Multivariate Data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1), pp. 109–122.
32. Kaspi, O., Israelsohn-Azulay, O., Yigal, Z., Rosengarten, H., Krmpotić, M., Gouasmia, S., Bogdanović Radović, I., Jalkanen, P., Liski, A., Mizohata, K., Räisänen, J., Kasztovszky, Z., Harsányi, I., Acharya, R., Pujari, P. K., Mihály, M., Braun, M., Shabi, N., Girshevitz, O., & Senderowitz, H. 2023. Toward Developing Techniques—Agnostic Machine Learning Classification Models for Forensically Relevant Glass Fragments. *Journal of Chemical Information and Modeling*, 63(1), pp. 87–100. https://doi.org/10.1021/acs.jcim.2c01362
33. Almirall, J. R., 2003. *Glass Examination and Comparison with a Focus on Refractive Index Measurements, Elemental Analysis and Interpretation of Data*, A Workshop for Practicing Forensic Scientists: Denver, Colorado.
34. Newbury, D. E., & Ritchie, N. W. 2013. Is Scanning Electron Microscopy/Energy Dispersive X-ray Spectrometry (SEM/EDS) Quantitative? *Scanning*, 35(3), 141–168. https://doi.org/10.1002/sca.21041
35. Trejos, T., Castro, W., & Almirall, J. R., 2006. *Elemental Analysis of Glass and Paint Materials by Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS) for Forensic Application*. U.S Department of Justice.
36. Zadora, G., 2009. Classification of Glass Fragments Based on Elemental Composition and Refractive Index. *Journal of Forensic Sciences*, 54(1), pp. 49–59.
37. Corzo, R., Hoffman, T., Ernst, T., Trejos T., Berman T., Coulson, S., Weis, P., Stryjnik, A., Dorn, H., et al., 2021. An Interlaboratory Study Evaluating the Interpretation of Forensic Glass Evidence Using Refractive Index Measurements and Elemental Composition. *Forensic Chemistry*, 22(100307), pp. 1–25.
38. Carriquiry, A., Heike, H., Tai, X. H., & Vanderplas, S., 2019. Machine Learning in Forensic Applications. *Significance*, 16(2), pp. 29–35.
39. Rigano, C., 2019. *Using Artificial Intelligence to Address Criminal Justice Needs*. National Institute of Justice, Issue 280.
40. Thurzo, A., Kosnáčová, H. S., Kurilová, V., Kosmeľ, S., Beňuš, R., Moravanský, N., Kováč, P., Kuracinová, K. M., Palkovič, M., & Varga, I. 2021. Use of Advanced Artificial Intelligence in Forensic Medicine, Forensic Anthropology and Clinical Anatomy. *Healthcare (Basel, Switzerland)*, 9(11), p. 1545. https://doi.org/10.3390/healthcare9111545
41. Ahmed Alaa El-Din, E., 2022. Artificial Intelligence in Forensic Science: Invasion or Revolution?. *Egyptian Society of Clinical Toxicology Journal*, 10(2), pp. 20–32. doi: 10.21608/esctj.2022.158178.1012
42. Garcia A, 2023 AI in Forensic Investigation and Crime Detection- Glass as Forensic Evidence. [Online] Available at: https://medium.com/@blinx/ai-in-forensic-investigation-and-crime-detection-glass-as-forensic-evidence-5e2e730c8303

43. Livingston, M., 2020. Preventing Racial Bias in Federal AI. *Journal of Science Policy & Governance*, 16(2), pp. 1–7.
44. Christopher M., 2023. The Role of AI in Forensic Science. [Online] Available at: https://medium.com/predict/the-role-of-ai-in-forensic-science-
45. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* (New York, N.Y.), 366(6464), pp. 447–453. https://doi.org/10.1126/science.aax2342
46. Gupta S., Sharma V., Johri P., 2020. Artificial Intelligence in Forensic Science. *International Research Journal of Engineering and Technology (IRJET)*, 7(5), pp. 7181–7184.
47. Frąckiewicz, M., 2023. The Future of Forensic Science: How AI Is Transforming the Field. [Online] Available at: https://ts2.space/en/the-future-of-forensic-science-how-ai-is-transforming-the-field/#gsc.tab=0
48. Zaffir, M. a. B. M., Nuwantha, P., Arase, D., Sakurai, K., & Tamura, H. (2021). Comparison of Deep Neural Network Models and Effectiveness of EMG Signal Feature Value for Estimating Dorsiflexion. *Electronics*, 10(22), p. 2767. https://doi.org/10.3390/electronics10222767

Potential Applications of Machine Learning in Forensic Questioned Document Examination

3

SURBHI MATHUR, SUMIT KUMAR CHOUDHARY, PARVESH SHARMA, KRITIKA SOOD AND VINAY ASERI

Introduction

While writing started as a medium of communication, handwriting also evolved as an important biometric tool for individualization with the passage of time. And thus the concepts of forgery and document fraud were understood and defined in due course. Detection of forgery became a prevalent practice in most countries across the world from very early times. Handwriting identification and forgery detection dates back to 539 AD and as such it predates several other forms of forensic science disciplines by centuries. With more people practising and doing research, the discipline became more formalized and treatises were published for examination and identification parameters, protocols and benchmarks. In 1910, Albert S Osborn mentioned how the examination of handwriting had become more scientific and had slowly done away with empiricism. However, the science and practice of document examination still largely rested on the knowledge, skill and experience of the handwriting experts or forensic document examiners (FDE) [1].

Technology-led disruptive innovation has brought about significant transformation in all walks of life. The tale of technology has left humans speechless and coveting more technological innovations to achieve effortless operation of complicated tasks. The remarkable contribution of technology can be envisaged with the technological progress that we have made during the past century, which has led to substantial development for mankind. The introduction of 3D printers, robotic machinery to automate manual operations at factories and self-driven cars are some of the revolutionary examples of technologies developed by man which has enhanced the quality of human life and the processes they undertake.

DOI: 10.4324/9781003449164-4

The dawn of the creation of intelligent machines began with the introduction and boom of artificial intelligence which works on the model of manufacturing intelligent machines that can mimic human intelligence, thereby improving its overall efficiency and accuracy. A subset of artificial intelligence is 'machine learning'. This deals with enabling a system to make decisions based on previous interactions and patterns. Therefore, machine-learning algorithms work on training the system using various types of structured and unstructured data/datasets, to equip the system with the intelligence to solve problems and learn from the same. The applications of AI–ML-driven technologies extend to various sectors and domains worldwide. From AI-powered assistants that work on the models of natural language processing to the development of sophisticated machines for aiding medical assistance (detection and reporting of cancer cells), the implementation of AI–ML technologies has established itself as an indispensable technology of the future [2].

It has impacted the ways of crime commission and has also, at the same time, forced the investigators to adopt counter-technological tools to detect such crimes. The advent of artificial intelligence (AI) and machine learning (ML) are big capability boosters that have led to several innovations and promises to continue their dominance in almost all spheres of human intervention including forensics and questioned documents examination [3]. In the realm of forensic questioned document examination, where the delicate interplay between science and law seeks to unravel the mysteries embedded in handwritten and printed artifacts, the advent of machine learning presents a transformative paradigm shift. This chapter extensively addresses the fast-paced growth of machine learning applications in the forensic examination of questioned documents. This includes the immense possibilities that machine-learning techniques bring along with them in terms of improvement in the speed of examination, its precision, as well as reduced bias while examining the documents [4]. It fundamentally refers to a blend of technology, forensic science and law to enable the analysis of documents and their admissibility in the legal framework. The applications of machine learning will continue to improve and be finetuned with human intelligence and its increasing interference in augmenting artificial competences. While capability augmentation takes place and may appear to revolutionize policy and practices, it is important to understand and strike a balance between human intelligence, expertise and computing prowess for probing the forensic truth or facts under question [5].

Forensic document examination, until about a decade ago, primarily focused on issues related to the examination and identification of forgery, handwriting, signatures, altered documents, erasures, and the analysis of paper and ink. This field required a deep understanding of the physical and

chemical properties of paper, ink, toner, and handwriting characteristics to effectively differentiate between genuine and forged documents. Mechanical impressions such as typewritten documents, seals, stamps, etc., were another crucial class of documents examined in larger volumes through manual endeavours. However, with the increasing volumes of digital and physical documents, there is an urgent need for sophisticated tools that can handle the scale and complexity of forensic examinations [6, 7].

The science of machine learning is an artificial intelligence field that enables computers to learn from data patterns and make predictions or decisions without having to be given specific instructions [8]. Forensic document examination works in harmony with machine learning, poised to transform the field by augmenting traditional methods with advanced computational capabilities [9]. Machine learning trains forensic experts on algorithms and computational models as tools that can speed up document analysis, avoid vagueness and identify hidden details that could be missed by human eyes.

Machine Learning Fundamentals

Machine learning is a part of artificial intelligence where systems can learn and enhance their performance through experience without direct programming. The core idea of machine learning involves using algorithms and statistical models, allowing a system to execute tasks without the need for explicit programming, as mentioned by Biswas and Shreemoy in 2019.

While human expertise in identifying handwriting and document characteristics is crucial for distinguishing genuine documents from forgeries, machine learning algorithms can be trained to detect even the minutest details, map patterns, and notice inconspicuous anomalies and features that are beyond the manual limits of detection. This advancement leads to improved accuracy, efficiency, and reliability in forensic findings related to document analysis.

Applications of Machine Learning in Forensic Document Examination

Handwriting Analysis: The handwriting of every individual is unique and forms the basis of individual identification or fixing authorship. From pictorial appearance to class characteristics to individual characteristics, there are a lot of features which can be observed and compared between questioned and standard handwriting samples to arrive at a genuineness or authorship

conclusion. Machine-learning algorithms can also be trained on a large number of datasets of a variety of handwritings to pick up the forensic markers for the distinction of forged from genuine and identify the author of a given handwriting sample. The machine-learning capabilities can differentiate between natural variations and fundamental differences for identifying potential forgeries and can also differentiate one writer from another or identify one writer from among a group of writers based on important forensic markers of individuality in handwriting.

Authorship attribution with deep learning models and especially neural networks and recurrent neural networks (RNNs) presents unparalleled capabilities. Through the processing of extensive textual data these models can discern subtle nuances in writing styles, contributing to more accurate determinations of document authorship. The ability of deep-learning models to learn intricate patterns in the way individuals express themselves in text enhances their effectiveness in authorship attribution tasks, making them powerful tools in forensic document analysis.

Signature Verifications: Signature verification is another very important and voluminous forensic task in document examination. Signatures are examined for specific class and individual characteristics, natural variations and inherent signs of forgery to establish whether it is a genuine or forged signature. Machine-learning models can be trained on a large and diverse group of genuine and forged signatures to help them learn the specificities and distinct characteristics of either type so that they not only examine but bring about higher accuracy, speed and efficiency in forensic examination of questioned signatures.

Convolutional neural networks (CNNs) have demonstrated high effectiveness in signature verification tasks. When trained on datasets that include both authentic and forged signatures and these models can automatically identify and extract relevant features. This capability enhances the accuracy of distinguishing between genuine and fraudulent signatures. CNNs excel in capturing spatial hierarchies and patterns and making them well suited for tasks such as signature verification in forensic applications.

Text Analysis: Besides handwriting and signature examination sometimes the forensic examination of text in a document itself can reveal a lot of important facts about its originality, authorship or possible tampering. The pattern, writing habits, choice of words, misspellings, margin, start and finish, etc., can be helpful under the umbrella of examination of forensic stylistics of the text or content. Machine-learning algorithms, if trained for such examinations on extensive datasets, can be extremely helpful in figuring out the consistency or unusual departures from the norm.

Recurrent neural networks (RNNs) and their specialized variant and long short-term memory (LSTMs) networks and play a significant role in

textual content analysis. These models excel in analyzing the sequential nature of language and allowing them to identify patterns and anomalies that may indicate document tampering or forgery and as highlighted by the work of [10]. Their ability to capture dependencies in sequential data makes them valuable tools in forensic applications for uncovering linguistic patterns that may suggest alterations or inconsistencies in documents.

Ink and Paper Analysis: Several cases in questioned documents call for analysis of ink as well as writing surface. ML algorithms can be trained to tackle and examine problems pertaining to ink and paper such as their physical properties, age of documents, type of material used, etc. Answers to these questions would be crucial in determining forgery, tampering, ageing, etc. [11].

Forgery Detection Leveraging Deep-Learning Models

Deep-learning models are capable of learning and processing complex and difficult patterns, which can be crucial in detecting forgeries in documents encompassing handwriting characteristics, personal styles, ink properties or other document-related specificities which can be instrumental in reaching the conclusion of whether a particular document has been forged or tampered or not. Their ability to discern subtle details and patterns enhances the accuracy and efficiency of identifying fraudulent documents and making them valuable tools in forensic document examination.

Machine Learning Techniques for Signature Validation

One of the most frequent tasks in forensic document analysis is signature verification. Its goal is to ascertain whether a signature under suspicion corresponds with known signature samples. One way to look at the process from the perspective of automating it is as a machine-learning problem based on a population of signatures [12, 13].

Machine Learning tasks fall into two categories: Person-dependent (also known as special) learning and person-independent (also known as general) learning. A population of authentic and counterfeit signatures belonging to several persons provides general learning about the distinctions between authentic and counterfeit signatures for all individuals. Comparing a questioned signature to one authentic signature is possible using the general-learning model. Through several samples of just that person's signature, special learning is able to identify within-person similarities in a person's signature [14].

Validating the genuineness of signatures is the most common activity in the field of forensic document analysis The inquiry about the validity of a signature—does this questioned signature (Q) match the known, real signatures (K) of this subject is the one that comes before a document examiner the most frequently. When rendering a determination in forensic casework, a forensic document examiner, often referred to as a questioned document (QD) examiner, draws on years of experience analysing signatures [15]. This can be lengthy, challenging and prone to human bias. Thus, the task of automating the verification of signatures as a machine-learning problem makes sense. The ability of a program to learn from examples improves as the number of examples rises, this is considered its machine-learning capability. Determining the authenticity of a questioned signature is the performance task of signature verification [16].

The key is to distinguish between variations that are authentic and those that are counterfeit [17]. To put it simply, the learning challenge is to learn a two-class classification issue where the input is the difference between two signatures. Comparing the disputed signature to every known signature completes the verification operation. One way to conceptualize the general learning issue is as a learning process whereby near misses serve as counter-examples [18].

Learning from authentic examples of a specific individual is the main goal of special learning. Finding out how each genuine signature in the class differs from the others is the main goal. Determining whether or not the questioned signature belongs to that class is the verification job, which is effectively a one-class problem.

The available information on automated techniques for verifying signatures is sporadic. It's also important to consider automatic writer verification techniques, which include figuring out if a handwriting sample rather than a signature was indeed written by the specified person [19]. Identification is the process of figuring out who, among a particular group of people, could have written the material in question. The challenges of handwriting identification and verification are similar to those of biometric identification and verification, about which a substantial body of literature exists. Recently, there has been talk of using a machine-learning approach for biometrics.

Feature Extraction and Similarity Computational Approach

Since every human being develops distinct penmanship tendencies that serve as a representation of their signature, signatures are trusted for identification. Therefore, two algorithms, one for extracting characteristics and the other for comparing the similarities between two signatures based on

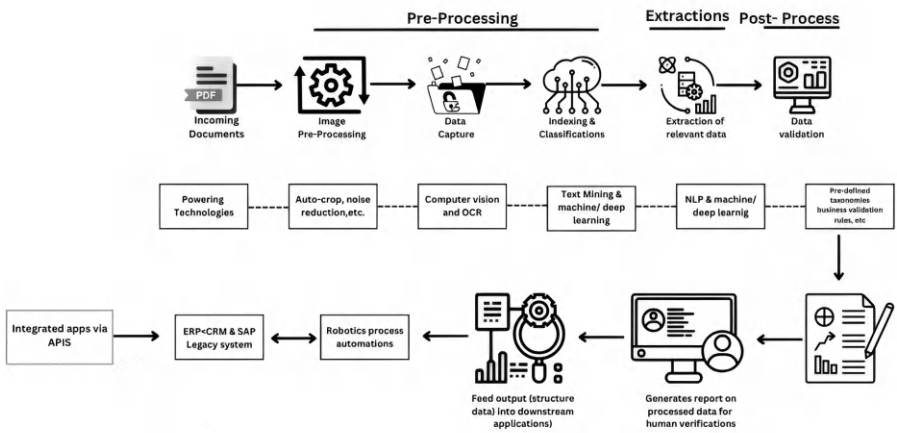

Figure 3.1 Machine-learning techniques helping QDE computational data management process.

those features, are at the core of every automatic signature verification system as brief in Figure 3.1. Features are the components that make something special. Such features are known as discriminating elements or elements of comparison in QD literature. The quantity of components in a specific person's samples may vary, and the combination of elements has a higher discriminating power [20].

Ticks, curve smoothness, pressure change smoothness, positioning, expansion and spacing, writing top and base, angulation/slant, overall pressure, pressure change patterns, macro forms, variations, connective forms and micro-forms are examples of these elements used by a human document examiner. Higher-level features such as rhythm, shape and balance are determined by basic characteristics such as speed, proportion, pressure and design.

The literature's descriptions of automatic signature verification techniques make use of a wholly distinct set of attributes. While some, such as wavelets, are focused on the texture of the picture, others concentrate on the geometry and topology of the signature image. Wavelet descriptors, projection distribution functions, extended shadow code and geometric features are some of the feature types employed for signature verifications [21].

GSC Features

Gradient structural and concavity (GSC) features assess an image's local scale characteristics; structural features measure an image's intermediate

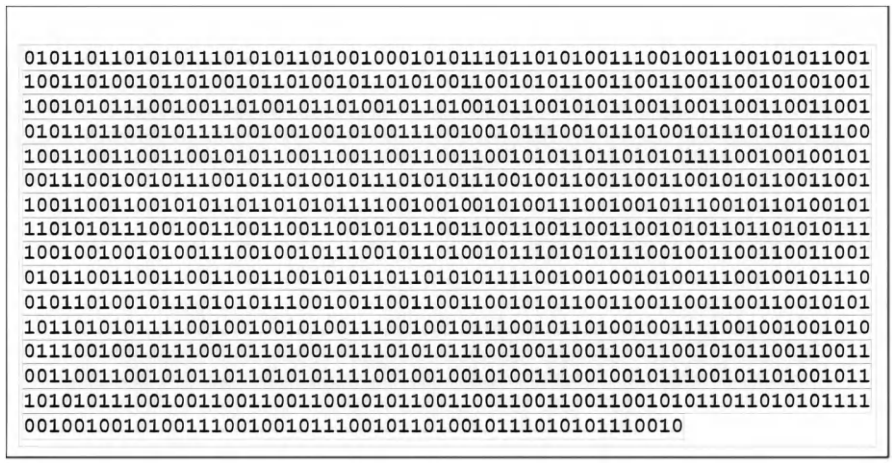

Figure 3.2 1024-bit binary features vector.

scale characteristics; and concavity can measure an image's properties throughout its whole scale. In line with this theory, three different feature maps are created, and each cell's corresponding local histograms are quantized into binary features. An example of a signature is displayed with a 4×8 grid superimposed on it to extract GSC features; the grid's rows and columns are created using the distributions of black pixels in the horizontal and vertical directions. Global word form features [22, 23] comprising 1024 bits that are produced by concatenating 384 gradient bits, 384 structural bits and 256 concavity bits have been recovered from these grids, as seen in Figure 3.2.

The strength of the match between two signatures is indicated by a score that is calculated using a similarity or distance metric. The paired data is transformed from feature space to distance space using the similarity measure.

Algorithm Formula

$$D(X,Y) = \frac{1}{2} - S_{11}S_{00} - S_{10}S_{01} / 2\sqrt{(S_{10}+S_{11})(S_{01}+S_{00})(S_{00}+S_{10})}$$

Below is a Python algorithm for signature grid method verification:

```
from sklearn import svm
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import accuracy_score
    from sklearn.preprocessing import StandardScaler
    # Assuming 'X' is your feature matrix and 'y' is the corresponding
labels (0 for genuine, 1 for forged)
```

```
    # X_train, X_test, y_train, y_test = train_test_split(X, y, test_
size=0.2, random_state=42)
    # Feature scaling
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)
    # SVM classifier
    clf = svm.SVC(kernel='linear')
    clf.fit(X_train_scaled, y_train)
    # Prediction
    y_pred = clf.predict(X_test_scaled)
    # Evaluation
    accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Global Research Advancing the Application of ML in QD Examination

Some of the better outcomes of leading research in the field of AI–ML convergence in QD analysis are listed in the following subsections.

Conventional Machine Learning–Based Classification

The incorporation of traditional ML models into forensics document examination has advanced significantly, particularly in the utilization of bit number models for classification purposes. The integration of bit number models within the conventional ML frameworks can offer advantages in terms of document classifications if the scientific intricacies behind this integration are properly understood [24, 25].

Support Vector Machines (SVMs) with Bit Number Models

In a binary classification SVM, the decision boundary is represented as a hyperplane: =0+11+2+2+ … +f(X)=β0+β1X1+β2X2+ … +βnXn.

The objective is to find the β coefficients that maximize the margin between classes.

Support vector machines (SVMs) are renowned for their ability to manage intricate data and can be improved by incorporating bit number models. In bit number models all features are represented in a binary format, where

each bit functions as a distinct feature. This methodology enables SVMs to effectively analyze and categorize documents by leveraging detailed bit-level patterns. This becomes particularly valuable in situations where a high-dimensional feature representation is necessary.

Decision Trees Utilizing Bit Number Models

Decision trees make decisions by splitting data based on features. The formula for predicting the target variable Y in a decision tree is based on a series of conditions:

$$Y = f(X1, X2,\ldots,Xn)$$

Each condition represents a split based on a specific feature, guiding the tree to a final prediction.

The decision trees have special abilities towards spontaneous classification of data and this can be further augmented by using the binary features of bit number models through integration of both. This added capability of employing binary features at the bit level makes the decision trees capable of accurately mapping and discriminating the patterns and characteristics within the questioned document at the time of examining it. Consequently, more precise and accurate results can be obtained through this integrated model of decision tree and bit number model [26].

Random Forests Enhanced by Bit Number Models

The unitary decision tree models can collectively generate the random forest model. As the word signifies, many trees together create a forest. In this case, prediction from each of the individual trees is also collated to create a combined prediction under the model of random forest. To perform the task of classification, the model considers the majority vote as shown below mathematically:

$$\text{mode } Y = \text{mode } \{Y1, Y2,\ldots, Yk\}$$

For regression, it could be an average.

As the decision tree model could be enhanced by integrating it with the bit number model, similarly, the random forests can also be enhanced by integrating it with bit number models for better learning and output. In such a situation, each decision tree component of a random forest model will facilitate processing the binary features at the bit level and later will be aggregated to give a collective insight. This integration, therefore, strengthens the random forest for dealing with and classifying complex, long and diverse data sets in suspected documents [27].

Naive Bayes with Bit Number Models: For the purpose of textual or document content analysis, naive Bayes classifiers can be used when aided by bit number models. This integrated model facilitates the representation of linguistic characteristics in the binary template and allows the naive Bayes model to recognize the linguistics or characteristic patterns at a granular level. This enables the collated model to detect forgery, tampering or authorship attribution.

K-Nearest Neighbours (KNN) with Bit Number Models: The task of document classification as well as mapping forensically important patterns and characteristics for addressing the problems pertaining to questioned documents examination can be effectively performed with the k-nearest neighbours (KNN) in combination with bit number models.

KNN performs this task by classifying the data points based on the majority class among their k-nearest neighbours. For a binary classification problem, the predicted class (Y) can be determined by the majority vote: = mode {1, 2,…,} Y = mode {Y1, Y2,…,Yk}

The use of bit-level representations empowers KNN to identify similarities in questioned documents based on discrete binary features. This approach facilitates efficient classifications [28].

Authorship Attribution with Bit Number Models: Disputed authorship is a common and voluminous problem in questioned documents. The machine-learning models, if trained on extensive datasets can be a valuable tool in the examination of handwriting or signatures and identification of the writer or author of the document as describe in Figure 3.3. This capability is enhanced further when combined with bit number models. The model can examine and map important characteristics and subtle patterns for reaching a conclusion regarding authorship [29].

Signature Verification Using Bit Number Models: Machine-learning models which can perform the task of signature verification are further enhanced, with the integration of bit number models. The binary representation processing of data enables to identification of forensically important characteristics through which genuine and forged signatures can be differentiated.

Textual Content Analysis with Bit Number Models: Machine-learning models are capable of performing signature verification tasks in the forensic arena. This capability is further augmented by bit number models. The important characteristic features are more accurately mapped and thus authentication of the signature of handwriting becomes easier and more acceptable. Similarly, textual analysis can also be performed with enhanced accuracy and pace.

Forgery Detection Leveraging Bit Number Models: Machine-learning algorithms have proved to be very skilled and automated systems for

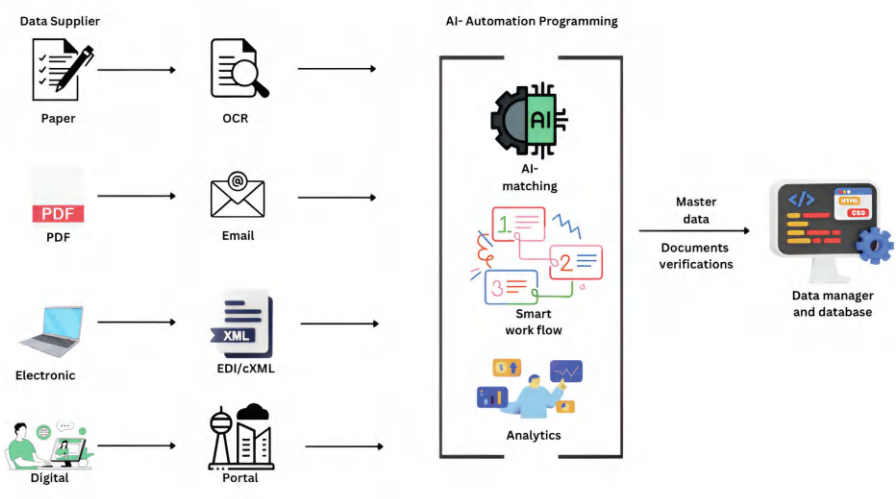

Figure 3.3 Extraction modelling and AI matching analytics for data manager.

decision-making in a variety of setups. Its efficiency and precision have been found to increase remarkably when it is integrated with bit number representations. ML models in combination with bit number representation have been very effective in categorically capturing the minutest details, handwriting characteristics, paper and ink characteristics, writing stylistics, etc, at the bit level, which makes it an excellent tool for the examination of questioned documents. It can be helpful in detecting forgeries, counterfeits, manipulations and tamperings, and thus can differentiate a genuine from a forged document.

Deep Learning–Based Classification

Deep-learning models have evolved tremendously over time owing to their encompassing neural networks and convolutional architectures which allow them to make data-driven decision-making. This model has seeped into almost all walks of life slowly and is having a profound impact including in the field of forensic science and forensic document examination as well. A deep-learning approach can enhance the efficiency as well as accuracy in the examination of suspected documents in addressing diverse kinds of problems related to questioned documents [30].

Neural Networks for Document Classification: Neural networks are the basic architectural units of deep-learning models. It can be in multiple layers, which are interconnected and through which it can learn simple

things to complex issues. The learning ability potential of neural networks is harnessed in document examination as well for a variety of applications. This can be helpful in authorship attribution, signature verification and detection of forgeries in questioned documents [30].

Convolutional Neural Networks (CNN) in Forensic Document Examination: The convolutional neural networks (CNNs) model has been extensively applied for image recognition tasks worldwide. The CNN is capable of independently learning and extracting peculiar patterns, which can be utilized for classification tasks. This quality can be usefully harnessed in the field of document examination by capturing the specific characteristics or forensic markers in a suspected document. The handwriting characteristics can be captured and used for signature verification or authentication.

Recurrent Neural Networks (RNN) for Content Analysis of Documents: Recurrent neural networks (RNNs) especially are capable of performing tasks where data is available in a sequential manner or in an order. It can perform textual or document content analysis in an effective manner. This capability of RNN models can be successfully exploited by forensic document examiners (FDEs) in examining and classifying the stylistics or linguistics pattern in a suspected document presented for examination. The analysis of orderliness or sequence of data or content can also enable an FDE to detect any potential alteration in the said document. Also, the stylistics or linguistic pattern can be individualized to the extent that authorship identification can be done.

Long Short-Term Memory (LSTM) Networks

Long short-term memory networks (LSTMs) are a specific type of recurrent neural networks (RNNs) which are especially capable of capturing and processing long-range dependencies in orderly data or data that are in a sequence. Using this capability, LSTMs can be extensively used to classify and analyze longer text data.

Advantages of the Approaches

AI–ML technologies, as they make headways in the field of forensics and scientific analysis of crime clue materials, there is a perceptible change in the effectiveness of analysis of such forensic exhibits received by a crime laboratory. Contemporary forensic tools/software used by forensic experts are based on AI–ML and help contribute to the timely investigation and analysis of evidence. The use of artificial intelligence learning algorithms in a

real-time investigation of video feeds gathered from closed circuit television cameras (CCTV) serves as an important acknowledgement of the enormous potential of AI–ML tools in forensic science. The AI–ML-based forensic tools for video investigation are endowed with the ability to classify and evaluate the video feeds/CCTV footage that are submitted for examination, thereby enabling the forensic expert to timely review a significant amount of video footage and investigate crucial elements. The attribute of the AI–ML forensic tools to recognize patterns is deployed for analyzing and detecting patterns in audio/texts/emails which have helped investigators deal with large complex data which are referred for examination. Considering the commendable mathematical and computational power of AI–ML-based forensic tools, investigators are now able to develop and build strong statistical evidence for augmenting the results drawn during the investigation. The above examples of the use of AI–ML-powered technologies in forensic science reinstate the promising capabilities of artificial intelligence [24].

The examination of questioned document evidence requires the forensic examiner to adopt scientific methods for examination, which are non-destructive in nature. There are various approaches that are employed for investigating different documentary evidence, the modern methods of examination involve the use of AI–ML-powered tools. The use of AI–ML tools has proven to be beneficial, particularly with regard to forensic document examination, as these tools are equipped with the proficiency to do more in less time. Due to the high processing power of the AI–ML tools, it is now easier to classify, collect and evaluate large sets of documentary data for analysis [25, 26].

The Supreme Court Portal for Assistance in Court's Efficiency (SUPACE) is an AI operational portal that exemplifies the need for AI–ML for the collection and evaluation of large documents. The AI portal was unveiled by the Chief Justice of India, S A Bobde, with the objective of processing large amounts of data that are submitted with regard to an individual case. The AI portal functions by collecting and filtering relevant facts from the documents pertaining to a case which can be used by judges for input and decision making. With the introduction of SUPACE, it is now easier for judges to conduct comprehensive research applicable to the case, in a short period of time allowing the judge to critically review all the aspects of the case before drafting the final opinion.

Smart document analysis can be achieved via AI–ML tools which can use AI-powered machine-learning algorithms to classify the contents of the document and produce structured data that can be further assessed for examination. The branch of forensic stylistics is associated with the application of science of language, in determining the authorship of the document in question. The document analysis done via AI–ML tools is capable

of determining the authorship of the document using machine learning that understands the unique stylistics pattern implemented by the author in making the document [25]

Machine-learning tools are also employed by leaders of the digital industry to flag emails and documents that might have restricted content with the potential to jeopardize the security of the organization. Emails and documents in cyberspace can now be filtered with regard to their content and can be flagged with the help of AI–ML-trained tools thereby playing a pivotal role in safeguarding the interests of the organization.

Sentimental analysis which is characterized by the potentiality to determine the emotional sentiment/tone with regard to the content of a document works on natural language processing which is again a subset of artificial intelligence. It is popularly carried out by forensic document examiners in the investigation of suicide notes, blogs, emails, etc., and also to profile the author of the document. The use of AI has aided in the categorization and profiling of documentary evidence based on sentimental analysis, which has helped document examiners to probe deeper to extract more evidence from the document in question [26].

AI algorithms are also used in the comparison of handwriting samples to determine the authorship of the handwriting found in the questioned document. The AI tool compares and analyzes the standards/admitted handwriting samples submitted with respect to the case of the handwriting present in the document in question. The standards/admitted writing serves as a dataset for the AI tool and trains the program to get acquainted with the master pattern of the writer. The same phenomenon is used to analyze and compare signatures that are submitted for examination [26].

With the exponential rise in AI technologies and their applications, the popularity of AI tools in forensic investigation is heading to linear growth. The AI approach has given birth to a transformational change in the way forensic investigations are conducted today, advocating its proficiency in carrying out forensic examinations of various pieces of evidence submitted for analysis.

Disadvantages of the Approaches

As forensic tools are now equipped with the ability to learn, adapt, reason and self-correct with time, the use of machine intelligence or artificial intelligence has simplified the assessment and analysis of forensic evidence pertaining to the investigation of a case. The AI approach in FDE has overall been beneficial in the investigation of documentary evidence for forensic investigators, but issues regarding its credibility and use still remain a lingering question

that remains to be addressed. The rate of accuracy of the AI tools is a very crucial factor in determining the success of the tool. It has been observed that in cases where large amounts of data are required to process and analyze, the AI tool fails to be accurate and may result in false positive or negative results. The lack of accuracy in investigating enormous documentary evidence may lead to ineffective investigation of all the facts of a case and ultimately result in compromising the evidentiary value of the same [27].

The AI tools are trained over time on large sets of data to identify patterns, make intelligent decisions and predict outcomes. Its efficiency and utility depend on the training process it is subjected to, which facilitates the AI–ML system to produce an output. The AI–ML tool has a limited understanding of the situation presented to it. It lacks the ability to deduce or comprehend the context independently outside the scope of its learning [28].

The handwriting of an individual is highly unique and is characterized by the presence of natural variations. Natural variations can be defined as the innate quality of a writer which is associated with the inability to execute the same manner of handwriting operation each and every time, giving birth to a range of variables in the handwriting. External factors such as intoxication, body posture, health, age, etc., also affect the degree of variation found in handwriting. While examining the handwriting samples the forensic examiner needs to assess the document keeping in mind all the possible variations that the writer may execute while writing. The AI–ML forensic tools rely on the dataset which has been fed to them for their programming, therefore lacking the normal ability to detect variations in handwriting which have been affected by extrinsic factors or disguise. Also, it is important to be conscious of the fact that the AI–ML software lacks the common sense of reasoning which becomes vital in decision-making and opinion framing [28].

AI–ML tools are also devoid of the capabilities to understand human emotions; all documents are logically analyzed and processed by the tool, therefore while examining documentary evidence and summarizing all the facts pertaining to a case, the tool fails to consider emotional angles of the case which might be relevant to acknowledge while furnishing a final opinion. The AI tool would remain as a supplement in such instances and would not conceive the same result and accuracy that Human Intelligence would serve [28, 29]

Conclusion and Future Scope

Document verification involves the process of authenticating a document or documents to ensure that they are accurate, genuine and fit for their intended purpose. Every year, instances of identity theft are reported in which personal

information and documents were used to open bank accounts, obtain loans, apply for debit or credit cards, etc. In order to fulfil legal requirements, criminals present forged documents that fool the verifier into believing they are legitimate by seeming extremely precise and correct. The creation of AI–ML-based systems in future that can automate the document verification process might be very beneficial in spotting fake documents submitted as legal proofs and reducing the pervasive issue of identity fraud.

Technology has given us the ability to remain anonymous, but criminals abuse this ability by using the internet to hide their tracks making it difficult for law enforcement to track them down. Nowadays, criminals try to hide their identities by sending threatening emails, messages and posts from anonymous online IDs in order to avoid disclosing their handwriting, which can be used as hard-core evidence in establishing their identities. The development of machine-learning tools that can use scientific parameters in assessing the language structure and formation from threatening emails, messages and posts can result in the successful identification of the suspect. The introduction and implementation of such tools can solve the challenging problem of decoding the identity of an individual, from the content sent anonymously. The aforementioned examples provide compelling evidence for the tremendous potential of AI–ML techniques in forensic documentary analysis and emphasize the need for the development of more tools with similar functionality in order to revolutionize current procedures and techniques for document analysis.

The revolutionary technologies backed by machine-learning applications guarantee their ticket to the future of limitless possibilities. The use of AI–ML in the field of forensic investigations would result in expediting the process of investigation as the employment of AI–ML would automate the investigation of tedious and challenging forensic evidence submitted for examination. Currently, forensic science laboratories are burdened with a large number of cases, which has resulted in slowing the process of justice delivery. The implementation of cutting-edge tools that are powered by machine learning can come to the rescue.

The process of FDE can be challenging because the examiner has to thoroughly study and dissect the handwriting/signature/discourse/language style, etc., of the author to reach an opinion. With the propitious development of AI–ML technologies in the near future, the tool can serve as an intelligent means to understand the variations that occur due to external factors or intentional disguises, with high accuracy to opine on several documents that are submitted for authorship analysis.

Chatbots also known as chatterbots have gained immense popularity recently. These can be defined as programs that utilize artificial intelligence and natural language processing to mimic human conversations and are

used to generate responses based on the queries or interactions initiated by the user. Advancements in chatbots are seen with the use of natural language processing understanding (NLPU) which is based on machine learning and deep learning to effectively respond to and accomplish the request of the user. According to a research study conducted by Nadia Zlate, AI chatbots can serve as future undercover investigators that can uncover potential criminals and their crimes by entering into a conversation with them. The AI chatbot can pose as a real human and can record the details of the crime narrated by the criminal thereby gathering necessary evidence for the same. The written texts to the chatbot can also serve for profiling the criminal with the help of forensic stylistics or the science of language.

References

1. McInnes L., Healy J., Melville J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29), 861.
2. Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., Fei-Fei L. (2015). Imagenet large scale visual recognition challenge, *International Journal of Computer Vision,* 115, 211–252.
3. Srinivasan H., Srihari S. N., Beal M. J. (2006). Machine learning for signature verification. *Computer vision, graphics and image processing: 5th Indian conference, ICVGIP 2006, Proceedings* (pp. 761–775). Springer.
4. Neupane S., Pyakurel M., Sinha K., Sharma B. A. (2024). GraphoMatch: Forensic handwriting analysis using machine learning. *International Journal of Science and Research Archive*, 11(2), 1526–1537.
5. Tan M., Le Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning* (pp. 6105–6114). PMLR.
6. Souza V. L. F., Oliveira A. L. I., Cruz R. M. O., Sabourin, R. (2020). An investigation of feature selection and transfer learning for writer-independent offline handwritten signature verification. In *Proceedings of the international conference on pattern recognition* (pp. 1475–1482). IEEE.
7. Slyter S. A. (1995). *Forensic signature examination*. Charles C. Thomas Publisher.
8. Sauvola J., Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition,* 33(2), 225–236.
9. Gupta D., Bag, S. (2020). A local-to-global approach for document image binarization. In Das, A., Nayak, J., Naik, B., Pati, S., Pelusi, D. (Eds.), *Computational intelligence in pattern recognition. Advances in intelligent systems and computing, vol 999.* Springer, Singapore. https://doi.org/10.1007/978-981-13-9042-5_60

10. Sehad A., Chibani Y., Hedjam R., Cheriet M. (2019). Gabor filter-based texture for ancient degraded document image binarization. *Pattern Analysis and Applications*, 22(1), 1–22.
11. European Network of Forensic Science Institutes. (2023). *The European Network of Forensic Science Institutes (ENFSI)*. Retrieved November 21, 2023, https://www.enfsi.eu/about-enfsi
12. Shobha R. N., Nair B. J. B., Chandrajith M., Hemantha Kumar G., Fortuny J. (2022). Restoration of deteriorated text sections in ancient document images using a tri-level semi-adaptive thresholding technique. *Automatika,* 63(2), 378–398.
13. Shobha R. N., Manohar N., Hariprasad M., Pushpa B. R. (2022). Robust recognition technique for handwritten Kannada character recognition using capsule networks. *International Journal of Electrical and Computer Engineering,* 12(1), 383–391.
14. Bird J. J. (2022). Robotic and generative adversarial attacks in offline writer-independent signature verification. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2204.07246
15. Chauhan M., et al. (2025). Vision-language model based handwriting verification. *IET Conference Proceedings*, 2024(10). https://doi.org/10.1049/icp.2024.3329
16. Drotár P., et al. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. *Artificial Intelligence in Medicine,* 67, 39–46.
17. Hazra A., Maity S., Pal B., Bandyopadhyay A. (2024). Adversarial attacks in signature verification: A deep learning approach. *Computer Science and Information Technologies,* 5(3), 215–226.
18. Zhang H., Guo J., Li K., Zhang Y. (2024). Offline signature verification based on feature disentangling aided variational autoencoder. https://doi.org/10.48550/arXiv.2409.19754
19. Arabio A. (2024). Quantifying writer variance through rainbow triangle graph decomposition. *Center for statistics and applications in forensic evidence*. https://dr.lib.iastate.edu/handle/20.500.12876/ywAbZ7Wv
20. Baek Y., Lee B., Han D., Yun S., Lee H. (2019). Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9365–9374). IEEE.
21. Jagtap A. B., Hegadi R. S., Santosh K. C. (2019). Feature learning for offline handwritten signature verification using convolutional neural network. *International Journal of Technology and Human Interaction*, 15(4), 54–62.
22. Breci E., Guarnera L., Battiato S. (2024). Innovative methods for non-destructive inspection of handwritten documents. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 4825–4829. https://doi.org/10.1109/ICASSP48485.2024.10448383
23. Marcinowski M. (2023). Evaluation of neural networks applied in forensics; handwriting verification example. *Australian Journal of Forensic Sciences,* 55(6), 745–754.

24. Adedayo O. M., Olivier M. S. (2025). Examination of customized questioned digital documents. *Journal of Forensic Sciences*, 70(2), 550–565.
25. Tageldin L., Venter H. (2023). Machine-learning forensics: State of the art in the use of machine-learning techniques for digital forensic investigations within smart environments. *Applied Sciences,* 13(18), 10169.
26. Srihari S. N., Cha S. H., Arora H., Lee, S. (2002). Individuality of handwriting. *Journal of Forensic Sciences*, 47(4), 856–872.
27. Bulacu M., Schomaker L. (2007). Text-independent writer identification and verification using textural and allographic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4), 701–717.
28. Found B., Rogers D. (2005). The probative character of handwriting in forensic document examination. *Science & Justice*, 45(2), 65–73.
29. Malik M. I., Liwicki M., Dengel A. (2013). Writer identification for historical documents using convolutional neural networks. *In 2013 12th International conference on document analysis and recognition* (pp. 1387–1391).
30. Farrahi Moghaddam R., Cheriet M. (2010). A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recognition,* 43(6), 2186–2198.

Application of Machine Learning in the Field of Forensic Medicine

4

NIHA ANSARI, VAISHALI,
DIVYANT KATARIA AND
YASASVIKUMAR VALA

Introduction

The field of forensic science, which is located on the border between scientific study and the legal profession, holds a major function in crime solving and punishment, ensuring a fair outcome for victims [1]. It includes many scientific fields, using various systematic methods to analyze evidence in relation to legal inquiries and cases. Forensic science consists of the major tasks of collecting, analyzing and comparing physical evidence in order to provide factual data or present expert opinion in legal matters. It includes several issues, such as the evaluation of autopsy data, finding a suspect and supporting an objective legal system. Initially, forensic science was not well developed, but over the years it has developed into a complex and branched field that is critical in today's criminal investigations as well as judicial systems.

The roots of forensic medicine can be traced back to the early ages when postmortem examinations and other related methods were used to determine the cause of death and to punish the guilty. Leading luminaries such as Rudolf Virchow and Sir Bernard Spilsbury historically advanced the development of this field. Postmortem examinations, commonly referred to as autopsies, are vital for explaining suspicious deaths and helping forensic pathologists during their investigations [2]. Specialist doctors carry out these tests called forensic pathologists to identify the reason and mechanism of death. [3]

Forensic medicine, a part of the medical profession, evaluates and determines medical realities within criminal and civil law matters [4]. Medical jurisprudence is derived from the Latin word 'jurisprudential' in which the word 'juris' refers to 'law' and the word 'prudence' denotes 'knowledge', meaning 'the knowledge or discipline of law', frequently defined as applying medical knowledge to legal concerns, covers integrating medical facts into legal matters [5, 6]. The topic of forensic medicine encompasses various

DOI: 10.4324/9781003449164-5

subfields. Forensic pathology involves examining deceased individuals to assess injuries and discover reasons for death, such as asphyxiation. Forensic psychiatry investigates the legal aspects of mental diseases. Forensic odontology uses dental knowledge for actions such as identifying and evaluating bite marks and employing specific dental traits for identification reasons [5]. Forensic anthropology investigates body traits and skeletal structures, generally to identify persons in court circumstances [7]. Forensic anthropologists can analyze skeletal remains to ascertain features such as age, sex, ancestral lineage and height, vital in identifying unidentified humans. Forensic medicine, commonly known as forensic pathology, is crucial in the vast field of forensic scientific investigations. Its major objective is to uncover the puzzles surrounding the reason and method of death, offering vital insights required for legal actions [1]. Forensic medicine is crucial to criminal investigations by supplying evidence and expert views admissible in court. Furthermore, it entails examining real persons to detect injuries, identify chemicals or assess mental states, underscoring its relevance in legal sectors.

Artificial intelligence (AI) is a basic technical innovation that enables automated robots to accomplish jobs that traditionally need human intelligence. In the sphere of forensic medicine, machine learning holds significant potential, altering how forensic experts research and solve challenging situations. The application of machine learning is a creative drift in the field of forensics, especially medicine, and a likely breaking point for the whole forensic field [8]. Experts in conventional forensic identification thoroughly gather data and provide identification opinions by integrating their professional experience with information from fundamental disciplines such as biology and medicine [8]. This method is both labour- and time-intensive, and it is impacted by unpredictable components that are difficult to regulate. Artificial intelligence (AI) holds the capability of transforming the field of forensic medicine by improving and automating the identification process, which will reduce the level of human intervention and the probability of introducing bias. These technologies might help to increase the accuracy and efficiency of forensic identification methods since large quantities of data can be objectively and quickly analyzed. The area of forensic medicine encompasses a large number of activities, such as postmortem examinations and medical record reviews. Also, it can assist forensic experts with activities such as dental profiling, face recognition and forensic photography [9]. Moreover, machine learning has a strong application in forensic pathology, where it can help in the very accurate identification of disease or injury by using photographs taken during an autopsy. The AI tools can help forensic pathologists diagnose diseases, compute tomography and identify the cause of death more accurately with the help of tissue samples and histopathological

images [10]. Machine learning can be defined as a subfield of AI that empowers machines to learn from data and make decisions or draw conclusions based on that data [11].

Significance of Forensic Medicine

Forensic medicine, sometimes known as forensic pathology, plays a vital role in several domains of society. Some important characteristics of its relevance are as follows:

1. To help in the quest for justice, forensic medicine delivers essential scientific evidence for use in court cases and legal inquiries.
2. Autopsy has been utilized as an essential aspect of forensic medicine. Using procedures such as DNA analysis, dental records and fingerprinting, aids in identifying deceased people in conditions of mass catastrophes, accidents or crimes against humanity [12].
3. Forensic medicine helps to prevent crimes by evaluating patterns and trends in injuries and fatalities, together with other forensic data.
4. In cases in which death has been precipitated by a weapon, such as a firearm, the wound could provide crucial information to the forensic pathologist. During the investigation, a forensic pathologist may often detect not merely the sort of weapon used but also provide necessary information. For example, in the context of a gunshot wound, they can reasonably identify the range and angle at which the firearm was fired [13].
5. A forensic pathologist's principal role is to employ their medical knowledge to address legal situations regarding death. They accomplish this by analyzing the corpse and obtaining information to discover how and why someone died. The objective is to give exact and trustworthy data that may be employed in court to identify the cause of death and bring justice to the case [14].
6. To identify the situations surrounding a death, forensic pathologists analyze crime scenes to assess the body's position, the existence of any injuries and other significant details.
7. The results of forensic pathology can be exploited in criminal cases by the prosecution as well as the defence to buttress their claims and give light on the circumstances surrounding a death.
8. It can provide speedy and exact proof that may be applied to determine suspects to establish or disprove their innocence or guilt.

Problems Faced during Forensic Examination

1. One of the most problematic issues includes establishing a clear and conclusive causal relationship between the trauma felt by the victim and the final cause of death [15]. For example, if the victim suffered several injuries, it might be difficult to identify which particular injury or combination of injuries caused the person's death. The kind and degree of the injuries, the individual's overall condition and past medical history and the circumstances surrounding the trauma all need to be thoroughly evaluated and weighed.
2. Decomposed bodies offer challenges for forensic pathologists, as decomposition may disguise injuries and make it challenging to identify the reason for death due to the loss of physical features along with evidence [16]. For instance, in a case when a person is found weeks after death in a sweltering, humid area, decomposition might be severe, rendering it hard to detect injuries or distinguish if they are antemortem or postmortem injuries.
3. Baldino et al. discussed the problems experienced by forensic pathologists when detecting dramatically transformed cadavers and diagnosing the cause of death in such circumstances. They underline the relevance of integrating standard procedures with particular forensic branches to strengthen the investigative process [17].
4. External factors outside the body could alter its physiological condition and the accuracy of the forensic analysis. Environmental factors, such as temperature, humidity and element exposure, can all speed up the breakdown process.
5. Some of the postmortem changes, including rigor mortis, livor mortis and decomposition, affect the state of the body and its capacity to determine the cause and manner of death.
6. Evaluating traumatic injuries, particularly in cases of blunt trauma, could be challenging. Separating a case that resulted from a fall, a car accident or an assault requires assessment and knowledge of the situation.
7. Evaluating gunshot injuries may not be easy, as the type of weapon, distance and angle of shooting have to be considered. For example, when a person is shot dead identifying the type of firearm and the range from which the shot was fired requires the help of a professional and may involve an analysis of the gunshot residue and the trajectory of the projectile.
8. Documentation and reporting are critical aspects of forensic pathology; however, they can be challenging since the work requires a

high level of accuracy and adherence to legal requirements [18]. For instance, it is critical to record injuries and findings in a simple language that is legal and could include using technical terms and illustrations.

9. At other times, there will be a lack of tangible evidence left for examination, and this hampers the ability to determine accurate conclusions on the cause and manner of death [18].
10. A lack of detailed medical history or documents for the deceased would limit the pathologist's ability to accurately assess prior ailments or drugs that may have contributed to the death.

Machine Learning in Forensic Medicine

In the domain of legal and medical investigation, forensic medicine is a major subject that focuses on the complex relationship between justice, law and medicine [16]. It has several sub-branches, and each of them has a specific responsibility in the process of finding the cause of death in cases of unexpected or unexplained deaths. Forensic pathology, for instance, is a crucial sub-branch that acts as the foundation of any death investigation by determining the manner, circumstances and cause of death [16, 19].

Traditionally, forensic pathology was focused on manual examination and interpretation of evidence such as autopsy results, medical records and toxicological tests [20]. These procedures were of the utmost significance, but human subjectivity, time restrictions and the large quantity of data involved hampered them. However, the development of machine learning (ML) has launched a new age in forensic pathology, presenting significant chances to increase the efficiency, accuracy and objectivity of death investigations. In forensic medicine, machine-learning algorithms can evaluate huge quantities of forensic data, identify hidden linkages and extract useful insights that individuals may fail to detect [21]. The combination of cutting-edge technology and forensic experience has significant potential to revolutionize forensic pathology and transform death investigations.

Fundamental Operating Mechanisms of Machine Learning in Forensic Medicine

Machine learning is a sort of artificial intelligence that focuses on creating mathematical models and algorithms that permit computers to acquire knowledge from data and make predictions or judgements without direct programming [22]. The basic operational principles of machine learning comprise numerous key steps:

- **Data Collection:** The primary stage in machine learning is to obtain appropriate data to train the algorithm. This data may comprise input–output pairs (supervised learning), unlabelled data (unsupervised learning) or a combination of the two (semi-supervised learning). For example, in forensic pathology, data gathering may entail acquiring autopsy reports, medical records, toxicology reports and other significant information about a death inquiry.
- **Data Preprocessing:** After the data has been obtained, it needs to be preprocessed to ensure it is in an acceptable format for analysis. Preprocessing may include fine-tuning medical images to a standard resolution, reducing artifacts and enhancing image quality.
- **Feature Extraction:** It is the process of obtaining usable qualities or properties from treated data. These qualities serve as input for the machine-learning algorithm. For example, in medical picture analysis, features include the size and form of structures within the body, the vibrancy of pixel values and the occurrence of certain patterns or textures.
- **Model Selection:** The third step is to choose an acceptable machine-learning model that fits the task under concern. Common machine-learning models include decision trees, support vector machines, neural networks and ensemble techniques. Andrej Thurzo et al. applied a 3D convolutional neural networks (CNN) model in their forensic investigation in the fields of age, sex, face and development determination [20].

Machine Learning in Forensic Pathology

Forensic pathology is a critical subject within the criminal justice system, giving vital information to help solve crimes and provide justice to victims. Over the years, the inclusion of machine-learning methods in forensic pathology has achieved promising results illustrating its ability to transform the ways forensic evidence examinations are conducted and analyzed. Due to the availability of higher order and efficient algorithms based on computational theory, machine-learning algorithms can potentially help forensic pathologists with some of these tasks as identifying patterns in large data, identifying the cause of death or even predicting other likely suspects based on the evidence. The application of technological advances in forensic pathology presents a lot of potential for enhancing the accuracy and effectiveness of investigations and consequently expediting precise solutions to criminal cases [23].

The concepts of machine learning and artificial intelligence are transforming patient management and transplant medicine in liver transplantation. Researchers are employing the application of artificial neural networks such as deep neural networks (DNNs) to enhance the accuracy of the prognosis of patient and organ survival and ascertain risk determinants that affect organ transplantation. These sophisticated machine-learning algorithms process large datasets that are characteristic of chronic diseases, assess transplant risks, distribute organs and manage patients after surgery. Different types of ML algorithms including extreme gradient descent boosting and logistic regression with least absolute shrinkage and selection operator (LASSO) are employed to give prognosis for long-term illness patients such as liver disease. These models employ many predictor factors such as demography, clinical history and laboratory results to enhance the projections and consequently the treatment of patients in the liver transplantation (LT) setting. The research focuses on the issues of model interpretability and trust in AI systems and proposes explainable AI frameworks such as Shapley values to help explain which features are being used for predictions. In the highly specialized area of care that is long-term medicine, AI, in the form of machine-learning models, has emerged as a critical enabler to help hasten transplant eligibility assessment, donor/recipient matching and post-transplant patient management. Using multiple data inputs and extracting the most important predictors, ML algorithms offer valuable information on the patient's prognosis and contribute to the personalized clinical management of LT. Although the work acknowledges the positive impact of AI and ML in LT treatments, the study also notes some of the challenges that will have to be addressed in the future such as regulation, model interpretability and efficacy in practice as illustrated in Figures 4.1 and 4.2 [24].

Significance of Forensic Pathology

- **Accurate Cause of Death Determination:** Forensic pathology is vital for accurately diagnosing the cause of death in circumstances of sudden, unexpected or suspicious fatalities. We can use machine learning to accurately forecast the cause of death based on the patient's most recent medical examination. By conducting postmortem examinations and examining tissue samples, forensic pathologists can uncover underlying disorders, injuries or toxicological causes contributing to death. Machine-learning algorithms can assist in the interpretation of complex pathological findings, helping forensic pathologists make more accurate diagnoses and rulings regarding the cause of death along with an accurate prediction for the same [25–27].

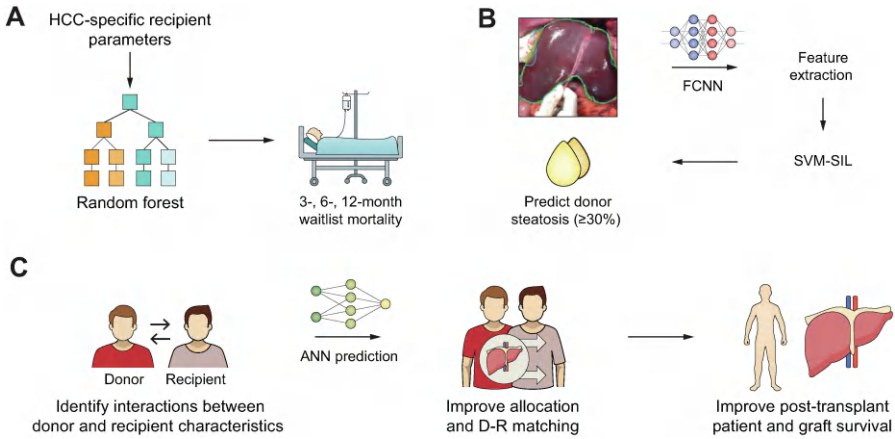

Figure 4.1 ML applications in the pre-transplant setting. (A) Using a random forest model, the risk of 3-, 6- and 12-month waitlist mortality was predicted to better prioritize HCC candidates for liver transplantation. (B) Determine organ quality using smartphone images using a combination of FCNN and SVM approaches. (C) Improving donor pathology assessment using CNNs to identify steatosis could outperform pathologists. Identifying best donor-recipient matches by uncovering hidden nonlinear relationships between demographic, clinical and laboratory data, leading to improved organ allocation and optimized transplant outcomes. FCNN, fully convolutional neural network; SVM, support vector machine [24].

- **Enhanced Forensic Investigations:** Machine-learning technologies in forensic pathology have the potential to enhance the efficiency and effectiveness of forensic investigations. By automating certain aspects of postmortem analysis, such as tissue classification, organ measurements and histopathological interpretation, machine-learning algorithms can expedite the forensic process and provide forensic pathologists with valuable insights to aid in their assessments [28–30].
- **Improved Disease Diagnosis:** Forensic pathology contributes to the diagnosis and identification of numerous diseases and medical conditions that may have led to an individual's death. Machine-learning techniques applied to medical imaging, such as computed tomography (CT) scans, as shown in Figure 4.3 [31], which is for the diagnosis of COVID-19 based on CT scans, and magnetic resonance imaging (MRI), can assist forensic pathologists in the detection and characterization of pathological conditions, enabling more accurate disease diagnoses and forensic assessments [29, 31, 32].

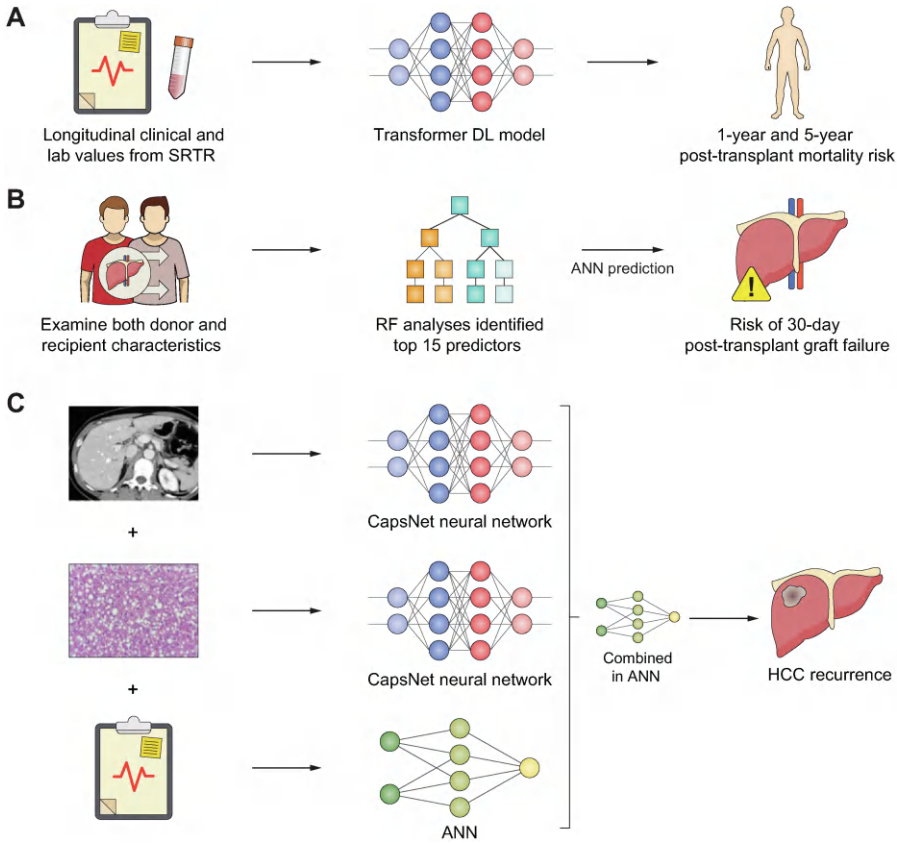

Figure 4.2 Machine learning can improve post-LT management. (A) Using longitudinal clinical and laboratory data from the SRTR database, patient specific 1-year and 5-year mortality risk can be predicted using a transformer model. (B) Incorporating both the top transplant and recipient characteristics can reliably predict graft failure using an ANN. (C) Combining medical imaging, histopathological, and clinical data in multiple models can predict the risk of HCC recurrence. ANN, artificial neural network; DL, deep learning; HCC, hepatocellular carcinoma; LT, liver transplantation; RF, random forest; SRTR, Scientific Registry of Transplant Recipients [24].

- **Advanced Data Analysis:** Machine-learning techniques offer advanced data analysis of forensic pathology findings, allowing for the detection of patterns, trends, and correlations in vast datasets of postmortem exams. By applying machine-learning approaches, forensic pathologists can reveal hidden insights from complex pathological data, leading to more thorough forensic assessments and better-informed decision-making.

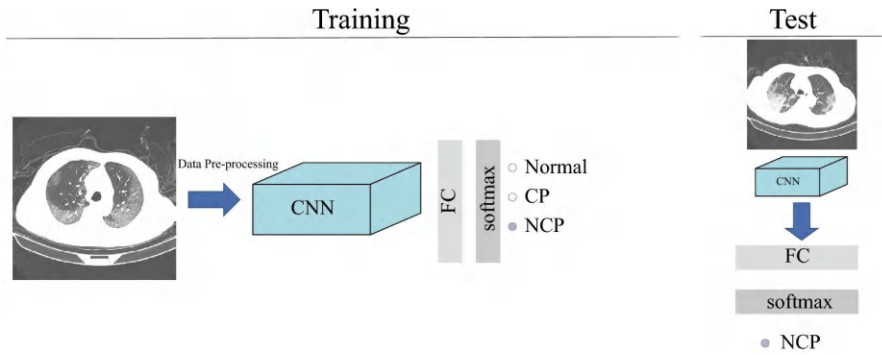

Figure 4.3 A general flowchart of a deep learning-based COVID-19 diagnosis system [31].

Problems Faced in Forensic Pathology

- **Subjectivity in Interpretation:** One of the primary challenges in the field of forensic pathology is the interpretation of postmortem results because of the subjectivity involved in the whole process. The judgement and experience of the examiner play an important role during the process of examining tissue samples for the purpose of finding out the cause of death and making forensic decisions based on that information. There could be some inconsistencies in the forensic conclusions provided by the machine-learning algorithms if they cannot accurately replicate the process of decision-making performed by the trained forensic pathologists.
- **Variability in Forensic Practices:** Different jurisdictions, laboratories and individual examiners might have different processes and practices for cases related to forensic pathology, and that might produce some inconsistencies in the interpretation of results and forensic studies related to pathology. An applicable solution for this issue should include combinations of various databases in order to integrate regional variances for the training of machine-learning algorithms to create heterogeneity in the model. In order to achieve accurate and reliable results in forensic examinations, it's very important to have standards for every process and methodology used during the examination.
- **Data Quality and Availability:** The training of machine-learning algorithms used in forensic pathology requires high-quality and comprehensive databases. However, these forensic pathology databases used for training machine-learning models can include sample biases and missing data, and another limitation is that they could include heterogeneous data. Because of that, forensic pathologists

should monitor the process of training these machine-learning algorithms so they provide accurate and reliable results at the end of the process.

- **Complexity of Pathological Data:** Data obtained from pathological results may be ambiguous and even diverse, involving different tissue specimens, histological changes and diagnostic outcomes of postmortem examinations. Such complex data may not be easy to understand by machine-learning algorithms, and this could be a problem, especially when the results are rare or unusual clinically. Machine learning is useful to forensic pathologists, but the results need to be thoroughly checked and validated to ascertain their accuracy in forensic examinations [26, 33, 34].

Steps for Performing Analysis in Machine Learning in Forensic Pathology

- **Conceptualization and Problem Formulation:** This step involves defining the particular forensic pathology problem or task that is going to be solved with the help of machine learning. Scientists define the goal, research questions and evaluation criteria to guide the development of machine learning models. For instance, in pathology, the problem could be to diagnose the probability of a tissue sample being cancerous or benign by examining histological characteristics.
- **Data Acquisition and Preparation:** During this step, researchers obtain relevant datasets from hospitals, research organizations, or a public database for the purpose of training ML algorithms. They preprocess the data to eliminate errors, inconsistencies and noise to make it suitable for analysis and more reliable. In pathology, this may involve obtaining digital histopathology slides and the related clinical metadata, including patients' demography and diagnosis.
- **Feature Extraction and Representation:** This stage involves transforming the raw data into features that are useful in the process of training the machine-learning algorithms. In pathology, features could mean the shape, size or location of certain cell types, the location of tissues in the body, or a measure of the intensity of staining of a tissue. More complex forms of machine-learning models, such as convolutional neural networks (CNNs), do not require features to be extracted as they can automatically be learned from digital pathology images.
- **Model Development and Evaluation:** Scientists usually develop machine-learning models with the help of techniques such as logistic regression models, random forests or deep-learning models.

These models are built based on supervised learning, where a set of input attributes is mapped to a set of goal outputs with the help of labelled data. Relevant evaluation parameters such as accuracy, sensitivity, specificity and the AUC–ROC are used in the assessment of machine-learning models. For example, in pathology, a model of diagnosis may assess the probability of cancer reoccurrence depending on the data on gene expressions.

- **Interpretation and Validation:** This step is related to the interpretation of outcomes from the machine-learning models to obtain insights into biological systems or pathological processes. The model performances are also checked with other datasets or by applying cross-validation techniques to check the stability of the models. In pathology, for example, the model interpretation may involve identifying features or biomarkers associated with disease progression or response to treatment, which may help guide a clinician's decisions.
- **Integration and Deployment:** These models are then deployed in clinical workflows or diagnostic applications to assist pathologists in practice environments after having been validated. It may include designing interfaces that can be easily used, integrating decision supports or models with current laboratory information systems, etc. For instance, in pathology, a risk assessment model for cancer could be integrated with digital pathology tools employed by pathologists in diagnosing cancer.
- **Ethical Considerations and Societal Impact:** Researchers should discuss the ethical issues and possible consequences of the integration of machine learning with pathology. This encompasses protecting the patient's identity, minimizing bias in the training and deploying algorithms and promoting accountability when developing and implementing the models. In pathology, some of the ethical issues may include consent in data usage, fair distribution of diagnostic tools and reporting the limitations of the models.
- **Continuous Improvement and Adaptation:** In the last step, refining and enhancing the models with feedback from the pathologists, new data and advancements in technology are carried out cyclically. Model assessment is done by researchers to check, modify or even redesign the current model to improve the accuracy and reliability of the results for a specific field. In the field of pathology, continual improvement may include modifying the models because of emergent biomarkers, revising the diagnostic criteria or responding to changes in clinical practice guidelines [35–37].

Machine Learning in Forensic Anthropology

Forensic anthropology is crucial for identifying human remains and understanding the events surrounding death. Traditionally, forensic anthropologists employ morphological studies of skeletal remains to assess variables such as age at death, sex, ancestry and stature [38, 39]. However, this examination can be complex and time-consuming, often requiring expertise and subjective interpretation. In recent years, the incorporation of machine-learning (ML) techniques has emerged as a viable strategy to boost the efficiency and accuracy of forensic anthropological analyses. Machine-learning algorithms, capable of processing enormous and diverse datasets, offer novel options for the automated study of skeletal remains and the prediction of biological profiles. By integrating computational approaches and statistical models, ML helps forensic anthropologists derive useful insights from skeletal data and hasten the identification process in forensic cases.

Significance of Forensic Anthropology

- **Identification of Human Remains:** Forensic anthropology plays a significant role in the identification of human remains, especially in cases when the remains are decomposed, fragmented or otherwise difficult to identify using standard procedures. Forensic anthropology involves the study of the skeletal and dental profiles of the deceased, thus assisting in the identification of the individual's identity, age, sex, ancestry and stature, in addition to the information that can be derived from machine-learning algorithms [38, 40, 41].
- **Understanding Trauma and Damage Patterns:** Forensic anthropologists know what the bones and patterns of a particular injury look like, and they are also aware of what information it could provide to help the investigation. Bone injuries such as fractures, gunshot injuries and other skeletal injuries are the areas of specialization of forensic anthropologists, and they aid forensic pathologists and investigators in their evaluations throughout crime scene investigations [42, 43].
- **Assessment of Postmortem Changes:** Forensic anthropologists are knowledgeable and experienced in studying changes that occur to the skeletal remains after death, such as decomposition of the body, carnivore damage and other taphonomic effects. This type of knowledge is valuable in circumstances where the state of human remains has been altered, as forensic experts are able to accurately distinguish

between antemortem, perimortem and postmortem injuries in order to reconstruct circumstances leading to death. [42, 44–46].

- **Reconstruction of Biological Profiles:** Forensic anthropologists can reconstruct the biological attributes of the deceased from unidentified bones and bring out aspects such as age, sex, ancestry and stature. This information is crucial for selecting the likely candidates in the databases of missing people and contributing to their identification, which corresponds with the objectives of the project to use machine learning for the identification of individuals in forensic practice [47–49].
- **Collaboration with Machine-Learning Technologies:** This field of forensic anthropology is likely to gain from working with machine-learning technologies since it is possible to use computational techniques to analyze large data sets of skeletal remains. The combination of machine-learning algorithms with conventional anthropological approaches during forensic analysis will enhance the effectiveness of the identification and reconstruction of events in forensic investigations [50–52].

Problems Faced in Forensic Anthropology

- **Incomplete or Fragmented Remains:** The first, and one of the most crucial problems of forensic anthropology, is working with the fragmented or incompleteness of the skeletal remains. Sometimes, forensic anthropologists are only able to recover fragmentary bones, and this may lead to several difficulties in the construction of the biological profile and identification of the cause of death; therefore, this leads to insufficient forensic studies.
- **Lack of Standardization:** Some analyses in forensic anthropology have no standard operating procedures and methods; therefore, the approaches used by practitioners are diverse. This lack of uniformity leads to variations in the forensic assessments and interpretations conducted, thereby jeopardizing the overall reliability and accuracy of forensic judgements.
- **Limited Access to Resources:** Other limitations are related to restricted access to resources such as skeletal collections, reference databases and equipment in a forensic anthropologist's work. This means that where there is an absence of better resources, forensic anthropologists may be unable to make comprehensive and accurate identifications of unknown individuals, thereby slowing the progress of forensic investigations.
- **Time and Cost Constraints:** The methodologies used in forensic anthropology can consume a lot of time and may also be expensive,

especially when handling complex skeletal structures or in the case of long forensic investigations. Limited money and resources may impede the ability of forensic anthropologists to undertake full analysis within realistic timescales, thereby delaying the conclusion of forensic cases [53].

Application in the Realm of Forensics

Machine Learning and Deep-Learning Methods for Sex Estimation of Infant Individuals: The research of machine-learning techniques, particularly deep learning, for sex estimation in infant skeletons. A recent study compared the effectiveness of machine-learning algorithms with expert visual assessment in estimating the sex of child skeletons from ilium pictures. The researchers utilized photos of 135 infant individuals aged between 5 months of gestation to 6 years from the University of Granada collection. The study applied deep-learning techniques such as VGG16 and ResNet50 pre-trained on ImageNet, obtaining an accuracy of 59%, which was near to expert evaluation. Classic ML approaches such as support vector machines (SVMs), random forests (RFs) and AdaBoost with HOG features were also tested, yielding an accuracy of 49%, which was less successful than deep learning. The results of the study suggested that deep learning approaches provided competitive outcomes compared to expert assessment, underlining the promise of AI techniques in forensic anthropology. The study also proposed exploring 3D models and software tools to further develop AI applications in forensic anthropology [54].

Pediatric Bone Age Assessment Using Deep Learning: Generalizability and Limitations: Here we examine the issues of generalizing pediatric bone age assessment using deep learning models, which is an important concern in this field. It was considered intriguing that the algorithm was evaluated on several datasets without segmentation and the impact of demographic characteristics such as sex was underlined. Several studies provide useful information for future studies to improve the clinical usability of automated bone age assessment. It also underlines the necessity to examine cross-institutional generalizability and demographic considerations for strong and equitable medical uses of deep learning in bone age assessment [55].

Advantages of the Study

- **Comprehensive Exploration:** The chapter presents a thorough assessment of the interaction between machine learning and forensic medicine, spanning numerous subfields such as forensic pathology,

anthropology and odontology. This comprehensive method offers useful insights into the possible uses of machine learning across many domains of forensic science.

- **Practical Implications:** By showing the practical applications of machine learning in forensic investigations, the study offers tangible benefits for forensic practitioners and legal experts. The application of modern algorithms and computational approaches has the potential to increase the efficiency, precision, and objectivity of forensic analyses, thereby strengthening the criminal justice system.
- **Future Directions:** This chapter presents a literature review on the contemporary advancements and the anticipated development in the field of forensic medicine particularly with the integration of the different machine-learning algorithms. Moreover, this forward-looking perspective presented in this chapter leads to further research and development that defines the course for the constant enhancement of forensic science.
- **Interdisciplinary Collaboration:** It is crucial to recognize that such work requires integrated cooperation between forensic experts, data scientists and technology specialists. This approach of breaking professional boundaries fosters creativity and efficiency in the processes of providing solutions in forensic medicine since different professions are involved.

Disadvantages of the Study

- **Limited Scope:** However, one might consider that this chapter is relatively generalized but limitations include lack of detailed information regarding certain subfields of forensic medicine or detailed explanation about applications within the field of forensic medicine. Certain areas of forensic science may receive less attention or remain undiscovered, perhaps disregarding vital parts of the subject.
- **Ethical Considerations:** The study may raise ethical questions surrounding the use of machine-learning algorithms in forensic investigations, specifically for privacy, prejudice and accountability. The potential for algorithmic errors or exploitation could have major ramifications for those participating in court procedures.
- **Technological Dependency:** Relying significantly on machine-learning technology may generate a dependency on advanced tools and algorithms, which could provide issues in terms of accessibility, pricing, and sustainability for forensic laboratories and practitioners.
- **Human Skill:** While machine-learning algorithms offer essential aid in forensic analysis, they should not replace the skill and judgement of human forensic professionals. The study should emphasize

the complementary nature of machine learning and human intelligence in forensic medicine, rather than pushing a wholly automated method.

Future Scope of Machine Learning

Advancements in Forensic Pathology

Machine-learning models, such as convolutional neural networks (CNNs) and decision trees, could be deployed to assess histological pictures, postmortem data and damage patterns. Here's how machine learning could be applied:

- **Automated Histopathological Analysis:** CNNs trained on vast datasets of histopathological images would be used to automatically identify tissue samples and find anomalies indicative of specific diseases or injuries. Transfer learning techniques could adapt pre-trained CNN models to forensic pathology datasets, providing a speedy and reliable diagnosis.
- **Predictive Modeling of Postmortem Changes:** Time-series analysis approaches, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, would be applied to anticipate postmortem interval based on characteristics such as body temperature, rigor mortis, and decomposition rates. These models would learn temporal patterns from prior postmortem data to anticipate the period since death with greater accuracy.
- **Pattern Recognition in Injury Assessment:** Decision tree algorithms, assisted by expert knowledge and feature selection approaches, would be applied to identify patterns of injuries based on their location, morphology and associated clinical aspects. Ensemble approaches such as random forests could incorporate many decision trees to boost classification performance and interpretability.

Example Result: A CNN-based model obtained an accuracy of over 90% in identifying histological images of traumatic injuries, enabling forensic pathologists to reliably diagnose and describe patterns of trauma in forensic investigations.

Innovations in Forensic Medicine:

Machine-learning approaches, including support vector machines (SVMs), deep-learning architectures and Bayesian networks, would be applied to

examine medical evidence, genetic data and forensic photographs. Here's how machine learning could contribute:

- **Facial Recognition and Age Progression:** Deep-learning models, such as generative adversarial networks (GANs) or variational autoencoders (VAEs), would be trained on facial imaging data to reconstruct facial features and simulate age progression. These models would learn latent representations of facial morphology to build realistic facial reconstructions and forecast age-related changes.
- **Genetic Profiling and Ancestry Prediction:** Bayesian networks or probabilistic graphical models would be used to examine genetic markers and infer ancestral origins, demographic affinities and genetic relationships. By merging genomic data with demographic information and reference databases, these models would probabilistically estimate individual ancestry and phenotype.
- **Trauma Analysis and Injury Reconstruction:** Machine-learning algorithms, such as k-nearest neighbours (KNN) or support vector regression (SVR), would be trained on forensic imaging data to rebuild injury patterns and predict impact dynamics. These models would learn from labelled datasets of traumatic injuries to predict damage features, weapon kinds and biomechanical forces.

Example Result: A deep learning-based age progression model correctly replicated face ageing in forensic photographs with a mean absolute error of less than two years, allowing the identification of missing persons and suspects based on long-term changes in appearance.

Transformations in Forensic Anthropology

Machine-learning approaches, such as geometric morphometrics, Bayesian statistics, and ensemble learning, would be applied to assess bone measurements, facial landmarks and 3D image data. Here's how machine learning could be integrated:

- **Morphometric Analysis and Ancestral Affiliation:** Geometric morphometric methods, coupled with unsupervised learning algorithms such as principal component analysis (PCA) or clustering techniques, would be applied to skeletal measurements and cranial features to identify ancestral clusters and population affinities. These

models would minimize dimensionality and highlight patterns of morphological variation among populations.

- **Age Estimation and Growth Modelling:** Regression-based approaches, such as support vector regression (SVR) or Gaussian process regression (GPR), would be trained using age-at-death data and skeletal growth trajectories to estimate biological age and developmental stage from skeletal remains. These algorithms would learn nonlinear correlations between skeletal characteristics and chronological age to forecast age-related changes properly.
- **Virtual Reconstruction and 3D Visualization:** Machine-learning algorithms, such as iterative closest point (ICP) algorithms or deep learning-based image registration approaches, would be utilized to align and recreate fractured skeletal remains from forensic imaging data. These models would repeatedly enhance geometric alignments and provide 3D reconstructions of skeletal structures for anatomical examination.

Example Result: A geometric morphometric study discovered discrete morphological clusters in cranial measures belonging to different ancestral populations, enabling forensic anthropologists to establish the ancestry of unidentified skeletal remains with high accuracy.

References

1. Houck, M. M. (2017). *Forensic pathology*. San Diego, CA: Academic Press.
2. DiMaio, V. M. (2021). *DiMaio's forensic pathology*. Boca Raton, FL: CRC Press. https://doi.org/10.4324/9780429318764
3. Richard, S. (2016). *Forensic science from the crime scene to the lab*. New York, NY: Pearson.
4. Wagner, S. A. (2009). *Death scene investigation—A field guide*. Boca Raton, FL: CRC Press. https://doi.org/10.1201/9781420086775
5. e-PG Pathshala. (2015). *Introduction to forensic medicine*. India. Retrieved from https://epgp.inflibnet.ac.in/Home/ViewSubject?catid=eCJfy23Kjy3c0vICLa6VYg==
6. Pandey, A. (2022, August 28). Medical jurisprudence and related laws in India. *Brain Booster Articles*. Retrieved from https://www.brainboosterarticles.com/post/medical-jurisprudence-and-related-laws-in-india
7. Saferstein, R. (2017). *Criminalistics: An introduction to forensic science* (12th ed.). New York: Pearson.
8. Fang, Y. L. (2020). New opportunities and challenges for forensic medicine in the era of artificial intelligence technology. *Fa Yi Xue Za Zhi*, *36*(1), 77–85. https://doi.org/10.12116/J.ISSN.1004-5619.2020.01.016

9. Rigano, C. (2019). Using artificial intelligence to address criminal justice needs. *National Institute of Justice*, *280*, 1–10. Retrieved from https://nij.ojp.gov/topics/articles/using-artificial-intelligence-address-criminal-justice-needs
10. Yogesh Kumar, A. K. (2023). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework, and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z
11. Tucci, L. (2023, September). *Machine learning (ML)*. Retrieved from https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML
12. Soni, V. (2017). Forensic medicine: A source and pathway of recognition in disaster victim identification. *International Journal of Forensic Medicine and Toxicological Sciences*, *2*(1), 2–7. https://doi.org/10.18231/J.IJFMTS.2017.002
13. Britannica, The Editors of Encyclopaedia. (2023, October 13). *Forensic medicine*. Encyclopædia Britannica. Retrieved from https://www.britannica.com/topic/forensic-medicine
14. Eriksson, A. (2016). Forensic pathology. In M. D. Zeegers (Ed.), *Forensic epidemiology principles and practice* (pp. 151–177). San Diego, CA: Academic Press.
15. M, N. (1989). Problems regarding the examination in forensic medicine. *The Japanese Journal of Legal Medicine*, *43*(5), 364–376.
16. Pinheiro, J. C. (2006). Forensic investigation of corpses in various states of decomposition: A multidisciplinary approach. In A. C. Schmitt (Ed.), *Forensic anthropology and medicine* (pp. 159–195). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-59745-099-7_7
17. Baldino, G. M. (2023). Multidisciplinary forensic approach in "complex" bodies: Systematic review and procedural proposal. *Diagnostics*, *13*(2), 310. https://doi.org/10.3390/diagnostics13020310
18. Jumbelic, M. I. (2005). Mass disasters. In *Encyclopedia of forensic and legal medicine* (pp. 197–207). Amsterdam, Netherlands: Elsevier. https://doi.org/10.1016/B0-12-369399-3/00232-9
19. Williams, D. J. (1998). *Forensic pathology: Colour guide*. Edinburgh, Scotland: Churchill Livingstone.
20. Thurzo, A., & Stanojevic, H. (2021). Use of advanced artificial intelligence in forensic medicine, forensic anthropology and clinical anatomy. *Healthcare (Basel)*, *9*(11), 1545.
21. Utilities One. (2023, November 10). *Forensic engineering and chemical analysis techniques*. Retrieved from https://utilitiesone.com/forensic-engineering-and-chemical-analysis-techniques
22. Brown, S. (2021, April 21). *Machine learning, explained*. MIT Sloan School of Management. Retrieved from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
23. Cui, M., & Zhang, D. Y. (2021, April 1). *Artificial intelligence and computational pathology*. Retrieved from https://www.nature.com/articles/s41374-020-00514-0.pdf
24. Bhat, M., Rabindranath, M., Chara, B. S., & Simonetto, D. A. (2023). Artificial intelligence, machine learning, and deep learning in liver transplantation. *Journal of Hepatology*, *78*(6), 1216–1233. https://doi.org/10.1016/j.jhep.2023.01.006

25. Peña-Solórzano, C. A., Albrecht, D., Bassed, R., Burke, M., & Dimmock, M. (2020, November 1). Findings from machine learning in clinical medical imaging applications: Lessons for translation to the forensic setting. *Forensic Science International*, *316*, 110538. https://doi.org/10.1016/j.forsciint.2020.110538
26. Dettmeyer, R. B. (2011). *Forensic histopathology: Fundamentals and perspectives*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20659-7
27. Zhao, W., Jiang, W., & Qiu, X. (2021). Deep learning for COVID-19 detection based on CT images. *Scientific Reports*, *11*, 14353. https://doi.org/10.1038/s41598-021-93832-2
28. Dobay, A., Ford, J., Decker, S., Ampanozi, G., Franckenberg, S., Affolter, R., Sieberth, T., & Ebert, L. C. (2020). Potential use of deep learning techniques for postmortem imaging. *Forensic Science, Medicine, and Pathology*, *16*(4), 671–679. https://doi.org/10.1007/s12024-020-00307-3
29. Nichols, J. A., Chan, H. W. H., & Baker, M. A. B. (2018). Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, *11*(1), 111–118. https://doi.org/10.1007/s12551-018-0449-9
30. Kim, C., You, S. C., Reps, J. M., Cheong, J. Y., & Park, R. W. (2021). Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. *Journal of the American Medical Informatics Association: JAMIA*, *28*(6), 1098–1107. https://doi.org/10.1093/jamia/ocaa277
31. Abrol, V. (2024). *Forensic science: Revealing the clues*. IntechOpen. https://doi.org/10.5772/intechopen.1003870
32. Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, *10*(3), 541. https://doi.org/10.3390/healthcare10030541
33. Spencer, A. G., Ross, W. K., & Domen, R. E. (2017). Forensic pathology education in pathology residency. *Academic Pathology*, *4*, 2374289517719503. https://doi.org/10.1177/2374289517719503
34. Tümer, A. R., Eskicioğlu, E., Sökmensüer, C., & Findikoğlu, T. (2022). Problems in postmortem pathology training. *Turkish Journal of Pathology*, *38*(1), 47–51. https://doi.org/10.5146/tjpath.2022.01569
35. Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023, June 8). Machine learning in digital forensics: A systematic literature review. *arXiv*. https://doi.org/10.48550/arxiv.2306.04965
36. Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: The present landscape of supervised methods. *Academic Pathology*, *6*, 2374289519873088. https://doi.org/10.1177/2374289519873088
37. Wiersema, J. M. (2016, September 5). Evolution of forensic anthropological methods of identification. Retrieved from https://journals.sagepub.com/doi/10.23907/2016.038
38. İşcan, M. Y. (2001). Global forensic anthropology in the 21st century. *Forensic Science International*, *117*(1–2), 1–6. https://doi.org/10.1016/s0379-0738(00)00433-3

39. Jayakrishnan, J., Reddy, J., & Kumar, R. (2021). Role of forensic odontology and anthropology in the identification of human remains. *Journal of Oral and Maxillofacial Pathology*, *25*(3), 543. https://doi.org/10.4103/jomfp.jomfp_81_21
40. Mesejo, P., Martos, R., Ibáñez, Ó., Novo, J., & Ortega, M. (2020). A survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification. *Applied Sciences*, *10*(14), 4703. https://doi.org/10.3390/app10144703
41. Blau, S. (2016, May 26). How traumatic: A review of the role of the forensic anthropologist in the examination and interpretation of skeletal trauma. Retrieved from https://www.tandfonline.com/doi/full/10.1080/00450618.2016.1153715
42. Adserias-Garriga, J. (2019). A review of forensic analysis of dental and maxillofacial skeletal trauma. *Forensic Science International*, *299*, 80–88. https://doi.org/10.1016/j.forsciint.2019.03.027
43. Holz, F., Birngruber, C. G., & Verhoff, M. A. (2015). [Pre- and perimortem bone trauma vs. postmortem damages – Principles of differentiation]. *Rechtsmedizin*, *236*(1–2), 51–63. Retrieved from https://pubmed.ncbi.nlm.nih.gov/26399122/
44. Almulhim, A. M., & Menezes, R. G. (2022, May 8). Evaluation of postmortem changes. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK554464/
45. Christensen, A. M., & Passalacqua, N. V. (2018, November 5). *Analysis of skeletal trauma*. Retrieved from https://www.sciencedirect.com/science/article/pii/B978012812201300013X
46. Austin, D., & King, R. E. (2016, September 1). The biological profile of unidentified human remains in a forensic context. *Academic Forensic Pathology*, *6*(3), 370–390. https://doi.org/10.23907/2016.039
47. Mk, S. (2016, September 1). Metric methods for the biological profile in forensic anthropology: Sex, ancestry, and stature. *Academic Forensic Pathology*, *6*(3), 391–399. https://doi.org/10.23907/2016.040
48. Mahfouz, M. R., Mustafa, A., Fatah, E. E. H. A., Herrmann, N. P., & Langley, N. R. (2017, June). Computerized reconstruction of fragmentary skeletal remains. *Forensic Science International*, *275*, 212–223. https://doi.org/10.1016/j.forsciint.2017.03.017
49. Ionescu, V., Teletin, M., & Voiculescu, E. (2016, May 1). Machine learning techniques for age at death estimation from long bone lengths. https://doi.org/10.1109/saci.2016.7507421
50. Thurzo, A., Kosnáčová, H., Kurilová, V., Kosmeľ, S., Beňuš, R., Moravanský, N., Kováč, P., Kuracinová, K. M., Palkovič, M., & Varga, I. (2021, November 12). Use of advanced artificial intelligence in forensic medicine, forensic anthropology and clinical anatomy. *Healthcare*, *9*(11), 1545. https://doi.org/10.3390/healthcare9111545
51. Olver, P. J., Coil, R., Melton, J. A., Olver, P. J., Tostevin, G., & Yezzi-Woodley, K. (2022). Use and misuse of machine learning in anthropology. *IEEE BITS*, 1–13. https://doi.org/10.1109/mbits.2022.3205143

52. Christensen, A. M., Passalacqua, N. V., & Bartelink, E. J. (2014). *Contemporary issues in forensic anthropology*. https://doi.org/10.1016/b978-0-12-418671-2.00015-x
53. Ortega, R., Irurita, J., Campo, E. J. E., & Mesejo, P. (2021, July 16). Analysis of the performance of machine learning and deep learning methods for sex estimation of infant individuals from the analysis of 2D images of the ilium. *International Journal of Legal Medicine*, *135*(6), 2659–2666. https://doi.org/10.1007/s00414-021-02660-6
54. Valliani, A., Schwartz, J., Arvind, V., Taree, A., & Kim, J. (2020, September 21). Multi-site assessment of pediatric bone age using deep learning. https://doi.org/10.1145/3388440.3412429
55. Kim, C., You, S. C., Reps, J. M., Cheong, J. Y., & Park, R. W. (2021). Machine-learning model to predict the cause of death using a stacking ensemble method for observational data. *Journal of the American Medical Informatics Association: JAMIA*, *28*(6), 1098–1107. https://doi.org/10.1093/jamia/ocaa277.

Application of Machine Learning in the Field of Forensic Biology and Serological Evidence Identification

5

SATISH KUMAR AND
JENNIFER JOHNSON

Introduction: Forensic Biological Evidence

The term 'biological' represents living organisms, life processes and their interactions. Evidence, on the other hand, is information that is presented in court to support an argument that has been legitimately stipulated as evidence of that claim to an authorized tribunal. Blood, saliva, semen, hair etc., all are referred to as forensic biological evidence, that is commonly encountered at a crime scene [1]. This biological evidence holds specific characteristics that aid forensic experts and investigators in carrying out forensic investigations judicially. During any forensic investigation, this evidence is of paramount significance for many reasons. It can connect a suspect to a victim, an object or a crime scene. This link can further be crucial in assembling evidence against the accused [2]. Also, forensically relevant biological evidence has efficiently helped in exonerating the innocent who were wrongfully convicted of a crime. DNA analysis is a powerful investigative technique useful for identifying individuals and establishing familial relationships. Moreover, forensic biological evidence is an essential part of criminal investigations and court procedures since it independently validates and supports the key elements of the case, serving as corroborative evidence. The scientific and objective aspect of such evidence lends credibility and dependability to the case as a whole.

Biological evidence refers to samples or specimens containing biological material, usually collected at a crime scene. It can be categorized into various forms based on its origin, which can be human or non-human (animal or plant). As illustrated in Figure 5.1, forensic biological evidence is broadly grouped into two categories, that is, body fluids (saliva, blood, urine, sweat, semen, tears, vaginal and nasal secretion) and non-body fluids (teeth, nails,

DOI: 10.4324/9781003449164-6

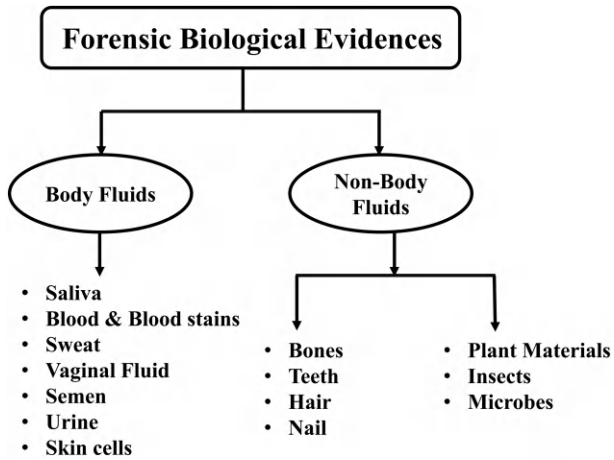

Figure 5.1 Categorization of forensic biological evidence.

bones, hair, tissues, botanical and microbial evidence) [3]. In serology analysis, biological fluids are identified through the use of both presumptive as well as confirmatory testing. Presumptive testing relates to a rapid kind of testing that indicates the presence of bodily fluid under investigation. On the other hand, confirmatory tests are relatively specific and take longer time for analysis as compared to the presumptive tests [4].

The quality and reliability of forensic analysis may be majorly impacted by several challenges faced while analyzing such biological evidence during forensic investigations. External factors (temperature, humidity, flora and fauna of the surrounding area) tend to degrade these biological samples rapidly. Body fluid identification in forensic science holds significant prominence as it provides the capability to differentiate them individually. However, there are enormous challenges faced in dealing with these body fluids during forensic investigations. Low sample quantity or sample mixtures majorly contribute to any hindrance during the analysis. Therefore, there is always a demand for significant advancements in forensics so as to overcome these challenges. Computational models coupled with human abilities will venture new avenues in the field of forensics by offering promising measures to uplift scientific assessment and thereby ensure accuracy and precision in forensic investigations.

Machine-Learning Approaches

Machine learning (ML) encompasses a broader collection of varied algorithms aimed at attempting to identify computational data patterns and then utilizing them to derive mathematical models which are capable of generalizing

the learned rules on unseen data. It is a subfield of artificial intelligence (AI) dealing with the implementation of computational algorithms and patterns to improve the overall performance of human labour. Over the years, the application of machine-learning algorithms has significantly advanced and attained wider practical approaches. It has become the most popular choice for creating useful software for speech recognition, language processing, computer vision and other related fields.

Different knowledge acquisition methods used under the umbrella of varied ML tools are supervised and unsupervised learning methods, semi-supervised learning and reinforcement learning methods as shown in Figure 5.2. Supervised machine-learning algorithm refers to when input data is matched with corresponding output labels, thereby producing precise predictions on unseen data [5]. Supervised learning includes the following approaches: decision tree–based learning, linear modelling and deep learning. Unsupervised learning includes ordination, clustering and anomaly detection [6]. In unsupervised learning, patterns are to be recognized by the learning system without any prior labels or specifications. On the other hand, in a reinforcement learning system, a machine-learning algorithm learns to achieve the goal or provide a list of actions based on the current system on its own by maximising a reward through trial and error via interaction with the given environment [7, 8].

Large-scale data is becoming more prevalent across all spheres of human endeavour, which has led to a surge in new requirements for the underlying machine-learning algorithms. Multivariate data analysis techniques can be classified into two main groups: (i) regression or calibration models and (ii) pattern recognition techniques. Further, the pattern recognition techniques

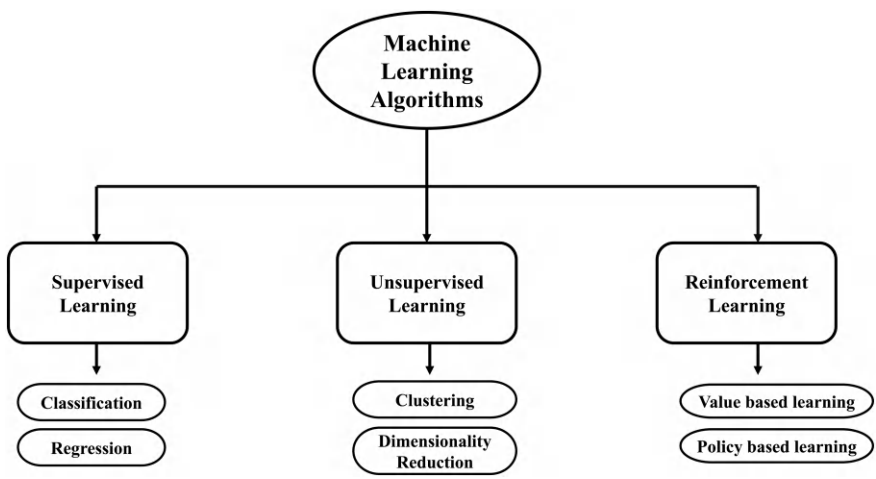

Figure 5.2 Classification of Machine-learning algorithms.

are subgrouped into supervised and unsupervised models. Principal component analysis (PCA) is one of the unsupervised models, whereas partial least square–discriminant analysis (PLS–DA) represents supervised models or classification [9].

Application of Machine Learning in Forensic Biology

Forensic investigations have the potential to be completely transformed by the integration of machine-learning algorithms in forensic biology. Forensic scientists may enhance the accuracy as well as efficiency of their investigations by using these machine-learning techniques in their analysis of DNA evidence. Figure 5.3 showcases the role of machine-learning algorithms in different areas of forensic biology that help in improving the overall accuracy and efficiency. Novel opportunities and challenges have been brought forth by recent research on machine learning algorithms in forensic domains. These algorithms have been utilized to address the limitation of human bias specifically in the field of forensics [10].

Advanced DNA Analysis

In forensic biology, machine-learning algorithms can identify genetic mutations in DNA that may go undetected through manual analysis of vast DNA datasets. Additionally, such algorithms can also assist in automating the DNA analysis procedure, which will further reduce manual time and effort. Six machine-learning algorithms namely random forest (RF), logistic

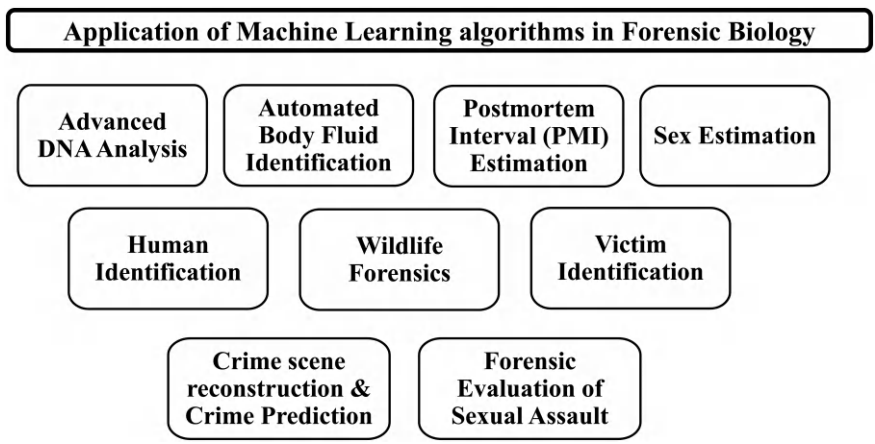

Figure 5.3 Role of machine learning in different areas of forensic biology.

regression (LR), k-nearest neighbour (KNN), stochastic gradient descent (SGD), Gaussian Naïve Bayes (GNB) and support vector machine (SVM) were applied to available DNA mixtures and were assessed for four matrices. It resulted in 95% accuracy using the ML tools [11]. In another study, artificial neural networks (ANN), a subset of machine-learning algorithms were explored along with regression analysis to evaluate DNA methylation patterns for age estimation [12].

Automated Body Fluid Identification

Identification of various body fluids during any forensic investigation is crucial to know about their origin and comprehend the timeline of events at the crime scene. Machine-learning algorithms aid in categorizing different body fluids based on their specific chemical and biomolecular characteristics. Researchers have used various ML models to identify specific body fluids. Specific microRNA (miRNA) biomarkers were reported to identify menstrual blood using logistic regression models [13]. Multi-class random classifier probabilistic models were also developed to predict the source of single or mixed body fluids respectively [14]. A novel model based on a random forest algorithm was developed to assess the type and source of different body fluids [15].

Human Identification

Over the years, human identification during any forensic investigation is achieved by determining the genetic profiles, utilizing the autosomal short tandem repeats (STRs) and mitochondrial DNA (mtDNA). However, owing to the complexity and large datasets, forensic scientists have considered machine-learning algorithms as a novel approach to predicting genetic relatedness [16]. In another study, three machine learning classifiers—random forest (RF), support vector machine (SVM) and artificial neural network (ANN) were used to predict externally visible characteristics (EVCs) obtained from DNA sources [17]. Using a customized convolutional neural network (CNN), the study created an automatic human identification system (DENT-net) that can accurately identify individuals from panoramic dental radiographs (PDRs) [18].

Postmortem Interval (PMI) Estimation

During forensic investigations, the estimation of time since death, that is, postmortem interval (PMI) is crucial yet challenging. Researchers aimed to

construct an accurate postmortem interval (PMI) prediction model utilizing the microbial sequencing data from internal organs in mouse remains. Regression models such as random forest (RF) and support vector (SV) helped in identifying potential microbial biomarkers for PMI estimation [19]. Alternatively, varied multivariate approaches such as generalized additive models (GAMs) and support vector machines (SVMs) allow various indicator components to be included in the models, hence making the PMI prediction more accurate [20]. Other machine-learning tools such as linear discriminant analysis (LDA) and logistic regression (LR) have also been widely used in forensic science for PMI estimation [21]. These approaches to machine learning utilize the positive aspects of multiple base classifiers and the prediction outcomes, integrating them to deliver accurate predictions. This further increases the diversity and complexity of the models while simultaneously enhancing prediction accuracy and robustness.

Sex Estimation

Ortega et al. [22] had earlier reported the efficacy of various machine-learning (ML) tools in accordance with an expert's visual examination to determine the gender of baby skeletons based on ilium images. In another study, convolutional neural network (CNN) models were developed for sex estimation based on pelvic regions of the body [23]. Based on photo-anthropometric indexes, specific sex and age estimations were earlier reported, wherein an artificial neural network enabled the classification of individuals from a Brazilian population [24]. Venema Javier et al. [25] analyzed different learning algorithms for male and female estimation via humerus bone images. The results obtained from this study attained the highest level of accuracy at 91.03% when compared to the ones obtained by a human expert with 83.33%.

Wildlife Forensics

In wildlife forensics, machine learning plays a pivotal role as it utilizes advanced computational techniques to assess and understand data related to crimes involving wildlife. Investigating illicit activities such as poaching, trafficking and the illegal trade in endangered species is the primary objective of wildlife forensics. Large datasets, such as involving acoustic recordings or camera traps, can be processed over time by these algorithms to track migration patterns, monitor changes in wildlife distribution and estimate wildlife population respectively. An automated approach was developed for species identification on massive matrix-assisted laser desorption/ionization (MALDI) datasets employing machine learning with minimal human input.

The study utilized semi-supervised machine-learning algorithms for species identification by collagen peptide fingerprinting [26]. In another study, machine-learning algorithms were used to differentiate tooth scores and tooth pits of carnivore species with high precision [27].

Crime Scene Prediction and Reconstruction

The incorporation of algorithms based on machine learning (ML) can improve crime scene reconstruction in numerous ways by providing insightful information that helps law and enforcement agencies solve crimes. Machine-learning algorithms, utilizing advanced deep-learning strategies on CCTV camera images, fraud detection systems and spam image recognition, have been employed in both the identification and prevention of criminal acts. Palanivinayagam et al. [28] proposed an algorithm to forecast the probability of specific crimes in a particular area. The study also reported a summary of different works where advanced machine-learning algorithms were used for the prediction of crime for a given area. Mandalapu et al. [29] reported a systematic review of the application of different machine-learning algorithms for the early detection of crime. A similar review was reported by Dakalbab et al. [30] on the application of various machine-learning algorithms for crime prediction. Therefore, advanced machine-learning algorithms have the potential to foresee future criminal activity, reconstruct complex crime scenes and track significant bits of evidence.

Victim Identification

Machine learning has the potential to significantly contribute to victim identification processes by identifying victim's facial characteristics or analyzing behavioural patterns associated with victims, that may help in understanding their habits, routines or interactions. A machine-learning algorithm was introduced to identify single nucleotide polymorphisms (SNPs) for victim identification from skin microbes [31]. Another research study was reported to evaluate the efficacy of machine learning algorithms in identifying panoramic radiograph pairs for individual identification [32]. It was observed that by utilizing a deep convolutional network an accuracy of 85% was obtained and precise prediction for personal identification was achieved. In order to predict the outcomes of criminal trials based on the evidence provided, researchers [33] utilized different machine-learning tools, respectively. The algorithm was able to accurately classify historical guilty verdicts

and thereby demonstrate the potential of using machine learning to assist in decision-making based on evidence in criminal trials.

Forensic Evaluation of Sexual Assault

Fernandes et al., in their work, have reported state-of-the-art deep-learning algorithms that were evaluated for the forensic assessment of sexual assault [34]. A framework of key steps involved in the automated detection of genital injuries and forensic assessment utilizing machine learning tools was proposed.

Other Areas of Forensic Biology

Recent developments in the field of forensic pathology have led to the application of various machine-learning algorithms to determine the nature of death causes, such as: (1) drowning by the identification of diatoms [35], (2) classification of microscopic and gross postmortem images [36] and (3) recognition of fatal brain injuries [37, 38]. In another study, convolutional neural networks (CNN) were utilized to classify bloodstain patterns, wherein a 99.73% success rate was achieved to classify the patterns respectively [39]. Similarly, in the field of forensic genetics, a machine-learning algorithm was applied for forensic STR allele extraction [40].

Case studies—Machine Learning in Forensic Biology

In accordance with US Patent number 10,957,421, Marciano and Adelman [41] established a completely novel hybrid machine-learning technique (MLA) to analyze the DNA mixture of different contributors. In comparison with current methods, their machine-learning algorithms potentially enable faster and more accurate deconvolution (separation) of DNA mixtures with minimal financial and computational capacity.

In January 2024, Delhi Police was able to crack a murder case with the help of an artificial-intelligence algorithm, wherein the AI algorithm was used to reconstruct the victim's deceased face by correcting facial discolouration and enhancing the eyes. Thus, such algorithms are often helpful for victim identification and to predict potential crime hotspots.

Therefore, the integration of machine-learning algorithms into the field of forensic biology will bring profound advancements that will aid in the accurate analysis and interpretation of biological evidence. Figure 5.4

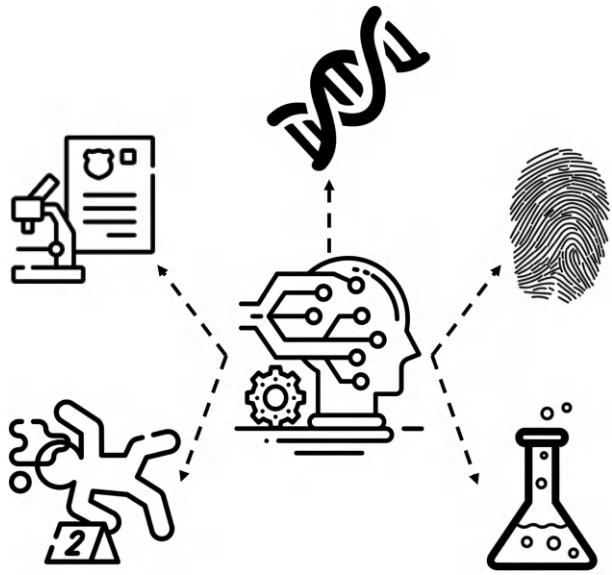

Figure 5.4 Schematic representation of the amalgamation of machine learning in the field of forensic biology and serological evidence identification.

illustrates a schematic representation of the amalgamation of machine learning in the field of forensic biology and serological evidence identification.

Advantages of machine learning approaches in forensic biology:

Machine learning algorithms hold significant promise for forensic biology with their ability to assist in the analysis and interpretation of intricate biological evidence. Some of the advantages of utilizing machine-learning algorithms in the field of forensic biology are as follows:

For biological evidence, such as DNA sequences, body fluid and hair characteristics and blood spatter patterns, machine-learning algorithms can assist forensic investigators in deciphering intricate patterns and linkages. This can help enhance the accuracy and ease in forensic analysis.

Machine-learning algorithms can minimize the possibility of human errors by automating time-consuming and repetitive tasks.

The accuracy and speed of DNA profiling can be enhanced through machine-learning algorithms, which can further help in identifying suspects and linking them to the victim or crime scene.

Analyzing probable suspects based on their geographic profiles, criminal behavioural patterns etc., these machine-learning algorithms can provide aid to different law enforcement agencies or organizations to enhance their overall investigations.

Disadvantages of Machine-Learning Approaches in Forensic Biology

Machine-learning algorithms have the ability to bring out better advancements in different areas of forensic biology. However, there are enormous challenges and disadvantages to using such complex tools and algorithms in this field.

Data Quantity and Quality: Machine-learning algorithms generally demand large, robust data sets for analysis. In forensic biology, obtaining large and diverse datasets is challenging due to the limitation of real-time crime scene samples. This may be an obstacle in providing the appropriate accuracy of the model.

Biasness: These algorithms are sometimes influenced by biased training, thereby producing false or inaccurate results.

Ethical Issues: Individual rights and privacy may be compromised if sometimes certain data sets (such as genomic data) are used for investigation purposes without prior consent.

Cost and Expertise: Applying machine learning tools to use in forensic biology requires an enormous financial and human investment in resources. Also, to generate, validate datasets and manage these tools, specialized knowledge is required.

Technology Advancements: Forensic experts and law enforcement organizations will need to update and adapt to recent developments in machine learning tools on a regular basis.

Future Aspects of Machine Learning in Forensic Biology

Machine-learning algorithms promise to hold great prominence in forensic biology and its varied aspects such as DNA analysis, serology and blood pattern analysis in future. These computational algorithms and learning tools can profoundly elevate the accuracy and competence of all these forensic biology techniques, that will overall boost the forensic investigations greatly. Machine learning can aid in quality control by identifying false or misleading forensic data, thereby maintaining overall integrity.

References

1. Virkler, K., & Lednev, I. K. (2009). Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene. *Forensic Science International*, *188*(1–3), 1–17. https://doi.org/10.1016/j.forsciint.2009.02.013

2. Avinash, P. C. (2021). Comparative role of serology and DNA profiling in forensics. *Research Journal of Forensic Sciences*, *12*, 19–21.
3. Rao, P. K., Pandey, G., & Tharmavaram, M. (2020). Biological evidence and their handling. In D. Rawtani & C. M. Hussain (Eds.), *Technology in forensic science* (1st ed., pp. 35–53). Wiley. https://doi.org/10.1002/9783527827688.ch3
4. Gefrides, L., & Welch, K. (2011). Forensic biology: Serology and DNA. In A. Mozayani & C. Noziglia (Eds.), *The forensic laboratory handbook procedures and practice* (pp. 15–50). Humana Press. https://doi.org/10.1007/978-1-60761-872-0_2
5. Babcock University, F.Y, O., J.E.T, A., O, A., J. O, H., O, O., & J, A. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/IJCTT-V48P126
6. Jovel, J., & Greiner, R. (2021). An introduction to machine learning approaches for biomedical research. *Frontiers in Medicine*, *8*, 771607. https://doi.org/10.3389/fmed.2021.771607
7. Sharma, S., & Chaudhary, P. (2023). Chapter 4 machine learning and deep learning. In P. Raj, A. Kumar, A. K. Dubey, S. Bhatia, & O. Manoj S (Eds.), *Quantum computing and artificial intelligence* (pp. 71–84). De Gruyter. https://doi.org/10.1515/9783110791402-004
8. Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine learning* (1st ed.). CRC Press. https://doi.org/10.1201/9781315371658
9. Alladio, E., Poggiali, B., Cosenza, G., & Pilli, E. (2022). Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Scientific Reports*, *12*(1), 8974. https://doi.org/10.1038/s41598-022-12903-0
10. Galante, N., Cotroneo, R., Furci, D., Lodetti, G., & Casali, M. B. (2023). Applications of artificial intelligence in forensic sciences: Current potential benefits, limitations, and perspectives. *International Journal of Legal Medicine*, *137*(2), 445–458. https://doi.org/10.1007/s00414-022-02928-5
11. Alotaibi, H., Alsolami, F., Abozinadah, E., & Mehmood, R. (2022). TAWSEEM: A Deep-learning-based tool for estimating the number of unknown contributors in DNA profiling. *Electronics*, *11*(4), 548. https://doi.org/10.3390/electronics11040548
12. Vidaki, A., Ballard, D., Aliferi, A., Miller, T. H., Barron, L. P., & Syndercombe Court, D. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Science International. Genetics*, *28*, 225–236. https://doi.org/10.1016/j.fsigen.2017.02.009
13. Hanson, E. K., Mirza, M., Rekab, K., & Ballantyne, J. (2014). The identification of menstrual blood in forensic samples by logistic regression modeling of miRNA expression. *ELECTROPHORESIS*, *35*(21–22), 3087–3095. https://doi.org/10.1002/elps.201400171
14. Iacob, D., Fürst, A., & Hadrys, T. (2019). A machine learning model to predict the origin of forensically relevant body fluids. *Forensic Science International: Genetics Supplement Series*, *7*(1), 392–394. https://doi.org/10.1016/j.fsigss.2019.10.025

15. Tian, H., Bai, P., Tan, Y., Li, Z., Peng, D., Xiao, X., Zhao, H., Zhou, Y., Liang, W., & Zhang, L. (2020). A new method to detect methylation profiles for forensic body fluid identification combining ARMS-PCR technique and random forest model. *Forensic Science International: Genetics*, *49*, 102371. https://doi.org/10.1016/j.fsigen.2020.102371
16. Govender, P., Fashoto, S. G., Maharaj, L., Adeleke, M. A., Mbunge, E., Olamijuwon, J., Akinnuwesi, B., & Okpeku, M. (2022). The application of machine learning to predict genetic relatedness using human mtDNA hypervariable region I sequences. *PLoS One*, *17*(2), e0263790. https://doi.org/10.1371/journal.pone.0263790
17. Katsara, M.-A., Branicki, W., Walsh, S., Kayser, M., & Nothnagel, M. (2021). Evaluation of supervised machine-learning methods for predicting appearance traits from DNA. *Forensic Science International: Genetics*, *53*, 102507. https://doi.org/10.1016/j.fsigen.2021.102507
18. Fan, F., Ke, W., Wu, W., Tian, X., Lyu, T., Liu, Y., Liao, P., Dai, X., Chen, H., & Deng, Z. (2020). Automatic human identification from panoramic dental radiographs using the convolutional neural network. *Forensic Science International*, *314*, 110416. https://doi.org/10.1016/j.forsciint.2020.110416
19. Liu, R., Gu, Y., Shen, M., Li, H., Zhang, K., Wang, Q., Wei, X., Zhang, H., Wu, D., Yu, K., Cai, W., Wang, G., Zhang, S., Sun, Q., Huang, P., & Wang, Z. (2020). Predicting postmortem interval based on microbial community sequences and machine learning algorithms. *Environmental Microbiology*, *22*(6), 2273–2291. https://doi.org/10.1111/1462-2920.15000
20. Muñoz Barús, J. I., Febrero-Bande, M., & Cadarso-Suárez, C. (2008). Flexible regression models for estimating postmortem interval (PMI) in forensic medicine. *Statistics in Medicine*, *27*(24), 5026–5038. https://doi.org/10.1002/sim.3319
21. Lu, X.-J., Li, J., Wei, X., Li, N., Dang, L.-H., An, G.-S., Du, Q.-X., Jin, Q.-Q., Cao, J., Wang, Y.-Y., & Sun, J.-H. (2023). A novel method for determining postmortem interval based on the metabolomics of multiple organs combined with ensemble learning techniques. *International Journal of Legal Medicine*, *137*(1), 237–249. https://doi.org/10.1007/s00414-022-02844-8
22. Ortega, R. F., Irurita, J., Campo, E. J. E., & Mesejo, P. (2021). Analysis of the performance of machine learning and deep learning methods for sex estimation of infant individuals from the analysis of 2D images of the ilium. *International Journal of Legal Medicine*, *135*(6), 2659–2666. https://doi.org/10.1007/s00414-021-02660-6
23. Cao, Y., Ma, Y., Yang, X., Xiong, J., Wang, Y., Zhang, J., Qin, Z., Chen, Y., Vieira, D. N., Chen, F., Zhang, J., & Huang, P. (2022). Use of deep learning in forensic sex estimation of virtual pelvic models from the Han population. *Forensic Sciences Research*, *7*(3), 540–549. https://doi.org/10.1080/20961790.2021.2024369
24. Porto, L. F., Correia Lima, L. N., Pinheiro Flores, M. R., Valsecchi, A., Ibanez, O., Machado Palhares, C. E., & De Barros Vidal, F. (2019). Automatic cephalometric landmarks detection on frontal faces: An approach based on supervised learning techniques. *Digital Investigation*, *30*, 108–116. https://doi.org/10.1016/j.diin.2019.07.008

25. Venema, J., Peula, D., Irurita, J., & Mesejo, P. (2023). Employing deep learning for sex estimation of adult individuals using 2D images of the humerus. *Neural Computing and Applications*, *35*(8), 5987–5998. https://doi.org/10.1007/s00521-022-07981-0
26. Gu, M., & Buckley, M. (2018). Semi-supervised machine learning for automated species identification by collagen peptide mass fingerprinting. *BMC Bioinformatics*, *19*(1), 241. https://doi.org/10.1186/s12859-018-2221-3
27. Courtenay, L. A., Yravedra, J., Huguet, R., Aramendi, J., Maté-González, M. Á., González-Aguilera, D., & Arriaza, M. C. (2019). Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *522*, 28–39. https://doi.org/10.1016/j.palaeo.2019.03.007
28. Palanivinayagam, A., Gopal, S. S., Bhattacharya, S., Anumbe, N., Ibeke, E., & Biamba, C. (2021). An optimized machine learning and big data approach to crime detection. *Wireless Communications and Mobile Computing*, *2021*, 1–10. https://doi.org/10.1155/2021/5291528
29. Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. *IEEE Access*, *11*, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344
30. Dakalbab, F., Abu Talib, M., Abu Waraga, O., Bou Nassif, A., Abbas, S., & Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open*, *6*(1), 100342. https://doi.org/10.1016/j.ssaho.2022.100342
31. Sherier, A. J., Woerner, A. E., & Budowle, B. (2022). Determining informative microbial single nucleotide polymorphisms for human identification. *Applied and Environmental Microbiology*, *88*(7), e00052-22. https://doi.org/10.1128/aem.00052-22
32. Ortiz, A. G., Soares, G. H., Da Rosa, G. C., Biazevic, M. G. H., & Michel-Crosato, E. (2021). A pilot study of an automated personal identification process: Applying machine learning to panoramic radiographs. *Imaging Science in Dentistry*, *51*(2), 187. https://doi.org/10.5624/isd.20200324
33. Mitchell, J., Mitchell, S., & Mitchell, C. (2020). Machine learning for determining accurate outcomes in criminal trials. *Law, Probability and Risk*, *19*(1), 43–65. https://doi.org/10.1093/lpr/mgaa003
34. Fernandes, K., Cardoso, J. S., & Astrup, B. S. (2018). A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, *21*(3), 629–640. https://doi.org/10.1007/s10044-018-0694-3
35. Yu, W., Xue, Y., Knoops, R., Yu, D., Balmashnova, E., Kang, X., Falgari, P., Zheng, D., Liu, P., Chen, H., Shi, H., Liu, C., & Zhao, J. (2021). Automated diatom searching in the digital scanning electron microscopy images of drowning cases using the deep neural networks. *International Journal of Legal Medicine*, *135*(2), 497–508. https://doi.org/10.1007/s00414-020-02392-z
36. Garland, J., Hu, M., Kesha, K., Glenn, C., Morrow, P., Stables, S., Ondruschka, B., & Tse, R. (2021). Identifying gross post-mortem organ images using a pretrained convolutional neural network. *Journal of Forensic Sciences*, *66*(2), 630–635. https://doi.org/10.1111/1556-4029.14608

37. Zinnel, L., & Bentil S. A. (2023). Convolutional neural networks for traumatic brain injury classification and outcome prediction. *Health Sciences Review, 9*, 100126.
38. Courville, E., Kazim, S. F., Vellek, J., Tarawneh, O., Stack, J., Roster, K., Roy, J., Schmidt, M., & Bowers, C. (2023). Machine learning algorithms for predicting outcomes of traumatic brain injury: A systematic review and meta-analysis. *Surgical Neurology International*, *14*, 262. https://doi.org/10.25259/SNI_312_2023
39. Bergman, T., Klöden, M., Dreßler, J., & Labudde, D. (2022). Automatic classification of bloodstains with deep learning methods. *KI - Künstliche Intelligenz*, *36*(2), 135–141. https://doi.org/10.1007/s13218-022-00760-y
40. Liu, Y.-Y., Welch, D., England, R., Stacey, J., & Harbison, S. (2020). Forensic STR allele extraction using a machine learning paradigm. *Forensic Science International: Genetics*, *44*, 102194. https://doi.org/10.1016/j.fsigen.2019.102194
41. Marciano, M., & Adelman, J. (n.d.). *(54) Systemand method for inter-species DNA mixture interpretation. Patent.*

A Machine Learning Approach in Toxicological Studies and Analysis of Forensic Exhibits

6

AKANKSHA SINGH KACHHAWAHA, VIJETA KHARE, AHLAD KUMAR AND ATHULYA RAJAN

Introduction

Toxicology is a multidisciplinary field that studies the adverse effects of chemicals, drugs and other substances on living organisms [1, 2]. It involves the identification, characterization and understanding of the mechanisms of the toxicants, diagnosis, management and treatment of individuals who have been exposed to toxic substances [3]. Toxicologists also aim to assess and manage risks associated with exposure to toxic agents development [4, 5] and implementation of regulations and guidelines to protect human health and the environment [6–8]. As shown in Figure 6.1, there are various branches of toxicology, each dealing with a different aspect of toxicants [9]. Environmental toxicology is about understanding the adverse effects of pollutants and toxic substances on the environment and its inhabitants to ultimately mitigate and prevent environmental pollution and its effects on ecosystems and human health [10, 11]. Predictive toxicology develops tools and techniques [12] to predict the toxic effects of chemicals or substances on living organisms in order to develop methods for the pre-evaluation of toxicity in drug development processes, industrial chemical production or exposure to environmental contaminants [13]. Clinical toxicology is about the diagnosis and treatment of poisons [14]. Along with the above, toxicology also finds its utility in the analysis of forensics, for the analysis of exhibits in criminal investigation procedure by combining principles of forensic and analytical toxicology [15, 16].

DOI: 10.4324/9781003449164-7

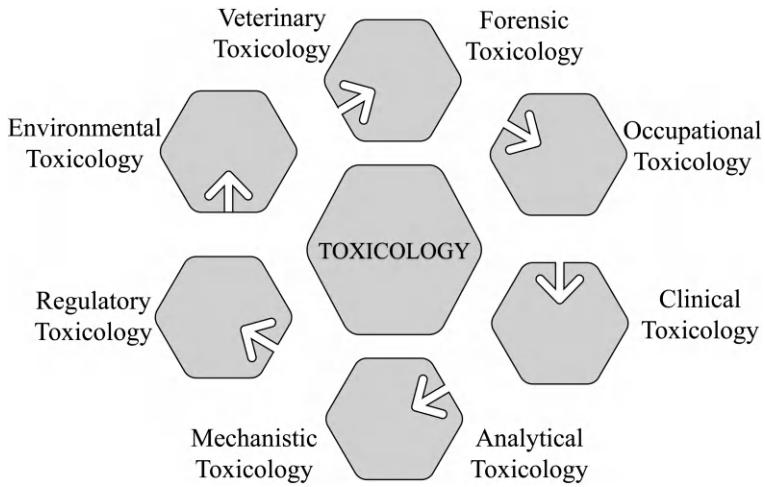

Figure 6.1 Branches of toxicology.

Traditionally, toxicologists have relied on animal testing and empirical studies to assess toxicity [17, 18]. Analysis of toxicants has predominantly relied on extensive sample preparation techniques and chromatographic analysis [19, 20]. However, the advent of machine-learning (ML) and deep-learning (DL) techniques has revolutionized the field of toxicology and facilitated the analysis of forensic exhibits [21, 22]. These advanced computational methods have enabled more accurate predictions of toxicity, faster risk assessment and the identification of novel toxicants [23]. Machine learning is, basically, a subfield of artificial intelligence (AI) that involves the development of algorithms and models that train computers to make predictions or decisions without being exclusively programmed for a particular task [24]. Machine learning is comprised of supervised learning (multiple linear regression, naive Bayes classifier, decision trees, support vector machines, ensemble learning, artificial neural networks, deep neural networks), unsupervised learning (principal component analysis, Kohonen's self-organizing maps) and reinforcement learning [25]. Machine learning techniques have found varied applications in various fields of toxicology [25–27]. In this chapter, we briefly summarize the applications of ML in different branches of toxicology and explore in detail the utility of ML and DL especially in the field of forensic toxicology and analysis of forensic exhibits. In later sections, various classical and modern methods of sample preparation, the general flow of analysis of toxicological exhibits and the role of ML in forensic analysis of the exhibits for qualitative and quantitative purposes have been discussed.

Machine Learning in Predictive Toxicology

ML and DL models have the ability to predict toxicity with high accuracy and subsequently fill the research gaps, as demonstrated in Figure 6.2 [21]. These models are trained on large datasets of chemical and biological data, allowing them to recognize patterns and relationships that might not be apparent through traditional experimentation. Predictive models can estimate toxicity endpoints, such as median lethal dose (LD50) and median inhibitory concentration (IC50) which are crucial for risk assessment and regulatory decisions [28]. It can also predict the harmful effects of substances on living organisms [29] and help in the structure–activity relationship (SAR) by establishing a relationship between the molecule and its biological activity. For such purposes, a variety of ML and DL models are available for use depending on the nature of the data and the problem at hand. A few of the most commonly used models in predictive toxicology are given below [30]:

Random Forests (RF): Random forest is one algorithm which has been widely used in predictive toxicology [31]. For instance, in a recent study, it was used to build classification models for mouse liver toxicity [32]. This algorithm is a combination of multiple decision trees that are merged to put out a single result and is suitable for both classification and regression problems [33].

Support Vector Machines (SVMs): SVMs are used in toxicology for binary classification problems, especially when working with limited samples. They can distinguish toxic from non-toxic compounds [34]

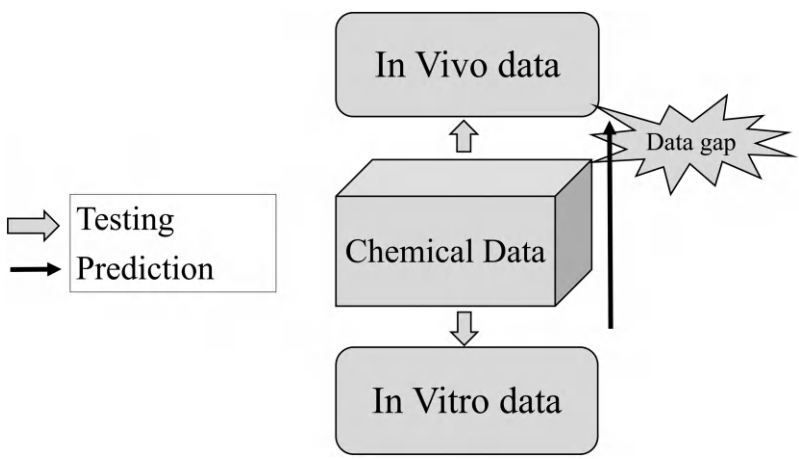

Figure 6.2 Predictive toxicology.

and have been found to be useful in the identification of aquatic toxicity mechanisms of various organic compounds [35].

Gradient Boosting Machines (GBM): Gradient boosting algorithms such as XGBoost and light GBM have become increasingly popular as they are able to handle complex data and provide good results. Gradient boosting combines various base models to form a strong ensemble model, often used for regression and classification problems in toxicology. In a study, GBM along with other methods was used to build a model for predicting the half-life of organic chemicals in humans [36].

Neural Networks: Deep-learning models are also increasingly used for the toxicity prediction and classification of various chemical substances. These models have the ability to learn complex patterns and extract information from the given data. The different types of neural networks commonly used in predictive toxicology include:

- **Feedforward Neural Networks (FNN):** An FNN is one of the most basic types of neural networks wherein the data moves in only one direction through input, hidden and output layers and is used for various toxicity prediction tasks [37].
- **Convolutional Neural Networks (CNN):** CNNs are usually used for image recognition purposes and are thus useful when dealing with image-based toxicity prediction such as molecular structures or biological assays [38].
- **Recurrent Neural Networks (RNN):** RNNs are used when working with sequential data and are suitable for predicting time-dependent toxicity or molecular sequences [39].
- **Graph Neural Networks (GNN):** GNNs are developed specially for graph data and hence, ideal for problems involving molecular graphs or chemical compounds [40].

DeepChem: DeepChem is a library that was created specifically for the discovery of drugs [41] as well as predictive toxicology. It provides pre-built deep learning models and tools tailored for chemical and biological data analysis [42].

Chemoinformatics Models: Chemoinformatics-based models, such as quantitative structure–activity relationship (QSAR) models and pharmacophore modelling are also commonly used for toxicity prediction alongside traditional ML and DL models. These models rely on chemical properties, molecular descriptors and structural features [43].

Ensemble Models: Often, a combination of different models is used to improve the model performance [44]. Some examples of ensemble techniques such as stacking and bagging can be used to combine the predictions from multiple models in order to achieve greater accuracy [45, 46].

Transfer Learning: Through transfer learning, a pre-trained model that has been fine-tuned for a specific toxicology task is re-used on a second related problem and can be particularly advantageous when dealing with limited data and computational resources [47].

Explainable Artificial Intelligence Models (XAI)

As the name suggests, explainable AI models not only make predictions but also give the rationale behind the model's decision. This improves the transparency of the models and makes the predictions more trustworthy, which is crucial in toxicological studies [48].

Ultimately, the choice of the model depends on factors such as the nature of the toxicological task, data availability as well as the desired level of interpretability. Nevertheless, different models can be experimented with to assess which one works best for the problem at hand.

Structure–Activity Relationship

Structure–activity relationship (SAR) studies are used to determine the relationship between the structural characteristics of a chemical compound and its biological activity, such as toxicity. With the increase in chemical libraries and datasets, ML and DL models have been widely used in SAR-based toxicity predictions [49, 50]. This can help identify features that are associated with toxicity, which is especially necessary when developing drugs or chemicals for use [51]. ML and DL techniques are currently utilized in SAR analysis in the following ways:

Feature Extraction and Selection

- **Molecular Descriptors:** Molecular descriptors are numerical values that represent chemical compounds. These descriptors are derived from various properties of the molecules from simple ones such as molecular formula and topology to more complex ones such as 3D structure [52]. These descriptors are used as input variables for the ML model in order to make predictions.

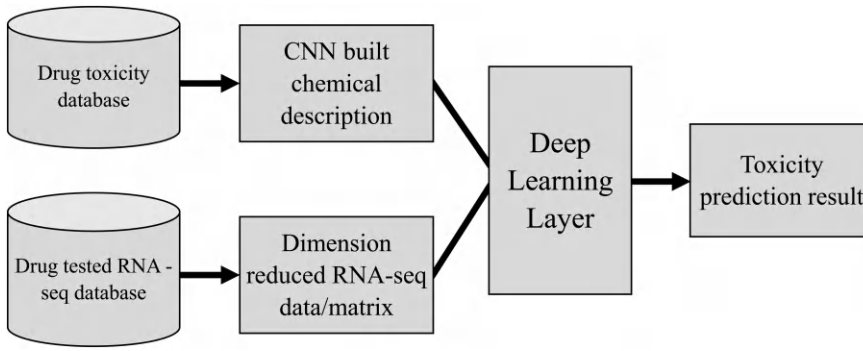

Figure 6.3 Deep learning–based predictive toxicity prediction.

- **Deep Learning for Feature Extraction:** DL models have the ability to learn and extract relevant features from the input data, such as molecular structures, without requiring manual feature engineering. Techniques such as graph convolutional neural networks (GCNs) and recurrent neural networks (RNNs) can process molecular graphs or sequences directly, maintaining valuable structural information (Figure 6.3) [53].

Data Preparation and Augmentation

- **Data Preprocessing:** The most important part of this initial step includes cleaning and standardizing molecular data, handling missing values, normalizing features and ensuring that the data is consistent.
- **Data Augmentation:** In case of limited data or samples, data augmentation techniques can be considered for generating new data points with slight variations, thus increasing the size and diversity of the dataset.

Model Training

- **Traditional ML Models:** ML algorithms such as random forests and support vector machines, among others, are trained on datasets of compounds with known activity levels. Thereafter, these models learn to map the relationship between molecular descriptors and the corresponding biological activities [54].
- **Deep Learning Models:** SAR models based on DL techniques such as GNNs can be used to predict values on various chemical and toxicological properties such as activity labels or continuous activity values [55].

Predictive Modeling

- **Binary Classification:** ML and DL models can be used for binary classification tasks, such as predicting whether a molecule is active or inactive against a specific target [56].
- **Regression Analysis**: Regression models are used for quantitative SAR (QSAR) analysis to predict the activity levels of chemical compounds and are able to assess the contribution of various features with respect to certain parameters [57].

Model Interpretability

- **Feature Importance:** As stated earlier, ML and DL models are capable of evaluating the contribution of each feature towards making the final prediction [58].
- **Attention Mechanisms:** Some DL models, for instance, GNNs use attention mechanisms to point out specific substructures or atoms within molecules, responsible for bioactivity [59, 60].

Model Validation

- **Cross-Validation:** Cross-validation techniques are used to evaluate the performance of the model on the portion of the dataset that was not used to train the model. This is done in order to measure the model's ability to generalize new unseen data.
- **External Validation:** Another type of validation that can be done is external validation in which the model is validated against similar external datasets [61].
- **Chemoinformatics Databases:** There are various databases available, such as ChEMBL and PubChem [62], that are utilized by ML and DL models for gathering chemical and biological data which is useful for SAR analysis.

Virtual Screening: ML and DL models today have employed numerous existing chemical libraries for screening drugs and toxic compounds, which in turn, has been a huge improvement for SAR analysis and predictions. Overall, these techniques have increased the speed of such analysis through automatic feature extraction, capturing complex patterns [63] and also improving the understanding of how chemical compounds work [64].

Machine Learning in Clinical Toxicology

In the pharmaceutical industry, predicting adverse events associated with drugs is critical. ML and DL models can analyze large-scale clinical and pharmacovigilance data to identify patterns that suggest potential adverse events. This helps in postmarketing surveillance and can lead to safer drug usage. A study evaluated machine-learning algorithms for predicting seizure due to acute tramadol poisoning using routine demographic, clinical and paraclinical data, and important predictive variables [65]. A support vector machine–based method along with a feature selection technique that can be applied to accurately predict paraquat poisoning toxicity is also reported [66].

Machine Learning in Environmental Monitoring and Toxicology

ML and DL are also applicable in environmental toxicology. These techniques can help in predicting the environmental impact of chemicals, assessing their potential for bioaccumulation and understanding their effects on ecosystems [67]. This knowledge aids in the regulation of pollutants and further helps in reducing their effects on the ecosystem as well as human health [68]. ML and DL are used in environmental toxicology specifically for the purpose of predicting the toxicity of various chemicals to aquatic organisms (Figure 6.4). As a result, they are able to identify potentially harmful substances and prioritize the same for further testing. They are also a great tool for predicting bioavailability, considering factors such as soil properties, climate and microbial activity [69]. In the case of environmental monitoring,

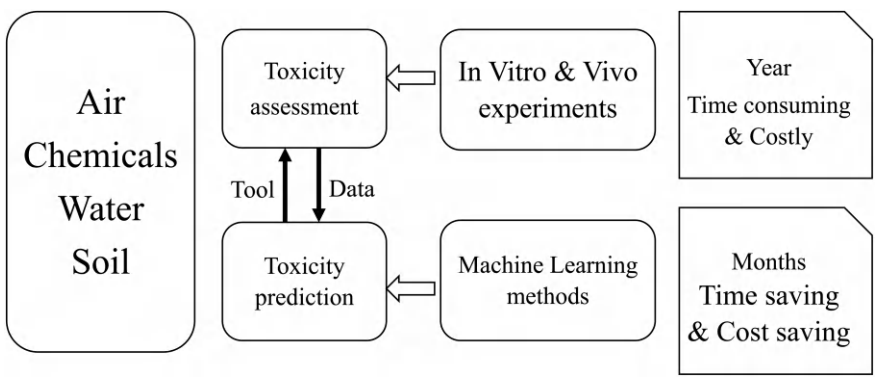

Figure 6.4 Predictive environmental toxicology [20].

these techniques help with (a) sensor data analysis: as more and more environmental sensor data are made available, ML algorithms can be used for processing such data from environmental sensors to assess and predict factors such as air quality [70]. They can also be used for real-time monitoring of water quality by detecting anomalies, which further aids in early warning systems [71], (b) spatial analysis: a GIS (geographic information system) combined with ML can analyze spatial data to understand the distribution of pollutants and assess contamination levels in specific areas [72, 73], (c) species sensitivity distributions (SSDs): ML models can analyze SSDs to predict the concentration of a pollutant that can harm a certain percentage of species in an ecosystem, for instance, the harmful effects of pesticides to aquatic life [74]. In general, ML can aid in the assessment of ecological risks by considering multiple stressors and their effects on the ecosystem [75].

Machine learning in Biological Monitoring and Toxicogenomics

Bioindicator Identification: With the help of ML models, the reaction of bioindicators to changes in the environment, or rather anomalies can be detected. These anomalies can then be used to spot environmental problems early on [76].

Metagenomics and Metatranscriptomics: DL techniques can analyze metatranscript data to understand the composition of microbial communities and their response to environmental stressors [77].

Gene Expression Analysis: DL models can also examine gene expression data to find which genes and biological pathways are reactive to environmental toxins [78].

Omics Integration: The integration of different omics, namely, genomics, transcriptomics, proteomics and metabolomics data with the help of ML can provide a comprehensive understanding of how organisms respond to environmental stressors [79].

Data Integration and Fusion

- **Multi-Source Data Fusion**: By integrating data from different sources such as environmental sensors and chemical databases, ML techniques can create broad environmental models [80].
- **Time-Series Analysis:** ML models can analyze time-series data to detect long-term trends and seasonal variations in environmental parameters.

Machine Learning in Risk Assessment

ML and DL play a crucial role in risk management in the field of environmental assessment and toxicology. As these techniques continue to be increasingly used for the detection and identification of chemicals or pollutants, they have also become instrumental in analyzing the risks imposed by such substances. Additionally, the output of such assessments acts as guidance for decision-makers and can even be used to notify authorities about potential environmental emergencies such as harmful algal blooms [81, 82] or air pollution episodes [83]. In the case of environmental toxicology, these techniques are also helping researchers and environmental agencies protect ecosystems and human health [84, 85]. They are particularly valuable for handling the complexity and volume of data associated with environmental assessments. In a recent study, a deep neural network (DNN) model was developed for the risk assessment of a drive-off scenario involved in an oil and gas drilling rig. It was an attempt at addressing the challenges of industrial risk assessment through machine learning and was able to achieve considerable accuracy [86]. Regarding climate change risk assessment, a variety of ML algorithms have been applied. Among them, the most commonly used are decision trees, random forests and artificial neural networks. These algorithms are usually applied for the assessment of flood and landslide risk events. Also, the application of ML to deal with remote sensing data is consistent and effective [87].

Forensic Toxicology and Analysis of Forensic Exhibits

Forensic toxicology focuses on the analysis of substances in biological specimens to aid criminal investigations for justice delivery. The primary goal of forensic toxicologists is to identify and quantify drugs, chemicals or other toxic substances in biological samples obtained from individuals involved in legal cases. This field plays a crucial role in various legal contexts, including criminal investigations, postmortem examinations and workplace- or traffic-related incidents. Forensic toxicologists use various sample preparation and analytical techniques (Figure 6.5), including chromatography, mass spectrometry and immunoassays, to identify and quantify effective analysis of the toxicants when eventually the extracts are quantified using instrumental methods. Based on the physicochemical characteristics of the compounds, there are different extraction techniques. However as non-volatile organic compounds are predominantly found in the visceral samples, we will restrict ourselves to discussing methods pertaining to the category of non-volatile organic compounds.

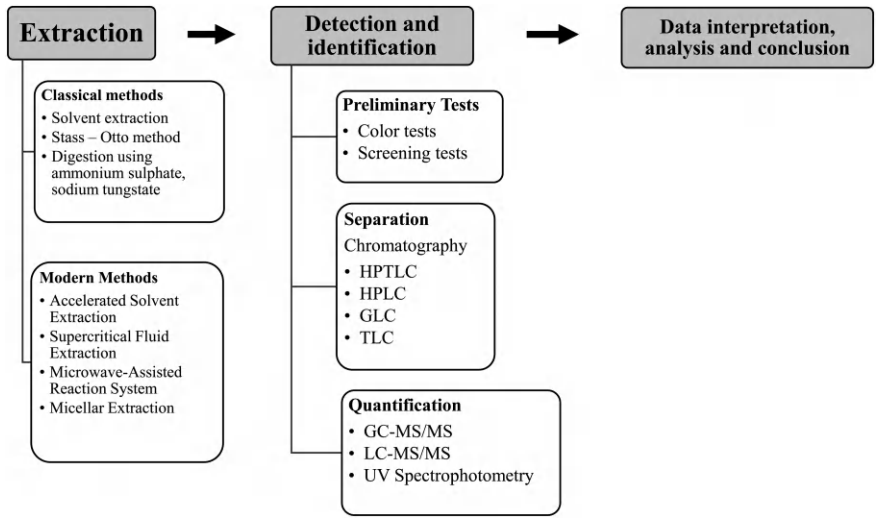

Figure 6.5 Analytical flow of toxicants.

Classical and Modern Methods of Sample Preparation

- **Stas–Otto Method:** A method utilized to extract and identify alkaloids. It has the names of the two chemists, Frederik Stas and Carl Otto, who created it. This approach is frequently used in forensic and toxicological laboratories for the detection and identification of alkaloids in diverse samples. It has historically been used for the examination of alkaloids from plant materials, such as medicinal herbs [88].
- **Solvent Extraction:** It is based on the distribution of alkaloid bases between acid or aqueous solution and immiscible organic solvent [89]. Distribution of a solute between two immiscible liquid phases in contact with each other, that is, two-phase distribution of a solute. The common solvents used for solvent extraction are n-hexane, toluene and acetonitrile.

Solvent extraction is one of the most common methods used in forensic toxicology related to biological matrices [16]. Drugs and metabolites are separated between the two phases based on their polarity. Buffering or pH adjustment of the specimen allows for selectivity in extraction techniques for acidic or basic drugs.

- **Solid Phase Extraction:** Solid phase extraction (SPE) was introduced as an alternative to solvent extraction in the mid-1970s. It was used for separation, preconcentration and solvent exchange of solutes for

solution [90]. This technique enables extraction, clean up and concentration of analytes before their quantification. SPE uses an adsorbent contained in a cartridge device or on a disk to adsorb selective species from the solution [91]. Different separation chemistries can be used with SPE techniques to extract various target analytes such as normal phase, reverse phase and ion exchange (both anion and cation) modalities [92]. Commercial cartridges used for SPE have 1–10 ml capacities and are discarded after use [93]. Analytes concentration can be achieved better with SPE than solvent–solvent extraction and usage of an SPE column provides a quicker and safer alternative [16]. SPE is reliable, cost-effective and environmentally friendly. SPE can be automated desirable in high throughput laboratories. Low solvent consumption can be achieved by miniaturization of SPE columns to mini-columns/cartridges [92].

- **Solid Phase Microextraction:** SPME, developed in 1990, is defined as a miniature version and equilibrium technique where the extraction phase is small in comparison to the sample volume [94]. It is an extraction method where analytes are directly adsorbed from the sample onto a fused silica fibre that has been coated with the suitable stationary phase. This method is used to extract organic chemicals from aqueous samples [16]. SPME can be used for volatile substances, drugs in biological specimens, etc. It uses a relatively smaller sorptive surface area and the sorptive fibre adsorbs analytes from gases, liquid or semi-solid samples upon exposure. Fibre coating is selected depending on the properties of the analytes. The commonly used are polydimethylsiloxane (PDMS), polyacrylate (PA), carboxen/polydimethylsiloxane (CAR–PDMS), etc.

By controlling the thickness and polarity of the fibre coating, having consistent sampling time and other parameters, SPME ensures high throughput and quantifiable results even at low analyte concentrations. SPME is solvent-free, easy to automate and its fibres are reusable and inexpensive.

- **Accelerated Solvent Extraction (ASE):** ASE is an extraction technique used for treating organic compounds from solids as well as semi-solid samples using liquid solvents. Liquid solvents are organic solvents used at high pressures and temperatures above their boiling points [95]. ASE requires less time and fewer solvents and gives better analyte recovery than traditional methods of extraction. The entire extraction process is automated and conducted in minutes for fast and easy extraction of multiple samples with very little solvent consumption.

Instrumental Analysis

Analytical instruments such as gas chromatography and liquid chromatography when combined with mass spectrometry are used for the accurate identification and quantification of target analytes. These provide high sensitivity and selectivity, enabling the detection of trace amounts of compounds.

Machine Learning in Analytical Aspects of Forensic Toxicology

Traditional toxicology tests are time-consuming and costly. ML and DL algorithms can be applied to high-throughput screening assays, enabling the rapid assessment of thousands of chemicals or compounds simultaneously. This not only reduces the need for animal testing but also expedites the identification of toxic substances. A recent study utilized ML in the simultaneous detection of urine sample manipulation and prohibited drugs in a single run using retention time alignment within progenesis QI and artificial neural network [96]. In a proof of concept study, a group clearly demonstrated the potential of machine learning in the analysis of high-resolution mass spectrometry (HRMS) enabling data-independent acquisition (DIA) data. A machine-learning model was evaluated using training, validation and test sets of solvent and whole blood samples containing drugs (of abuse) common in forensic toxicology using feedforward neural network model architecture [97]. In another study metabolomics and machine learning were used for the determination of sudden cardiac death, which in forensic practice is one of the difficult tasks by combining metabolic characteristics from specimens of cardiac blood and cardiac muscle [98].

Challenges and Future Directions

While ML and DL hold great promise in the field of toxicology, there are several challenges to overcome. These include the need for high-quality, standardized data, interpretability of models and ethical considerations surrounding the use of AI in toxicological research. In the future, toxicologists will likely see the integration of multi-omics data (genomics, proteomics, metabolomics) into ML and DL models, enabling a more holistic understanding of toxicity mechanisms. Additionally, the development of explainable AI techniques will help address model interpretability concerns.

References

1. C. D. Klaassen, J. B. Watkins, L. J. Casarett, and J. Doull, Eds., *Casarett & Doull's Essentials of Toxicology*, 4th ed. New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, New Delhi, Milan, Singapore and Sydney, Toronto: McGraw Hill, 2021.
2. J. Timbrell and F. A. Barile, *Introduction to Toxicology*. Boca Raton: CRC Press, 2023.
3. T. L. Guidotti, "Toxicology," in *Essentials of Medical Geology*, O. Selinus, Ed. Dordrecht: Springer Netherlands, 2013, pp. 597–609. doi: 10.1007/978-94-007-4375-5_26.
4. J. V. Rodricks, *Calculated Risks: The Toxicity and Human Health Risks of Chemicals in Our Environment*. New York: Cambridge University Press, 2006.
5. P. L. Williams, R. C. James, and S. M. Roberts, Eds., *Principles of Toxicology: Environmental and Industrial Applications*, 2nd ed. New York: Wiley, 2000.
6. J. Descotes and F. Testud, "Toxicovigilance: A new approach for the hazard identification and risk assessment of toxicants in human beings," *Toxicol. Appl. Pharmacol.*, vol. 207, pp. 599–603, Oct. 2005, doi: 10.1016/j.taap.2005.02.019.
7. U. Gundert-Remy, H. Barth, A. Bürkle, G. H. Degen, and R. Landsiedel, "Toxicology: A discipline in need of academic anchoring—the point of view of the German Society of Toxicology," *Arch. Toxicol.*, vol. 89, no. 10, pp. 1881–1893, 2015, doi: 10.1007/s00204-015-1577-7.
8. "What does a toxicologist do?" Accessed: Jan. 18, 2024. [Online]. Available: https://publichealth.tulane.edu/blog/what-does-a-toxicologist-do/
9. C. Winder and N. H. Stacey, *Occupational Toxicology*. Boca Raton: CRC Press, 2004.
10. I. S. Chadwick J., *Principles of Environmental Toxicology*. London: CRC Press, 2017. doi: 10.1201/9781315273785.
11. W. Landis, R. Sofield, M.-H. Yu, and W. G. Landis, *Introduction to Environmental Toxicology: Impacts of Chemicals Upon Ecological Systems*, 3rd ed. Boca Raton: CRC Press, 2003.
12. C. Helma, *Predictive Toxicology*. Boca Raton: CRC Press, 2005.
13. B. W. Brooks et al., "Toxicology advances for 21st century chemical pollution," *One Earth*, vol. 2, no. 4, pp. 312–316, Apr. 2020, doi: 10.1016/j.oneear.2020.04.007.
14. P. K. Gupta, *Fundamentals of Toxicology: Essential Concepts and Applications*. India, Amsterdam and Boston: BS Publications; Elsevier/AP, Academic Press, 2016.
15. "Forensic toxicology and its relevance with criminal justice delivery system in India," *Forensic Res. Criminol. Int. J.*, vol. 4, no. 4, Apr. 2017, doi: 10.15406/frcij.2017.04.00121.
16. "Toxicology manual 2021.pdf." Accessed: Jul. 29, 2023. [Online]. Available: http://dfs.nic.in/pdfs/toxicology%20manual%202021.pdf
17. "Animals-chapter-9-animal-use-in-toxicity-studies.pdf." Accessed: Jan. 18, 2024. [Online]. Available: https://www.nuffieldbioethics.org/wp-content/uploads/Animals-Chapter-9-Animal-Use-in-Toxicity-Studies.pdf

18. G. A. Van Norman, "Limitations of animal studies for predicting toxicity in clinical trials," *JACC Basic Transl. Sci.*, vol. 4, no. 7, pp. 845–854, Nov. 2019, doi: 10.1016/j.jacbts.2019.10.008.
19. O. H. Drummer, "Chromatographic screening techniques in systematic toxicological analysis," *J. Chromatogr. B. Biomed. Sci. App.*, vol. 733, no. 1, pp. 27–45, Oct. 1999, doi: 10.1016/S0378-4347(99)00265-0.
20. V. Pérez-Fernández, L. Mainero Rocca, P. Tomai, S. Fanali, and A. Gentili, "Recent advancements and future trends in environmental analysis: Sample preparation, liquid chromatography and mass spectrometry," *Anal. Chim. Acta*, vol. 983, pp. 9–41, Aug. 2017, doi: 10.1016/j.aca.2017.06.029.
21. W. Guo et al., "Review of machine learning and deep learning models for toxicity prediction," *Exp. Biol. Med.*, vol. 248, no. 21, pp. 1952–1973, Nov. 2023, doi: 10.1177/15353702231209421.
22. T. D. Wankhade, S. W. Ingale, P. M. Mohite, and N. J. Bankar, "Artificial intelligence in forensic medicine and toxicology: The future of forensic medicine," *Cureus*, Aug. 2022, doi: 10.7759/cureus.28376.
23. R. J. Kavlock et al., "Computational toxicology—a state of the science mini review," *Toxicol. Sci.*, vol. 103, no. 1, pp. 14–27, May 2008, doi: 10.1093/toxsci/kfm297.
24. J. M. Helm et al., "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Curr. Rev. Musculoskelet. Med.*, vol. 13, no. 1, pp. 69–76, Feb. 2020, doi: 10.1007/s12178-020-09600-8.
25. Z. Lin and W.-C. Chou, "Machine learning and artificial intelligence in toxicological sciences," *Toxicol. Sci.*, vol. 189, no. 1, pp. 7–19, Jul. 2022, doi: 10.1093/toxsci/kfac075.
26. M. Rigatti, S. Carreiro, and E. W. Boyer, "Chapter 42 - Machine learning applications in toxicology," in *Artificial Intelligence in Clinical Practice*, C. Krittanawong, Ed. Academic Press, 2024, pp. 377–382. doi: 10.1016/B978-0-443-15688-5.00005-X.
27. M. W. H. Wang, J. M. Goodman, and T. E. H. Allen, "Machine learning in predictive toxicology: Recent applications and future directions for classification models." *Chemical Research in Toxicology,* vol. 34, no. 2, pp. 217–239, 2020.
28. C. N. Cavasotto and V. Scardino, "Machine learning toxicity prediction: Latest advances by toxicity end point," *ACS Omega*, vol. 7, no. 51, pp. 47536–47546, Dec. 2022, doi: 10.1021/acsomega.2c05693.
29. Y. Zhou, Y. Wang, W. Peijnenburg, M. G. Vijver, S. Balraadjsing, and W. Fan, "Using machine learning to predict adverse effects of metallic nanomaterials to various aquatic organisms," *Environ. Sci. Technol.*, vol. 57, no. 46, pp. 17786–17795, Nov. 2023, doi: 10.1021/acs.est.2c07039.
30. G. Xu, X. Teng, X.-H. Gao, L. Zhang, H. Yan, and R.-Q. Qi, "Advances in machine learning-based bacteria analysis for forensic identification: Identity, ethnicity, and site of occurrence," *Front. Microbiol.*, vol. 14, 2023, Accessed: Jan. 18, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1332857
31. P. Mistry, D. Neagu, P. R. Trundle, and J. D. Vessey, "Using random forest and decision tree models for a new vehicle prediction approach in computational toxicology," *Soft Comput.*, vol. 20, no. 8, pp. 2967–2979, Aug. 2016, doi: 10.1007/s00500-015-1925-9.

32. X.-W. Zhu, Y.-J. Xin, and Q.-H. Chen, "Chemical and in vitro biological information to predict mouse liver toxicity using recursive random forests," *SAR QSAR Environ. Res.*, vol. 27, no. 7, pp. 559–572, Jul. 2016, doi: 10.1080/1062936X.2016.1201142.
33. A. Koutsoukas, J. St. Amand, M. Mishra, and J. Huan, "Predictive toxicology: Modeling chemical induced toxicological response combining circular fingerprints with random forest and support vector machine," *Front. Environ. Sci.*, vol. 4, Mar. 2016, doi: 10.3389/fenvs.2016.00011.
34. "Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines," *J. Biomed. Inform.*, vol. 68, pp. 132–149, Apr. 2017, doi: 10.1016/j.jbi.2017.03.002.
35. O. Ivanciuc, "Support vector machine identification of the aquatic toxicity mechanism of organic compounds," . *Internet Electronic Journal of Molecular Design*, vol. 1, pp. 157–172, 2002.
36. J. Lu et al., "Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine," *Biochim. Biophys. Acta BBA - Gen. Subj.*, vol. 1860, no. 11, Part B, pp. 2664–2671, Nov. 2016, doi: 10.1016/j.bbagen.2016.05.019.
37. Y. Meng and B.-L. Lin, "A feed-forward artificial neural network for prediction of the aquatic ecotoxicity of alcohol ethoxylate," *Ecotoxicol. Environ. Saf.*, vol. 71, pp. 172–86, Aug. 2007, doi: 10.1016/j.ecoenv.2007.06.011.
38. E. Asilar, J. Hemmerich, and G. F. Ecker, "Image based liver toxicity prediction," *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1111–1121, Mar. 2020, doi: 10.1021/acs.jcim.9b00713.
39. H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discov. Today*, vol. 23, no. 6, pp. 1241–1250, Jun. 2018, doi: 10.1016/j.drudis.2018.01.039.
40. M. E. Markou, "Explainable AI for predictive toxicology." Master thesis, University of Oslo, 2022.
41. H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS Cent. Sci.*, vol. 3, no. 4, pp. 283–293, Apr. 2017, doi: 10.1021/acscentsci.6b00367.
42. "Deepchem/deepchem," *deepchem*, Jan. 24, 2024. Accessed: Jan. 24, 2024. [Online]. Available: https://github.com/deepchem/deepchem
43. F. Fan, D. Toledo Warshaviak, H. K. Hamadeh, and R. T. Dunn, "The integration of pharmacophore-based 3D QSAR modeling and virtual screening in safety profiling: A case study to identify antagonistic activities against adenosine receptor, A2A, using 1,897 known drugs," *PLOS ONE*, vol. 14, no. 1, p. e0204378, Jan. 2019, doi: 10.1371/journal.pone.0204378.
44. I. I. Baskin, "Machine learning methods in computational toxicology," in *Computational Toxicology*, vol. 1800, O. Nicolotti, Ed., in *Methods in Molecular Biology*, vol. 1800. New York, NY: Springer, 2018, pp. 119–139. doi: 10.1007/978-1-4939-7899-1_5.
45. L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.
46. L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, Jul. 1996, doi: 10.1007/BF00117832.

47. X. Zhen et al., "Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: A feasibility study," *Phys. Med. Biol.*, vol. 62, no. 21, pp. 8246–8263, Oct. 2017, doi: 10.1088/1361-6560/aa8d09.
48. I. R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?," *Comput. Methods Programs Biomed.*, vol. 212, p. 106415, Nov. 2021, doi: 10.1016/j.cmpb.2021.106415.
49. G. Idakwo, S. Thangapandian, J. Luttrell, Z. Zhou, C. Zhang, and P. Gong, "Deep learning-based structure-activity relationship modeling for multi-category toxicity classification: A case study of 10K Tox21 chemicals with high-throughput cell-based androgen receptor bioassay data," *Front. Physiol.*, vol. 10, 2019, Accessed: Jan. 24, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2019.01044
50. L. K. Vora, A. D. Gholap, K. Jetha, R. R. S. Thakur, H. K. Solanki, and V. P. Chavda, "Artificial intelligence in pharmaceutical technology and drug delivery design," *Pharmaceutics*, vol. 15, no. 7, p. 1916, Jul. 2023, doi: 10.3390/pharmaceutics15071916.
51. M. Staszak, K. Staszak, K. Wieszczycka, A. Bajek, K. Roszkowski, and B. Tylkowski, "Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship," *WIREs Comput. Mol. Sci.*, vol. 12, no. 2, p. e1568, 2022, doi: 10.1002/wcms.1568.
52. A. Cherkasov et al., "QSAR modeling: Where have you been? Where are you going to?," *J. Med. Chem.*, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/jm4004285.
53. J. Mao et al., "Comprehensive strategies of machine-learning-based quantitative structure-activity relationship models," *iScience*, vol. 24, no. 9, p. 103052, Sep. 2021, doi: 10.1016/j.isci.2021.103052.
54. P. Carracedo-Reboredo et al., "A review on machine learning approaches and trends in drug discovery," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, Jan. 2021, doi: 10.1016/j.csbj.2021.08.011.
55. H. Wang, W. Liu, and J. Chen, "Chapter 11 - QSAR modeling based on graph neural networks," in *QSAR in Safety Evaluation and Risk Assessment*, H. Hong, Ed. Academic Press, 2023, pp. 139–151. doi: 10.1016/B978-0-443-15339-6.00012-6.
56. S. K. Kwofie, K. Agyenkwa-Mawuli, E. Broni, W. A. Miller III, and M. D. Wilson, "Prediction of antischistosomal small molecules using machine learning in the era of big data," *Mol. Divers.*, vol. 26, no. 3, pp. 1597–1607, Jun. 2022, doi: 10.1007/s11030-021-10288-2.
57. C. W. Yap, H. Li, Z. L. Ji, and Y. Z. Chen, "Regression methods for developing QSAR and QSPR models to predict compounds of specific pharmacodynamic, pharmacokinetic, and toxicological properties," *Mini Rev. Med. Chem.*, vol. 7, no. 11, pp. 1097–1107, Nov. 2007, doi: 10.2174/138955707782331696.
58. G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 5, p. e1379, Sep. 2020, doi: 10.1002/widm.1379.

59. "Sci-Hub | A review on the attention mechanism of deep learning. Neurocomputing, 452, 48–62 | 10.1016/j.neucom.2021.03.091." Accessed: Jan. 24, 2024. [Online]. Available: https://sci-hub.se/https://doi.org/10.1016/j.neucom.2021.03.091
60. Z. Xiong et al., "Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism," *J. Med. Chem.*, vol. 63, no. 16, pp. 8749–8760, Aug. 2020, doi: 10.1021/acs.jmedchem.9b00959.
61. P. Gramatica, "External evaluation of QSAR models, in addition to cross-validation: Verification of predictive capability on totally new chemicals," *Mol. Inform.*, vol. 33, no. 4, pp. 311–314, Apr. 2014, doi: 10.1002/minf.201400030.
62. T. R. Lane, D. H. Foil, E. Minerali, F. Urbina, K. M. Zorn, and S. Ekins, "Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery," *Mol. Pharm.*, vol. 18, no. 1, pp. 403–415, Jan. 2021, doi: 10.1021/acs.molpharmaceut.0c01013.
63. I. Zafar et al., "Reviewing methods of deep learning for intelligent healthcare systems in genomics and biomedicine," *Biomed. Signal Process. Control*, vol. 86, p. 105263, Sep. 2023, doi: 10.1016/j.bspc.2023.105263.
64. A. Srinivas Reddy, S. Priyadarshini Pati, P. Praveen Kumar, H.N. Pradeep, and G. Narahari Sastry, "Virtual screening in drug discovery - A computational perspective," *Curr. Protein Pept. Sci.*, vol. 8, no. 4, pp. 329–351, Aug. 2007, doi: 10.2174/138920307781369427.
65. B. Behnoush, E. Bazmi, S. Nazari, S. Khodakarim, M. Looha, and H. Soori, "Machine learning algorithms to predict seizure due to acute tramadol poisoning," *Hum. Exp. Toxicol.*, vol. 40, no. 8, pp. 1225–1233, Aug. 2021, doi: 10.1177/0960327121991910.
66. H. Chen et al., "An effective machine learning approach for prognosis of paraquat poisoning patients using blood routine indexes," *Basic Clin. Pharmacol. Toxicol.*, vol. 120, no. 1, pp. 86–96, 2017, doi: 10.1111/bcpt.12638.
67. T. H. Miller, M. D. Gallidabino, J. I. MacRae, S. F. Owen, N. R. Bury, and L. P. Barron, "Prediction of bioconcentration factors in fish and invertebrates using machine learning," *Sci. Total Environ.*, vol. 648, pp. 80–89, Jan. 2019, doi: 10.1016/j.scitotenv.2018.08.122.
68. S. Zhong et al., "Machine learning: New ideas and tools in environmental science and engineering," *Environ. Sci. Technol.*, p. acs.est.1c01339, Aug. 2021, doi: 10.1021/acs.est.1c01339.
69. S. Cipullo, B. Snapir, G. Prpich, P. Campo, and F. Coulon, "Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models," *Chemosphere*, vol. 215, pp. 388–395, Jan. 2019, doi: 10.1016/j.chemosphere.2018.10.056.
70. The Department of Computer Engineering, San Jose State University, USA, G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *Int. J. Environ. Sci. Dev.*, vol. 9, no. 1, pp. 8–16, 2018, doi: 10.18178/ijesd.2018.9.1.1066.
71. E. El-Shafeiy, M. Alsabaan, M. I. Ibrahem, and H. Elwahsh, "Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique," *Sensors*, vol. 23, no. 20, Art. no. 20, Jan. 2023, doi: 10.3390/s23208613.

72. S. Bhattacharyya et al., "Microplastics, their toxic effects on living organisms in soil biota and their fate: An appraisal," 2022, pp. 405–420. doi: 10.1007/978-3-031-09270-1_17.
73. "IJERPH | Free full-text | Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques." Accessed: Jan. 24, 2024. [Online]. Available: https://www.mdpi.com/1660-4601/8/1/75
74. "Sci-Hub | Environmental risk assessment of pesticides in the River Madre de Dios, Costa Rica using PERPEST, SSD, and msPAF models. *Environ. Sci. Pollut. Res.*, 25(14), 13254–13269 | 10.1007/s11356-016-7375-9." Accessed: Jan. 24, 2024. [Online]. Available: https://sci-hub.se/https://doi.org/10.1007/s11356-016-7375-9
75. C. Simeoni et al., "Evaluating the combined effect of climate and anthropogenic stressors on marine coastal ecosystems: Insights from a systematic review of cumulative impact assessment approaches," *Sci. Total Environ.*, vol. 861, p. 160687, Feb. 2023, doi: 10.1016/j.scitotenv.2022.160687.
76. A. N. Grekov, A. A. Kabanov, E. V. Vyshkvarkova, and V. V. Trusevich, "Anomaly detection in biological early warning systems using unsupervised machine learning," *Sensors*, vol. 23, no. 5, Art. no. 5, Jan. 2023, doi: 10.3390/s23052687.
77. J. Asante and J. Osei Sekyere, "Understanding antimicrobial discovery and resistance from a metagenomic and metatranscriptomic perspective: Advances and applications," *Environ. Microbiol. Rep.*, vol. 11, no. 2, pp. 62–86, Apr. 2019, doi: 10.1111/1758-2229.12735.
78. "Nanomaterials | Free full-text | Transcriptomics in toxicogenomics, part III: Data modelling for risk assessment." Accessed: Jan. 24, 2024. [Online]. Available: https://www.mdpi.com/2079-4991/10/4/708
79. J. Manochkumar, A. K. Cherukuri, R. S. Kumar, A. I. Almansour, S. Ramamoorthy, and T. Efferth, "A critical review of machine-learning for 'multi-omics' marine metabolite datasets," *Comput. Biol. Med.*, vol. 165, p. 107425, Oct. 2023, doi: 10.1016/j.compbiomed.2023.107425.
80. S. Salcedo-Sanz et al., "Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources," *Inf. Fusion*, vol. 63, pp. 256–272, Nov. 2020, doi: 10.1016/j.inffus.2020.07.004.
81. I. Busari, D. Sahoo, R. Harmel, and B. Haggard, "A review of machine learning models for harmful algal bloom monitoring in freshwater systems," *J. Nat. Resour. Agric. Ecosyst.*, vol. 1, pp. 63–76, Nov. 2023, doi: 10.13031/jnrae.15647.
82. J. H. Kim et al., "Machine learning-based early warning level prediction for cyanobacterial blooms using environmental variable selection and data resampling," *Toxics*, vol. 11, no. 12, p. 955, Nov. 2023, doi: 10.3390/toxics11120955.
83. J. Wang, W. Xu, Y. Zhang, and J. Dong, "A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization," *Chaos Solitons Fractals*, vol. 158, p. 112098, May 2022, doi: 10.1016/j.chaos.2022.112098.
84. T. H. Miller et al., "Machine learning for environmental toxicology: A call for integration and innovation," *Environ. Sci. Technol.*, vol. 52, no. 22, pp. 12953–12955, Nov. 2018, doi: 10.1021/acs.est.8b05382.

85. C. Schür, L. Gasser, F. Perez-Cruz, K. Schirmer, and M. Baity-Jesi, "A benchmark dataset for machine learning in ecotoxicology," *Sci. Data*, vol. 10, no. 1, Art. no. 1, Oct. 2023, doi: 10.1038/s41597-023-02612-2.
86. N. Paltrinieri, L. Comfort, and G. Reniers, "Learning about risk: Machine learning for risk assessment," *Saf. Sci.*, vol. 118, pp. 475–486, Oct. 2019, doi: 10.1016/j.ssci.2019.06.001.
87. F. Zennaro et al., "Exploring machine learning potential for climate change risk assessment," *Earth-Sci. Rev.*, vol. 220, p. 103752, Sep. 2021, doi: 10.1016/j.earscirev.2021.103752.
88. R. Singh, "Chronology of preceding medico-legal practices with reference to post-mortem forensic toxicology," *Forensic Sci. Int. Rep.*, vol. 5, p. 100275, Jul. 2022, doi: 10.1016/j.fsir.2022.100275.
89. S. Mistry, "Extraction method of alkaloids," *Solution Parmacy*. Accessed: Jul. 29, 2023. [Online]. Available: https://solutionpharmacy.in/extraction-method-of-alkaloids/
90. R. Urkude, V. Dhurvey, and S. Kochhar, "Pesticide residues in beverages," in *Quality Control in the Beverage Industry*, Elsevier, 2019, pp. 529–560. doi: 10.1016/B978-0-12-816681-9.00015-1.
91. "Solid-phase extraction," *Chemistry LibreTexts*. Accessed: Jul. 30, 2023. [Online]. Available: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Supplemental_Modules_(Analytical_Chemistry)/Analytical_Sciences_Digital_Library/Contextual_Modules/Sample_Preparation/03_Solid-Phase_Extraction
92. B. S. Levine and S. Kerrigan, Eds., *Principles of Forensic Toxicology*, Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-42917-1.
93. "Solid phase extraction - an overview | ScienceDirect Topics." Accessed: Jul. 30, 2023. [Online]. Available: https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/solid-phase-extraction
94. B. Buszewski and M. Szultka, "Past, present, and future of solid phase extraction: A review," *Crit. Rev. Anal. Chem.*, vol. 42, no. 3, pp. 198–213, Jul. 2012, doi: 10.1080/07373937.2011.645413.
95. B. E. Richter, B. A. Jones, J. L. Ezzell, N. L. Porter, N. Avdalovic, and C. Pohl, "Accelerated solvent extraction: A technique for sample preparation," *Anal. Chem.*, vol. 68, no. 6, pp. 1033–1039, Jan. 1996, doi: 10.1021/ac9508199.
96. G. L. Streun, A. E. Steuer, L. C. Ebert, A. Dobay, and T. Kraemer, "Interpretable machine learning model to detect chemically adulterated urine samples analyzed by high resolution mass spectrometry," *Clin. Chem. Lab. Med. CCLM*, vol. 59, no. 8, pp. 1392–1399, Jul. 2021, doi: 10.1515/cclm-2021-0010.
97. G. L. Streun, M. P. Elmiger, A. Dobay, L. Ebert, and T. Kraemer, "A machine learning approach for handling big data produced by high resolution mass spectrometry after data independent acquisition of small molecules – Proof of concept study using an artificial neural network for sample classification," *Drug Test. Anal.*, vol. 12, no. 6, pp. 836–845, Jun. 2020, doi: 10.1002/dta.2775.
98. "Forensic identification of sudden cardiac death: A new approach combining metabolomics and machine learning," *Anal. Bioanal. Chem.* Accessed: Jan. 17, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s00216-023-04651-5

Application of Machine Learning in the Field of Forensic Fingerprint Sciences

7

ASHISH BADIYE, NEETI KAPOOR
AND MUSKAN SINGAL

Introduction

Machine learning is the branch of artificial intelligence and can be defined as the process by which a machine is learned or trained to work more efficiently. The sample is trained by feeding the input–output data, or input data alone, and the machine's behaviour is studied. This means how the machine behaves or responds to all possible inputs, or we can say that machine learning involves training machines to recognize patterns, extract insights and make predictions or decisions based on data. Therefore, we can define machine learning as the equation or algorithm used to improve some performance measures while executing some tasks through training (Jordan and Mitchell, 2015). In simple words, machine learning teaches the machine how to handle extensive data efficiently. In cases of large data, humans sometimes cannot interpret the data or extract useful information from the data. In those cases, machine-learning algorithms are used to interpret data and give conclusions based on previously fed data (Mahesh 2020).

Machine learning has revolutionized how machines learn, adapt and make decisions. Different types of machine-learning algorithms are used in a wide variety of data and to solve different machine-learning problems. For example, linear regression algorithms are used to understand the relation between two variables. It draws the straight line that best fits the variables and predicts one value based on the value of another variable. The decision tree algorithm uses a series of questions to reach a conclusion. The random forest algorithm combines the result of many decision tree algorithms and predicts the conclusion. Neural network algorithms are like a simplified model of the human brain. They work by adjusting connections between neurons to understand the pattern (Hastie et al., 2009; Murphy 2012; Mahesh 2020).

DOI: 10.4324/9781003449164-8

In modern forensic investigations, machine learning has become an essential tool due to its capacity to unravel complex patterns, handle large amounts of data and uncover concealed insights. For instance, imagine a puzzle and each piece of the puzzle has specific information that is useful in solving the case. Traditional forensic investigation methods include manually examining each piece of the puzzle, which could be time-consuming, or some minute detail can be overlooked. However, machine learning is the digital puzzle solver that can solve the most complex puzzle in less time (Lefèvre 2018; Metcalf et al., 2017).

Fingerprints are considered one of the dominant biometric traits due to their various characteristic traits such as acceptability, high security, reliability and low cost (Awad 2012, Jain et al., 1999). Therefore, there are lots of fingerprint identification systems, but they lack identification time and accuracy in different stages of fingerprint examination such as fingerprint acquisition, fingerprint processing and enhancement, fingerprint extraction, fingerprint matching and fingerprint classification as described in Figure 7.1 (Egawa et al., 2012, Jain et al., 2011). Machine learning is used for fingerprint/pattern examination to overcome these problems.

From identifying minute details in fingerprints or deciphering the link between different pieces of evidence, machine learning is also used in DNA forensics to decode complex DNA sequences, genetic profiling or identifying markers. Besides these, it is used in financial fraud investigation, pattern

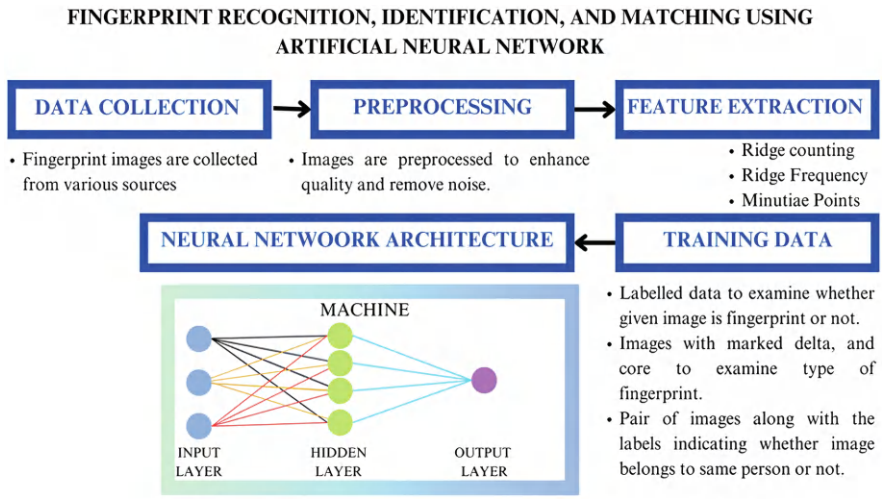

Figure 7.1 Fingerprint recognition, identification and matching using artificial neural networks is the process that involves training neural networks to analyze fingerprint images and make decisions about their identity or similarity (Awad, 2012; Hambalík, 2016; Shehu. et. al., 2018).

evidence analysis, digital investigations, medicine and toxicology and many more (Padma and Don, 2022; Haroon et al., 2020).

Machine Learning Techniques in Fingerprint/ Pattern Recognition/Identification/Matching

In biometric security and pattern recognition, using machine-learning algorithms has enhanced the security and other parameters in the fingerprint/pattern recognition/identification/matching field. Machine-learning algorithms contribute to fingerprint identification and matching to meet the growing demand for accurate and robust human identity verification methods. From classical algorithms to advanced deep learning models, these techniques have enhanced the precision of fingerprint analysis and enabled automation and scalability in various applications (Awad 2012).

1. **Artificial Neural Network**

It is a model inspired by the human brain's neural network, that is, it works similarly to the biological neural network in the human brain. Neural network algorithms are used for various tasks such as classification, regression, pattern recognition and management (Krogh 2008). The neural network algorithm works on three layers: input, hidden and output (Walczak 2019).

- **Input Layer:** The starting point receives the raw input data. Each neuron in the input layer corresponds to the specific feature on the attribute, and its value represents the value of that feature. Each neuron passes the value to the neuron of the next level.
- **Hidden Layer:** It is the intermediate layer responsible for learning and representing complex pattern relationships in data. It is the layer where the network learns to recognize features and combinations of features that are relevant to making predictions. The number of hidden layers and the number of neurons in each hidden layer are hyperparameters that need to be chosen based on the complexity of the problem.
- **Output Layer:** It produces the final results of neural network computation.

Components

- **Neural (Node):** It is the basic unit of the neural network algorithm. Each neuron works on different layers; therefore, each receives input, processes it and produces an output (Walczak 2019).

- **Connection (Synapses):** Neurons are connected through weighted connections called synapses (Walczak 2019).
- **Propagation:** This is how input data is passed through the network. There are two types of propagation methods: (i) feed forward propagation method: in this method, input data is passed through the network to generate prediction or output and (ii) back propagation method: this method follows the feed forward propagation method. This method adjusts the network's connection weight to minimize the error between predicted and target values (Krogh 2008).
- **Training Data:** This is the supervised machine-learning algorithm; training data consists of input features and corresponding target values.

1. **Support Vector Machine**

Support vector machine (SVM) is one of the most powerful machine-learning algorithms for linear and non-linear classification, regression, and outliner detection. In linear classification, the decision boundary that separates different classes is a single line or hyperplane, whereas non-linear classification involves more complex decision boundaries that a single straight line or hyperplane can not define. SVM utilizes the 'Kernel trick' technology to excel in non-linear classification. The classes might have intricate and curved feature space. Moreover, the SVM algorithm can be extended to classify data into more than two groups with the help of a technique known as 'multi-class-classification' in combination with various methods such as one-versus-all and one-versus-one (Suthaharan et al., 2016; Pisner et al., 2020).

Components

- **Data Points and Features:** It refers to a specific instance or example from the data set that you are using to train, validate or test the SVM model. Each data point is linked with a set of features that describes the characteristics (Guenther and Schonlau, 2016).
- **Hyperplane:** Hyperplane is the plane or line that separates different classes in the feature space (feature space is the collection of all feature vectors from space). In 2D cases, the hyperplane is the line, and in higher dimensions, it is the hyperplane. The algorithm aims to find the hyperplane that maximizes the margin (Pisner and Schnyer, 2020).
- **Support Vector:** Support vectors are the closest data points to the hyperplane. These data points determine the optimal position of the hyperplane (Meyer and Wien, 2015).

- **Margin:** Distance between the hyperplane and nearest support vector. The higher the margin, the better the generalization of the new data will be.
- **Kernel Function:** This technology handles the non-linear relationship between features (Pisner Schnyer, 2020).

The SVM algorithm is used in fingerprints to classify fingerprint images in different categories such as arches, loops, whorl, etc. (Awad, 2012). Fingerprint images were classified into one of five categories: whorl, right loop, left loop, arch and tented arch, using multi-class classification (Yao et al., 2001). The SVM algorithm was designed to classify the fingerprint images in different categories like whorl, right loop, left loop, arch and tented arch with 92.5% accuracy (Alias and Radzi, 2016). The SVM algorithm was designed and trained to recognize the finger code. Moreover, the research compares two methods that are SVM and RBF. And, concluded that SVM has better recognition rates (Elmir et al., 2012). The fingerprint classification method based on the twin support vector machine was studied in which the Gabor filter is used to extract texture features and the experimental results show that applying the twin support vector machine algorithm gives good classification results (Ding et al., 2020). SVM classifier was used to separate fingerprint images into one of the three quality classes: good, medium, and poor (Liu et al., 2008).

Convolutional Neural Network

It is the extended version of the artificial neural network algorithm. This algorithm is mainly used to extract the features from grid-like matrix datasets. Convolutional neural networks have revolutionized computer tasks by automatically learning hierarchical features from input data, making them particularly effective in tasks such as image classification, object detection, etc. A convolutional neural network (CNN) works on the filters, and filters keep sliding on the image and capture the desired features (O'Shea and Nash, 2015).

*Architecture/Workings (*Albawi et al., 2017; O'Shea and Nash, 2015)

- **Input Layer:** The input layer to CNN is typically an image, represented as a grid of pixel values.
- **Convolutional Layer:** These are the core building blocks of CNN. They apply a set of learnable filters to input images. Each filter is a small, square matrix that slides across the input image in horizontal and vertical directions. Multiple filters are applied in parallel to

create various feature maps, capturing different aspects of the input layer.
- **Pooling Layer:** This layer reduces the spatial dimensions of the feature maps while retaining their most important information.
- **Flattening:** After several convolutional and pooling layers, the resulting maps are flattened into a dimensional vector.

This step connects the CNN to a fully connected layer.

- **Fully Connected Layer:** This layer consists of multiple neurons (nodes) that are fully connected to neurons in the previous layer and tell about the global pattern and relationship obtained from convolutional and pooling layers.
- **Output Layer:** This layer provides the final prediction and classification.
- **Training:** CNNs are trained using labelled data and an optimization algorithm. The network parameters, including filter weights and biases, are adjusted during training to minimize a loss function. The loss function measures the difference between predicted and actual values.
- **Backpropagation:** Errors are propagated backwards through the network during training, and gradients are computed for each layer. And errors are updated.
- Analyzes: After training the data, CNN can be used to make predictions on new and unseen data.

In digital fingerprint classification, the most significant challenge faced by researchers is the low-quality images. A model consisting of several preprocessing stages such as edge enhancement, data resizing and data augmentation was made to overcome this challenge. Initially, poor quality original raw fingerprint images were processed using previt and lapace filters to enhance the edge and, in this research, the classification accuracy varied from 67.6% to 98.7% for the validation set and from 70.2% to 75.6% for the test set (Dincă Lăzărescu et al., 2022). Moreover, a CNN autoencoder is used to reconstruct the fingerprint images, where autoencoder is the technique that can replicate the data into images. CNN autoencoder resulted in a more than 90% accuracy rate for fingerprint identification in different databases (Saponara et al., 2021). Damaged fingerprints can be recognized using the CNN algorithm in MATLAB® software. A comparison was made between the traditional point-matching recognition method, the traditional CNN recognition method and the improved CNN recognition method. It was found that CNN algorithms had higher recognition rates and lower false acceptance rates than traditional

point matching. However, the enhanced CNN recognition method shows the best result, with the highest recognition and lowest false acceptance rates (Li, 2021). The advanced CNN consists of 15 layers and is classified into two stages. The first stage is the preparation stage, which includes fingerprint collection, augmentation and preprocessing. While the second stage deals with feature extraction and matching. It was found that this model gives 100% accurate results for both training and validating datasets (Althabhawee and Alwawi, 2022).

Generative Adversarial Network

A generative adversarial network is a deep-learning algorithm that consists of two neural networks competing against each other in a zero-sum framework. The two neural networks are the generator and discriminator, which are trained simultaneously through a competitive process. The generator is the neural network that takes random noise as input and attempts to generate data samples that resemble real data. Then, a random noise vector is passed through multiple layers of the neural network; then, the network gradually transforms the noise into data that becomes indistinguishable from real data. Therefore, the output of the generator is an image sample that resembles the real image. At the same time, the discriminator is the neural network that serves as the binary classifier that takes the input data and tries to distinguish between real and fake data generated by the generator. And it outputs the result in terms of probability, indicating how likely the input data is to be real (close to 1) or fake (close to 0) (Creswell et al., 2018; Wang et al., 2017).

The generative adversarial network (GAN) algorithm enhances the quality of low-quality fingerprint images. Super-resolution GAN is trained to enhance the spatial resolution of fingerprint images and generate a high-resolution version of the same fingerprint image that can provide more detailed fingerprint feature extraction; therefore, for this, the Fdeblur–GAN model was made for deblurring fingerprint images using conditional GAN (Joshi et al., 2021). Moreover, a powerful progressive GAN mode, consisting of two stages, progressive offline training and interactive online testing, makes the model focus on minutiae and the orientation field and achieve better recognition accuracy (Huang et al., 2020). A GAN algorithm with certain modifications forces the model to generate three additional maps to the ridge maps to ensure that the generation process must consider orientation and frequency information and to force the generator to reserve ID information during the reconstruction process. This modified model was applied and tested on different databases, and it was observed that a rank 10 accuracy of 88.02% was achieved on the IIIT-Delhi latent fingerprint database and rank 50 accuracy of 70.89% on the IIIT-Delhi MOLF database (Dabouei et al., 2018).

K-Means Clustering Method

It is the popular unsupervised machine-learning algorithm used for partitioning a dataset into distinct groups or clusters based on the similarities of data points. 'k' represents the pre-defined value, that is, the number of groups (Sinaga and Yang, 2020).

Workings

- Decide the number of clusters you want to make.
- Randomly select centroid. The centroid represents data points that represent the average position of all data points within the clusters.
- Each data point is assigned to the nearest cluster centroid based on the distance metric.
- Now, centroids are re-assigned within the cluster, and new data points are added based on the distance.
- The same steps are repeated until a stopping criterion is met (Sinaga Yang, 2020; Hossain et al., 2019).

Components

- **Data Points:** The individual items or observations you want to cluster.
- **Number of Clusters (k):** This is the pre-defined number of clusters you want to make in the data set (Pham et al., 2005).
- **Cluster Centroid:** These are representative points at the centre of each cluster.
- **Distance Metric:** It measures the distance between two data points.

In the field of fingerprint examination, k-means algorithms are used for different purposes than in the research (Wang et al., 2002); datasets of fingerprints are classified into different classes/clusters such as arch, left loop, right loop and whorl as depicted in Figure 7.2. A hierarchical k-means clustering algorithm has been utilized to classify fingerprint images into four quality classes: good, dry, normal and wet (Munir et al., 2012). Moreover, k-means clustering is also used in image segmentation (Mehidi et al., 2019, Cherrat et al., 2019).

Decision Tree

The decision tree is a type of supervised-learning algorithm that is used for problems majorly related to classification and regression. This algorithm is widely used because it is easy to understand and interpret. Moreover, it can be used for categorical and continuous data (Navada et al., 2011).

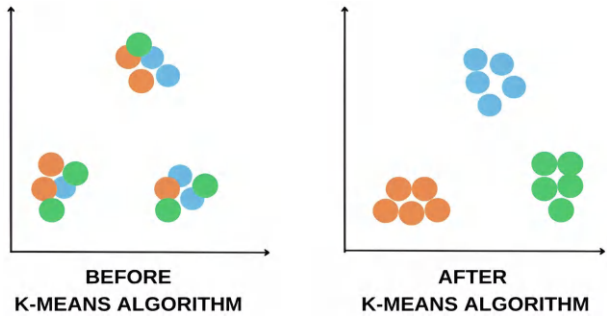

Figure 7.2 All the data points are mixed; after applying the k-means algorithm, data points are partitioned into distinct groups or clusters based on similarities between data points. For example, each colour represents a fingerprint pattern class. Before applying the k-means algorithm, all the fingerprint patterns are mixed. However, after the application of the algorithm, all the patterns are divided into different groups.

Workings

- The algorithm starts with the entire dataset, which consists of input features and their corresponding target labels. The goal is to create a tree structure that can predict based on input features (Charbuty Abdulazeez, 2021).
- The algorithm selects the best feature to split the dataset into subsets. It evaluates different features based on different criteria.
- After selecting the best feature, the dataset is split into subsets.
- Similarly, these steps are repeated recursively for each created subset and split until the stopping criteria are met (Navada et al., 2011).
- Once a stopping criteria is met, the leaf node represents the decision tree's final predictions.

Components

- **Root Node:** The topmost mode of the decision tree, representing the entire dataset at the beginning of the tree-building process. It is the node where the first feature is selected for splitting the data (Delibasic et al., 2011).
- **Internal Node:** These nodes represent the decision points in the tree. Each internal node is associated with the feature and conditions based on that value.
- **Branches:** Branches emanate from an internal node that leads to child nodes or leaf nodes
- **Leaf Node:** It is the terminal node of a decision tree. It does not have any child nodes. The leaf node represents the final decision or prediction made by the tree (Delibasic et al., 2011).

The J48 decision tree algorithm has been utilized in fingerprint-based gender classification. The accuracy of this approach is approximately 96.28% for the four fingerprint features namely, ridge count, ridge density, ridge thickness to valley thickness ratio and white line count (Abdullah et al., 2016).

Machine Learning in Palm Print and Hand Print Analysis

Besides fingerprints, biometric parameters such as handprints and palm prints are also majorly used in a person's identification. Due to the rapid growth and technological advancement, a reliable and secure biometric identification method was required. Therefore, it increases accuracy, reliability and security. Several models based on machine learning were introduced for palmprint/handprint identification, recognition and examination (Zhang et al., 2003). Handprint and palm print identification are primarily based on geometrical and texture characteristics, which are essential for stable recognition features (Kong et al., 2008; Zhao and Zhang, 2020). Palm prints have several unique features such as principal lines (flexion creases), minutiae points, wrinkle ridges (secondary creases), deltas, patterns and singular points that can be used in individual identification (Kong et al., 2009).

Measurements of handprints were used to determine the gender of the Sinhalese population in Sri Lanka. The results concluded that the classification and regression tree (CART) algorithm can differentiate gender with 91.67% accuracy with hand lengths, handbreadth and palm length. Also, other algorithms such as SVM and naive Bayes showed results with low accuracy, 83.33% (Dayarathne et al., 2021). PalmNet is the method used to extract highly discriminative palmprint-specific descriptors. This method uses the application of the Gabor filter in CNN. It was tested on databases captured using touchless acquisition procedures and heterogeneous devices. It was found that this method gives high accuracy (Genovese et al., 2019). Alexnet is also a structure based on CNN. It is used in palmprint recognition and identification using the PRelu activation function and recognition accuracy was found to be 99.9% (Gong et al., 2019).

Advantages and Disadvantages of Machine Learning

Humans have been using different machines or techniques to simplify their lives. Machine learning is one of the techniques that humans use to make their work easier (Mahesh, 2020). Forensic science deals with large databases that are reference databases for standard values. Therefore, a significant

Table 7.1 Use of Different Algorithms for Examination of Different Parameters in Fingerprint Sciences

S.No.	Year	Algorithm	Parameters	Result	Reference
1	2008	Back propagation neural network (BP), SOM neural network in inclusion with Gabor filtering and Zernike moments	Hands' geometry and texture features	This two-stage neural network algorithm resulted in personal identification with an accuracy of 97.6%	Kong et al., 2008
2	2009	Binary orientation co-occurrence vector (BOCV)	Local orientation features of palm print	BOCV outperforms the CompCode, POC and RLOC by reducing the equal error rate (EER) significantly	Guo et al., 2009
3	2010	Principal component analysis (PCA) algorithm in combination with Back Propagation neural network (BP), SOM neural network	Hands' geometry and texture features	For effective personal identification, the system accuracy can reach above 95.4% accuracy rate	Lin, 2010
4	2017	K-nearest neighbour algorithm, BSIF codes, Gabor filter	Extracting all the features from the region of interest.	Higher accuracy	Younesi and Amirani, 2017
5	2017	Complete direction representation (CDR)	All the values in the direction line	Palmprint matching speed of the CDR algorithm is very fast, which is about 10 times faster than that of representative spatial coding-based methods such as CompC and Ordinal Code	Jia et al., 2017
6	2019	Convolutional neural network and Prelu activation function	Selection of the region of interest, main line, wrinkles, triangulation, and detail points.	The accuracy of recognition was found to be 99.9%	Gong et al., 2019

Table 7.1 (Continued)

S.No.	Year	Algorithm	Parameters	Result	Reference
7	2020	Discriminative deep convolutional networks (DDR)	Deep discriminative features	DDR produces the best recognition performance in generic palmprint recognition compared to other state-of-the-art methods	Zhao and Zhang, 2020
8	2020	K-means algorithm, classical linear and quadratic discriminant analyses	All hand measurement characteristics.	Finger 1 and the palm measurements are the best hand part classifier. The breadth and circumference measurements showed better sex discrimination than length measurements. The best individual features for gender identification are the circumference of the palm followed by the breadths of the thumb and the index	Hida et al., 2020
9	2021	CART, SVM and Naïve Bayes	hand lengths, handbreadth, and palm length	The CART algorithm can differentiate gender with 91.67% accuracy. SVM and naïve Bayes showed results with low accuracy, i.e. 83.33%	Dayarathne et al., 2021
10	2021	Adjusted binary classification (ABC) algorithm	hand lengths, handbreadth, and palm length	In the cross-validated sample, ABC provided 95% accuracy. In the testing sample, ABC models achieved 97.3–100% accuracy	Jerković et al., 2021

Table 7.1 (Continued)

S.No.	Year	Algorithm	Parameters	Result	Reference
11	2023	linear discriminant analysis and logistic regression methods	hand length, handbreadth, maximum handbreadth, palm length, thumb finger length, index finger length, middle finger length, ring finger length, and little finger length	Logistic regression algorithm demonstrated maximum accuracy (91.10%) for sex identification. Meanwhile, the linear discriminant analysis algorithm showed a maximum accuracy of 91.40%	Islam et al., 2021
12	2023	Siamese network-based approach consisting of two convolutional neural network		Palm prints were recognized with more than 95% accuracy in different databases	AlShemmary and Ameen, 2023

problem arises in the management of that database. Machine learning successfully organizes the data into systematic form (Wuest et al., 2016).

Moreover, sometimes, the data is large enough that it is difficult to extract the information of interest even after visualising the whole data, or if possible, it takes more time (Mahesh, 2020). For example, the examiner has to identify whether the print 'matches' or 'does not match' from the database containing 10,000 samples. In this case, it will be tough for the examiner who will match each print manually. Here, using different machine-learning algorithms will accelerate the identification process. Even it reduces the chances of manual error. Many times, different conditions such as residual composition, impression condition, skin condition, disability and acquisition devices result in poor-quality prints. These poor-quality prints are challenging to examine manually. However, machine-learning algorithms such as the Gabor and GAN algorithms are used for fingerprint enhancement. That enhances the print, thus making it more compatible for the examination (Win et al., 2020; Hong et al., 1998). According to the different studies of the J48 decision tree algorithm in fingerprint gender classification problem, the accuracy of the approach is approximately 96.28% for the four fingerprint features namely, ridge count, ridge density, ridge thickness to valley thickness ratio and white line count (Abdullah at al., 2016) and CNN autoencoder resulted in a more than 90% accuracy rate for fingerprint identification in different databases (Saponara et al., 2021) this demonstrates that fingerprint identification, classification and examination is done more effectively by the machine-learning process rather than a manual process.

Besides this, the identification, classification and examination of fingerprints by machine-learning algorithms has some challenges. The algorithm's reliability highly depends on the quality and quantity of the training data. If the training data is not proper or is of a low amount, then the reliability of the result might be affected. Moreover, the result depends on the instructions provided; any biases or inaccuracies in the instructions will result in inaccurate conclusions (Singh et al., 2016). As machine-learning algorithms depend entirely on the training data, sometimes, results remain inconclusive due to erroneous fitting of suspected samples and training data (Anguita et al., 2010). Lack of interpretability and explanation of the result is one of the significant issues the examiners face. That makes it difficult for the examiners to verify the result or check the accuracy and reliability of the result (Tyagi et al., 2022).

Moreover, machine-learning models may be prone to vulnerable attacks where malicious actors can manipulate the input data to deceive the algorithms. Addressing these challenges requires a balanced approach that prioritizes machine-learning algorithms' transparency, accuracy and reliability.

Future Aspects of Machine Learning

With the increase in machine learning and artificial intelligence usage in every sector, it is highly predictable that the use of machine learning algorithms and artificial intelligence will keep increasing (Tyagi et al., 2022). Machine-learning techniques are now widely used in different domains of forensic science. Its use in the fields of fingerprint, handprint and palmprint examination is discussed above. Further, more advancements can be made in the examination of spoofed fingerprints. Spoofed fingerprints are artificial or fabricated fingerprints used to disguise the fingerprint recognition system, that is, spoofed prints are used to bypass the biometric authentication system, smartphone unlocking or other security features. Therefore, more advancements can be made in machine-learning algorithms to distinguish between genuine and spoofed prints (Adam and Sathesh, 2021). Fingerprints can also reveal the person's daily routine or lifestyle by examining the chemical composition of the fingerprints by mass spectrometry. Thus, utilizing machine-learning algorithms in the determination of the compositional profile of the prints will be of great use (Hinners et al., 2018; Win et al., 2020). Several types of research in the field of fingerprint examination deal with determining gender from fingerprints (Thakar et al., 2018; Iloanusi and Ejiogu, 2020; Jayakala, 2021; Ibitayo et al., 2022) or determining age from fingerprints (Hinners et al., 2020; Basavaraj and Rafi, 2015; Falohun et al., 2016) or determination of the age of fingerprints (Popa et al., 2010; Chen et al., 2021; Girod et al., 2016). These research areas are of great importance in forensic investigation. Therefore, advancement can be made in this area by utilizing machine-learning algorithms to investigate the determination of age from fingerprints, the determination of the age of fingerprints and the determination of gender from fingerprints. This will not only help the examiner narrow down the suspect list but also help the examiner determine these parameters with higher accuracy, reliability and minimal human error. Machine learning can be used not only in the examination of fingerprints, handprints and palmprints but also in other domains of forensic sciences. More research can be conducted to determine the test's accuracy, reliability, and admissibility in a court of law.

References

Abdullah, S. F., Rahman, A. F. N. A., Abas, Z. A., & Saad, W. H. M. (2016). Fingerprint gender classification using univariate decision tree (J48). *International Journal of Advanced Computer Science and Applications*, *7*(9), 217–221.

Adam, D. E., & Sathesh, P. (2021). Evaluation of fingerprint liveness detection by machine learning approach—a systematic view. *Journal of IoT in Social, Mobile, Analytics, and Cloud*, *3*(1), 16–30.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Alias, N. A., & Radzi, N. H. M. (2016, May). Fingerprint classification using support vector machine. In *2016 Fifth ICT International Student Project Conference (ICT-ISPC)* (pp. 105–108). IEEE.

AlShemmary, E., & Ameen, F. A. (2023). Siamese network-based palm print recognition. *Journal of Kufa for Mathematics and Computer, 10*(1), 108–118. http://dx.doi.org/10.31642/JoKMC/2018/100116

Althabhawee, A. F. Y., & Alwawi, B. K. O. C. (2022). Fingerprint recognition based on collected images using deep learning technology. *IAES International Journal of Artificial Intelligence, 11*(1), 81. https://doi.org/10.11591/ijai.v11.i1.pp81-88

Anguita, D., Ghio, A., Greco, N., Oneto, L., & Ridella, S. (2010, July). Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. In *The 2010 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Awad, A. I. (2012). Machine learning techniques for fingerprint identification: A short review. In *Advanced Machine Learning Technologies and Applications: First International Conference, AMLTA 2012, Cairo, Egypt, December 8–10, 2012. Proceedings 1* (pp. 524–531). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35326-0_52

Basavaraj Patil, G. V., & Rafi, M. (2015). Human age estimation through fingerprint. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(4), 3530–3535.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(1), 20–28. https://doi.org/10.38094/jastt20165

Chen, H., Shi, M., Ma, R., & Zhang, M. (2021). Advances in fingermark age determination techniques. *Analyst*, *146*(1), 33–47.

Cherrat, E. M., Alaoui, R., & Bouzahir, H. (2019). Improving of fingerprint segmentation images based on K-means and DBSCAN clustering. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(4), 2425–2432.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, *35*(1), 53–65. https://doi.org/10.1109/MSP.2017.2765202

Dabouei, A., Kazemi, H., Iranmanesh, S. M., Dawson, J., & Nasrabadi, N. M. (2018, October). ID preserving generative adversarial network for partial latent fingerprint reconstruction. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (pp. 1–10). IEEE. https://doi.org/10.1109/BTAS.2018.8698580

Dayarathne, S., Nawarathna, L. S., & Nanayakkara, D. (2021). Determination of gender using foot, footprint, hand and hand print measurements in a Sinhalese population in Sri Lanka using supervised learning techniques. *Computer Methods and Programs in Biomedicine Update*, *1*, 100017. https://doi.org/10.1016/j.cmpbup.2021.100017

Delibasic, B., Jovanovic, M., Vukicevic, M., Suknovic, M., & Obradovic, Z. (2011). Component-based decision trees for classification. *Intelligent Data Analysis*, *15*(5), 671–693. https://doi.org/10.3233/IDA-2011-0489

Dincă Lăzărescu, A. M., Moldovanu, S., & Moraru, L. (2022). A fingerprint matching algorithm using the combination of edge features and convolution neural networks. *Inventions*, *7*(2), 39. https://doi.org/10.3390/inventions7020039

Ding, S., Shi, S., & Jia, W. (2020). Research on fingerprint classification based on twin support vector machine. *IET Image Processing*, *14*(2), 231–235. https://doi.org/10.1049/iet-ipr.2018.5977

Egawa, S., Awad, A. I., & Baba, K. (2012). Evaluation of acceleration algorithm for biometric identification. In *Networked Digital Technologies: 4th International Conference, NDT 2012, Dubai, UAE, April 24–26, 2012, Proceedings, Part II* (Vol. 4, pp. 231–242). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30567-2_19

Elmir, Y., Elberrichi, Z., & Adjoudj, R. (2012). Support vector machine-based fingerprint identification. In *CTCI 2012 Conference*.

Falohun, A. S., Fenwa, O. D., & Ajala, F. A. (2016). A fingerprint-based age and gender detector system using fingerprint pattern analysis. *International Journal of Computer Applications*, *136*(4), 0975–8887.

Genovese, A., Piuri, V., Plataniotis, K. N., & Scotti, F. (2019). PalmNet: Gabor-PCA convolutional networks for touchless palmprint recognition. *IEEE Transactions on Information Forensics and Security*, *14*(12), 3160–3174. https://doi.org/10.1109/TIFS.2019.2911165

Girod, A., Ramotowski, R., Lambrechts, S., Misrielal, P., Aalders, M., & Weyermann, C. (2016). Fingermark age determinations: Legal considerations, review of the literature and practical propositions. *Forensic Science International*, *262*, 212–226.

Gong, W., Zhang, X., Deng, B., & Xu, X. (2019, September). Palmprint recognition based on convolutional neural network–AlexNet. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 313–316). IEEE. https://doi.org/10.15439/2019F248

Guenther, N., & Schonlau, M. (2016). Support vector machines. *The Stata Journal*, *16*(4), 917–937. https://doi.org/10.1177/1536867X1601600407

Guo, Z., Zhang, D., Zhang, L., & Zuo, W. (2009). Palmprint verification using binary orientation co-occurrence vector. *Pattern Recognition Letters*, *30*(13), 1219–1227. https://doi.org/10.1016/j.patrec.2009.05.010

Hambalık, P. M. A. (2016). Fingerprint recognition system using artificial neural network as feature extractor: Design and performance evaluation. *Tatra Mountains Mathematical Publications*, *67*, 117–134. https://doi.org/10.1515/tmmp-2016-00

Haroon, M., Tripathi, M. M., & Ahmad, F. (2020). Application of machine learning in forensic science. In *Critical Concepts, Standards, and Techniques in Cyber Forensics* (pp. 228–239). IGI Global.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Hida, N., Abid, M., & Lakrad, F. (2020). Supervised and unsupervised machine learning for gender identification through hand's anthropometric data. *International Journal of Biometrics*, *12*(3), 337–355. https://doi.org/10.1504/IJBM.2020.108485

Hinners, P. M. A., O'Neill, K. C., & Lee, Y. J. (2018). Revealing individual lifestyles through mass spectrometry imaging of chemical compounds in fingerprints. *Scientific Reports*, *8*(1), 5149.

Hinners, P. M. A., Thomas, M., & Lee, Y. J. (2020). Determining fingerprint age with mass spectrometry imaging via ozonolysis of triacylglycerols. *Analytical Chemistry*, *92*(4), 3125–3132.

Hong, L., Wan, Y., & Jain, A. (1998). Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 777–789.

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *13*(2), 521–526.

Huang, X., Qian, P., & Liu, M. (2020). Latent fingerprint image enhancement based on progressive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 800–801). IEEE.

Ibitayo, F. B., Olanrewaju, O. A., & Oyeladun, M. B. (2022). A fingerprint based gender detector system using fingerprint pattern analysis. *International Journal of Advanced Research in Computer Science*, *13*(4), 35–47.

Iloanusi, O. N., & Ejiogu, U. C. (2020). Gender classification from fused multi-fingerprint types. *Information Security Journal: A Global Perspective*, *29*(5), 209–219.

Islam, M. Z., Supto, M. R., Asadujjaman, M., & Chakrabortty, R. K. (2021, February). Sex identification from hand measurements using machine learning. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 536–540). IEEE. https://doi.org/10.1109/ICCCIS51004.2021.9397222

Jain, A. K., Bolle, R., & Pankanti, S. (Eds.). (1999). *Biometrics: Personal Identification in Networked Society* (Vol. 479). Springer Science & Business Media.

Jain, A. K., Ross, A. A., & Nandakumar, K. (2011). Fingerprint recognition. In *Introduction to Biometrics* (pp. 51–96). Springer. https://doi.org/10.1007/978-0-387-77326-1_2

Jayakala, G. (2021). Gender classification based on fingerprint analysis. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(10), 1249–1256.

Jerković, I., Kolić, A., Kružić, I., Anđelinović, Š., & Bašić, Ž. (2021). Adjusted binary classification (ABC) model in forensic science: An example on sex classification from handprint dimensions. *Forensic Science International*, *320*, 110709. https://doi.org/10.1016/j.forsciint.2021.110709

Jia, W., Zhang, B., Lu, J., Zhu, Y., Zhao, Y., Zuo, W., & Ling, H. (2017). Palmprint recognition based on complete direction representation. *IEEE Transactions on Image Processing*, *26*(9), 4483–4498. https://doi.org/10.1109/TIP.2017.2705424

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joshi, A. S., Dabouei, A., Dawson, J., & Nasrabadi, N. M. (2021, August). FDeblurGAN: Fingerprint deblurring using generative adversarial network. In *2021 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCB52358.2021.9484406

Kong, A., Zhang, D., & Kamel, M. (2009). A survey of palmprint recognition. *Pattern Recognition*, *42*(7), 1408–1418. https://doi.org/10.1016/j.patcog.2009.01.018

Kong, J., Lu, Y., Wang, S., Qi, M., & Li, H. (2008). A two stage neural network-based personal identification system using handprint. *Neurocomputing*, *71*(4–6), 641–647. https://doi.org/10.1016/j.neucom.2007.08.020

Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, *26*(2), 195–197. https://doi.org/10.1038/nbt1386

Lefèvre, T. (2018). Big data in forensic science and medicine. *Journal of Forensic and Legal Medicine*, *57*, 1–6. https://doi.org/10.1016/j.jflm.2017.08.001

Li, H. (2021). Feature extraction, recognition, and matching of damaged fingerprint: Application of deep learning network. *Concurrency and Computation: Practice and Experience*, *33*(6), e6057. https://doi.org/10.1002/cpe.6057

Lin, L. (2010, September). Palmprint identification using PCA algorithm and hierarchical neural network. In *International Conference on Intelligent Computing for Sustainable Energy and Environment* (pp. 618–625). Springer. https://doi.org/10.1007/978-3-642-15615-1_73

Liu, E., Zhao, H., Li, X., Fu, C., Liu, X., & Wang, F. Y. (2008). Palm-line detection and extraction. In *2008 International conference on computer science and software engineering* (Vol. 1, pp. 245–248). IEEE.

Mahesh, B. (2020). Machine learning algorithms–A review. *International Journal of Science and Research (IJSR)*, *9*(1), 381–386.

Mehidi, I., Belkhiat, D. E. C., & Jabri, D. (2019). An improved clustering method based on K-means algorithm for MRI brain tumor segmentation. In *2019 6th International conference on image and signal processing and their applications (ISPA)* (pp.1–6). IEEE.

Metcalf, J. L., Xu, Z. Z., Bouslimani, A., Dorrestein, P., Carter, D. O., & Knight, R. (2017). Microbiome tools for forensic science. *Trends in Biotechnology*, *35*(9), 814–823. https://doi.org/10.1016/j.tibtech.2017.03.006

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in Package e1071*, *28*(20), 597.

Munir, M. U., Javed, M. Y., & Khan, S. A. (2012). A hierarchical k-means clustering based fingerprint quality classification. *Neurocomputing*, *85*, 62–67. https://doi.org/10.1016/j.neucom.2012.01.002

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011, June). Overview of use of decision tree algorithms in machine learning. In *2011 IEEE Control and System Graduate Research Colloquium* (pp. 37–42). IEEE. https://doi.org/10.1109/ICSGRC.2011.5991826

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. https://doi.org/10.48550/arXiv.1511.08458

Padma, K. R., & Don, K. R. (2022). Artificial neural network applications in analysis of forensic science. In *Cyber Security and Digital Forensics* (pp. 59–72). https://doi.org/10.1002/9781119795667.ch3

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103–119. https://doi.org/10.1243/095440605X8298

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Andrea Mechelli and Sandra Vieira (Edsa.), *Machine Learning* (pp. 101–121). Academic Press.

Popa, G., Potorac, R., & Preda, N. (2010, June 1). Method for fingerprints age determination. *Romanian Journal of Legal Medicine*, *18*(2), 149–154.

Saponara, S., Elhanashi, A., & Zheng, Q. (2021). Recreating fingerprint images by convolutional neural network autoencoder architecture. *IEEE Access*, *9*, 147888–147899. https://doi.org/10.1109/ACCESS.2021.3124746

Shehu, Y. I., Ruiz-Garcia, A., Palade, V., & James, A. (2018). Detection of fingerprint alterations using deep convolutional neural networks. In *Artificial Neural Networks and Machine Learning – ICANN 2018* (pp. 51–60). Springer. https://doi.org/10.1007/978-3-030-01418-6_6

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Singh, A., Thakur, N., & Sharma, A. (2016, March). A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310–1315). IEEE.

Suthaharan, S. (2016). Support vector machine. In *Machine Learning Models and Algorithms for Big Data Classification* (pp. 207–235). Springer. https://doi.org/10.1007/978-1-4899-7641-3_9

Thakar, M. K., Kaur, P., & Sharma, T. (2018). Validation studies on gender determination from fingerprints with special emphasis on ridge characteristics. *Egyptian Journal of Forensic Sciences*, *8*(1), 1–7. https://doi.org/10.1186/s41935-018-0070-z

Tyagi, A., Kukreja, S., Meghna, M. N., & Tyagi, A. K. (2022). Machine learning: Past, present and future. *NeuroQuantology*, *20*(8), 4333.

Walczak, S. (2019). Artificial neural networks. In *Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-computer Interaction* (pp. 40–53). https://doi.org/10.4018/978-1-5225-7368-5.ch004

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., & Wang, F. Y. (2017). Generative adversarial networks: Introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, *4*(4), 588–598. https://doi.org/10.1109/JAS.2017.7510583

Wang, S., Zhang, W. W., & Wang, Y. S. (2002, October). Fingerprint classification by directional fields. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces* (pp. 395–399). IEEE. https://doi.org/10.1109/ICMI.2002.1167027

Win, K. N., Li, K., Chen, J., Viger, P. F., & Li, K. (2020). Fingerprint classification and identification algorithms for criminal investigation: A survey. *Future Generation Computer Systems*, *110*, 758–771. https://doi.org/10.1016/j.future.2019.10.008

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Research*, *4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Yao, Y., Frasconi, P., & Pontil, M. (2001, June). Fingerprint classification with combinations of support vector machines. In *International Conference on Audio-and Video-Based Biometric Person Authentication* (pp. 253–258). Springer. https://doi.org/10.1007/3-540-45344-X_37

Younesi, A., & Amirani, M. C. (2017). Gabor filter and texture based features for palmprint recognition. *Procedia Computer Science*, *108*, 2488–2495. https://doi.org/10.1016/j.procs.2017.05.157

Zhang, D., Kong, W. K., You, J., & Wong, M. (2003). Online palmprint identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1041–1050. https://doi.org/10.1109/TPAMI.2003.1227981

Zhao, S., & Zhang, B. (2020). Deep discriminative representation for generic palmprint recognition. *Pattern Recognition*, *98*, 107071. https://doi.org/10.1016/j.patcog.2019.107071

A Machine Learning Approach for the Digital Forensics

8

ANKIT SRIVASTAV, UJAALA JAIN AND TANURUP DAS

Introduction

The digital forensic research workshop (DFRWS) states digital forensics is:

> The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations [1, 2].

In simple terms, digital forensics directly deals with digital information for legal proceedings. It includes the use of scientific technologies and methods to process the data created by multiple digital objects [3]. It deals with digital devices and data in the context of cybercrimes or other digital investigations [4].

The National Institute of Standards and Technology (NIST) has designed a four-step procedure to be used in the digital forensics investigation process as mention in Figure 8.1. However, they can be performed in a flexible manner depending upon the case complexity [3].

Forensic Significance

Digital forensics investigation recovers, collects and analyses data, helping investigators to identify and prevent any unauthorized access to the gathered information [3]. It is significant in not just criminal but civil litigation as well. It extracts the data for the determination of the root cause of system failures or any breach of security. It is useful in achieving multiple goals, some of which are mentioned below [5]:

DOI: 10.4324/9781003449164-9

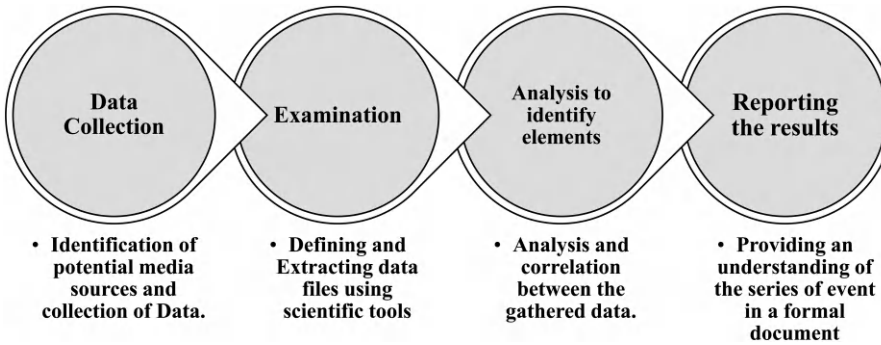

Figure 8.1 Phases of digital investigation.

- **Imaging:** It is the process of making a bit-by-bit copy of a device such as a USB drive, hard drive or CD-ROM to preserve the integrity of the original device. This copy is used by the examiner to analyze and reconstruct the recovered, deleted and damaged data.
- **Data Recovery:** It is the process of getting back data that was erased by accident or lost as a result of hardware malfunction.
- **Timeline Analysis:** This creates a chronological order for the events related to an instance to govern what actions were taken at what time and by which parties.
- **Network Forensics:** It identifies the network traffic and connections including its source and destination of data. It is helpful for the identification of any malicious or unauthorized activities along with the reconstruction process for the network activities.
- **Memory Analysis:** It is the examination of a computer's volatile memory for the recognition of open network connections, running processes, system malware and other system information.
- **Steganography Detection:** This involves the identification of hidden messages or data embedded in different files. It detects the relevant data from the hidden communications or data for investigation purposes.
- **Multimedia Forensics:** It involves the identification and authentication of the audio, video and images as evidence using specific tools.

Problems Encountered

Digital forensics has become one of the biggest critical areas of security with the growing threat of cyber attacks [3]. There are substantial challenges faced during the collection and examination of digital evidence due to the increased volume and diversity of data along with the complex nature of

the hardware and software platforms that use encryption [3, 6]. There is a lack of updated knowledge about the recent digital forensic tools threatening the development of the industry along with a question on data accuracy. Issues relating to the correlation between the data from different sources and consistency in the results could also affect the efficiency of digital investigations [3]. The investigation procedure requires a large amount of human intervention, resulting in slow investigations in comparison to the pace of digital crimes [2, 3]. Human powers along with other existing resources are still unable to solely investigate digital crimes [2]. Another obstacle faced is due to the increasing amount of file formats and operating systems which prevent the International Journal of Organisation and Collective Intelligence from developing standardized digital forensic tools and methods. Therefore, a lack of standardization in the process, formatting and storage has become a significant issue [3].

The availability of software tools used for any manipulation of digital media or conduction of a crime has again raised the difficulty of the detection and identification of digital evidence. For example, editing software along with mobile photography devices interfere in the detection and identification of computer-generated images (CGI), manipulated and recaptured images including splicing, double-compression, copy-move, removal, sharpening, resampling, resizing, filter applications and many more [6].

Machine-Learning Approach for the Examination of Digital Evidence

Machine learning has revolutionized digital forensic investigations with the introduction of new tools. It is a branch of artificial intelligence using algorithms to detect the patterns and relationships between data. The algorithms are designed in such a way that they automatically improve their work while performing a specific task. They are trained to work on a set of labelled data, where the labels describe the desired output for the input data. The simplest example is a spam folder in emails, where the algorithms label the emails as spam or not spam [4]. This technology is useful for security and fraud detection [3]. There is a certain procedure for developing machine-learning models which can be summed up in six steps as shown in Figure 8.2:

- **Selection of Machine Learning Approach:** There must be a clear understanding of what needs to be accomplished and the selection of algorithm must be based on that.

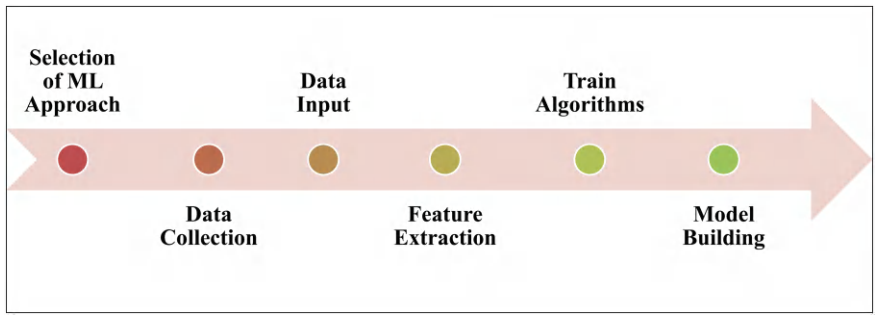

Figure 8.2 Development of ML models.

- **Data Collection:** Data available in different forms such as text files, spreadsheets or even on paper needs to be collected in electronic form [7].
- **Data Input:** The accuracy and performance of the model increase with the increase in the amount of data.
- **Feature Extraction:** It studies the behaviour of the customer and the service provider to decide their intentions. It includes features such as a customer's identity, location, time, network, mode of payment, etc.
- **Train Algorithms:** A set of rules based on which actions, such as acceptance (legitimate) or rejection (fraudulent), are taken is known as an algorithm. Data is input for the model to work efficiently [1]. This is an important step as the model must be trained with relevant data to prevent unbiased or false results [4].
- **Model Building:** The model after training is all set to be implemented with greater accuracy. However, the models have to be constantly updated with the new implementation of new plans by the criminals [1].

Machine-learning approaches including text analysis, optical recognition, image processing, voice recognition and character recognition have become the state of the art [6]. The concept covers applications such as speech recognition, fraud detection systems, automatic text classification, object recognition and many more [8]. These techniques gather information from huge volumes of data by matching conceptual models which enables data mining, hunting anomalies, identification of patterns and knowledge discovery in digital forensic investigation [6]. With the help of various tools and methods, experts can obtain important data and proof that will strengthen their investigations and shield their organisations from any mishaps in the future [5]. The tools which can be used in digital forensics include:

- **Software Tools:** Software such as FTK, EnCase and X-ways Forensics are designed for the analysis of electronic data.
- **Hardware Devices:** These comprise items such as forensic duplicators and write blockers, which create precise duplicates of electronic devices and stop data from being altered during the gathering process.
- **Cloud Forensics:** Examining cloud data storage, infrastructure and data which was moved to or from the cloud are all part of this process.
- **Tools for Forensic Imaging:** These programs create a bit-by-bit copy of the whole storage device for the prevention of data loss or any alteration while the inquiry is underway. Tools for forensic imaging counts in FTK Imager, dd and EnCase [9].
- **Data Recovery Tools:** They seek information that is still on the storage device but has been removed from the file system. File carving techniques are used by data recovery tools to locate and retrieve erased data. For example, EaseUS, Recuva and PhotoRec Data Recovery.
- **Tools for File Analysis:** Recovery of file, identification and file identification with modified timestamps or other metadata can all be accomplished with the use of file analysis tools. Tools for file analysis include Forensic Explorer, X-ways and Autopsy.
- **Tools for Network Forensics:** These tools are capable of interception and examination of network packets in order to determine the kind, timing and source and destination of data. Tools for network forensics include tcpdump, NetworkMiner and Wireshark [10].
- **Memory Analysis Tools:** Tools for memory analysis can be used to find malware, hidden processes and several potential security risks on the system. It includes Redline, Rekall and Volatility.
- **Tools for Steganography Detection:** Tools for detecting steganography or concealed data can be used to find communications or secret data that could be important for an inquiry including StegSolve, StegDetect and StegHide.
- **Tools for Keyword Searching:** These tools allow you to look for particular words or phrases in a lot of data. They include FileLocator Pro, dtSearch and GRAP.

In digital forensics, the umbrella of machine learning covers deep-learning models which are useful in multiple domains such as adversarial image forensics, computer forensics, image temper detection and many more [6]. Digital forensics has numerous investigation models, such as the abstract digital forensics model (ADFM), digital forensics research workshops model

Table 8.1 Models of Digital Forensics

Models Phases	DFRWS	ADFM	IDIP	EEDIP
Acquisition	Collection and storage	Collection and storage	Collection, storage, deployment, documentation and image acquisition	Crime scene security, collection, storage, deployment, detection, documentation and image acquisition
Examination	Examination	Examination	Tracing	Tracing
Analysis	Analysis	Analysis	–	–
Presentation	Report	Report	Presentation	Presentation
Review	Proof	Return of rvidences	Review	Review

(DFRWS), integrated digital investigation process model (IDIP), and end-to-end digital investigation process model (EEDIP) [3].

Algorithm for Digital Investigation and Examination

Algorithms are developed for the digital investigation process as a whole and for individual examinations of digital evidence. These must be transparent, interpretable and explainable to understand and verify the results [4]. The applicable algorithms are:

Support Vector Machine

It is used in the detection of computer-generated images. First, the extraction of a histogram and multifractal spectrum structures from the residual images and regression model fitness features and then measuring the binary similarity of PRNU (photo response non-uniformity) [6]. It has the ability to handle both regression and classification problems. It can perform complex tasks on structured as well as unstructured data depending on the kernel function and uses the samples from the training data set for the classification of objects. It basically identifies the features in each data set and decides a hyperplane based on which the data is classified into two categories. The error rate here is minimised with the increasing marginal distance between the two categories [3]. It is a supervised learning algorithm used for the classification of phishing emails as it separates data which is non-linear [1]. SVM is a state-of-the-art method used in machine-learning theory. It is designed for multi-class and binary classification. Its main aim is to identify the hyperplane with maximum distance between data of different classes for further processing [11]. It is viewed as an algorithm that learns from training using particular data sets to provide correct predictions or outputs. It is

counted under the supervised learning category as it performs tasks for data analysis, pattern recognition, regression analysis and classification. The SVM algorithm is trained by the example datasets to build a model that will define the categories of the input data. It can be used in both linear and non-linear classification [12]. SVM-based techniques are considered appropriate to solve binary problems in image forensics [6].

Convolution Neural Network (CNN)

CNN works based on observational data. However, it is a complex model due to its interconnectivity with large neurons. It addresses image processing tasks including pattern recognition. In this, feature extraction is a data-driven process which can classify images based on their shapes [11]. It is a variant of neural networks, popularized recently with the availability of computational power to train the models. It works tremendously on large datasets with great accuracies for image classification, object detection and face identification. The datasets on which it has worked are Pascal visual object challenge (VOC) 2007, VOC 2012 and Image Net 2012 [8].

Decision Tree Algorithm

It can be used for both classification and regression of tasks. The interpretation is easier in this type [3] and can be used as a statistical model as well [12, 13]. It can communicate the output from the tests to the classification of data items. It forms a flow chart–type structure for instance a tree for the classification of data into classes [12]. It considers decision logic framing them into a tree-like structure. The topmost node is the root node, internal nodes are the input variables, the child nodes are branched by the classification algorithms after the test completion and the process continues until the leaf node is ready for the decision [3]. The decision tree algorithm works in a bottom up approach, starting from the roots and going to the leaves which demonstrates it is class one. The roots become the primary attribute while the final class is the leaves [12].

K-Nearest Neighbour Algorithm

It is a non-generalizing or non-parametric learning method which does not focus on developing general models [3, 13]. It can handle both regression and classification tasks to provide accurate data depending on its quality. The training data is stored in an n-dimensional space [3].

Naive Bayes

These are the probabilistic classifiers derived from the application of naive Bayes theorem [12]. The classification or clustering tasks can be performed by this unsupervised learning algorithm. A small amount of data is needed

as the training data for the estimation of necessary parameters. It can be simply implemented as a technique for making clusters and needs no specification for an outcome. It can also work as a supervised learning technique in cases where it relies on both targets and input variables [3]. They are essential in high-dimension datasets because they become a baseline for the classification problem by being a rapid algorithm based on its naïve assumption behaviour about data. They are classified into two types: Gaussian and multinominal naive Bayes. Gaussian naive Bayes works on the hypothesis that in every label, the data is pinched from a Gaussian distribution. In multinominal naive Bayes, the features are assumed to be generated by a simple multinominal distribution [12].

K-Means Algorithm

It is a simple yet efficient technique that classifies datasets into k-centres. It becomes more proficient when there are significant variables and thus can be compared to hierarchical clustering as well. Thus, in this algorithm, the efficient elements are the implementation and data interpretation [3].

Principal Component Analysis Algorithm

It takes into account the observation of various possible correlated variables and converts them into linearly uncorrelated values. Implementation of orthogonal transformation helps in the quick performance of the tasks. This also eliminates the need for any prior knowledge with the computation of the model. It also provides features such as data feature classification and estimation [3].

Logistic Regression Algorithm

This algorithm solves the classification problems by identifying which class is associated with which instance. It can be used as a binary classifier as it gives outcomes between 0 and 1 [3].

Singular Value Decomposition Algorithm

It is a concept of factorization used in matrixes that converts a high-dimensional dataset into a low-dimensional representation considering the dominant patterns. Invariance features can be extracted using singular values, and the decomposition method can be used from an image or a signal [3].

Apriori Algorithm

It is widely used for data mining and discovers the relationship shared between multiple data sets. It is designed in such a way that it performs well for the database where there have been several transactions. However, its performance can be impacted by factors such as the requirement of 'n' numbers

Table 8.2 ML and Digital Forensics [3]

Forensic Application	ML Algorithm	Investigation Phase
Image and video forensics	SVM, KNN, RF, PCA, SVD	Examination, analysis
Network forensics	DT, SVM, k-means, KNN and naive Bayes	Examination, analysis
Memory forensics	K-means	Analysis
Malware forensics	LR, DT, SVD, PCA	Analysis
Email forensics	LR, SVM, RF, DT	Analysis
Mobile forensics	Apriori, k-means	Analysis

of frequent item sets in the database scans [3]. It is highly efficient for the criminal analysis [7].

These algorithms are useful in the development of ML approaches for the proper investigation of digital crimes. Tables 8.1 and 8.2 show the summary of different models and algorithms applied in digital forensics. Such approaches for the digital evidence have turned out to be highly critical for the analysis of digital evidence as discussed below:

Images Analysis

Deep-learning models, mainly CNN-based models are useful for the automatic multidimensional feature extraction, training and classification for great accuracy recognition [6]. It works tremendously on large datasets with great accuracy for image classification, object detection and face identification. The datasets on which it has worked are Pascal visual object challenge (VOC) 2007, VOC 2012, and Image Net 2012 [8]. However, in situations of blind detection, the performances of such models are reduced [6]. Metadata studied for the evidence contains immense information about the source, camera maker, model, time and location [4]. Singular value decomposition algorithm: invariance features can also be extracted using singular values, and the decomposition method can be used from an image, or a signal using the singular value decomposition model [3]. SVM is used in the detection of computer-generated images. First, the extraction of the histogram and multifractal spectrum components from the residual images and regression model fitness features and then measuring the binary similarity of PRNU (photo response non-uniformity) [6].

Video Analysis

CNN-based models are a well-known criterion for the detection of fake videos or for the extraction of recompression errors. Model hybrids such as the

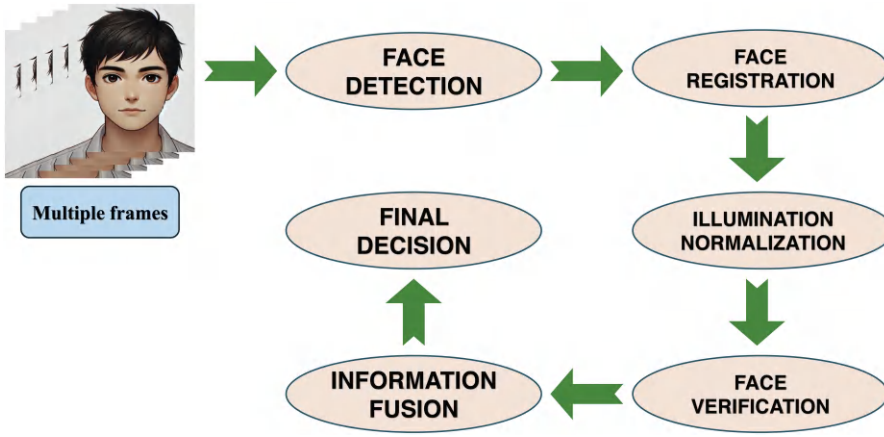

Figure 8.3 Face recognition system [14].

CNN–LSTM model can also be utilized due to enhanced results for facial motion differences between the original and fake videos [6]. Video analysis also includes biometric analysis. Biometric is the authentication technique used in forensic science that deals with physiological as well as behavioural factors. One such example is the face recognition technique which works by feature extraction and matching process as briefed in Figure 8.3.

Audio and Speech Analysis

It extracts speaker-specific features for their identification. The logistic regression algorithm can be used as a binary classifier as it gives outcomes between 0 and 1 [3]. These analyses are highly useful in cases of telephonic conversations or media exchanges [4].

E-Commerce Fraud Prevention

It is utilized to provide safety to Internet users from criminals using different tools such as Subono and Riskified. Subono allows viewing the customer's address, validation of the email address used and many such functions while the Riskified tool ensures real-time insights to avoid any time lapse in fraud detection considering IP location, browser fingerprinting, proxy detection, etc [1]. It also identifies any unusual financial transaction as fraud detection [4] Even the Big Four accountancy firms namely, KPMG, PWC, EY and Deloitte have all made significant investments in AI research [15].

Fraud Detection

Fraud is any activity that is done with the intention to deceive. One such example is computer network intrusions which mean manipulation of network service offered by some other host. Battling fraud demands a wide collection of algorithms, approaches and most importantly data sources which are effective in such scenarios [12]. There are also certain data sets available for the analysis of forgery detection methods, such as the Dresden Image Database, Image Manipulation Dataset and CASIA Tampered Image Detection Evaluation Database [16].

Network Traffic Analysis

It helps in the identification of the communication patterns between the devices helping in the identification of sources or origin of crime. It uses unsupervised learning to recognise any anomalous patterns or behaviours such as future threats, unusual communications, fraudulent acts and large data transfers. It works on real-time scenarios and flags the traffic which matches anomalous patterns [4].

Social Media Data Analysis

It is essential to aid in building a picture of the social networks of a suspect. It includes the language, sentiments or relationships between individuals. Techniques such as natural language processing (NLP) are beneficial to studying the content of a post and comments [4].

Meta Data Analysis

File authentication using the hashing technique. The fingerprinting of data is done to authenticate its integrity based on mathematical calculation, giving a unique hexadecimal value. It is a technique used for the identification and verification of the file. The task is performed by cryptographic hashing algorithms such as SHA-1 and MD5 [7].

Text Analysis

It includes the analysis of text or emails using attributes such as structure and linguistic trends giving promising results with 84% accuracy. The use of clustering algorithms and SVM are considered the best algorithms for both texts and email analysis [17]. It can be performed using information extraction, information retrieval as well as natural language processing tools. Text

analytics also referred to as text mining is the extraction of meaningful information from some textual data using a mixture of algorithms. It analyses structured and unstructured as well as semi-structured data and identifies the unknown information present in it. It can be an extractive, inductive or even deductive type of forensic technique depending on the type of algorithms used for the knowledge extraction. It performs several tasks such as text processing, cleaning up unwanted information, tokenization by splitting into slices, removal of stop words (the, is, a, etc.), stemming the words into base words (bikes to bike) and, most importantly, document-term or term document matrix analysis. Tools useful for text mining are Attensity, dtSearch, Forensic Toolkit from AccessData Apache OpeNLP, Natural Language Tool Kit, WEKA, etc. [7].

Network Tracing

Tools such as Iris, Net Witness, Xplico, Net Intercept and SoleraDS5150 help in examining the network traffic to detect intrusion and determine how the crime had occurred [17]. It is beneficial for network surveillance and security analysis.

Link Analysis

It establishes the connection between the evidence gathered and the suspects. By converting the information into a collection of linked entities or objects, it is possible to ascertain the structure and content of a body of knowledge [12]. It is an extractive technique which uncovers the hidden associations between entities buried under a large amount of data. It is based on graph theory, a branch of mathematics. It builds an immediate visual picture of communications for the investigator to help them understand the puzzle. It frames the case data gathered from multiple devices with mutual device users on a single map. Tools such as Analyst's Notebook, Maltego, Netmap, Centriguge and marketVisual analytics are popular for use in performing link analysis [7].

Malware Analysis

Malware analysis and behaviour identification can be accomplished by machine-learning methods. This can assist investigators in figuring out the malware's origin and mode of operation. New instruments and methods for identifying and evaluating malware can also be created using machine learning algorithms [4].

Advantages of the Approaches

The automation sourced through the machine-learning approach brings invaluable support to the investigators as well as researchers by speeding up the procedure and the processing capacity [6]. With the use of machine-learning approaches, the efficiency of the entities will surely increase along with the reduction in revenue losses which may occur due to digital crimes [1]. It is the potential of a machine-learning approach that can deliver high performance, better collaborative functions with more sophistication to perform efficient investigations with massive datasets [7]. Machine learning helps in the recognition of any criminal actions or risks by data analysis and segmentation. Additionally, it enables investigators to examine the widely distributed data sets located in wired and social networks. To put it simply, it can offer a well-structured repository containing the cleaned data from digital investigations with well-known characteristics and outcomes [2]. Understanding machine learning is greatly important as with its increasing scope, criminals tend to use such techniques to commit crimes. Thus, machine learning is not only used in the analysis of a digital crime but is equally important to be identified in cases where these technologies are used [18].

Disadvantages of the Approaches

The applications are restricted because of their use in security in terms of image forensics and large data. When there is a difference in the output results during the test phase, a need to develop a new class is raised [11]. CNN models in image analysis bring challenges to the network designing for learning features through weak traces related to precise manipulations. Compressed images are tedious to analyse because of their quality aspects or various compressions as available methods function under certain conditions [6]. Machine-learning algorithms require massive data storage which becomes really difficult for professionals to handle [19].

Conclusion

This chapter focused on highlighting the way machine learning has completely transformed digital investigations. Digital forensics, with its rapid advancements, requires tools that are capable of running faster or at least match its pace; hence machine learning is highly valuable in cases of digital

evidence. Whether it is image, audio, video, malware, cloud storage or any other digital evidence, machine learning brings solutions to all. It can be in terms of certain tools for network, software, hardware, recovery, etc. Not only has this but DFRWS and ADFM models could act as an entire investigation model. Moreover, these machine-learning tools, software, and models act as complementary tools that can help in processing large volumes of information faster and more accurately.

Future Aspects of Machine Learning

Several areas in digital forensics need the introduction of machine learning for better and more effective processes. The development of more advanced machine-learning models is needed for protection and prevention from cyber threats. More intelligent methods and tools are needed to be developed to bring in automatic investigations of malicious activities and the suspect's machines for the determination in accurate time [2]. Machine learning is used for evidence collection and detection but needs more attention in the evidence reconstruction and analysis phase. One of the major problems of adjustment of relevant settings especially in image analysis/enhancement is affecting the examination, demanding identification of the best settings for the investigation [6]. More work on the security of data involved or used in machine learning must be focused. One of the most important and focused areas has turned out to be cloud forensics. It is flooded with data and logs. Examining all this can be vulnerable to errors and is even a labour-intensive method. Humans are limited due to their capacity and flexibility but a computer can examine the entire data in a single day. With each day, the algorithms will also be able to learn from their experience and will improve their results with time [17].

References

1. Palit, S., & Roy, C. S. (2022). Machine learning in digital forensic investigation. *International Journal of Advances in Computer Science and Cloud Computing*, 10(2). ISSN(p): 2321–4058, ISSN(e): 2321–4392.
2. Iqbal, S., & Alharbi, S. A. (2020). Advancing automation in digital forensic investigations using machine learning forensics. *Digital Forensic Science*, *3*, 1–15.
3. Al Balushi, Y., Shaker, H., & Kumar, B. (2023, January). *The use of machine learning in digital forensics*. 1st International Conference on Innovation in Information Technology and Business (ICIITB 2022). Atlantis Press, pp. 96–113.

4. Sampath, K. K. (n.d.). *How machine learning is transforming digital forensics investigations*.
5. Patel, R. B. (2023). *The use of artificial intelligence in digital forensics*. Authorea Preprints.
6. Nayerifard, T., Amintoosi, H., Bafghi, A. G., & Dehghantanha, A. (2023). Machine learning in digital forensics: A systematic literature review. *arXiv preprint* arXiv:2306.04965.
7. Mariyanna, S. (2017). *Machine learning for cyber forensics and judicial admissibility*. https://doi.org/10.13140/RG.2.2.32426.16327.
8. Oladipo, F., Ogbuju, E., Alayesanmi, F. S., & Musa, A. E. (2020). *The state of the art in machine learning-based digital forensics*. SSRN 3668687.
9. Nelson, B., Phillips, A., & Steuart, C. (2016). *Guide to computer forensics and investigations*. Cengage Learning.
10. Rouse, M. (2018). *Network forensics*. TechTarget.
11. Passi, A. (2021). Digital image forensics based on machine learning approach for forgery detection and localization. In *Journal of physics: Conference series* (Vol. 1950, No. 1, p. 012035). IOP Publishing.
12. Qadir, A. M., & Varol, A. (2020). *The role of machine learning in digital forensics*. 2020 8th International Symposium on Digital Forensics and Security (ISDFS). IEEE, pp. 1–5.
13. Reddy, K. (2022). Use of machine learning in digital forensics. *Nasscom Community*. https://community.nasscom.in/communities/machine-learning/use-machine-learning-digital-forensics
14. Amin, R., Gaber, T., ElTaweel, G., & Hassanien, A. E. (2014). Biometric and traditional mobile authentication techniques: Overviews and open issues. In Aboul Ella Hassanien, Tai-Hoon Kim, Janusz Kacprzyk, & Ali Ismail Awad (Eds.), *Bio-inspiring cyber security and cloud services: Trends and innovations* (pp. 423–446). Springer.
15. Brown, S. (2018, April 9). Driving faster, more accurate and more beneficial tax decisions. *IBM*.
16. de Carvalho, T. J., Pedrini, H., & de Rezende Rocha, A. (2015). Visual computing and machine learning techniques for digital forensics. *Revista de Informática Teórica e Aplicada*, *22*(1), 128–153.
17. Khan, H., Hanif, S., & Muhammad, B. (2021). A survey of machine learning applications in digital forensics. *Trends in Computer Science and Information Technology*, *6*(1), 020–024.
18. Bhatt, P., & Rughani, P. H. (2017). Machine learning forensics: A new branch of digital forensics. *International Journal of Advanced Research in Computer Science*, *8*(8), 217–222.
19. Mohmed, A. L., & Palanivel, K. (2022). Digital forensics triage classification model using hybrid learning approaches. *International Journal of Innovative Research in Computer Science & Technology*, *10*(3), 29–39.

From Teeth to Technology Exploring AI's Role in Forensic Odontology

9

DHWANI PATEL AND
SANTHIYA RAGHAVAN

Introduction

Interestingly, the development of robotic technologies in many ways has become a kind of alternative and a parody of the human brain which has been quite developed in the field of dentistry [1]. The era of computerization began some time ago with the digitization of archives and medical devices for maintaining records; however, cloud servers have emerged as the true enablers of integration, serving as repositories for centralized data management. The use of computerized dental records was useful in the identification of victims in the aftermath of the World Trade Center attacks and the Indian Ocean earthquake and tsunami [2]. Human remains have become severely incinerated in some of these cases, and apart from the oral cavity, which is incredibly durable, there are no other means of identification [3].

Together with these factors, artificial intelligence (AI) has developed tremendously over the past decade and has been incorporated into a multitude of industries without complications [1, 4]. As the scenarios change, the field of healthcare has been more emphasized as compared to other time periods, and issues such as AI should be able to help physicians detect and diagnose a variety of dental issues. [5].

Forensic odontology, or FO, looks at both teeth and teeth remains, mainly for human identification purposes. [2, 3] The use of artificial intelligence more or less resolved the challenges faced by the existence of professional and personal bias when dealing with odonatological information. [6] In the case of FO artificial intelligence technology, this is not the only function, as it can also be utilized in solving crimes through the identification of bite marks, estimating the anatomy of the teeth and the mandible, sexing and ageing among other things. [7] All of these artificial-intelligence models are based

DOI: 10.4324/9781003449164-10

on two types of neural networks, convolutional neural networks (CNNs) and artificial neural networks (ANNs). [4] After numerous research studies and investigations, it has been proven that these models provide the highest level of accuracy compared to other means, proving the newer technology to be encouraging and promising in nature. [4] These AI networks and models can be utilized in different stages of the crime as well as in documenting the consequences of the crime to assist in legal issues. [5, 8, 9]

As forensic odontology has to do with the identification of disaster victims if forensic odontologists had not previously worked in this area, the rate of error here would be very high as the identification processes rely on speculation and pictorial representation. [10] In any case, antemortem dental records are high-end every time searching for the person or estimating the age. [11, 12]. The dental tissues left by the victim are used to build a 'match' from the person's earlier existing dental tissues for comparison purposes [3, 11, 13]. As it is quite common with mass disasters, where the identity is in doubt for a number of reasons such as burnt matter, mutilation or something else, the first questions should always be age and gender estimates. This, however, is not without its own challenges as dental age estimation and chronological age estimation differ [10]. It is feasible to establish the biological age of the deceased or living person utilizing cutting-edge technology such as artificial intelligence [14]. Forensic odontology includes the use of oral records, x-rays, plaster models and other tools in the field for the application of AI models for several comparative research [1, 5, 7–9].

The uncertainty projected through human vision and thought processes is to a great extent mitigated through AI-based models [5, 9]. In various criminal cases, including sexual attacks and abuse cases, homicide, rape and child abuse, bite marks and other saliva-related evidence can be one of the strongest pieces of evidence [15, 16]. A human bite mark and an injury are occasions that can be used as evidence; however, if not reviewed in due time, they distort making the whole process of providing evidence useless.

Gender Determination and Artificial Intelligence

When identifying an individual in the aftermath of a mass disaster, when remains are discovered, or in medicolegal cases, determining the individual's gender is of the utmost importance [17]. Skeletal bone examination plays a very important role in determining the gender of an individual. It is possible to measure human teeth, which are a component of the human cranium, in both living people and in the bones of those who have passed away [18]. Hormonal changes can influence the sizes and shapes of the teeth,

making them different from each other in the permanent dentition of different genders. This makes the permanent dentition different for men and women [18, 19]. When it comes to sexual dimorphism in humans, canine teeth are known to have the highest degree of differences [20]. The most preserved tooth in the oral cavity than any other teeth is canines as they are least affected by dental caries and periodontal illnesses. [21]. Fidya et al. studied a new AI technology and determined that the multi-layer perceptron resulted in the highest accuracy in the gender determination of canine teeth. [19, 21–23].

A study by Patil et al. used panoramic radiographs to determine gender, showing a high level of accuracy using the AI-based model. When it came to determining the gender of an individual based on panoramic radiographs, this model demonstrated an exceptionally high level of accuracy [20]. Discriminant and logistic analysis are two types of comparative analysis which were performed on this model in comparison to the traditional and conventional approaches. Both of these approaches have frequently demonstrated outstanding outcomes in gender determination [18, 19].

Age Estimation and Artificial Intelligence

Age estimation is vital to human identification; this includes criminal cases involving unknown skeletal remains, in addition to mass fatality and accident scenes. [11, 17]. AI and neural networks can be trained to autonomously and effectively improve the accuracy of age estimation using teeth [2, 13, 24, 25]. As the strongest and most durable structures in the human body, teeth can typically survive most environmental conditions (fire, water, etc.) and are thus frequently present and available where other identifying evidence may not be [12, 10]. Teeth are crucial in advancing the identification of unknown bodies and skeletal remains found at crime scenes, disaster sites and road traffic accidents [17]. They significantly contribute to both comparative and reconstructive identity analyses. To enhance the precision of dental age estimation, various machine-learning techniques have been introduced. The advancement of artificial intelligence has led to the creation of several programmed neural networks, enabling computers to autonomously estimate age with improved accuracy [23, 25, 26]. Different kinds of programs have been developed which enable or are trained to calculate age, with the help of neural networks.

Estimation of age is principally necessary for the purpose of determining the chronological age of an individual, particularly in the context of medicolegal matters, particularly in situations when a legal confirmation is required [26]. In most cases, the estimation of an individual's age is carried

Figure 9.1 Automated bone age assessment by BoneXpert. Once left hand and wrist radiographs are sent to the BoneXpert artificial intelligence software server, the software applies an active appearance model to analyze the 13 bones. Following this, the left hand and wrist radiographs marked with the final bone age are sent to the picture of archiving and communication system [37].

out by professionals with the use of panoramic x-rays of jawbones and hand–wrist radiographs [25, 27] (as mentioned in Figure 9.1). The application of an artificial intelligence–based model for staging the development of the lower third molar on panoramic radiographs was described in a study that was carried out by Tobel et al. In a scenario including fivefold cross-validation, the performance was evaluated using a variety of validation metrics, including accuracy, rank-N recognition rate, mean absolute difference and linear kappa coefficient. The deep learning convolutional neural network approach yielded superior results compared to all other approaches that were evaluated.

Cheiloscopy

Lip prints are distinctive and unique to every person. An identification procedure that is often utilized in forensic investigations is called cheiloscopy, which is the examination of lip prints [2, 3, 28, 29] (as mentioned in Figure 9.2). The human lips are a biometric modality that is currently in development [28]. For the purpose of lip-based biometric verification, a probabilistic neural network is utilized together with a novel biometric system that is only based on lip shapes and new lip geometrical measures [28, 30]. It is different from other parameters such as the texture of the lip surface.

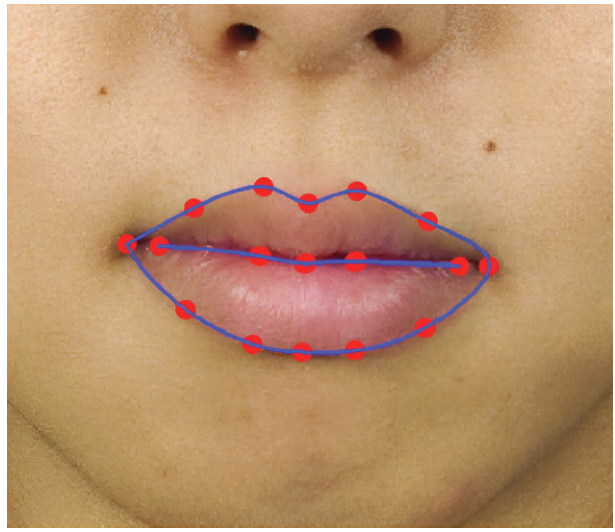

Figure 9.2 Face alignment (green) and useful landmarks for lip alignment (red dots).

Facial Reconstruction

The process of forensic facial approximation or reconstruction involves the rebuilding of the face of an individual whose identity is unknown from their skeletal remains [13]. It is a combination of sculpturing, osteology, anatomy and anthropology. It is a technique that is utilized in the field of forensics when there are unidentifiable remains involved in a crime [31]. The computerized facial reconstruction approach makes use of a laser video camera that is either interfaced with a computer or with CT scanning [32]. The application of artificial neural networks allows for the diagnosis of a person's gender based on their skeletal features with an accuracy rate of 95% [14]. When used for determining the gender of skeletal remains, approaches that are based on artificial intelligence will eliminate human bias, do not require any specialised skill, and offer results in a short amount of time [13, 33].

Artificial Intelligence in Mandibular Morphology

Facial reconstruction is a key forensic tool that can assist greatly in mandibular morphology and mandibular prediction. This can be particularly vital in instances of mass fatality and disaster scenarios in which a facial reconstruction is necessary but may need to be performed without the availability of a

mandibular bone.[34] As Khanagar et al. writes on the use of AI to help assist in this process:

> Sandoval et al. reported using an AI-based model for predicting the mandibular morphology through craniomaxillary variables on lateral radiographs in patients with skeletal class I, II and III, using automated learning techniques, such as Artificial Neural Networks and Support Vector Regression. [18, 35] The results of the study were quite promising. The ANN model demonstrated high predictability, and this model could play a key role in facial reconstruction. [17, 18]

Application of Artificial Intelligence in Forensic Dentistry

AI is increasingly being used in forensic dentistry applications and provides potential applications in facial reconstruction, age estimation, sex estimation, dental identifications (in matching to individuals) and bite mark analysis. It can also be used to search dental records. Task automation and AI and AI technologies' ability to synthesize massive amounts of data and scan reams of records rapidly, make it an appealing option. [4, 14–16, 21, 26, 36]

However, its shortcomings and certain drawbacks may limit its application. AI models may be trained on faulty or biased data, thus skewing results. Privacy, security and ethical considerations in decision-making and data use are likewise concerns. Any possible benefits of AI usage must be carefully weighed against these possible negatives associated with the technology. [1, 4, 7]

Conclusion

In order to maximize the benefits of AI in forensic dentistry, its implementation must be guided by robust regulations and ethical guidelines. While AI has the potential to transform forensic dentistry, it should serve as a supplementary tool rather than a replacement for traditional forensic methods.

References:

1. Corbella S, Srinivas S, Cabitza F. Applications of deep learning in dentistry. *Oral Surg Oral Med Oral Pathol Oral Radiol*. 2021;132:225–38.
2. Chowdhry A, Kapoor P, Juneja A, Chawla K, Bablani Popli D, Jasuja OP, et al. *Handbook of forensic odontology*; *An Indian Perspective, Century Publication*. 2018.

3. Forensic odontology – the forensics library. https://aboutforensics.co.uk/forensic-odontology/ (accessed October 10, 2020).
4. Rajaraman V. John McCarthy: Father of artificial intelligence. *Resonance*. 2014:198–207.
5. Journal IJSREM. Artificial intelligence: An odyssey in forensic odontology. *Int J Sci Res Eng Manag*. 2022;6. https://doi.org/10.55041/IJSREM16597.
6. Chinnikatti SK. Artificial intelligence in forensic science. *Forensic Sci Add Res*. 2018;2(5):182–3.
7. Khanagar S, Naik S, Sarode S, Patil S, Vishwanathaiah S, Bhandi S. Application and performance of artificial intelligence technology in forensic odontology – A systematic review. *Leg Med (Tokyo)*. 2020;46:101826. https://doi.org/10.1016/j.legalmed.2020.101826.
8. Ahmed O, Saleem S, Khan A, Daruwala S, Pettiwala A. Artificial intelligence in forensic odontology – A review. *Int Dent J Stud Res*. 2023;11:54–60. https://doi.org/10.18231/j.idjsr.2023.012.
9. Pathak J, Swain N, Pathak D, Shrikanth G, Hosalkar R. Role of various stakeholders in application of artificial intelligence to forensic odontology – A potential perspective. *Ann Dent Spec*. 2021;9:47–52. https://doi.org/10.51847/CBwpxBuRc0.
10. Blau S, Briggs CA. The role of forensic anthropology in disaster victim identification (DVI). *Forensic Sci Int*. 2011;205(1–3):29–35. https://doi.org/10.1016/j.forsciint.2010.07.038.
11. Prabhakar S. Recent trends in forensic odontology: An overview. *Int J Community Dent*. 2021;9:82–6.
12. Kavitha B, Einstein A, Sivapathasundharam B, Saraswathi T. Limitations in forensic odontology. *J Forensic Dent Sci*. 2009;1(1):8–10. https://doi.org/10.4103/0974-2948.50881.
13. Fernandes M, Kichler A, Rosa G, Sakaguti N, Franco A, Oliveira R. The role of forensic odontology in the quantification of dental and facial aesthetic impairment: A case report from the civil jurisprudence. *Cuad Med Forense*. 2017;23:41–45.
14. Galante N, Cotroneo R, Furci D, Lodetti G, Casali MB. Applications of artificial intelligence in forensic sciences: Current potential benefits, limitations and perspectives. *Int J Legal Med*. 2023;137(2):445–58. https://doi.org/10.1007/s00414-022-02928-5.
15. Mahasantipiya JS, Yeesarapat U, Suriyadet T, Thaiupathump T. Bite mark identification using neural networks: A preliminary study. In: *Proceedings of the international multiconference of engineers and computer scientists*. Hong Kong; 2011. http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp65-68.pdf (accessed July 7, 2020).
16. Mahasantipiya PM, Yeesarapat U, Suriyadet T, Sricharoen J, Dumrongwanich A, Thaiupathump T, et al. Bite mark identification using neural networks: A preliminary study. In: *World congress on engineering*. International Association of Engineers; 2010. p. 65–8.
17. Hill AJ, Hewson I, Lain R. The role of the forensic odontologist in disaster victim identification: Lessons for management. *Forensic Sci Int*. 2011;205(1–3):44–7. https://doi.org/10.1016/j.forsciint.2010.08.013

18. Patil V, Vineetha R, Vatsa S, Shetty DK, Raju A, Naik N, et al. Artificial neural network for gender determination using mandibular morphometric parameters: A comparative retrospective study. *Cogent Eng*. 2020;7(1). https://doi.org/10.1080/23311916.2020.1723783
19. Fidya F, Priyambadha B. Automation of gender determination in human canines using artificial intelligence. *Dent J (Majalah Kedokt Gigi)*. 2018;50:116–20. https://doi.org/10.20473/j.djmkg.v50.i3.p116-120
20. More CB, Vijayvargiya R, Saha N. Morphometric analysis of mandibular ramus for sex determination on digital orthopantomogram. *J Forensic Dent Sci*. 2017;9:1–5. https://doi.org/10.4103/jfo.jfds_25_15
21. Reddy VM, Saxena S, Bansal P. Mandibular canine index as a sex determinant: A study on the population of western Uttar Pradesh. *J Oral Maxillofac Pathol*. 2008;12(2):56. https://doi.org/10.4103/0973-029X.44577
22. Heng D, Manica S, Franco A. Forensic Dentistry as an Analysis Tool for Sex Estimation: A Review of Current Techniques. *Res Rep Forensic Med Sci*. 2022;12:25–39. https://doi.org/10.2147/RRFMS.S334796
23. Bewes J, Low A, Morphett A, Pate FD, Henneberg M. Artificial intelligence for sex determination of skeletal remains: application of a deep learning artificial neural network to human skulls. *J Forensic Leg Med*. 2019;62:40–3.
24. Shah JS, Ranghani AF, Limdiwala PG. Age estimation by assessment of dentin translucency in permanent teeth. *Indian J Dent Res*. 2020;31(1):31. https://doi.org/10.4103/ijdr.IJDR_428_18
25. Vila-Blanco N, Carreira MJ, Varas-Quintana P, Balsa-Castro C, Tomas I. Deep neural networks for chronological age estimation from OPG images. *IEEE Trans Med Imaging*. 2020;39(7):2374–84. https://doi.org/10.1109/TMI.2020.2968765
26. Gupta S, Chandra A, Agnihotri A, Gupta OP, Maurya N. Age estimation by dentin translucency measurement using digital method: an institutional study. *J Forensic Dent Sci*. 2017;9:42–8. https://doi.org/10.4103/jfo.jfds_76_14
27. De Back W, Seurig S, Wagner S, Marre B, Roeder I, Scherf N. Forensic age estimation with Bayesian convolutional neural networks based on panoramic dental X-ray imaging. MIDL 2019 Conference Abstract Pappers, 2019. https://openreview.net/forum?id=SkesoBY49E (accessed July 8, 2020).
28. Caldas IM, Magalhães T, Afonso A. Establishing identity using cheiloscopy and palatoscopy. *Forensic Sci Int*. 2007;165(1):1–9. https://doi.org/10.1016/j.forsciint.2006.04.010
29. Mun S, Ahn I, Lee S. The association of quantitative facial color features with cold pattern in traditional East Asian medicine. *Evid Based Complement Alternat Med*. 2017;2017:9284856. https://doi.org/10.1155/2017/9284856
30. Mishra G, Ranganathan K, Saraswathi TR. Study of lip prints. *J Forensic Dent Sci*. 2009;1(1):28. https://doi.org/10.4103/0974-2948.50885
31. Khanagar SB, Al-Ehaideb A, Maganur PC, Vishwanathaiah S, Patil S, Baeshen HA, Sarode SC, Bhandi S. Developments, application, and performance of artificial intelligence in dentistry - A systematic review. *J Dent Sci*. 2021;16(1):508–22. https://doi.org/10.1016/j.jds.2020.06.019
32. Zain-Alabdeen E, Felemban D. Artificial intelligence and skull imaging advancements in forensic identification. *Saudi J Health Sci*. 2023;12:171–7.

33. Vodanovic M, Subasic M, Milošević D, Galic I, Brkić H. Artificial intelligence in forensic medicine and forensic dentistry. *J Forensic Odonto-Stomatol*. 2023;41:30–41.
34. Khanagar SB, Vishwanathaiah S, Naik S, Al-Kheraif AA, Divakar DD, Sarode SC, et al. Application and performance of artificial intelligence technology in forensic odontology - A systematic review. *Leg Med (Tokyo)*. 2021;48:101826.
35. Sandoval TCN, Sonia Victoria GP, Gonzalez FA, Robinson AJ, Contreras CI. Use of artificial neural networks for mandibular morphology prediction through craniomaxillar variables. *Univ Odontol Bogotá*. 2016;35:1–6.
36. Katne T, Kanaparthi A, Srikanth Gotoor S, Muppirala S, Devaraju R, Gantala R. Artificial intelligence: demystifying dentistry–the future and beyond. *Int J Contemp Med Surg Radiol*. 2019;4(4):D6–9.
37. Vila-Blanco N, Carreira M J, Varas-Quintana P, Balsa-Castro C, Tomás I. Deep neural networks for chronological age estimation from OPG images. In *IEEE Transactions on Medical Imaging*, 2020;39:2374–2384.

Potential Application of Machine Learning in Forensic Anthropology

10

VINEETA SAINI AND
ARUNIMA DUTTA

Forensic anthropologists can provide deep insights into the analysis and identification of skeletal remains in criminal proceedings using detailed knowledge about bone development and morphological variations [1]. In cases of mutilated, decomposed and charred bodies, where DNA or antemortem dental records of the deceased are not available for comparison purposes, anthropological examination can pose as a helpful tool for such a challenging task. It also forms an integral part of analyzing the cause, manner, mode and time since death of the deceased. Forensic anthropologists also render their skills in the identification of victims of mass murders, wars, genocides and mass disasters such as building fires, building collapses, train accidents, ship sinkings and airplane accidents [1]. This personal identification can be established by the compilation of successional comprehensive information regarding the sex, stature, ethnicity, age and congenital or traumatic deformities, followed by craniofacial reconstruction and facial superimposition of the victim's skeletal remains. This database can be created by collecting the measurements of different areas of the human skeleton directly or by imaging techniques. Many studies conducted previously on the human skeleton have shown promising accuracies for the pelvis, skull, mandible, sternum and long bones in establishing a biological profile [2–4].

Challenges Faced by Forensic Anthropologists

Even though forensic anthropologists aid in personal identification, they frequently face obstacles when conducting their investigations.

- **Insufficient Victim Information and Lack of Context:** Forensic anthropologists often face a shortage of information about acquired bones and skeletal remains, such as the location where the remains were located or any objects found with the deceased. This leads to reduced chances of correct biological profiling, time since death estimate and cause of death.

DOI: 10.4324/9781003449164-11

- **Damaged Skeletal Evidence and Taphonomical Changes:** Fragmented skeletal remains or injuries caused during autopsies can cause a hindrance in personal identification and determination of bone trauma. Further, the taphonomical changes, weathering and scavenging activities may result in fragmentation, brittleness, crumbling, colour change or mineralization of bones and hinder the identification process.
- **Population Intermixing:** The most difficult challenge faced by forensic anthropologists is the intermingling of populations. With the immigration of people from their native regions in search of better opportunities, it becomes difficult to obtain population-specific data.
- **Ethical Management:** Since forensic anthropologists have to deal with accidental, homicidal and mass disaster victims, ethical management and a humanitarian approach have to be ensured while dealing with the family of the deceased. When the remains are discovered in isolated or politically insecure locations, access and authorization to carry out an investigation may be restricted, making this difficult.
- **Objectivity in Opinion:** Forensic anthropologists are required to be more objective while ascertaining the details rather than being biased and subjective.

Although forensic anthropologists face a multitude of challenges, their expert opinion garners immense importance in establishing the admissibility of such evidence in the court of law, ensuring justice for the family of the victim.

During skeletal examination, forensic anthropologists use a range of techniques such as radiology, visual assessment of morphological characteristics and measurements between particular anatomical landmarks but these methods require technical skill and time.

Using artificial intelligence (AI) in forensic anthropology speeds up the analysis process and improves accuracy. The term 'artificial intelligence' refers to the replication of human knowledge by machines to accomplish tasks easily. It uses a detailed database to perceive information, process it and provide subsequent action accordingly [5]. AI possesses the capacity to swiftly, accurately and efficiently analyze large volumes of data. It can also recognize patterns and make predictions, which can be helpful in forensic investigations.

These days, AI is extensively used in forensic anthropology, specifically its applications in facial recognition, skeletal analysis and the identification of

human remains. However, these estimations are based on data extracted by the scientists through mathematical formulas and professional experiences. Therefore, they can be time-consuming and inaccurate especially while processing large amounts of data [6]. Hence incorporating AI and its different branches has shown some promising advancements in the field of personal identification. With the advent of technology, various imaging techniques such as radiographs, computed tomography (CT) scans, magnetic resonance imaging (MRI) and cone bone computed tomography (CBCT) have eased the process of anthropological examinations. But still, some of the facets of artificial intelligence especially in detecting skull damage from CT scans and facial soft tissue prediction from the skull are yet to be explored by forensic anthropologists [7, 8].

AI also has various subfields one of which is machine learning (ML). ML algorithms are used to improve the classification performance of sex, age, stature and ancestry estimation based on skeletal remains. It's a segment of artificial intelligence capable of predicting without direct programming, utilizing mathematical models created from sample 'training' data.

It contains certain algorithms prepared to deliver specific responses based on imported databases and training. It enables the machine to understand human instructions and evolve to an extent where it can function without any human intervention. Machine learning includes various types of algorithms, which are frequently used to estimate sex, stature, age and ancestry, such as linear discriminant function analysis, support vector machines, artificial neural networks, decision trees, random forest, naive Bayes classification, binary logistic regression, multinomial and penalized multinomial logistic regression, multivariate adaptive regression splines and extreme gradient boosting (XGB), etc. (Figure 10.1) [9–15]. This is achieved by artificial neural networks that form a part of deep learning which is yet another subset of Machine Learning itself [16].

Discriminant Function Analysis (DFA)

Also called linear discriminant (LDA) or canonical discriminant analysis while quadratic discriminant analysis is a type of LDA, which is used to separate non-linear datasets. Due to its computational simplicity, discriminant function analysis (DFA) is widely used in anthropology to explore and understand patterns of biological variation among different groups or populations [17]. DFA classifies individuals into predetermined groups by analyzing measured variables such as skeletal or dental traits, along with cranial or postcranial measurements for sex estimation [18–20]. By analyzing

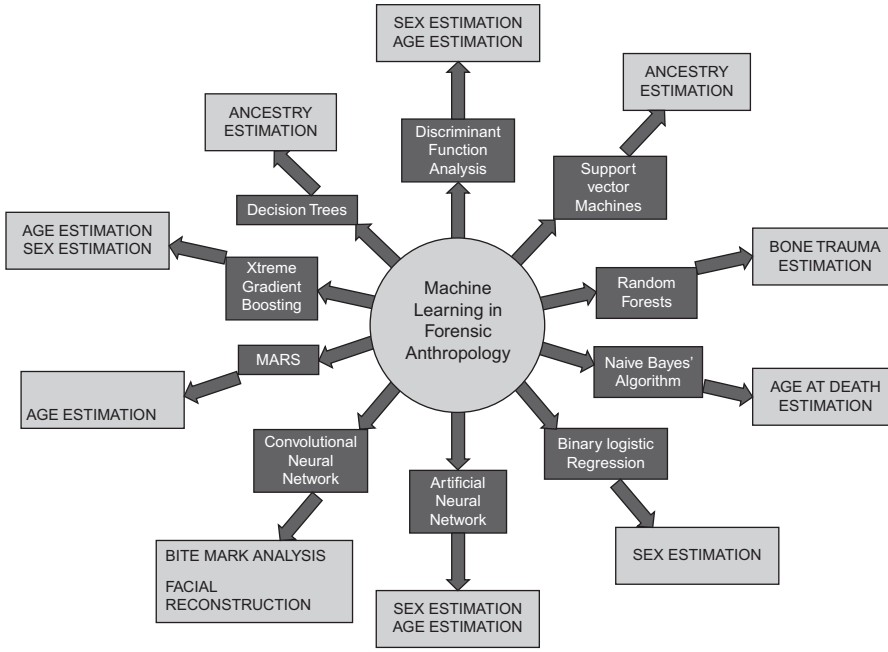

Figure 10.1 Diagrammatic representation of different branches of machine learning and their specific forensic anthropological applications.

the relationships between these variables and group membership, DFA can determine which variables contribute most significantly to the discrimination between groups. DFA finds its implications in studying sexual differences, population variations, secular changes, patterns of migrations and human evolution [18]. DFA provides the understanding of complex interrelationships among genetic and epigenetic factors that cause population-specific variations (Figure 10.2) [1].

Support Vector Machines (SVMs)

SVMs stand out as one of the most widely used and potent supervised-learning algorithms, capable of handling both classification and regression tasks related to sex, age and ancestry estimation. SVMs work well with small to medium-sized datasets and can handle high-dimensional feature spaces. SVMs operate on the principle of identifying an optimal hyperplane that maximizes the separation between data points belonging to distinct categories. The hyperplane is defined by support vectors, that is, a subset of training

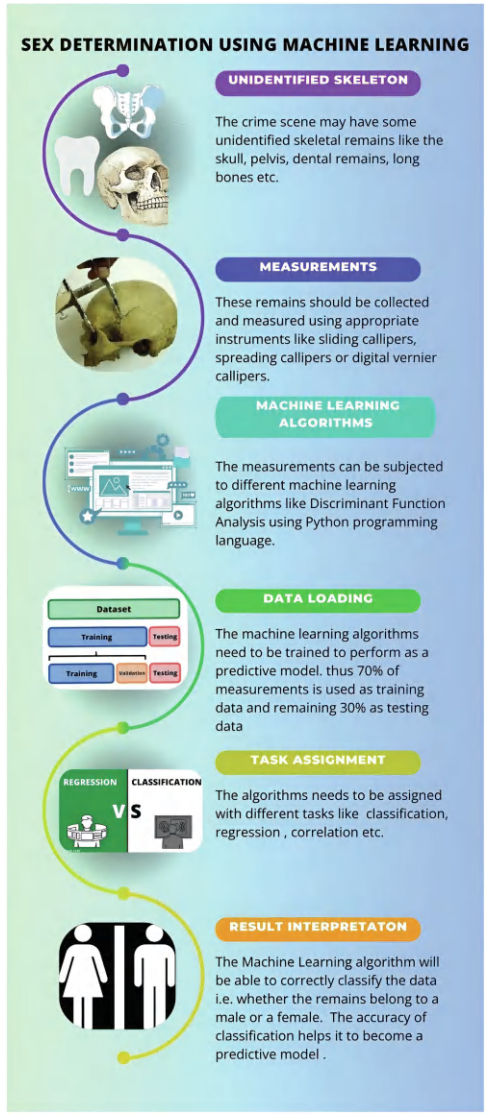

Figure 10.2 Infograph of sex estimation using machine learning DFA algorithms.

samples. SVMs strive to maximize the margin, which is the distance between the hyperplane and the data points closest to each class.

Sex estimation forms an important part of forensic anthropological examinations. SVMs have already been applied to obtain sexually dimorphic traits by training them on a data set containing measurements of the pelvis,

cranium, long bones and other skeletal remains [21–24]. Ancestry estimation also can be established by the SVMs by training it on known datasets of different racial groups and associated specific morphological or metrical features. In a study conducted by Spiros and Hefner (2020), ancestry estimation was carried out within American Black and White population groups with an accuracy of 88–92% using 8 cranial and 11 postcranial variables [15].

SVMs help in age estimation by examining the morphological and structural variations of the bones that take place with time such as the maxillary sutures, epiphyseal fusion and wearing of knee joints [25–27]. By training on a dataset of people with known ages and the accompanying skeletal traits, a support vector machine (SVM) can be trained to forecast the age of unidentified skeletal remains.

As it is a supervised learning algorithm, it can also be trained on known sample datasets beforehand to recognize comparable trauma patterns, skeletal diseases and anomalies in unknown remains.

In forensic anthropology, SVMs are only one of many machine-learning algorithms that can be utilized. Their efficacy is contingent on the quality and extent of the training dataset, as well as the selection of suitable features for analysis. Moreover, the expertise and comprehension of forensic anthropologists play a crucial role in interpreting results and making informed decisions based on the SVM's output (Figure 10.3).

Random Forest (RF)

It is a prevalent ML algorithm and belongs to the group of supervised-learning techniques, used for both classification and regression problems encountered related to ancestry estimation. This algorithm basically contains multiple decision trees and uses the average outputs of the various datasets to enhance the accuracy (Figure 10.3). The conventional method of ancestry estimation includes an examination of both morphoscopic traits and anthropometric measurements. However, a study shows that the application of a random forest model can combine both datasets and increase the accuracy of estimation. The results revealed that discriminant function analysis gave an accuracy of 75.4% whereas using random forest models the accuracy was increased to 89.6% [28]. AncesTrees is a novel algorithm proposed for ancestry estimation with randomized decision trees. The database used in designing this algorithm comprised 23 craniometric variables from 1734 individuals with an accuracy of 93.8% [29]. Forensic anthropology also deals with the assessment of bone trauma. Random forest algorithms are also found to have applications in classifying skeletal trauma as blows or falls. Research was performed on 400 anonymous CT scan images of patients with fractures from falls and

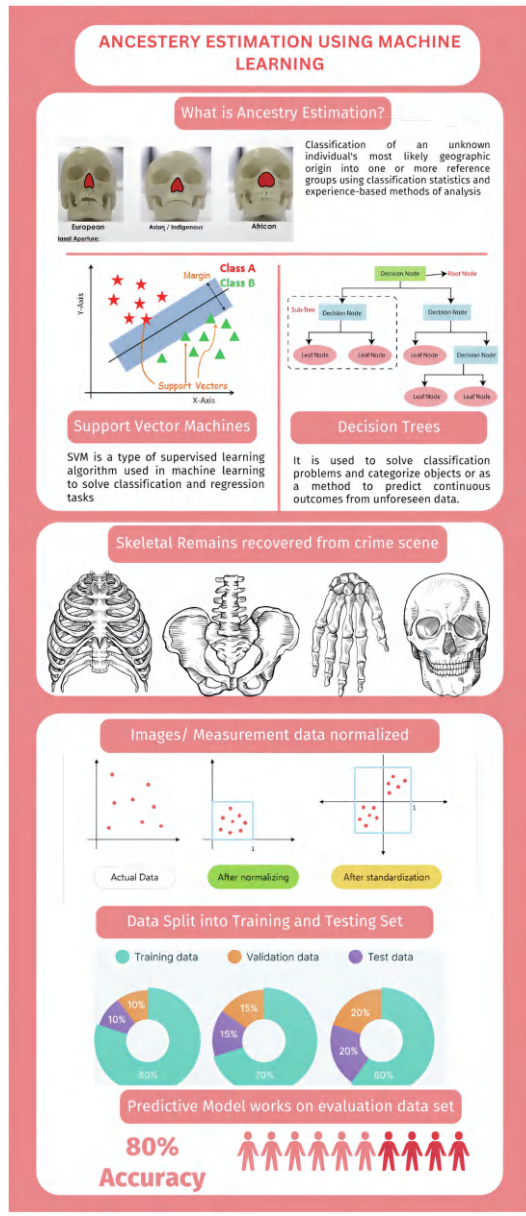

Figure 10.3 Workflow for ancestry estimation using SVM and DT.

blows. An accuracy rate of 83% was obtained, thereby revealing random forests to be a promising tool in differentiating between types of skeletal trauma [30]. Determining age is a key objective when establishing the identity of an unknown person. Random forests have been vital in establishing age from left-hand bone length. This study was conducted on the Asian population

for subjects aged from newborns up to 18 years. The results revealed that the random forest algorithm gave a better detection as compared to artificial neural network (ANN) models [31]. Random forest algorithm has been used in sex estimation from left-hand length. This study was also conducted in the Asian population in subjects aged between 16 to 18 years old. The results revealed an estimation accuracy of 91.67%. However, because of the specific age group, the results may not be applicable to a diverse population [31]. Age estimation can also be done from dental remains based on their mineralized morphology. A study done on 1477 panoramic dental radiographs of subjects aged between 2 to 18 years in a South China population revealed that random forests were best suited for age estimation in juveniles [32].

Naive Bayes Classification (NBC)

It is one of the simplest and fastest ML algorithms. It is a probability-based algorithm that functions on Bayes' theorem and is mostly used for classification tasks. This algorithm works on two main assumptions, that is, all the features within the dataset are independent of each other and each feature has an equal chance of being the outcome. The naive Bayes algorithm resulted in 90% accuracy in age estimation from the histomorphometric characteristics of 294 male corpses aged between 10 and 93 years [33]. In an alternative method, CT scan images of the pubic symphysis in the Korean population were scored using the Suchey–Brooks standard. The outcome depicted that Bayes' algorithm exhibited greater accuracy in comparison to traditional methods (Figure 10.4) [34]. Sex determination forms an integral part of personal identification. A novel probabilistic approach to estimating sex from the pelvis called 'CADOES' has been proposed in a study. This study aimed to propose a model that would consider pelvic variables of a certain population for sex estimation. The novel classification algorithm provided an accuracy of 85–97% with 38 pelvic variables [35].

Binary Logistic Regression (BLR)

Binary logistic regression (BLR) is a regression model that provides a binary variable as the output, that is, 0 or 1 and is prevalent in forensic anthropological examination for sex determination.

Verma and associates (2020) in a study on sexual dimorphism of upper and lower extremities on 344 Hamachi subjects found handbreadth to be most dimorphic followed by foot length and hand length with a sexing accuracy

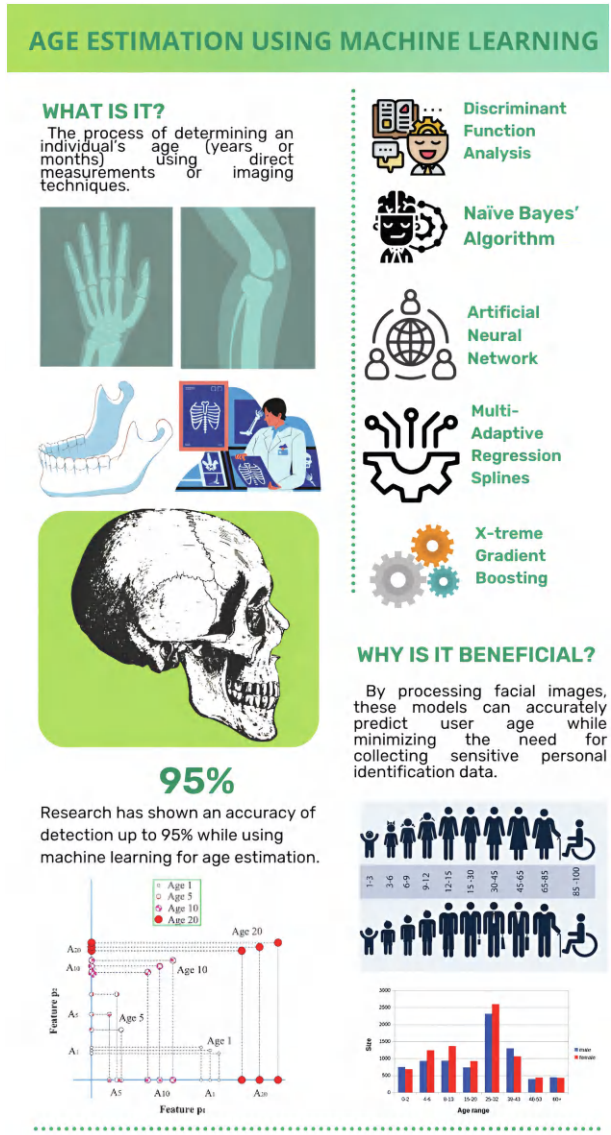

Figure 10.4 Infograph for age estimation using machine learning.

of 90.1% for males and 91.3% for females using BLR [36]. The other study was conducted to compare the accuracy of estimation between DFA and BLR using 12 anthropometric measurements of the ears from 497 subjects. The results revealed that the prediction rates of both algorithms were nearly the same [37]. Sex determination from craniometric measurements can be validated using statistical tools. A comparative study among LDA, BLR and

SVM was carried out which revealed that BLR performs slightly better than the other techniques [38]. Logistic regression allows greater flexibility in the datasets which may not be provided in linear regression. To predict the outcome, BLR uses a likelihood ratio rather than least squares, thereby making it a better fit for the final model. In a study performed to assess ancestry from five craniofacial variables, BLR gave an accuracy of estimation of up to 90% [39]. Sex estimation from long bones is imperative and binary logistic regression is more suitable than linear regression or discriminant function analysis. A study revealed that logistic regression provided the highest accuracy of sex estimation from measures of the articular surfaces and shaft of long bones [40]. In a study on sexual dimorphism of metatarsal bones of the South African population Bidmos and associates achieved a sexing accuracy of 79% and 84% using BLR and DFA respectively [41]. For stature estimation, linear regression analysis can be used. The long bones especially the femur, tibia, fibula, humerus, radius and ulna are found to provide the highest accuracy in stature estimation using simple and multiple linear regression methods (Figure 10.5) [42].

Artificial Neural Network (ANN)

ANN is a versatile machine-learning approach that consists of interconnected nodes (neurons) organized in layers. ANN can capture complex relationships between input features but may require larger datasets for optimal performance. It comprises computing systems which help in sensing and processing information similar to the functioning of biological neural networks. ANN contains three layers namely the input layer, the hidden layer and the output layer. These ANNs play a vital role in deep learning and are found to have prominent applications in forensic anthropology. ANN has a fascinating feature to identify structures and weights in any image which are important for their classification. However, these neural networks require a large collection of data before they can start functioning independently. Forensic imaging techniques have grown over the last years and have collectively formed a whole new branch called 'virtual anthropology'. CT or 'computed tomography' and CBCT or 'cone-based computed tomography' have shown very promising results in virtual anthropology by providing high-contrast images and visualization in both 2D and 3D planes. These CT images can be stored as DICOM data and processed for further image enhancement. Machine learning has helped in such enhancements and their 3D restructuring from DICOM images. The restructuring is processed by converting DICOM images into STL (stereo lithographic) files [43]. ANN has also been

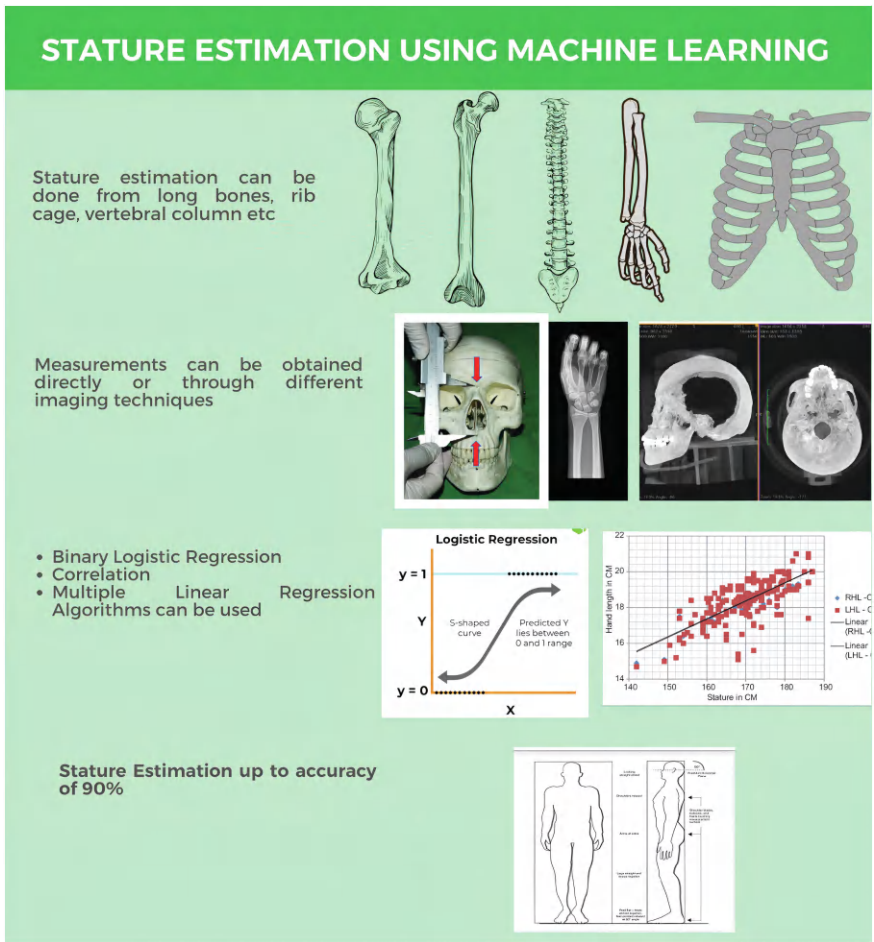

Figure 10.5 Stature estimation using machine-learning techniques regression analysis.

found to be useful in CD craniofacial annotation and superimposition for facial reconstruction and bite mark analysis. The use of landmark-based algorithms and face soft tissue datasets by ML can help forensic artists create more accurate and representative facial approximations. A study shows that an artificial intelligence algorithm named 'Artificial Immune Recognition System' based on genetics was used to perform superimposition of the skull and showed better identification of the craniofacial landmarks as compared to conventional methods [44]. Another eminent application of ANN has been in sex, age and ancestry estimation from various parts of the human skeleton. Knecht et al. approached four traditional statistical and two ML models (SVM and ANN) to compare sexual dimorphism from the greater

sciatic notch and reported the highest sex classification accuracy using ANN [45]. The machine-learning algorithms can be aggregated together to form a stacked model, which shows higher accuracy when compared to single algorithms. One such stacked ML algorithm was used for sex determination from the patella of the South African population and it showed an accuracy of 90.8% whereas the multivariate discriminant function showed an accuracy of 81.9–84% (Figure 10.4) [46].

Hefner and Ousley compared different machine-learning models of morphoscopic traits of the cranium to evaluate ancestry in African, European and Hispanic Americans. ANN has the greatest classification accuracy (87.9%) among ANN, SVM and RFM [47].

Convolutional Neural Networks (CNN)

CNNs are primarily used for image-based tasks, such as facial reconstruction and skeletal trauma analysis. They excel at capturing spatial relationships in images through convolutional layers and have been applied to various aspects of forensic anthropology. Unlike ANN, convolutional neural networks have multiple layers starting from the input layer, convolutional layer, pooling layer, fully connected layer and output layer. CNNs are known for their superior functioning in image, speech and audio input signals. It helps in computer vision which is a subfield of artificial intelligence that aids in image interpretation. 3D-CNN allows interpreting images using specialized 3D kernels that predict the segmentation in a volumetric patch of an image. A study aimed to establish sexual dimorphism from maxillofacial features from radiographs using CNN. Transfer learning (TL) CNN architecture provided a greater accuracy as compared to the other architecture [48]. Like other neural networks, CNN also requires a well-formed database. A study suggested age estimation of young children and adults using whole-body low-resolution x-ray images using CNN. The database used to train the CNN comprised 910 multispectral images and showed a minimum discrepancy in age calculation [49]. Facial reconstruction of damaged skulls is one of the most challenging tasks faced by forensic anthropologists. CNN can help in resolving this issue using special algorithms to not just provide a 3D reconstruction but also perform a biometric matching of the reconstructed skull [50, 51]. Age-at-death estimation is vital in forensic identification, but there is a lot of conflict in its accurate calculation. Convolutional neural networks have been shown to estimate the age precisely up to a certain extent. A reference dataset containing 500 individuals with ages between 19–101 years was used to train the CNN model. Also, a novel software DRNNAGE was developed to cater to age estimation within certain skeletal traits [52]. Most

age estimation methods from dental radiographs can be used for up to 25 years but after that age estimation in the elderly and adults becomes difficult. Hence, a novel CNN named 'Soft Stagewise Regression Network' SSR NET was developed to perform multi-stage age estimation from a single image. Based on SSR NET another novel software was developed named DENSEN to perform similar tasks but with a higher sensitivity [53]. Sex determination forms an important part of biological profiling and can also be done using convolutional neural networks. A study conducted using 1476 lateral cephalometric radiographs showed that CNNs can be used to obtain sexual dimorphism with an accuracy of 90% [54]. Ear biometrics are often used for somatometric examinations, and are easily found for personal identification. Ear biometrics remain unaffected by any facial expressions and hence deep learning convolutional networks can be used to study them. These neural networks must be trained with a known dataset before use. A study suggested that such a trained neural network was able to determine morphological landmarks accurately [50, 55].

Wen et al. used CNN for ancestry estimation from Chinese population groups (156 Yellow and 178 White skulls) as samples, in which 80% served as training sets and 20% as test sets. They obtained 95.88% accuracy on the training set and 95.52% accuracy [56].

Multivariate Adaptive Regression Splines (MARS)

MARS is a regression algorithm put forth by Jerome H. Friedman in 1991 and was an extension of the linear models. This algorithm works as a non-parametric technique and automatically models non-linear relationships between variables. Conventional juvenile age estimation methods often fall short of statistical validity (Figure 10.4). Hence a standard approach using MARS was used to predict age from iliac biometric variables. The study was conducted on 176 subjects aged between 0–12 years. The MARS predictive model depicted iliac width, module and area providing better age estimation [57]. Another study conducted on 1310 subjects aged between birth to 12 years, used MARS to predict age from diaphyseal dimensions. The results revealed that univariate models can be used only for younger children whereas multivariate diaphyseal length models generated better results for older children [11]. Another approach at sub-adult age estimation was made from the fifth lumbar vertebrae, clavicles of 534 males from the French population and iliac measurements of 244 subjects aged between 0–12 years. The MARS model was used to combine both sets of measurements. The model integrated the non-linear relationships among the variables and generated an accuracy of 92% [58].

eXtreme Gradient Boosting (Xgboost)

This machine-learning algorithm is an extension of gradient boosting in decision trees. Weights are associated to all the independent variables which are then assigned to the decision trees. The weight of the variables which are predicted incorrectly by the decision tree is increased and fed to the next decision tree. Hence, the total system now works together to provide a stronger and more accurate estimation. Age estimation among juveniles on the basis of mineralized dental morphology is an important task in forensic investigations (Figure 10.4). Shan et al. found *XGBoost* as the best predictive model for age estimation using the Demirjian method performed on 1477 panoramic radiographs of the South China population aged between 2–18 years [32]. SexEst is a free web application based on the extreme gradient boosting mathematical model designed for skeletal sex estimation from cranial and postcranial measurements. The models were optimized and gave a prediction accuracy of 80.8–89.5% for postcranial variables and 81.2–87.7% for cranial variables [59]. In a study conducted to predict age from permanent teeth in the Sri Lankan population, measurements from 3321 subjects were involved. The comparative results among machine-learning algorithms proved the extreme gradient boosting model (XGBoost) to be the best fit with a highest accuracy of 88% [60].

Decision Trees in Forensic Anthropology

Decision trees are a supervised learning technique used in machine learning. It basically provides a graphical representation of all the possible solutions to any problem. It contains two nodes namely a decision node and a leaf node. The decision node is responsible for all the decision-making and has numerous branches whereas the leaf node depicts the outcome of the decision. These decision trees are used for categorizing data and obtaining values based upon previous outcomes (Figure 10.4).

Advantages of Machine Learning

The main aim of forensic science is to solve a crime and bring faster justice to the victim. But, with the ever-growing number of crimes and varied 'modus operandii', the impending cases often hold back the justice delivery system. Machine learning with its quicker approach to identifying anatomical landmarks can help in reducing the count of pending cases. Also, the process

of analysing the skeletal evidence is a tiresome task, which can be solved easily if the ML algorithm is trained once. Since this field requires multitasking to ensure thorough analysis and reporting, neural networks can become an important tool while performing various identifications simultaneously. Manual estimation of the anthropological remains by professional knowledge and expertise can sometimes cause errors in minute calculations. These can be avoided using artificial intelligence. ML algorithms can identify trends and patterns which is useful while analysing large volumes of data. Also, with continuous development, these algorithms can function up to an extent where they no longer need human interference.

Best Performing Machine-Learning Method

The best AI method for personal identification problems related to skeletal remains may depend on various aspects including the quality and size of the available dataset, the specific identification attribute (age/sex/stature/racial affiliation), along with the experience and expertise of the forensic anthropologist [6]. Further, a combination of multiple ML techniques or employing hybrid approaches may yield better results by using the strengths of different algorithms.

There are several ML methods but each method is suitable for a specific biological attribute of skeletal remains. Some methods are good for sex prediction, some are good for ancestry estimation and some are good for age estimation and so on. Nikita et al. investigated the classification accuracy of statistical methods and machine learning algorithms including BLR, multinomial and penalised multinomial logistic regression (MLR, pMLR), LDA, NBC, DT, RF, ANN, SVM with linear, polynomial or radial kernels, MARS and *XGBoost* in the context of skeletal sex and ancestry estimation LDA and SVM perform best respectively, with high prediction accuracy and minimal bias in most tests (Figure 10.5) [6]. On the other hand, regression analysis and random forest suited well for stature estimation and for age estimation.

Limitations of Machine Learning in Forensic Anthropology

Despite becoming an indispensable tool in the field of forensic sciences, there are some specific shortcomings of machine learning. For the algorithm to become user-friendly and adjust to the training data sets it takes lots of trials. Therefore, it requires time and expenses to maintain such infrastructure. ML has immense promise in many domains, including forensic anthropology, but it has limits.

- **Limited Data Availability:** Forensic anthropology works with unique and complicated situations, making vast diverse datasets difficult to gather and the ML algorithms need plenty of data to learn and predict. If the data collected is incorrect, the results generated will be insignificant. Further manual or automated data acquisition is time-consuming and labour-intensive. Thus the scarcity of anthropological datasets limits the performance and generalization of the ML models for the biological attributes [61].
- **Lack of Interpretability:** Many ML algorithms, including deep-learning models, are 'black boxes' because they lack transparency and interpretability. In forensic anthropology, where expert views are respected, it might be difficult to explain how the algorithm produced a choice or prediction [61].
- **Bias and representativeness:** ML models depend on training data quality and representativeness.ML models may inherit and increase training data biases. This may impact forensic anthropological investigations by causing erroneous predictions or findings [5, 61]
- **Domain-Specific Problems:** In forensic settings, it is quite common to receive incomplete or damaged skeletal remains, which presents distinct obstacles. Such skeletal remains may restrict the features ML systems may use to derive meaningful patterns. When dealing with deteriorated or fragmentary remains, ML models may have trouble determining the age, sex, stature or ancestry estimation [5, 61].
- **Expertise and Human Involvement:** Machine-learning algorithms form a part of artificial intelligence which is yet again a simulation of human intelligence. Therefore, result interpretation and accuracy of results can only be achieved with human interference. Also, any error in the data or algorithm can ultimately lead to a wrong output. Forensic anthropology demands specialized knowledge and expertise. ML can automate processes and provide insights, but it cannot replace forensic anthropologists' experience and judgement [50].
- **Ethical Issues:** AI in forensic anthropology raises ethical issues as well. Researchers and practitioners must address the ethical and legal consequences of their work since any technology might be misused or have unforeseen repercussions in personal identification in different forensic situations. This involves responsibly obtaining and using data to train AI algorithms and assessing how AI-driven analysis may affect forensic investigators and their communities [50, 62].

Despite these drawbacks, ML may be useful in forensic anthropology. It might help forensic anthropologists analyze vast data sets, find trends, and

make educated conclusions. It should be utilized judiciously, with human competence, and with an understanding of its limits and biases. ML in forensic anthropology creates ethical difficulties. The findings from forensic anthropological examinations are admitted in the court as supporting evidence and hence ensure a fair trial and speedy delivery of justice.

Future Aspects of Machine Learning in Forensic Anthropology

Machine learning has widespread applications in the field of forensic anthropology especially pertaining to personal identification. The varied use of different supervised and unsupervised learning algorithms holds a promising future in anthropological examinations. These algorithms have enabled the analysis of high volumes of skeletal data with greater precision, accuracy and speed. These algorithms may also help in creating biological profiling along with facial reconstruction, time since death estimation, trauma and pathology identification [5]. Detailed analysis of the skeletal remains especially about population-specific morphometric variations can be easily accomplished using machine learning algorithms. It is imperative to incorporate an ethical approach while dealing with such confidential data which may be compromised by using these algorithms entirely. Hence, human intervention in the collection procedures and interpretation of data may be acknowledged in the usage of machine learning algorithms in forensic anthropological examinations.

References

1. Morrison, G. S., Weber, P., Basu, N., Puch-Solis, R., & Randolph-Quinney, P. S. (2021). Calculation of likelihood ratios for inference of biological sex from human skeletal remains. *Forensic Science International Synergy, 27*(3), 100202. https://doi.org/10.1016/j.jofri.2019.05.005
2. Bruzek, J. (2002). A method for visual determination of sex, using the human hip bone. *American Journal of Physical Anthropology, 117*(2), 157–168.
3. Fukuta, A., Kato, C., Biwasaka, H., Usui, A., Horita, T., & Kanno, S. (2020). Sex estimation of the pelvis by deep learning of two-dimensional depth images generated from homologous models of three-dimensional computed tomography images, *Forensic Science International: Reports, 2*, 100129.
4. Krenn, V. A., Webb, N. M., Fornai, C., & Haeusler, M. (2022). Sex classification using the human sacrum: Geometric morphometrics versus conventional approaches. *PLoS One, 17*(4), e0264770. https://doi.org/10.1371%2Fjournal.pone.0264770

5. Galante, N., Cotroneo, R., Furci, D., Lodetti, G., & Casali, M. B. (2023). Applications of artificial intelligence in forensic sciences: Current potential benefits, limitations and perspectives. *International Journal of Legal Medicine*, *137*(2), 445–458. https://doi.org/10.1007/s00414-022-02928-5
6. Nikita, E., & Nikitas, P. (2020). On the use of machine learning algorithms in forensic anthropology. *Legal Medicine, 47*, 101771. https://doi.org/10.1016/j.legalmed.2020.101771
7. Mangrulkar, A., Rane, S. B., & Sunnapwar, V. (2021). Automated skull damage detection from assembled skull model using computer vision and machine learning. *International Journal of Information Technology*, *13*, 1785–1790. https://doi.org/10.1007/s41870-021-00752-5
8. Ramanathan, N., Chellappa, R., & Biswas, S. (2009). Age progression in human faces: A survey. *Journal of Visual Languages & Computing*, *15*, 3349–3361.
9. Auyeung, T. W., Lee, J. S., Kwok, T., Leung, J., Leung, P. C., & Woo, J. (2009). Estimation of stature by measuring fibula and ulna bone length in 2443 older adults. *The Journal of Nutrition, Health & Aging*, *13*(10), 931–936. https://doi.org/10.1007/s12603-009-0254-z
10. Srivastava, R., Saini, V., Pandey, S. K, Rai, R. K., & Tripathi, S. K. (2012). A study of sexual dimorphism in the femur among North Indians. *Journal of Forensic Science, 57*(1), 19–23. 10.1111/j.1556-4029.2011.01885.x.
11. Stull, K. E., L'Abbé, E. N., & Ousley, S. D. (2014). Using multivariate adaptive regression splines to estimate subadult age from diaphyseal dimensions. *American Journal of Physical Anthropology*, *154,* 376–386. https://doi.org/10.1002/ajpa.22522
12. Mehta, M., Saini, V., Menon, S. K., & Patel, M. N. (2019). Applicability and reliability of foramen magnum for sex determination in contemporary Gujarati population: A computed tomographic study. *Journal of Forensic Radiology and Imaging*, *17*, 31–35. https://doi.org/10.1016/j.jofri.2019.05.005
13. Saini, V., Chowdhry, A., & Mehta, M. (2022). Sexual dimorphism and population variation in mandibular variables: A study on a contemporary Indian population. *Anthropological Science*, *130*(1), 59–70. https://doi.org/10.1537/ase.2108282
14. Kaur, S., & Saini, V. (2016). Secular changes on stature reconstruction from hand and foot dimensions among Sikhs of Delhi. *Journal of Forensic Research, 7(*2), 1000321. http://dx.doi.org/10.4172/2157-7145.1000321
15. Spiros, M. C., & Hefner, J. T. (2020). Ancestry estimation using cranial and postcranial macromorphoscopic traits. *Journal of Forensic Sciences, 65*(3), 921–929. https://doi.org/10.1111/1556-4029.14231
16. Pal, D., Ghosh, A., Majumdar, S., Pan, A. & Ghosh, D. (2023). Machine learning in healthcare: A review. *Methods, 12*(1), 60–66. https://doi.org/10.2174%2F1389202922666210705124359
17. Saini, V., Srivastava, R., Shamal, S. N., Singh, T. B., Kumar, V., Kumar, P., & Tripathi, S. K. (2014). Temporal variations in basicranium dimorphism of North Indians. *International Journal of Legal Medicine, 128*(4), 699–707. https://doi.org/10.1007/s00414-013-0957-x

18. Saini, V. (2019). Secular trends in cranial chord variables: A study of changes in sexual dimorphism of the North Indian population during 1954–2011. *Annals of Human Biology, 46*(6), 519–526. https://doi.org/10.1080/03014460.2019.1677773
19. Bytheway, J. A., & Ross, A. H. (2010). A geometric morphometric approach to sex determination of the human adult os coxa. *Journal of Forensic Sciences*, *55*, 859–864. https://doi.org/10.1111/j.1556-4029.2010.01374.x
20. Gonzalez, R. A. (2012), Determination of sex from juvenile crania by means of discriminant function analysis. *Journal of Forensic Sciences, 57*, 24–34. https://doi.org/10.1111/j.1556-4029.2011.01920.x
21. Musilova, B., Dupej, J., Veleminska, J., Chaumoitre, K., & Bruzek, J. (2016). Exocranial surfaces for sex assessment of the human cranium. *Forensic Science, International*, *269*, 70–77. https://doi.org/10.1016/j.forsciint.2016.11.006
22. Curate, F., Umbelino, C., Perinha, A., Nogueira, C., Silva, A. M., & Cunha, E. (2017). Sex determination from the femur in Portuguese populations with classical and machine learning classifiers, *Journal of Forensic Legal Medicine*, *52*(1), 75–81, https://doi.org/10.1016/j.jflm.2017.08.011
23. Fliss, B., Luethi, M., Fuernstahl, P., Christensen, A. M., Sibold, K., Thali, M., & Ebert, L.C. (2019). CT-based sex estimation on human femora using statistical shape modeling, *American Journal of Physical Anthropology, 169*(2), 279–286, https://doi.org/10.1002/ajpa.23828.
24. Imaizumi, K., Bermejo, E., Taniguchi, K., Ogawa, Y., Nagata, T., & Kaga, K. (2020) . Development of a sex estimation method for skulls using machine learning on three-dimensional shapes of skulls and skull parts. *Forensic Imaging*, *22*, 200393. https://doi.org/10.1016/j.fri.2020.200393
25. Wang, Y. H., Liu, T. A., Wei, H., Wan, L., Ying, C. L, & Zhu, G. Y. (2016). Automated classification of epiphyses in the distal radius and ulna using a support vector machine. *Journal of Forensic Sciences*, *61*(2), 409–414. https://doi.org/10.1111/1556-4029.13006
26. Sinthubua, A., Theera-Umpon, N., Auephanwiriyakul, S., Ruengdit, S., Das, S., & Mahakkanukrauh, P. (2016). New method of age estimation from maxillary sutures closure in a Thai population. *La Clinicaterapeutica*, *167*(2), 33–37. https://doi.org/10.7417/CT.2016.1918
27. Lei, Y. Y., Shen, Y. S., Wang, Y. H., & Zhao, H. (2019). Regression algorithm of bone age estimation of knee-joint based on principal component analysis and support vector machine. *Fa yi xue za zhi, 35*(2), 194–199. https://doi.org/10.12116/j.issn.1004-5619.2019.02.012
28. Hefner, J. T., Spradley, M. K., & Anderson, B. (2014). Ancestry assessment using random forest modelling. *Journal of Forensic Sciences, 59*(3), 583–589. https://doi.org/10.1111/1556-4029.12402
29. Navega, D., Coelho, C., Vicente, R., Ferreira, M. T., Wasterlain, S., & Cunha, E. (2015). AncesTrees: ancestry estimation with randomized decision trees. *International Journal of Legal Medicine, 129*(5), 1145–1153. https://doi.org/10.1007/s00414-014-1050-9

30. Henriques, M., Bonhomme, V., Cunha, E., & Adalian, P. (2023). Blows or falls? Distinction by random forest classification. *Biology*, *12*(2), 206. https://doi.org/10.3390/biology12020206
31. Darmawan, M. F., Abidin, A. F. Z, Kasim, S., Sutikno, T., & Budiarto, R. (2020). Random forest age estimation model based on length of left hand bone for Asian population. *International Journal of Electrical and Computer Engineering, 10*(1), 549. http://doi.org/10.11591/ijece.v10i1.pp549-558
32. Shan, W., Sun, Y., Hu, L., Qiu, J., Huo, M., Zhang, Z., & Yue, X. (2022). Boosting algorithm improves the accuracy of juvenile forensic dental age estimation in southern China population. *Scientific Reports, 12*(1), 15649. https://doi.org/10.1038/s41598-022-20034-9
33. Zolotenkova, G. V., Rogachev, A. I., Pigolkin, Y. I., Edelev, I. S., Borshchevskaya, V. N., & Cameriere, R. (2022). Age classification in forensic medicine using machine learning techniques. *Sovremennye tehnologii v Medicine*, *14*(1), 15–22. http://doi.org/10.17691/stm2022.14.1.02.
34. Kim, J., Lee, S., Choi, I., Jeong, Y., Woo, E. J. (2022). A comparative analysis of Bayesian age-at-death estimations using three different priors and Suchey-Brooks standards. *Forensic Science International, 336*, 111318. https://doi.org/10.1016/j.forsciint.2022.111318
35. d'OliveiraCoelho, J., & Curate, F. (2019). CADOES: An interactive machine—learning approach for sex estimation with the pelvis. *Forensic Science International, 302*, 109873. https://doi.org/10.1016/j.forsciint.2019.109873
36. Verma, R., Krishan, K., Rani, D., Kumar, A., Sharma, V., Shrestha, R., & Kanchan, T. (2020). Estimation of sex in forensic examinations using logistic regression and likelihood ratios. *Forensic Science International: Reports*, *2*, 100118. https://doi.org/10.1016/j.fsir.2020.100118
37. Rani, D., Krishan, K., & Kanchan, T. (2023). A methodological comparison of discriminant function analysis and binary logistic regression for estimating sex in forensic research and case-work. *Medicine, Science and the Law, 63*(3), 227–236. https://doi.org/10.1177/00258024221136687
38. Santos, F., Guyomarc'h, P., & Bruzek, J. (2014). Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines. *Forensic Science International, 245*, 204.e1–204.e2048. https://doi.org/10.1016/j.forsciint.2014.10.010
39. DiGangi, E. A., & Hefner, J. T. (2013). *Ancestry estimation. Research methods in human skeletal biology* (pp. 117–149). Academic Press, Elsevier.
40. Saunders, S. R., & Hoppa, R. D. (1997). Sex allocation from long bone measurements using logistic regression. *Canadian Society of Forensic Science Journal, 30*(2), 49–60. https://doi.org/10.1080/00085030.1997.10757086
41. Bidmos, M. A., Adebesin, A. A., Mazengenya, P., Olateju, O. I., & Adegboye, O. (2021). Estimation of sex from metatarsals using discriminant function and logistic regression analyses. *Australian Journal of Forensic Sciences*, *53*(5), 543–556. https://doi.org/10.1080/00450618.2019.1711180
42. Verma, R., Krishan, K., Rani, D., Kumar, A., & Sharma, V. (2020). Stature Estimation in forensic examinations using regression analysis: A likelihood ratio perspective. *Forensic Science International: Reports, 2*, 100069. https://doi.org/10.1016/j.fsir.2020.100069

43. Lo, M., Mariconti, E., Nakhaeizadeh, S., & Morgan, R. M. (2023). Preparing computed tomography images for machine learning in forensic and virtual anthropology. *Forensic Science International: Synergy, 6,* 100319. https://doi.org/10.1016/j.fsisyn.2023.100319
44. Yuvaraj, N., Kousik, N., Raja, R. A., & Saravanan, M. (2020). Automatic skull-face overlay and mandible articulation in data science by AIRS-Genetic algorithm. *International Journal of Intelligent Networks*, *1*, 9–16. https://doi.org/10.1016/j.ijin.2020.05.003
45. Knecht, S., Nogueira, L., Servant, M., Santos, F., Alunni, V., Bernardi, C., & Quatrehomme, G. (2021). Sex estimation from the greater sciatic notch: A comparison of classical statistical models and machine learning algorithms. *International Journal of Legal Medicine*, *135*(6), 2603–2613. https://doi.org/10.1007/s00414-021-02700-1
46. Bidmos, M. A, Olateju, O. I., Latiff, S., Rahman, T., & Chowdhury, M. E. H. (2023). Machine learning and discriminant function analysis in the formulation of generic models for sex prediction using patella measurements. *International Journal of Legal Medicine, 137*(2), 471–485. https://doi.org/10.1007/s00414-022-02899-7
47. Hefner, J. T., & Ousley, S. D. (2014). Statistical classification methods for estimating ancestry using morphoscopic traits. *Journal of Forensic Sciences, 59*(4), 883–890. https://doi.org/10.1111/1556-4029.12421
48. Franco, A., Porto, L., & Heng, D.(2022) . Diagnostic performance of convolutional neural networks for dental sexual dimorphism. *Scientific Reports 12*, 17279 . https://doi.org/10.1038/s41598-022-21294-1
49. Janczyk, K., Rumiński, J., Neumann, T., Głowacka, N., & Wiśniewski, P. (2022). Age prediction from low resolution, dual-energy x-ray images using convolutional neural networks. *Applied Sciences, 12*(13), 6608. https://doi.org/10.3390/app12136608
50. Thurzo, A., Kosnáčová, H. S, Kurilová, V., Kosmeľ, S., Beňuš, R., & Moravanský, N. (2021). Use of advanced artificial intelligence in forensic medicine, forensic anthropology and clinical anatomy. *Healthcare, 9*(11), 1545. https://doi.org/10.3390/healthcare9111545
51. Dubey, R. K., & Choubey, D. K. (2023). Deconstructive human face recognition using deep neural network. *Multimedia Tools Application, 82*, 34147–34162. https://doi.org/10.1007/s11042-023-15107-4
52. Navega, DCosta., E., & Cunha, E. (2022). Adult skeletal age-at-death estimation through deep random neural networks: A new method and its computational analysis. *Biology*, *11*(4), 532. https://doi.org/10.3390/biology11040532
53. Wang, X., Liu, Y., Miao, X., Chen, Y., Cao, X., Zhang, Y., Li, S., & Zhou, Q. (2022). DENSEN: A convolutional neural network for estimating chronological ages from panoramic radiographs. *BMC Bioinformatics, 23*, 426. 10.1186/s12859-022-04935-0.
54. Khazaei, M., Mollabashi, V., Khotanlou, H., & Farhadian, M. (2022). Sex determination from lateral cephalometric radiographs using an automated deep learning convolutional neural network. *Imaging Science in Dentistry, 52*(3), 239–244. https://doi.org/10.5624/isd.20220016

55. Cintas, C., Quinto-Sanchez, M., & Acuna, V. (2017). Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks. *IET Biometrics*, *6*(3), 211–223. https://doi.org/10.1049/iet-bmt.2016.0002
56. Wen, Y., Mingquan, Z., Pengyue, L., Guohua, G., Xiaoning, L., & Kang, L. (2020). *Ancestry estimation of skull in Chinese population based on improved convolutional neural network*. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South), pp. 2861-2867.
57. Corron, L., Marchal, F., Condemi, S., Chaumoître, K., & Adalian, P. (2017). A new approach of juvenile age estimation using measurements of the ilium and multivariate adaptive regression splines (MARS) models for better age prediction. *Journal of forensic Sciences*, *62*(1), 18–29. https://doi.org/10.1111/1556-4029.13224
58. Corron, L., Marchal, F., Condemi, S., Telmon, N., Chaumoitre, K., & Adalian, P. (2019). Integrating growth variability of the ilium, fifth lumbar vertebra, and clavicle with multivariate adaptive regression splines models for subadult age estimation. *Journal of Forensic Sciences*, *64*(1), 34–51. https://doi.org/10.1111/1556-4029.13831
59. Constantinou, C., & Nikita, E. (2022). SexEst: An open access web application for metric skeletal sex estimation. *International Journal of Osteoarchaeology*, *32*(4), 832–844. https://doi.org/10.1016/j.fsir.2023.100317
60. De Silva, H. H., Nawarathna, L. S., & Vithanaarachchi, V. S. N. (2022). Machine learning regression tree approach for age prediction from eruption status of permanent teeth in Sri Lankan children. *Journal of Advances in Mathematics and Computer Science, 37*(1), 1–7. https://doi.org/10.9734/jamcs/2022/v37i130425
61. Stewart, M. (2019). The limitations of machine learning. *Towards Data Science*. https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6
62. Frąckiewicz, M. (2023). AI in Robotic forensic anthropology. *Artificial Intelligence, TS2 Space*. https://ts2.space/en/ai-in-robotic-forensic-anthropology/

Potential Application of Machine Learning in Forensic Ballistics

11

POOJA AHUJA, KANICA CHUGH AND NIHA ANSARI

Introduction

Forensic ballistics, the study of firearms, ammunition and the effects of projectiles, is a crucial field within forensic science. It plays a pivotal role in criminal investigations by analyzing ballistic evidence to link firearms to specific crimes [1]. Machine learning (ML) is a rapidly evolving technology, which offers many opportunities for intelligence data analysis [2, 3]. ML and artificial-intelligence (AI) applications are ubiquitous and ever-changing in our society, many of which are illustrated in Figure 11.1. A brief classification of ML is also detailed in Figure 11.2. Despite this pervasiveness, machine-learning forensics (MLF) remains an emerging field within forensic science—and one that is still relatively underdeveloped. However, a real potential exists to leverage machine learning to identify criminal patterns, predict criminal activities (e.g., predict the location and timing of crimes) and automate investigative processes [4, 5]. As a result, there is a significant opportunity to bridge this gap and harness ML's capabilities for advancing forensic science [6].

The integration of machine learning (ML) in forensic ballistics specifically holds significant potential for enhancing the efficiency and accuracy of analyses. Here are several potential applications of machine learning in forensic ballistics:

- **Bullet and Cartridge Categorization:** Machine-learning algorithms can be trained to classify and categorize bullet and cartridge characteristics. By analyzing the markings on fired bullets and cartridge cases, ML models can identify unique patterns, such as rifling marks and firing pin impressions. This automated classification can significantly speed up the initial stages of investigations.
- **Firearm Identification:** ML can assist in the identification of the firearm used in a crime by analyzing the unique ballistic signatures left

DOI: 10.4324/9781003449164-12

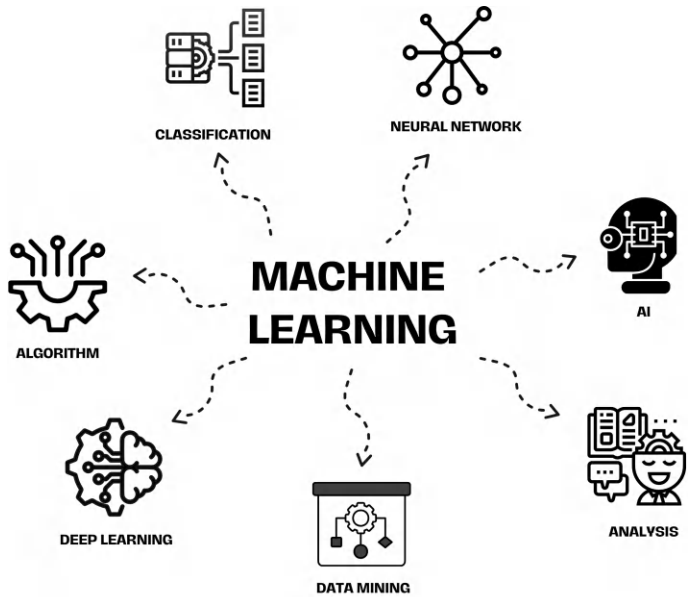

Figure 11.1 Applications of machine learning. Source: https://medium.com/hashmapinc/data-science-for-executives-bd6e766a6a19

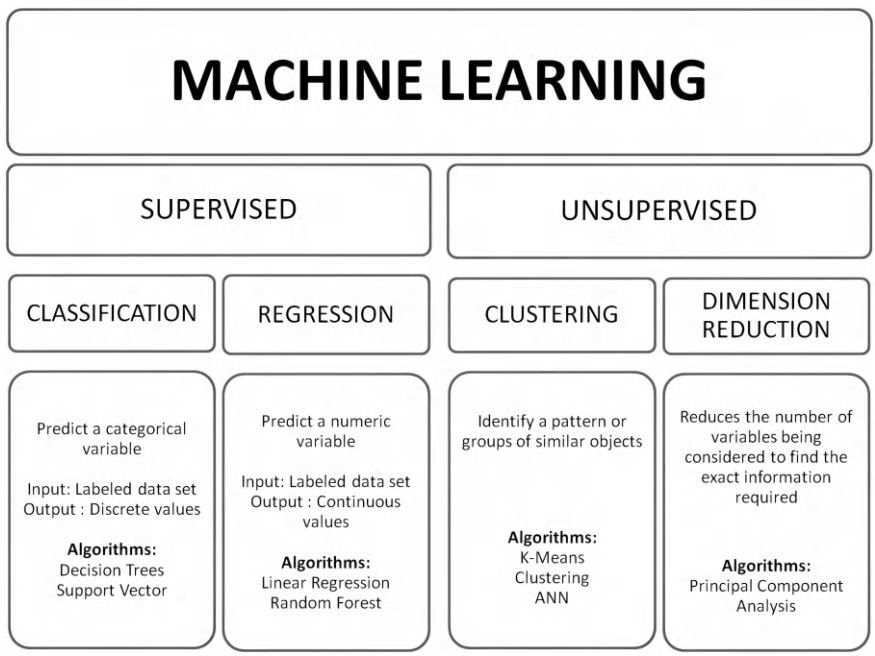

Figure 11.2 Classification of machine learning [35].

on projectiles. This involves creating a database of firearm signatures and training algorithms to match these signatures to the characteristics of recovered bullets or cartridge cases. Automated identification can reduce human error and enhance the speed of investigations.

- **Trajectory Analysis:** Machine learning can aid in reconstructing the trajectory of projectiles. By analyzing the impact patterns and angles, as well as considering environmental factors, ML algorithms can assist forensic experts in determining the likely path of a fired projectile. This information is crucial for understanding the dynamics of a shooting incident.
- **Shot Placement and Impact Analysis:** ML algorithms can analyze acoustic signals related to bullet impacts, helping forensic experts understand the trajectory and location of each shot. This contributes to the reconstruction of the event and the determination of the shooter's position.
- **Pattern Recognition in Bullet Striations:** Machine learning can be employed to analyze microscopic features on bullets, such as striations and markings left by the firearm's barrel. ML algorithms can identify unique patterns, facilitating the matching of recovered bullets to specific firearms.

Identification of firearms is one of the most crucial, complex and difficult aspects of a criminal investigation. Regarding the marks on fired bullets and cartridge cases, each firearm, regardless of size, manufacturer or model, has its own unique fingerprint. ML/AI will greatly simplify the task of determining the potential impact area of a projectile [7]. Using image processing, artificial neural networks can aid specialists in scanning the database for gunpowder and cartridge cases, and comparing bullet markings, firearm identity and other ballistic evidence without requiring direct intervention. A unique analytical technique for recognizing projectile samples utilizing line-scan imaging, based on the rapid Fourier transform [8]. Also, machine learning in forensic ballistics is used to examine the class characteristics such as rifling marks and individual characteristics of a bullet [9].

In the future, machine learning and artificial intelligence will also be utilized to assist military decision-makers [10]. Armed forces will receive augmented reality data via heads-up displays and weapon control systems, which will be provided by artificial intelligence. It will be used to identify and classify threats, prioritize targets and display the location of friendly forces and safe distances surrounding them. The AI system will use data from many sensors across the battlefield to construct a picture based on information that the current army is unaware of. In the near future, soldiers will handle the majority of military actions, but AI will provide analysis and

recommendations based on large datasets that are too massive for unaided humans to comprehend [7, 9]. Also in the future, miniature robots can be employed on the battlefield to deliver loaded magazines to individual soldiers as their guns' basic combat load runs out [10].

Applications of Machine Learning in Forensic Ballistics

Firearm Identification

The recordings of firearms utilized at the scene of an increasing number of crimes are accessible for investigative purposes. The objective of forensic investigation is to ascertain whether the item under scrutiny is, in fact, a firearm, as well as to specify its model, calibre and other relevant characteristics. The forensic analysis of cartridge discharge sounds (specifically, shock waves and muzzle blasts) is widely recognized [11–14] and is utilized in criminal investigations. The integration of acoustic evidence and exterior ballistic calculations can facilitate the determination of the shooter's distance and location [15, 16]. Moreover, an examination of the noise generated by a projectile throughout its trajectory and upon impact is possible [17–19]. An additional application of acoustic analysis of crime scene records is the determination of the sequence of shots fired by various shooters [10, 20].

Firearms produce various acoustic signals during their operation, including the sound of a shot, the sound of a flying bullet and the sound of the bullet's impact. These acoustic signals can be analyzed, compared, and identified to differentiate between different types of firearms and their mechanical actions. Forensic investigation of the sounds of cartridge discharge, combined with exterior ballistic calculations, can help determine the distance to the shooter and their position. By analyzing the acoustic signals made by firearms, including the sound of a shot, the sound of a flying bullet and the sound of the bullet's impact, it is possible to identify different types of firearms and their mechanical actions. Machine learning has been found to be the most promising method for analyzing and identifying these acoustic signals, which can be used in forensic identification.

Gun-Shot Residue Analysis

Quantitative profile–profile relationship (QPPR) modelling is a novel machine-learning technique that aims to establish associations and predictions regarding the chemical properties of unspent ammunition derived from gunshot residue (GSR) [21–23]. By leveraging the post-discharge GSR profiles, the methodology forecasts the pre-discharge chemical profiles of

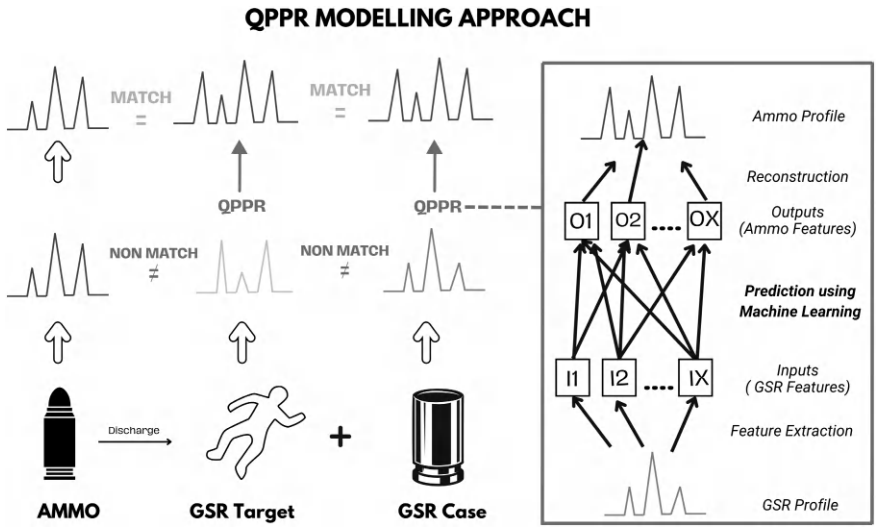

Figure 11.3 Machine learning predicts ammunition from gunshot residue [21].

specific ammunition components using contemporary machine-learning techniques. Comparing the predicted profiles to one another and to other profiles that have been measured enables forensic investigators to establish evidentiary connections. The efficacy of the QPPR modelling approach in forecasting GC–MS profiles of smokeless powders (SLPs) derived from organic GSR in spent cases has been demonstrated as described in Figure 11.3. The experimentally determined profiles closely resemble the predicted profiles [24] providing a quantitative method for associating GSR with particular types of ammunition, this novel approach has the capacity to accurately link evidence in a variety of forensic scenarios [25].

Bullet and Cartridge Categorization

Machine learning is helpful in the identification of bullets by providing a systematic approach to computing the similarity between two ballistic images and classifying between genuine matches and false matches. In the context of ballistic image matching, machine-learning techniques, specifically a supervised learning-based approach, have been shown to be superior to non-learning-based methods. By leveraging advances in computer vision and machine learning, a learning-based approach can address the limitations of prior work, particularly in the context of forensic practice [26]. The method employs a gentle boost-based learning scheme to select a discriminative subset of local cells in the spatial domain, with each cell constituting a weak classifier using the classic cross-correlation function (CCF) score [27]. The

proposed approaches for ballistic image matching involve a learning-based method to compute the similarity between two ballistic images with breech face impressions [26, 28]. The study compares the proposed approach with state-of-the-art methods on both controlled laboratory data (NIST (National Institute of Standards and Technology) dataset) and a newly collected operational forensic lab (OFL) dataset. The results show promising performances on the NIST dataset, with the proposed approach achieving perfect classification. It outperforms the NIST techniques in terms of expected error, indicating its effectiveness in ballistic image matching. On the more challenging OFL dataset, the proposed approach also performs better than the global cross-correlation method, especially at low false-positive rates [29]. The proposed approach demonstrates potential in addressing these challenges and highlights the need for a large operational benchmark ballistic image database to develop probabilistic models for ballistics matching [30, 31]. The approaches show promising performances and outperform state-of-the-art methods in terms of expected error, especially on the more challenging operational forensic lab dataset.

Overall, machine learning plays a crucial role in improving the accuracy and reliability of ballistic image matching, especially in operational forensic settings. It allows for the development of a systematic and scientifically sound method for comparing ballistic images, addressing concerns about the reliability and validity of tool mark-based forensic evidence [32].

Shooting Distance Estimation

Many factors come into play with shooting distance estimation. Take, for example, shotgun patterning, which Oura et al. notes is due to several factors [33]:

> Firstly, shotgun barrel length has a major effect on pattering and short barrels … tend to offer a wide spread pattern already from short shooting distances. Longer barrels in turn tend to provide tighter patterns, Secondly, the choke has a major influence on patterning …In addition to barrel length and selection of the choke, several factors such as bore size, pellet size and material (e.g., lead vs. steel) influence patterning.

These various factors combine to create a complex shotgun pattern, making it an important consideration in forensic analysis and the estimation of shooting distance.

Machine-learning techniques are increasingly used to interpret shooting distances, with some studies applying regression analysis and TinyResNet-based algorithms to estimate the range of fire for various firearms, including

shotguns, AK-47s and Karshinov rifles [34]. In such studies, regression analysis is instrumental for estimating shotgun pellet patterns, with shooting distances typically falling within the confidence intervals. The findings offer hope as to the usefulness of regression analysis in estimating shooting distance from shotgun patterns. Additionally, while another study explored the potential of neural network architectures to classify shotgun pattern images based on shooting distance, due to the complex nature of shotgun fire, future studies will no doubt be needed to develop reliable and applicable algorithms for such pattern interpretation. [34].

Considering factors such as choke, shell and pellet variation are crucial in future studies to develop more accurate algorithms. The ultimate goal is to develop robust and generalizable algorithms that will serve as a beneficial tool for forensic investigators, improving the accuracy of forensic shotgun pattern interpretation, particularly in scenarios with limited background information available [34]. The findings have implications for forensic investigators and law enforcement agencies, suggesting that deep learning algorithms could improve the accuracy and efficiency of forensic shotgun pattern interpretation in the future.

Advantages of Machine Learning in Forensic Science

- **Pattern Recognition and Classification:** Machine-learning algorithms demonstrate proficiency in the discernment of intricate patterns within ballistic evidence, thereby facilitating the classification of unique firearm-related characteristics. This capacity enhances the identification and differentiation of firearms based on nuanced markings and projectile trajectories.
- **Efficient Bullet Matching:** Automated systems underpinned by machine-learning methodologies markedly enhance the expeditious matching of bullets to specific firearms. Through the analysis of distinctive markings on bullets and casings, machine-learning algorithms streamline the identification process, thereby reducing manual workload and expediting investigative timelines.
- **Advanced Ballistic Imaging Analysis:** Machine learning contributes to the refinement of ballistic imaging analysis by automating the comparison and matching of intricate striations on bullets or cartridge cases. Automated image recognition algorithms serve to aid forensic experts in identifying and interpreting complex patterns, surpassing the capabilities of conventional methodologies.

- **Enhanced Database Matching:** The integration of machine learning into ballistic databases establishes a framework for expedited and precise matching of ballistic evidence with existing records. This capability streamlines the investigative process, facilitating the prompt identification of potential linkages to prior criminal incidents.
- **Reduction of Human Error:** The application of automation through machine learning mitigates the inherent risk of human error in forensic analysis. By automating routine and repetitive tasks, forensic experts can redirect their focus towards more intricate aspects of the investigation, culminating in heightened reliability and accuracy of results.
- **Predictive Modeling for Firearms and Ammunition:** Machine-learning models, when appropriately trained, exhibit the capacity to predict various characteristics related to firearms, ammunition or shooting incidents. This predictive capability empowers investigators with informed decision-making tools, thereby augmenting situational awareness during criminal investigations.
- **Holistic Data Fusion:** Machine learning facilitates the integration of information from diverse forensic sources, engendering a comprehensive understanding of a crime scene. This integrative approach, synthesizing ballistics data with inputs from other forensic disciplines, engenders a nuanced perspective, potentially revealing latent connections and insights.

The infusion of machine-learning methodologies into forensic ballistics confers a spectrum of advantages, ranging from increased efficiency to augmented analytical capabilities, thereby significantly contributing to the refinement of criminal investigative processes.

Disadvantages of Machine Learning in Forensic Science

- **Data Quality and Bias in Machine Learning Models:** The efficacy of machine-learning (ML) models in forensic ballistics is contingent upon the quality and representativeness of the training dataset. If the dataset exhibits bias or lacks comprehensive coverage, ML models may produce inaccurate or skewed results, particularly in the context of firearm and ammunition diversity.
- **Overfitting Challenges:** ML models are susceptible to overfitting, a phenomenon wherein models perform exceedingly well on the

training dataset but falter in generalizing to new, unseen data. In forensic ballistics, overfitting may lead to models overly tailored to specific ballistic evidence types, impeding adaptability across diverse scenarios.

- **Interpretability Concerns:** The interpretability of ML models, particularly complex architectures such as deep neural networks, poses a challenge in forensic ballistics. The opacity of decision-making processes may hinder forensic experts' ability to comprehend and explain the rationale behind a model's conclusions, raising questions about transparency and interpretability.
- **Human Expertise versus Machine Learning:** The intricate expertise and intuition inherent in human forensic analysts may not be fully encapsulated by ML models. The contextual understanding and experiential knowledge of forensic experts, crucial in the nuanced interpretation of complex ballistic evidence, may not be adequately represented by machine learning methodologies.
- **Dynamic Nature of Forensic Science:** Forensic science, including ballistics, evolves continually with advancements in methodologies and technologies. ML models may face challenges in adapting rapidly to these changes, necessitating frequent updates and retraining to maintain accuracy and relevance.

The application of machine learning in forensic ballistics demands careful consideration of these scientific challenges to ensure the robustness, fairness and ethical integrity of the investigative process. Collaboration between experts in machine learning and forensic analysis is essential for navigating these complexities.

Conclusion

Forensic ballistics is a crucial field in forensic science, analyzing firearms and ammunition to link them to specific crimes. Machine learning (ML) is a rapidly expanding discipline that provides intelligent data analysis skills. ML is used in various applications, including bullet and cartridge categorization, firearm identification, trajectory analysis and shot placement and impact analysis. However, the application of ML in forensic ballistics is still in its infancy due to the lack of awareness among forensic scientists and experts in ML and data mining. The integration of ML in forensic ballistics holds significant potential for enhancing efficiency and accuracy. Machine learning (ML) can be used in forensic ballistics to analyze microscopic features on bullets,

identifying unique patterns and matching recovered bullets to specific firearms. This technology simplifies the task of determining the potential impact area of a projectile, and can also be used to assist military decision-makers in identifying threats and prioritizing targets. Applications of ML in forensic ballistics include firearm identification, acoustic analysis and determining the sequence of shots fired by various shooters. In the future, AI systems will provide analysis and recommendations based on massive datasets.

References

1. Barash, M., McNevin, D., Fedorenko, V., & Giverts, P. (2024). Machine learning applications in forensic DNA profiling: A critical review. *Forensic Science International: Genetics*, 69(102994), 102994. https://doi.org/10.1016/j.fsigen.2023.102994
2. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)* [Internet], 9(1), 381–386.
3. Scaruffi, P. (2018). *Intelligence is not artificial - expanded edition: A history of artificial intelligence and why the singularity is not coming any time soon*. Createspace Independent Publishing Platform.
4. Qadir, A. M., & Varol, A. (2020). *The role of machine learning in digital forensics*. 2020 8th International Symposium on Digital Forensics and Security (ISDFS).
5. Tageldin, L., & Venter, H. (2023). Machine-learning forensics: State of the art in the use of machine-learning techniques for digital forensic investigations within smart environments. *Applied Sciences* (Basel, Switzerland), 13(18), 10169. https://doi.org/10.3390/app131810169
6. Scientific Working Group on DNA Analysis Methods (SWGDAM): Validation Guidelines for DNA Analysis Methods. (n.d.).
7. Kudonu, M., AlShamsi, M. A., Philip, S., Khokhar, G., Hari, P. B., & Singh, N. (2022). *Artificial intelligence: Future of firearm examination*. 2022 Advances in Science and Engineering Technology International Conferences (ASET). https://doi.org/10.1109/aset53988.2022.9735105
8. Khan, S., Divakaran, A., & Sawhney, H. S. (2010). *Weapon identification using acoustic signatures across varying capture conditions*. US Patent Application US20100271905A1.
9. Magic bullets: The future of artificial intelligence in weapons systems. www.army.mil. (n.d.). Retrieved August 21, 2022, from https://www.army.mil/article/223026/magic_bullets_the_future_of_artificial_intelligence_in_weapons_systems
10. Li, S.-T., Kuo, S.-C., & Tsai, F.-C. (2010). An intelligent decision-support model using FSOM and rule extraction for crime prevention. *Expert Systems with Applications*, 37(10), 7108–7119.
11. Giverts, P., Sofer, S., Solewicz, Y., & Varer, B. (2020). Firearms identification by the acoustic signals of their mechanisms. *Forensic Science International*, 306(110099), 110099. https://doi.org/10.1016/j.forsciint.2019.110099

12. Haag, L. C. (1979). A preliminary inquiry into the application of sound spectrography to the characterizations of gunshots. *AFTE Journal*, *11*, 61–63.
13. Hollien, H., & Hollien, K. A. (1994). Acoustic patterning of small-arms gunfire. *AFTE Journal*, *26*, 41–49.
14. Haag, L. C. (2003). Light and sound as physical evidence in shooting incidents. *AFTE Journal*, *35*, 317–321.
15. Paredes, D. M., & Apolinario, J. A. (2014). Shooter localization using microphone arrays on elevated platforms. 2014 IEEE Central America and Panama Convention (CONCAPAN XXXIV). IEEE, pp. 1–6. https://doi.org/10.1109/CONCAPAN.2014.7000419
16. McCombs, N. D., & Vargas, E. (2014). The sound of shots. *AFTE Journal*, *46*, 33–42.
17. Haag, L. C. (2016). The exterior and terminal ballistics of the model 1780 Girardoni air rifle carried by Meriwether Lewis during the voyage of discovery 1803–1806. *AFTE Journal*, *48*, 131–137.
18. Haag, L. C. (2002). The sound of bullets. *AFTE Journal*, *34*, 255–263.
19. Maher, R. C. (2016). Gunshot recordings from a criminal incident: Who shot first? *The Journal of the Acoustical Society of America*, *139*, 2024. https://doi.org/10.1121/1.4949969
20. Inman, K., & Rudin, N. (2001). *Principles and practices of criminalistics*. CRC Press.
21. Haag, M. G., & Haag, L. C. (2011). *Shooting incident reconstruction* (2nd ed.). Academic Press.
22. Heard, B. J. (2008). *Handbook of firearms and ballistics* (2nd ed.). John Wiley & Sons.
23. Gallidabino, M. D., Barron, L. P., Weyermann, C., & Romolo, F. S. (2019). Quantitative profile–profile relationship (QPPR) modelling: A novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from gunshot residue (GSR). *The Analyst*, 144(4), 1128–1139. https://doi.org/10.1039/c8an01841c
24. Rakhimbekova, A., Madzhidov, T. I., Nugmanov, R. I., Gimadiev, T. R., Baskin, I. I., & Varnek, A. (2020). Comprehensive analysis of applicability domains of QSPR models for chemical reactions. *International Journal of Molecular Sciences*, 21(15), 5542. https://doi.org/10.3390/ijms21155542
25. Aalbers, S. E. (2013). *The evidential value of gunshot residue composition comparisons*. Delft University of Technology.
26. Roth, J., Carriveau, A., Liu, X., & Jain, A. K. (2015). *Learning-based ballistic breech face impression image matching*. 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS).
27. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504
28. Biedermann, A., & Taroni, F. (2006). A probabilistic approach to the joint evaluation of firearm evidence and gunshot residues. *Forensic Science International*, 163(1–2), 18–33. https://doi.org/10.1016/j.forsciint.2005.11.001

29. Daugman, J. G. (1993). High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11), 1148–1161. https://doi.org/10.1109/34.244676
30. Weller, T. J., Zheng, A., Thompson, R., & Tulleners, F. (2012). Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides. *Journal of Forensic Sciences*, 57(4), 912–917. https://doi.org/10.1111/j.1556-4029.2012.02072.x
31. Vorburger, T., Yen, J., Bachrach, B., Renegar, T., Filliben, J., & Ma, L. (2007). *Surface topography analysis for a feasibility assessment of a national ballistics imaging database*. NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online].
32. Song, J.-F., & Vorburger, T. V. (2000). *Proposed bullet signature comparisons using autocorrelation functions*. Proceedings of National Conference of Standards Laboratories.
33. Oura, P., Junno, A., & Junno, J.-A. (2021). Deep learning in forensic shotgun pattern interpretation-A proof-of-concept study. *Legal Medicine*, *53*, Artikkeli 101960. https://doi.org/10.1016/j.legalmed.2021.101960
34. Alfonsi, A., Calatri, S., Cerioni, E., & Luchi, P. (1984). Shooting distance estimation for shots fired by a shotgun loaded with buckshot cartridges. *Forensic Science International*, 25(2), 83–91. https://doi.org/10.1016/0379-0738(84)90017-3
35. Tangirala S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications(IJACSA)*, *11*(2), 612–619. http://dx.doi.org/10.14569/IJACSA.2020.0110277

Application of Machine Learning in Big Data Analysis

12

SUMIT KUMAR CHOUDHARY, SURBHI MATHUR, PRAVESH SHARMA AND ANUBHAV SINGH

Introduction

Almost all human enterprises across diverse professional, cultural, geographical or social congregations have become data-centric and data-driven. The tendency and dependency on data usage is increasing and so is the data itself. The precipitation of more and more data has been created, which we now call 'big data'. Big data is an asset as well as a challenge for all modern-day enterprises. Big data is a great enabler if put to use in a meaningful manner. A careful review can present a wonderful correlation between your past actions and past outcomes. The mistakes are retrieved and opportunities for course correction for a better outcome in future are promulgated. An insight into what better work has been done in the past is also evidenced and gives a clue about how it can be further improved. Thus, several procedural steps can be amended, resurrected or augmented based on big data analysis. Figure 12.1 shows various techniques utilized for data mining. Resource management and its optimal applications can be ensured for enhanced efficiency. Reasons for failures can be determined and expunged. Trends and specific patterns can be intelligently discerned, which can give a glance into future trajectories. Risk–benefit balance can be reworked in the favour of the benefit. Red flags and green flags can be better understood for well-informed actions regarding what is to be dropped and what is to be pursued. In essence, predictive analytics and smart decisions can be made out of big data analytics.

It is essential that big data as a concept is well understood before we move on to applying machine-learning algorithms to conceive its utility and challenges. Figure 12.2 shows the three Vs of big data, variety, volume and velocity. These describe the essence of the task at hand and also give a glimpse of the complexity pertaining to data that needs to be handled. Managing and analyzing big data poses serious challenges. The challenge of big data lies in the speed with which it is expanding (velocity), the exponential amount

DOI: 10.4324/9781003449164-13

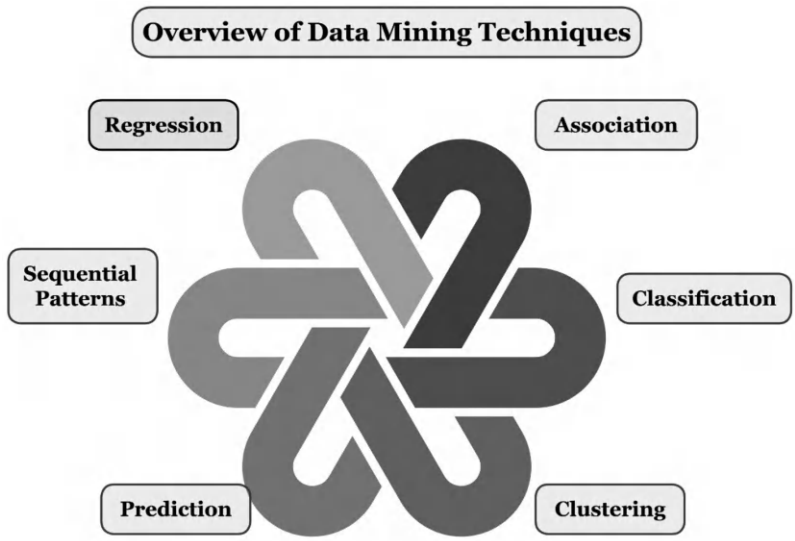

Figure 12.1 Collection of data mining techniques.

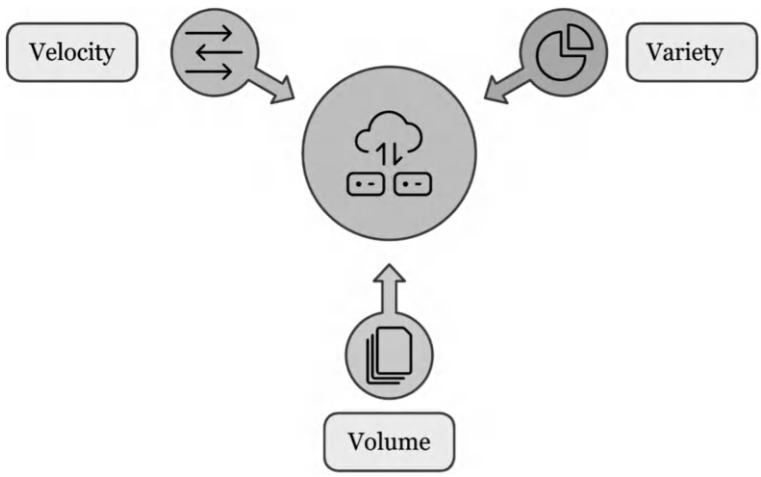

Figure 12.2 The three Vs of big data.

which is being constantly added (volume), the diversity and complexity of the incoming data (variety) and the multiplication of sources through which new data is making the big data larger every day. So, with big data amalgamated with so many challenging dimensions, the analysis of such data is an arduous task. The conventional data analytics approaches may fall short of expectations when confronted with the complexity of big data and this would only underline the maxim that modern problems need modern solutions.

Machine learning (ML) commonly referred to as a subfield of artificial intelligence is capable of processing large amounts of data based on its algorithm and can learn from these data to improve further data analytics, classification and pattern identification and finally predict things. Machine learning in big data analytics can help businesses, industries and even law enforcement agencies to identify patterns, forecast outcomes in a precise manner and gain valuable insights from vast amounts of data [1]. Thus, big data with its velocity, volume and variety traits can be effectively handled and necessitates the application of sophisticated methods such as machine learning for its analysis [2].

Basics of Machine Learning

Understanding the core fundamentals of machine learning is quintessential before applying it in the analysis of big data. Machine-learning algorithms enable a system to take input data, learn from it, produce meaningful insights or predict things and constantly improve its performance based on more data availability for perfecting its predictive attribute. Primarily there are three basic types of machine learning: (i) supervised learning, (ii) unsupervised learning and (iii) reinforcement learning.

When the system learns and gets trained on labelled datasets and accordingly makes predictions, it is called supervised learning [3]. While it may have a very high accuracy percentage in prediction as it is trained on a labelled dataset but may suffer from failing to identify unknown patterns which were not part of the training dataset. The system uses logistic regression and linear regression for binary classification and predicting continuous variables [4]. For improved output, these models are required to be trained on huge amounts of data allowing them to learn scalable implementations, parallel processing and distributed computing techniques [5].

Conversely, when the system is trained on unlabelled datasets and the system develops the ability to make predictions or generate outputs based on its ability to identify new patterns and structures, it is known as unsupervised learning. The system explores hidden patterns, relationships and similarities between the unlabelled data to come out with a specific probability for new input data [6].

Sometimes, a semi-supervised learning model is also found which works with partial similarity to both supervised and unsupervised learning by using both labelled and unlabelled data for training. Reinforcement learning is feedback-based learning, where after every output, the system enhances its learning through reward–penalty feedback. The surrounding has a definitive impact on the system learning in this case. The system improves every time after receiving positive or negative feedback on their outcomes.

Preprocessing and Feature Engineering

Big data by its nomenclature is characterized by large voluminous data which may be very complex as well as unstructured. Thus, to process such data through machine-learning tools, it is important to preprocess the raw data through feature engineering involving steps such as cleaning and altering the raw data. Cleaning typically refers to processing the data to reduce noise, manage missing values and resolve inconsistencies. For altering the data or transforming the data as we may call it, scaling normalization and encoding the variables should be done [7]. Feature engineering can further help to pick pertinent features from within the raw data for improving the overall performance and output of ML models.

Neural Networks and Deep Learning

The ability of machine learning to deal with large volumes and complexly structured data through deep learning has gained immense traction of late. Recurrent neural networks (RNNs) are frequently used for sequential data analysis such as voice and natural language processing. Similarly, convolutional neural networks (CNNs) find a lot of application in tasks pertaining to image identification [8]. The deep-learning models have to a great extent transformed the ability of systems towards language understanding and machine translation. The deep-learning models can be effectively trained on massive amounts of data with inherent complexity using distributed computing frameworks, parallel processing and specialized hardware accelerators such as tensor processing units (TPUs) and graphics processing units (GPUs).

Applications of Big Data

The use of machine learning for analyzing big data has several real-world applications in almost all intersections of human enterprise such as in industries, manufacturing units, cybersecurity measures and risk mitigation, marketing, the healthcare sector and the financial segment. The financial sector benefits from the deployment of ML models in evaluating credit risk, preventing fraud, trading activities, etc. It supports drug research, personalized therapy and the identification of diseases in the healthcare industry. Through customer segmentation, targeted advertising and churn prediction, the marketing industry benefits from machine learning. Machine learning is used in manufacturing for predictive maintenance, supply chain optimization and

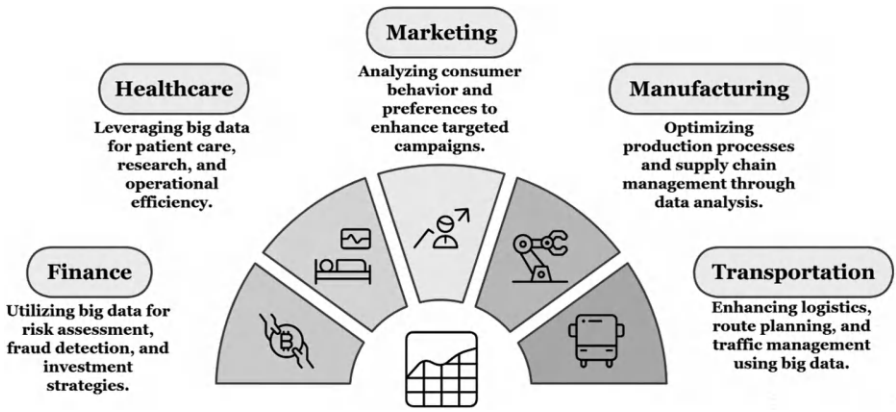

Figure 12.3 Applications of big data.

quality control. Machine learning is used in cybersecurity for network intrusion detection, threat intelligence and anomaly detection [9]. Each application emphasizes the particular problems that machine-learning algorithms solve as well as the advantages that result from their application.

Different industries have been revolutionized by big data, which has made it possible for businesses to learn useful lessons, make better decisions and spur innovation [10, 11].

Here are a few significant big data uses from various industries as represented in Figure 12.3.

Finance

- **Identifying and Preventing Fraud:** Big data analytics can find patterns and irregularities in financial transactions to quickly spot fraud.
- **Risk Assessment:** Financial organizations can more effectively analyze and manage risks by analyzing massive amounts of financial and market data.
- **Algorithmic Trading:** Complex trading algorithms that can make quicker and more educated investment decisions can be created thanks to big data analysis.

Healthcare

- **Personalized Medicine:** It is possible to customize treatments and drugs for specific individuals by analyzing vast amounts of patient data, such as medical records, genomic data and lifestyle data.

- **Disease Prediction and Early Detection**: Big data analysis from enormous healthcare data sources can suitably be used to learn trends for predicting early signs of onset of illness and what kind of interventions at a preliminary stage can yield better results.
- **Drug Discovery:** The process of new drug discovery can benefit immensely from the speedy analysis of a variety and large volumes of data such as genomic, chemical and biochemical and the data from clinical trials and their results.

Marketing and Customer Analytics

- **Customer Segmentation:** Marketing houses can proficiently classify and cluster their buyer base in view of their spending behaviour, socioeconomics and interests based on insights gained from big data analysis.
- **Targeted Advertising:** Organizations can customize their product-promoting endeavours and make the advertisements more target-oriented by giving specific messages of interest to specific crowds by examining client information and their online behaviour.
- **Churn Prediction and Customer Retention:** Big data analysis can assist organizations in minimising client saturation and devise methods to keep their most significant clients intact and their purchasing intent active.

Manufacturing and Supply Chain

- **Predictive Maintenance:** Data analysis can be used to foresee and predict maintenance schedules, support prerequisites and avert failures which may disrupt the manufacturing process abruptly.
- **Quality Control:** Data analytics can examine previous manufacturing and performance records to realize quality issues or shortcomings to address them and improve the quality of products.
- **Supply Chain Optimization:** Big data analytics can throw light on problems with inventory management in the past and how and where to improve it. Similarly, cost-cutting exercises, enhancing efficiency, etc., can also be achieved through data-based insights.

Transportation and Logistics

- **Route Optimization:** Route optimization can be achieved through big data analysis of ongoing traffic data, weather situations and past records and trends.

- **Demand Forecasting:** Big data analysis can help the sales industry manage inventories and production by gauging patterns and trends of demands, consumer behaviour, sales data and other influencing factors.
- **Fleet Management:** Big data analytics can help examine information from vehicles, such as fuel use, upkeep prerequisites and driver conduct to maintain flawless fleet operations and reduce operational costs.

Energy and Utilities

- **Smart Grid Management:** Smart meter, sensor and framework foundation information can be broken down utilizing big data analysis to improve energy dissemination, lower blackouts and boost productivity.
- **Energy Consumption Optimization:** Associations can track down opportunities to save energy and distribute assets all the more proficiently by dissecting meteorological information and patterns of energy utilization.
- **Renewable Energy Optimization:** To expand the understanding and use of environmentally friendly power sources, big data analysis can be utilized to predict weather conditions, energy yield measurements and demand.

The above applications are only indicative and not comprehensive of what big data analysis is capable of. There can be several more, encompassing almost all areas of human lives and enterprise. The future holds even more promise and compelling use of big data analysis as the data will continue to grow exponentially and technology-led innovations with ML models will keep on evolving.

Ethical Considerations and Challenges

As machine learning is increasingly used for big data analysis, ethical issues and difficulties are brought to the fore. It is important that these ethical challenges are addressed appropriately in order to freely apply ML tools for big data analysis. Some pertinent ethical dilemmas are discussed below [12]:

- **Data Privacy:** A lot of private data may form part of the big data that is to be analyzed and hence organizations need to put necessary safeguards in place to ensure that the privacy of data are respected and also the data are protected as per the provisions of data protection

laws in force. Data encryption, storing data in secure mode or anonymization methods can prove to be useful in reducing the threat of compromising data privacy.

- **Bias and Fairness:** ML algorithms can yield skewed or biased results if the algorithm even unintentionally supports any kind of biases or predispositions in the data used for training the model. It's thus important that such biases, wherever are pre-eliminated for the model to learn correctly and yield fair outcomes. The fairness in outcome for any ML model can be ensured and enhanced by implementing methods such as bias detection, model interpretability and fairness-aware learning. These together would enhance fairness and reduce prejudice in results.
- **Interpretability and Explainability:** With the increasing complexity of ML models and algorithms, it becomes difficult to precisely understand the workings of the model. Interpretation and explanation are important in understanding the process of decision-making and also accounting for it. Methods such as feature importance analysis, model visualization and rule extraction make it easier to understand the rationale behind the ML model making a particular decision.
- **Data Quality and Reliability:** The quality of big data to be processed generally suffers from quality issues, noise and other data gaps. The precision and reliability of ML models can be badly affected by poor quality of data. Thus to ensure the quality of data, data validation procedures, data cleaning approaches and data validation methods need to be done.
- **Scalability and Resource Constraints:** For handling and processing huge amounts of information as big data resource constraints such as appropriate infrastructure and algorithms need to be put in place. Similarly, with data getting bigger and bigger, the processing unit should be scalable to higher and higher limits. Distributed computing frameworks, parallel processing and cloud infrastructure can help the ML model implementation to overcome both the issues of scalability and resource constraints.
- **Algorithmic Transparency and Accountability:** It is critical to lay out responsibility and receptiveness since ML calculations regularly make significant decisions in many fields. Transparency and accountability inspire more confidence and trust in users in the system.
- **Human Oversight and Decision-Making:** Human intervention in the form of result verification, compliance with ethical standards

and deciphering intricate patterns are very important for the system to work for optimal and legitimate output by forging this human–machine collaborative decision-making system.

Big data analysis using machine learning offers numerous advantages and prospects for use in a variety of industries. However, it is crucial to address the problems and ethical issues that come with this integration. Organizations may utilize the power of machine learning while minimizing risks by giving data protection, fairness, interpretability and accountability a top priority. Machine learning for big data analysis will be used ethically and for the benefit of society if there is constant monitoring and evaluation of algorithms and their effects on society.

Analysis of Big Data Using Machine Learning

Machine learning is fundamental and indispensable for big data analysis owing to its capability of fast and accurate outputs helping users to gain insights on critical matters of their interest. ML can help in the following ways in big data analysis:

- **Pattern Recognition:** A vast amount of data can most of the time present complex patterns and intra-linkages which can be difficult to understand manually or with simple statistical tools. Machine learning offers the capability to decipher such relationships and patterns between the data even if the data is unstructured or with multiple variables [13].
- **Predictive Analytics:** Machine learning has this distinctive capability of making data-based predictions by analyzing the data patterns and relationships. This is the biggest strength of ML models for their wide acceptance as their forecasts are extremely valuable for decision-making in almost all human enterprises [13, 14].
- **Classification and Clustering:** Machine learning can very effectively and meaningfully classify and cluster voluminous data into smaller categories. These clusters can then be evaluated for trend forecasting to aid in decision-making [14].
- **Anomaly Detection:** A large amount of information can be utilized by ML algorithms to track down peculiarities or anomalies. These oddities could be indications of a red flag, fraud, irregular behaviour or information issues which should assist the user in making an informed decision in real-time [15].

- **Natural Language Processing (NLP):** There are several ML-enabled NLP applications that can perform sentiment analysis, text categorization, topic modelling and language translation, which helps in processing unstructured textual data to bring out gainful insights from such data. These data may include text information, including messages or e-mails, social media posts and customer reviews [16].
- **Recommendation Systems:** ML models can produce tailor-made recommendations for business enterprises by analyzing a huge amount of data pertaining to customer behaviour, choices, spending habits, purchasing time periods, item qualities purchased in the past, etc. These recommendations can guide businesses to further focus more on customer preferences, improve engagement, achieve higher customer satisfaction, customise advertisements, etc. [17].
- **Deep Learning:** A type of ML called deep learning involves preparing neural networks with various layers to naturally examine complex data relationships and retrieve patterns for forecasting information. They can process massive amounts of information and thus are suited for big data analysis [18].
- **Scalability and Efficiency:** The machine-learning algorithms can be programmed in a way which can be scaled up to handle more and more amounts of data also maintaining and capable of enhancing the efficiency with which this massive amount of data is being processed. The ML model can effectively use a distributed computing network for load distribution [19].
- **Data Preprocessing and Feature Engineering:** Machine-learning algorithms yield the best outcomes while dealing with organized and clean data. However, such data is a rarity. Thus, preprocessing and feature engineering of data are required to optimize the output. The preprocessing steps may include data cleansing, data normalization and feature optimization. On the other hand, feature engineering involves steps such as developing fresh features or selecting important features for better outcomes of ML models. Preprocessing and feature engineering are important as ML models often handle multidimensional data with a lot of variety [20].
- **Supervised Learning for Prediction and Classification:** ML models are prepared as supervised learning algorithms for handling big data to give accurate predictions and classifications of data. The patterns and relationships of a dataset are analyzed for making predictions or classification of the said data. This has wide applications in fraud detection, risk assessment, sentiment analysis and customer churn prediction [21].

- **Unsupervised Learning for Pattern Discovery and Clustering:** Unsupervised learning fittingly works on unlabelled data to discover hidden structures and patterns within the data. This has applications in customer segmentation, anomaly detection and pattern recognition. In cases of multidimensional data, the number of unwarranted dimensions is reduced without disturbing its structure and variability using techniques such as t-distributed stochastic neighbour embedding (t-SNE) and principal component analysis (PCA) [22].
- **Deep Learning for Complex Pattern Recognition:** To handle and analyze data with extreme complexities and intricate patterns, deep learning models have evolved which use neural networks of multiple layers. Deep-learning models are capable of automatically discovering these complexities within data and analysing it with accuracy for precise predictions. The two most common deep learning models are: recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Both these are frequently used in big data analysis including tasks related to natural language processing, picture, audio and video recognition and recommendation systems [23].
- **Real-Time Analysis and Streaming Data:** ML algorithms are also capable of analyzing real-time analysis of ongoing streaming of data making it possible for users to make instantaneous decisions on issues of urgency. Algorithms such as online learning, incremental learning and adaptive models facilitate real-time analysis and prediction possible through continuous analysis of streaming data [24].
- **Scalable Machine Learning:** Scalable ML models are essential for handling the volume, velocity and variety of data in big data analysis. This can be achieved by distributed computing networks such as Apache, Spark and Hadoop that distribute the computing load on several computers or clusters. Tensor processing units (TPUs) and graphics processing units (GPUs) are examples of dedicated hardware solutions that are helpful in parallel processing [25].
- **Optimization and Automation:** The performance of ML models is improved with steps such as optimization and automation. Optimization approaches such as gradient descent, genetic algorithms and Bayesian optimization can be performance enhancers. Similarly, automation of data cleaning, feature selection, model selection and hyperparameter tuning also plays an important role in performance enhancement. Thus, ML models provide users with the privilege of automating procedures, getting accurate predictions, meaningful insights, and advantageous decision-making [26].

Introduction to Deep Learning

Learning how to represent data hierarchically, deep learning relies on training neural networks with numerous layers. Deep learning makes it possible to automatically extract intricate features and patterns from raw data, in contrast to typical machine-learning techniques that rely on manually created features. Due to its capacity to handle difficult tasks including natural language processing, speech recognition and image recognition, it has drawn considerable attention and gained popularity.

Historical Background

The idea of artificial neural networks was first proposed in the 1940s and 1950s, which is where deep learning first emerged. Deep learning sprang to prominence, however, in the 1980s and 1990s because of the creation of techniques such as backpropagation that made it possible to train deep neural networks effectively [27]. Deep learning has advanced more quickly in recent years thanks to improvements in computer power, the accessibility of large-scale datasets and the creation of specialized hardware (such as GPUs).

To Understand Deep Learning, It Is Essential to Grasp Several Key Concepts

- **Neural Networks:** Deep learning is built on neural networks. They are made up of layers of neurons, which are interconnected nodes. Each neuron takes in information, uses an activation function and sends the output to the layer below. An input layer, one or more hidden layers and an output layer are possible in neural networks. The data can be represented in progressively more complicated ways thanks to the hidden layers.
- **Activation Functions:** The neural network can simulate intricate interactions between inputs and outputs thanks to the non-linearities introduced by activation functions. The sigmoid, tanh and ReLU (rectified linear unit) activation functions are frequently used. They use a neuron's input to determine its output value.
- **Backpropagation:** Deep neural networks are trained using the core algorithm of backpropagation. It involves propagating mistakes backwards from the output layer to the input layer to iteratively change the network weights. Backpropagation enhances the performance of the network by minimizing the discrepancy between expected and actual outputs.

- **Deep Neural Networks:** Deep neural networks may learn hierarchical data representations since they have several hidden layers. From the input data, each layer pulls features that are more and more abstract. The depth of the network makes it suitable for handling high-dimensional and unstructured data since it enables the learning of complicated patterns and complex relationships.
- **Deep Learning Architectures:** Different deep learning architectures have been created to handle various data and task types. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) make use of spatial correlations and temporal correlations respectively to analyze image and video and sequential textual data.
- **Deep Neural Networks:** The underlying innovation of deep learning is deep neural networks. They are composed of interconnected neurons or hubs organized in layers. Every neuron learns, actuates it and afterwards sends its result to the layer underneath. Deep neural networks can learn progressive information representation, which empowers them to perceive unpredictable patterns and connections.
- **Feedforward Networks:** It is also called multilayer perceptrons (MLPs), and can be considered as the most basic type of deep neural network. They comprise a hidden layer or layers, an output layer and an input layer. Data goes in a unidirectional manner from the input layer through the hidden layers to the output layer.
- **Convolutional Neural Networks (CNNs):** CNNs were created with the sole purpose of processing grid-like input, such as photos and movies. They make use of the idea of convolution, which entails applying a group of teachable filters on the incoming data. Convolutional layers are used by CNNs for feature extraction, and pooling layers are used for downsampling and lowering the number of spatial dimensions. After that, fully linked layers for classification or regression receive the collected features. In applications including image classification, object identification and image segmentation, CNNs have displayed excellent performance [28].
- **Recurrent Neural Networks (RNNs):** To handle issues such as time series, sequential data, speech and text, recurrent neural networks (RNNs) were developed. RNNs, as opposed to feedforward networks, have recurrent connections that enable data to remain and be processed throughout time steps. RNNs can detect temporal dependencies in the data as a result. An RNN uses its hidden state as input and memory at each time step, which affects predictions made later. These two well-liked RNN variations, long short-term memory (LSTM) and gated recurrent unit (GRU) help to solve

the vanishing gradient issue and make it easier to learn long-term dependencies [28].

- **Deep Belief Networks (DBNs):** A class of generative models known as deep belief networks (DBNs) consists of numerous layers of stochastic binary units. Using a method known as restricted Boltzmann machines (RBMs), DBNs are trained in an unsupervised way. The deep belief networks when trained can generate new samples, which resemble the training data set. DBN models can be widely used for dimensionality reduction, collaborative filtering and generative modelling [28].

Deep Neural Networks for Classification and Regression

Deep learning has proven to be an exceptionally well-performing model for recognizing complex correlations, capturing hierarchical representations and in number of categorization or classification tasks. Let us see how deep learning models work in three important classification tasks: sentiment analysis, text categorization, and image identification:

- **Deep Learning for Image Recognition:** Image recognition through deep learning models is a challenging task involving classifying patterns, and object or facial identification from a photograph or any form of visual data. However deep learning models such as convolutional neural networks (CNNs), capable of mining out important features from raw pixel data have powered the image recognition abilities tremendously. Pre-trained CNN models use their multiple interconnected layers, and pooling layers to extract the local features and patterns for obtaining high accuracy in recognising images and object identification from new datasets [29].
- **Deep Learning for Text Categorization:** Text classification involves characterizing text-based data into predetermined groups, such as papers or articles. Recurrent neural networks (RNNs) and transformer models have been found to be particularly promising with their applications in text categorization. RNNs are able to explore and establish contextual relevance and sequence of information in the given data so the RNN models can easily carry out document classification, sentiment analysis and topic modelling. Similarly, having trained on extensive textual data and by implementing self-attention methods, transformer models such as BERT (bidirectional encoder representation from transformers), have proved to be a valuable tool in natural language processing [29].

- **Deep Learning for Sentiment Analysis:** The objective of sentimental analysis is to recognize the emotion or polarity communicated in text, for example, whether it is positive or favourable, or negative or against or neutral or impartial. By gathering the semantic and context-oriented subtleties in the message, deep learning approaches have shown to be capable of sentimental analysis. Sentimental analysis tasks are effectively performed by recurrent neural networks (RNNs) and convolutional neural networks (CNNs). While RNN models such as LSTM and DRU excel in modelling the sequence of information in data and long-term contextual information, the CNN models equipped with recognizing local patterns and features perform well with the textual data comprising syntactic and compositional data [29].

Introduction to Deep Neural Network Architectures

The deep learning model of machine learning is built on deep neural network topologies. The deep neural network helps the model to identify complex patterns and intricate correlations within the data. Let us briefly discuss two of the most important deep neural network architectural models: (i) multi-layer perceptrons (MLPs) and (ii) convolutional neural networks (CNNs). While CNNs are explicitly intended for handling network-like information, making them especially useful for image classification tasks, MLPs succeed at distinguishing intricate non-linear relationships [30].

(i) **Multi-Layer Perceptrons (MLPs):** Multi-layer perceptrons (MLP), as the name suggests, are made up of multiple layers of neurons, each interconnected with its preceding and succeeding layer of neurons. The first layer represents the 'input layer', which is passed through several intermediary layers known as 'hidden layers' and finally the last layer which is referred to as 'output layer'. Due to this interconnectivity for processing the input data, it is also known as a feedforward neural network as each neuronal layer provides an activation function and passes on the output to the subsequent layer in the channel. The multiple layers of neurons are able to recognise even the non-linear correlations from the input data through progressive learning and thus are useful in handling data requiring complex decision-making and classification. MLPs are ordinarily trained through backpropagation, a technique for modifying the network's weights in view of the error occurring between expected and noticed results [31].

(ii) **Convolutional Neural Networks (CNNs):** Convolutional neural networks (CNNs) were made with the sole reason of handling

frameworks such as information, for example, photographs and movies. Convolution is a method utilized by CNNs, and it involves applying a progression of channels which can be trained to the information or input data. A strategy known as pooling is utilized to consolidate the results of the channels, which limits the spatial aspects while holding the main data. The channels or filters are premeditated to capture the local features and patterns such as surface texture and edge patterns, etc. Multiple pooling layers are an important characteristic of CNN models. The CNN models capture hierarchical depictions of data to extract the relevant features from the input data while the pooling layers downsample the feature maps, thereby reducing the computational load but still safeguarding the vital information within the data. All these retrieved features are then cumulatively processed through all the associated layers of CNN to make the final prediction [32]. CNNs are especially useful in classifying images, visual data, image recognition tasks, facial identification in forensics and object detection tasks for general and forensic applications owing to their ability to capture and work with local and spatial patterns and ability to take advantage of translational invariance and hierarchical structure.

Introduction to Regression Problems in Deep Learning

Regression analysis is based on a statistical approach to gauge the correlation between a dependent and independent variable to make predictions based on the various regression models. Regression problems focus on making predictions regarding continuous variables and recognizing the relationships and patterns within the data through deep learning techniques. Let us examine the application of deep learning models in forecasting continuous variable and time series:

- **Deep Learning for Predicting Continuous Variables:** Regression problems make use of deep learning models such as multi-layer perceptrons (MLPs) and variants of recurrent neural networks (RNNs), to gain effectiveness in predicting the continuous variables. The model can be trained on labelled data for the expected outcome. The MLPs can be modified by changing the output layer to have a single neuron without an activation function for applying it in regression challenges. The ability of MLPs to capture non-linear relationships between the input data and the target variable makes MLPs particularly useful in regression analysis when the relationship between the two is complex and non-linear [33, 34]. RNNs can be used for time-dependent regression tasks and are effective for modelling sequential

data, such as long short-term memory (LSTM) and gated recurrent unit (GRU).

- **Deep Learning for Time Series Forecasting:** Historical data points help in predicting forthcoming values pertaining to time series forecasting. Deep learning models such as RNNs and their types such as LSTM and GRU perform well in predicting time series. RNNs are good at recognizing temporal linkages and patterns in time series data. The problem of vanishing gradients is resolved by the LSTM and GRU RNN variations, enabling the learning of long-term relationships in time series data [35].

Deep Reinforcement Learning:

Deep Reinforcement Learning (DRL) is a very powerful AI-based tool which combines two very potent fields—the deep neural network and the reinforcement learning. This model consists of several important elements:

- **State Representation:** DRL models characteristically process the unencoded or encoded data from the environment and represent its current state.
- **Action Selection:** Depending upon the current state, the model decides action using a policy network which correlates states to probable actions.
- **Value Estimation:** Value networks are used by deep RL agents to calculate the worth of being in a certain state or taking a certain action.
- **Experience Replay:** The DRL technique learns more effectively through experience replay. Such experiential learning helps the model to make better choices and decisions.
- **Deep Q-Networks (DQNs):** DQNs constitute a sub-category of deep reinforcement learning (DRL). This utilizes a deep neural network to make an approximation of the action-value function Q (s, a), which is the fundamental idea behind DQNs. The Q-network receives the current state and outputs the Q-values for every possible course of action. The agent chooses actions based on the Q-values and iteratively changes the Q-network to improve performance [36].
- **Deep Generative Models:** These models allow for the synthesis of new examples and the development of fresh samples that mimic the training data. Deep generative models' capacity to produce realistic and varied samples across a range of domains, including images, text and audio, has attracted a lot of attention [36].
- **Variational Autoencoders (VAEs):** One kind of deep generative model that combines concepts from variational inference with

autoencoders is called a variational autoencoder (VAE). To encourage the encoder to create latent representations that adhere to a particular distribution, usually a Gaussian distribution, VAEs during training optimize a loss function [37]. VAEs have been effectively used in a variety of fields, such as molecular design, text production and the creation of images, and they offer a strong foundation for understanding complex data distributions.

- **Generative Adversarial Networks (GANs):** A generative adversarial network is a type of deep generative model having two networks—a generator network and a discriminator network, which together play an adversarial game. The discriminator network attempts to distinguish precisely between genuine data and created samples. The generator unit keeps on creating samples that very closely resemble the genuine data so that it goes unnoticed. This adversarial training process augments the learning process of the model so that it can create top-quality samples based on the training data [38]. GAN models have been very successful in creating highly realistic images, closely resembling human faces and artistic works.
- **Flow-Based Models:** As the name indicates flow-based models are based on the concept of flow. The data in this model is changed through a series of invertible mapping to facilitate accurate calculation of sample likelihood. These are deep generative models which have several applications such as anomaly detection, density estimation and image production [39].
- **Semantic Indexing:** Semantic indexing deals with **the** automatic extraction of important or meaningful data from within the voluminous unstructured data. It can be carried out with the help of deep learning–based ML models such as CNN and RNN.
- **Discriminative Tasks and Semantic Tagging:** Discriminative task deals with extracting any specific information which is possible as information is classified and stored in groups based on semantic tagging, etc. Text categorization, sentiment analysis, and named entity recognition (NER) are a few typical examples of discriminative tasks [39].
- **Sentiment Analysis:** Sentiment analysis helps in determining the sentiment of any text information. It may be a letter, a social media post or any such thing. The emotion or sentiment of the text is classified as positive, negative or neutral. This method can be extremely beneficial for business houses in analyzing customer feedback, brand progression, etc. In elections, public sentiments at large can be determined. For sentiment analysis, machine-learning methods such as supervised learning using labelled data, support vector machines

(SVMs), recurrent neural networks (RNNs) and transformer-based models are frequently utilized [39].

- **Named Entity Recognition (NER):** Recognition and classification of names of individuals, organizations, objects, places, dates, etc., within a text is called 'named entity recognition' (NER). ML techniques such as conditional random fields (CRFs), bidirectional LSTM-CRFs and transformer-based models are the most commonly used models for NER which performs information retrieval, question-answering, as well as text summarization [40].
- **Text Categorization:** Text categorization, also known as text classification is the method of organizing text information into separate groups. Text classification is crucial for content organization, information retrieval and prediction systems. Naive Bayes, support vector machines (SVMs) and deep learning models such as convolutional neural networks (CNNs) and transformer-based models are common ML models used for text categorization [41].

Introduction to Semantic Tagging and Understanding

Semantic tagging and understanding refers to the attaching of some kind of tag or label to the input textual information and mining out some insightful message from it. This enables the ML models to comprehend the textual information better and also to organize, index and analyze the text data. Semantic tagging and understanding applications can facilitate better information retrieval processes and natural language processing of data. This has several applications, some of which are: (i) named entity recognition (NER), (ii) part-of-apeech (POS) tagging by giving words in a sentence grammatical label, such as nouns, verbs, adjectives, etc., (iii) sentiment analysis, (iv) topic modelling and (v) question-answering. The two most powerful methods of semantic tagging and understanding are: Word2Vec and long short-term memory (LSTM) networks. The effectiveness of ML applications in all these text analysis-based applications is improved collectively by Word2Vec and LSTM networks.

- **Word2Vec:** Word2Vec helps to learn word embeddings, thereby helping in semantic tagging and understanding. This method treats words as dense vector representations in a continuous space and thus makes it feasible to understand the semantic correlation between the words and similarity indexing. The continuous bag-of-words (CBOW) and skip-gram architectures of Word2Vec help train the model to expect and recognize closely related words when applied to new input data of similar context words. Thus, tasks such as semantic

tagging, word similarity or analogies can be performed with the help of these models [42].

- **Long Short-Term Memory (LSTM) Networks:** RNN models of the LSTM type are able to correlate data sequences bearing distant relationships or linkage. LSTM models are functionally powered to explore the context of the information even distantly placed to establish the linkage and thus find a good amount of application in named entity identification, text classification and semantic tagging. LSTMs have demonstrated a history of performing well in discovering the semantic structure of textual information because they integrate and apply memory cells with gating mechanisms to select, store and retrieve any relevant information [42].

Predictive Modelling

Predictive modelling is a powerful tool and outcome of big data analytics wherein past data, patterns and trends are analyzed to make future trend predictions or projections. Mathematical models and algorithms are created to draw inferences from past actions or data and analyze them to forecast future actions, behaviours or probable events. A massive amount of data is being analyzed through a machine-learning approach to make credible predictions by predictive modelling [43]. The business houses benefit immensely from the predictive modelling in making advantageous decisions such as enhanced decision-making, improved efficiency and performance, improved personalized experiences for consumers, fraud detection and risk management, healthcare and medical applications, etc.

Natural Language Processing

Natural language processing (NLP) has revolutionized ML-based big data analysis of voluminous textual data and comprehension. Transformers and attention mechanisms are popular deep learning models which enable natural language processing when it comes to big data analysis. Both models enhance the NLP task with respect to its performance, scalability and language comprehension capabilities.

- **Transformers:** Transformers are powerful deep learning models capable of handling sequential text data in natural languages. While the recurrent neural networks (RNNs) process the text information in a sequential manner, the transformers allow parallel processing of the complete text input sequence. They employ a self-attention

mechanism for processing the textual information and very innovatively are also able to establish contextual relationships between words or phrases. The interdependencies and interactions of each word with the rest of the text data in sequence are analyzed to find the contextual link. One of the finest transformer-based models is 'BERT' (bidirectional encoder representations from transformers) which excels in sentiment analysis, item identification, question-answering and machine translation [44].

- **Attention Mechanisms:** Attention mechanisms facilitate extraction of textual data which serves the purpose based on relevance in the context to produce accurate predictions. The NLP task performance is tremendously improved with the attention mechanism and has also proved to be effective in the implementation of projects such as machine translation, which harnesses the capability of the attention mechanism in dealing with particular words in the source language to yield accurate translations [45].

Image and Video Analysis

Due to a number of tech-enabled factors such as the widespread availability of cameras, portable imaging devices, surveillance devices, autonomous smart devices, etc, the volume of images and videos has seen an exponential rise. If the past and present trends are any indication, this volume and velocity of image and video accumulation will further multiply manifolds. These images and videos contain very important information calling for their analysis for application in several fields be it e-commerce, healthcare, the tourism industry or law enforcement. Traditional computing systems encountered challenges while dealing with such image and video datasets as they relied on manually created features and rule-based algorithms [45]. The deep learning models are best suited for big data analysis concerning image and video as data. Deep learning models are extremely powerful tools with an inbuilt capacity to learn automatically the hierarchical representations right from the raw pixel data, which constitutes the visual data. Deep learning models are able to automatically extract complex and intangible information from the input visual data and thus the analysis is more trustworthy and reliable.

Deep learning models such as convolutional neural networks (CNNs) have demonstrated much better performance in image analysis such as image categorisation, picture segmentation, image synthesis and object identification. CNNs work on local spatial features to learn the distinctive characteristics of types, margins and textures. CNNs with stacking layers can learn

easily even the most complex and intricate visual data resulting in higher accuracy in identification.

Videos with a longer duration of time component are more difficult to analyze. However, deep learning models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are capable of identifying temporal relationships between the video data and analyzing it [45]. These models are particularly useful and suitable for analysis of data or clips of CCTV surveillance and in video summarization owing to their capability of accurate object identification, identifying movements and forecasting subsequent frames.

The advancement in ML has made it possible to create fully realistic images and synthesize video through generative modelling of deep learning models. Generative adversarial networks (GANs) and variational autoencoders (VAEs) are known to create real-looking pictorial and video clips closely resembling as genuine visual data. These models find wide applications in the field of visual or graphical content creation, preparing creative art forms, etc.

Anomaly Detection

Big data analytics heavily relies on anomaly detection since it identifies trends or instances that dramatically vary from expected behaviour. Anomaly detection becomes much more crucial in the context of big data, as enormous amounts of different data are produced, for a variety of reasons. First of all, big data analytics works with a variety of data sources and types, such as structured, unstructured and semi-structured data. Whether they are numerical values, textual data, network traffic, sensor readings or multimedia content, anomaly detection algorithms assist in finding anomalies across diverse data kinds [46].

Anomaly detection can also automatically point out systemic failures, cybersecurity infringement or breach, different types of fraud, equipment failure or any other critical malfunctioning. Big data analysis can mine out data of interest, critical to decision-making, from large chunks of noise, resulting in enhanced operational efficiency, reduced risk and detection of abnormalities and outliers which is normally not possible through routine statistical techniques [46].

Incremental Learning for Non-Stationary Data

Non-stationary data changing data or dynamic data presents a different set of challenges in big data analytics. In order to handle such dynamic data,

adaptive procedures, known as idea drift, such as online learning, concept drift detection and adaptive learning rates, should be used [47].

Online learning offers real-time learning opportunities for training the model. It is also referred to as incremental learning or streaming learning. This learning mode continuously updates the model due to a constant inflow of new. The online learning method trains the model's parameters as per new input data on an incremental basis to accordingly improve the output. This incremental learning allows the model to stay updated as it keeps on training itself on the most recent patterns and progressively most recent data sets.

The phenomenon of change in the data distribution owing to changes in statistical features of the data over time is known as 'concept drift'. Changes in user preferences, seasonal variations or environmental changes, etc., can lead to concept drift in big data analytics. Taking this into account, it is imperative that concept drift detection is essentially applied to ensure that the model remains accurate and the output trustworthy. The concept drift detection methods use a wide range of techniques. Some of the popular ones include: sliding window methods, ensemble methods, online change detection algorithms and statistical tests such as the Kolmogorov–Smirnov test and the CUSUM test.

The data stream may have different properties at different times and thus the learning parameter needs to be adjusted accordingly. Thus, adaptive learning rates are very important from a model training perspective. It shall ensure that the learning rate is modified dynamically as a function of data streaming and the model is updated appropriately. The AdaGrad algorithm is one such algorithm which allows for adjustment of the learning rate for each parameter depending on the previous gradient information.

Multi-Dimensional Data

Multi-dimensional data exhibiting a lot of variables and features can be challenging to analyze. The challenges can be in the form of increased computational complexity, overfitting and poor interpretability. Any of these can substantially impact the efficiency or outcome of the ML model. However, these can be tackled by certain methods such as feature selection, reducing dimensionality and regularisation.[47].

- **Feature Selection:** Through the feature selection method, features of relevance or interest are only selected and the rest are eliminated out of a multi-dimensional dataset. This is achieved through three approaches: filter methods, wrapper methods and embedded methods. In the filter method, features are chosen based on their

relevance and predictive ability which can be derived from statistical outputs of correlation or other statistical coefficients which measure its applicability in a particular context. The wrapper method selects the relevant features by approaching the process as a search problem and employing the algorithm as a 'black box'. Embedded methods make use of regularization methods such as L1 regularisation (Lasso), which results in automatic feature selection.

- **Dimensionality Reduction:** This is another method to control or reduce certain unwanted dimensions of any data for targeted processing and outcome. There can be two kinds of approaches to dimensionality reduction—linear methods and non-linear methods. The linear method performs dimensionality reduction by determining the variance and principal component analysis (PCA). The non-linear relationships in the data can be retrieved and the unwanted dimensionality reduced through non-linear techniques such as t-SNE (t-distributed stochastic neighbour embedding) and UMAP (uniform manifold approximation and projection).
- **Regularization:** The regularization technique helps to reduce the complexity and overcrowding of data by training the model to apply penalty terms during training when the model becomes too complex or fits the data too closely. These are achieved by L1 regularization (Lasso), and L2 regularization (Ridge).

Large-Scale Models

Implementation of large-scale deep learning models in big data analytics requires addressing several important issues and challenges as discussed below:

- **Computational Complexity:** The computational complexity involved in training massively scalable deep learning models is one of the main obstacles. Deep neural networks have a huge number of layers and millions or even billions of parameters, making training them extremely computationally intensive. Big data processing and analysis can be very resource-intensive, necessitating high-performance computing infrastructure, such as GPUs or distributed computing systems, to meet the computational demands [48].
- **Scalability:** When working with big data analytics and extensive deep learning models, scalability is a major difficulty. It becomes increasingly important to build systems that can manage the

scalability requirements as the volume, velocity and variety of data increase. Effective data storage and retrieval systems coupled with distributed processing frameworks and parallel computing methods can help in training deep learning models on large data sets thereby enabling the model to handle large amounts of complex data. [48].

- **Data Availability and Preprocessing:** Large volumes of good training data are essential for the success of deep learning models. But gathering, prepping and preprocessing huge amounts of data for deep learning may be a difficult proposition and may include data gathering, data cleaning, noise reduction and control, missing values, stitching data sources variability, data quality and consistency, etc. Storage of huge data and their accessibility may also pose difficulties [48].
- **Model Interpretability:** To overcome the difficulty of model interpretability in the context of big data, it is necessary to build methods and tools for deriving insightful justifications and knowledge from sizable deep-learning models [48].
- **Deployment and Real-Time Processing:** It can be difficult to deploy a deep learning model for real-time processing and inference in big data analytics after it has been trained. Large amounts of data must be processed in real-time, including streaming data and applications that must be responsive. To do this, effective deployment mechanisms, low-latency systems, and optimized inference pipelines are needed. Large-scale deep learning model deployment on edge devices or in contexts with limited resources also increases deployment difficulties [48].
- **Model Updates and Adaptability:** Big data analytics frequently work with dynamic and changing data, necessitating the ability to adapt and update models over time. Model retraining, incremental learning and sustaining performance and accuracy over time are problems posed by incorporating new data, managing idea drift and adjusting the taught models to emerging patterns.

The above challenges require a multipronged approach integrating deep learning knowledge, management of voluminous data, distributed computing and system optimization. The advancements happening across hardware infrastructure, data pre-treatment methods and distributed computer networks, along with others offer potential solutions to these problems.

The training and deployment of large-scale deep learning models is a challenge. However, it can be overcome with distributed computing platforms. A distributed computing platform as the name suggests distributes or spreads the computational workload between several devices leading to

efficient management and processing of big data. There are two very popular distributed computing frameworks widely used for big data analytics: (i) Apache Spark and (ii) TensorFlow's distributed mode.

Apache Spark provides a unified analytics engine along with a free and open-source distributed computing framework, which helps in managing a large amount of data. The TensorFlow is also an open-source deep learning computing framework which provides a distributed mode that enables processing and managing the voluminous data. Both these open-source systems have suitable architecture and provide scalable solutions coupled with distributed computing networks for handling large amounts of data and its computational needs [49].

Conclusion and Future Scope

The application of machine learning (ML) for big data analysis has a promising future and potential for growth and innovation. There is always a scope for further improvement and augmentation in the technology for tackling emerging and complicated issues pertaining to big data analytics. Big data is capable of providing useful insights for decision making and thus ML-based approaches shall hold the key for effective big data analysis. With the passage of time and further constant evolution of technology, the growing need and demands including transparency and trust shall be taken care of by new ML models. Deep learning techniques are capable of creating better models and architectures which can proficiently take up complicated data and its analysis. Transfer learning and pre-trained models can boost the speed of the analysis. Federated learning models can work effectively with distributed data sources. Automated machine learning (AutoML) approaches shall ensure automated learning at several points including model deployment which in turn will enhance the ease of usage for ordinary untrained people and thus expand the user base resorting to using ML-based methods for big data analysis. For the purpose of highly dynamic industries, modified ML algorithms would be handling big data for almost real-time analytics allowing quick decision-making. Integration of ML technologies with other high-end emerging technologies such as edge computing, blockchain, augmented reality (AR) and the Internet of Things (IoT) will further bolster the speed and accuracy of big data analytics.

There will be an emphasis on developing ML models and algorithms which are more user-friendly, less technical and easy to understand, learn and interpret. In such a scenario, more users will use it, trust it and can comprehend the decision-making process facilitated by the ML technique. The

near future will witness prominence being attached to developing ML models and frameworks where the biases are lessened and ethical considerations are given due importance.

The ML procedures should adopt and ensure fairness, ethics, accountability, transparency and judiciousness while analysing big and voluminous data. Each domain, such as banking, energy, healthcare, transportation, smart administration as well as forensics, shall benefit from domain-specific ML-based applications to transform the normal operations in these industries. ML can also be extremely useful in carrying out research in the areas of big data and can bring about breakthroughs and newer developments. Thus, ML shall continue to evolve and impact the future of big data analysis through research and innovation for the advantage of all human enterprises and society at large.

References

1. Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, *275*, 314–347.
2. Tarwani, K. M., Saudagar, S. S., & Misalkar, H. D. (2015). Machine learning in big data analytics: An overview. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 270–274.
3. Tarwani, K. M., Saudagar, S. S., & Misalkar, H. D. (2015). Machine learning in big data analytics: An overview. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(4), 270–274.
4. Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013, May). Addressing big data issues in scientific data infrastructure. In *2013 International conference on collaboration technologies and systems (CTS)* (pp. 48–55). IEEE.
5. O'Leary, D. E. (2013). Big data, the 'internet of things' and the 'internet' of signs. *Intelligent Systems in Accounting, Finance and Management*, *20*(1), 53–65.
6. Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, *57*(7), 86–94.
7. Jaswant, U., & Kumar, P. N. (2015). Big data analytics: A supervised approach for sentiment classification using mahout: An illustration. *International Journal of Applied Engineering Research*, *10*(5), 13447–13457.
8. Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2013). Sensing as a service and big data. arXiv preprint arXiv:1301.0159.
9. Turk, M. (2012). A chart of the big data ecosystem, take 2. https://mattturck.com/a-chart-of-the-big-data-ecosystem-take-2/.
10. Dean, J. (2014). *Big data, data mining, and machine learning: Value creation for business leaders and practitioners*. John Wiley & Sons.

11. Lee, K. M. (2014). Grid-based single pass classification for mixed big data. *International Journal of Applied Engineering Research*, *9*(21), 8737–8746.
12. Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social media mining: An introduction*. Cambridge University Press.
13. Wang, J., Tang, Y., Nguyen, M., & Altintas, I. (2014, December). A scalable data science workflow approach for big data bayesian network learning. In *2014 IEEE/ACM international symposium on big data computing* (pp. 16–25). IEEE.
14. Li, L., Wu, Y., & Ye, M. (2015). Experimental comparisons of multi-class classifiers. *Informatica*, *39*(1), 71–85.
15. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.
16. Karacapilidis, N., Tzagarakis, M., & Christodoulou, S. (2013). On a meaningful exploitation of machine and human reasoning to tackle data-intensive decision making. *Intelligent Decision Technologies*, *7*(3), 225–236.
17. Qian, H. (2014). PivotalR: A package for machine learning on big data. *R Journal*, *6*(1), 57–67.
18. Nasridinov, A., & Park, Y. H. (2014). Combining unsupervised and supervised machine learning to analyze crime data. *International Journal of Applied Engineering Research*, *9*(23), 18663–18669.
19. Wang, L., & Alexander, C. A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(2), 52–61.
20. Tiwari, S. (2022, October). *Approaches involving big data analytics (BDA) using machine learning, described*. 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT). IEEE, pp. 1–7.
21. Sukumar, S. R. (2014, August). *Machine learning in the big data era: Are we there yet*. Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Workshop on Data Science for Social Good (KDD), pp. 1–5.
22. Bu, Y., Borkar, V., Carey, M. J., Rosen, J., Polyzotis, N., Condie, T., ... Ramakrishnan, R. (2012). *Scaling datalog for machine learning on big data*. arXiv preprint arXiv:1203.0160.
23. Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM Sigmetrics Performance Evaluation Review*, *41*(4), 70–73.
24. Hido, S., Tokui, S., & Oda, S. (2013, December). *Jubatus: An open source platform for distributed online machine learning*. NIPS 2013 Workshop on Big Learning, Lake Tahoe.
25. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J. M. (2012). *Distributed graphlab: A framework for machine learning in the cloud*. arXiv preprint arXiv:1204.6078.
26. Baldominos, A., Albacete, E., Saez, Y., & Isasi, P. (2014, December). *A scalable machine learning online service for big data real-time analysis*. 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD). IEEE, pp. 1–8.
27. Vu, A. T., Morales, G. D. F., Gama, J., & Bifet, A. (2014, October). *Distributed adaptive model rules for mining big data streams*. 2014 IEEE International Conference on Big Data (Big Data). IEEE, pp. 345–353.

28. Hoi, S. C., Wang, J., Zhao, P., & Jin, R. (2012, August). *Online feature selection for mining big data*. Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, pp. 93–100.
29. Wang, L., & Alexander, C. A. (2016). Machine learning in big data. *International Journal of Mathematical, Engineering and Management Sciences*, *1*(2), 52–61.
30. Hindman, B., Konwinski, A., Zaharia, M., & Stoica, I. (2009, June). *A common substrate for cluster computing*. HotCloud.
31. Agarwal, A., Chapelle, O., Dudík, M., & Langford, J. (2014). A reliable effective terascale linear learning system. *The Journal of Machine Learning Research*, *15*(1), 1111–1133.
32. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... Ng, A. (2012). Large scale distributed deep networks. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, *1*, 1223–1231.
33. Kearns, M. (1998). Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, *45*(6), 983–1006.
34. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, *51*(1), 107–113.
35. Chu, C. T., Kim, S., Lin, Y. A., Yu, Y., Bradski, G., Olukotun, K., & Ng, A. (2006). Map-reduce for machine learning on multicore. *Advances in Neural Information Processing Systems*, *19*, 281–288.
36. Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010, June). *Pregel: A system for large-scale graph processing*. Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp. 135–146.
37. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., & Hellerstein, J. M. (2012). *Distributed graphlab: A framework for machine learning in the cloud*. arXiv preprint arXiv:1204.6078.
38. Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434.
39. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.
40. Dror, G., Koenigstein, N., Koren, Y., & Weimer, M. (2012, June). *The yahoo! music dataset and kdd-cup'11*. Proceedings of KDD Cup 2011. PMLR, pp. 3–18.
41. Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for big data: A review. *Big Data Research*, *2*(3), 87–93.
42. Condie, T., Mineiro, P., Polyzotis, N., & Weimer, M. (2013, June). *Machine learning for big data*. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 939–942.
43. Nti, I. K., Quarcoo, J. A., Aning, J., & Fosu, G. K. (2022). A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Mining and Analytics*, *5*(2), 81–97.
44. Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: A big data-ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347.

45. Ngiam, K. Y., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, *20*(5), e262–e273.
46. Hassan, M. M., Gumaei, A., Alsanad, A., Alrubaian, M., & Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. *Information Sciences*, *513*, 386–396.
47. Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., ... Ramkumar, P. N. (2020). Machine learning and artificial intelligence: Definitions, applications, and future directions. *Current Reviews in Musculoskeletal Medicine*, *13*, 69–76.
48. Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Viti, P., Mancini, A., & Frontoni, E. (2023). From knowledge-based to big data analytic model: A novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0. *Journal of Intelligent Manufacturing*, *34*(1), 107–121.
49. Nguyen, D. K., Sermpinis, G., & Stasinakis, C. (2023). Big data, artificial intelligence and machine learning: A transformative symbiosis in favour of financial technology. *European Financial Management*, *29*(2), 517–548.

Index