3-1-2024

Machine learning applications in forensic DNA profiling: A critical review

Mark Barash
San Jose State University, mark.barash@sjsu.edu

Dennis McNevin
University of Technology Sydney Centre for Forensic Science

Vladimir Fedorenko
Saratov State University

Pavel Giverts
Israel Police

Recommended Citation

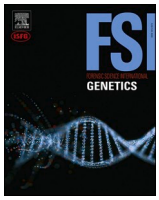
Mark Barash, Dennis McNevin, Vladimir Fedorenko, and Pavel Giverts. "Machine learning applications in forensic DNA profiling: A critical review" *Forensic Science International: Genetics* (2024). https://doi.org/10.1016/j.fsigen.2023.102994

Forensic Science International: Genetics 69 (2024) 102994

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



Machine learning applications in forensic DNA profiling: A critical review

Mark Barash [a,b,*,1], Dennis McNevin [b,2], Vladimir Fedorenko [c,3], Pavel Giverts [d,4]

[a] Department of Justice Studies, San José State University, San Jose, CA, United States
[b] Centre for Forensic Science, School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney, Broadway, Ultimo, NSW 2007, Australia
[c] The Educational and Scientific Laboratory of Forensic Materials Engineering of the Saratov State University, Russia
[d] Division of Identification and Forensic Science, Israel Police HQ, Haim Bar-Lev Road, Jerusalem, Israel

ARTICLE INFO

Keywords:
Machine learning
Forensic DNA profiling
Human identification
AI
STRs

ABSTRACT

Machine learning (ML) is a range of powerful computational algorithms capable of generating predictive models via intelligent autonomous analysis of relatively large and often unstructured data. ML has become an integral part of our daily lives with a plethora of applications, including web, business, automotive industry, clinical diagnostics, scientific research, and more recently, forensic science. In the field of forensic DNA, the manual analysis of complex data can be challenging, time-consuming, and error-prone. The integration of novel ML-based methods may aid in streamlining this process while maintaining the high accuracy and reproducibility required for forensic tools. Due to the relative novelty of such applications, the forensic community is largely unaware of ML capabilities and limitations. Furthermore, computer science and ML professionals are often unfamiliar with the forensic science field and its specific requirements. This manuscript offers a brief introduction to the capabilities of machine learning methods and their applications in the context of forensic DNA analysis and offers a critical review of the current literature in this rapidly developing field.

1. Introduction

Forensic DNA profiling is the backbone of forensic science and one of the most rigorous forensic disciplines. In the last four decades, this field has made significant progress and cemented its 'gold standard' reputation by incorporating extremely sensitive, accurate, robust and extensively validated methods of human identification [1,2]. In a similar fashion, machine learning (ML) is the fastest-growing field of computer science that offers extremely powerful capabilities for intelligent data analysis [3,4]. These learning algorithms dominate the field of artificial intelligence (a field which focuses on mimicking human abilities) and are considered the benchmark of data processing methods.

ML has penetrated almost every aspect of our lives. Many of us are used to the fact (although not always aware) that ML algorithms are used for producing fruitful web search results, targeting online advertisements, effective spam filters in our email boxes, photo tagging in social networks, predicting stock market trends, constructing self-driving cars, better understanding of the human genome structure and

Abbreviations: ANN, artificial neural networks; AT, analytical threshold; BN, Bayesian networks; CART, classification and regression trees; CE, capillary electrophoresis; CNN, convolutional neural network; DBSCAN, density-based spatial clustering of applications with noise; DeT, Decision Tree; DL, deep learning; DT, dynamic threshold; EPG, electropherogram; GAN, generative adversarial networks; GDA, generalised discriminant analysis; HID, human identification; k-NN, k-nearest neighbours; LDA, linear discriminant analysis; LR, likelihood ratio; MCMC, Markov Chain Monte Carlo; MAC, maximum allele count; MCA, multiple correspondence analysis; ML, machine learning; MLE, maximum likelihood estimation; MLP, multilayer perception; MLR, multinomial logistic regression; MPS, massively parallel sequencing; NB, Naive Bayes; NGS, Next Generation Sequencing; NoC, number of contributors; NT, no threshold; PCA, principal component analysis; PCoA, principal coordinates analysis; PG, probabilistic genotyping; PGS, probabilistic genotyping software; RF, random forest; SNP, single nucleotide polymorphism; ST, stochastic threshold; STR, short tandem repeat; SVM, support vector machine; TAC, total allele count; t-SNE, t-distributed stochastic neighbour embedding.

* Corresponding author at: Department of Justice Studies, San José State University, San Jose, CA, United States.
E-mail address: mark.barash@sjsu.edu (M. Barash).
[1] 0000-0002-5445-6316
[2] 0000-0003-1665-3367
[3] 0000-0002-3979-2602
[4] 0000-0002-4019-5055

https://doi.org/10.1016/j.fsigen.2023.102994
Received 29 July 2023; Received in revised form 6 November 2023; Accepted 26 November 2023
Available online 1 December 2023

many additional applications. In forensic science however, the application of ML is still in its infancy. This is primarily due to the fact that most forensic scientists are not aware of the possibilities of ML, while ML and data mining specialists are not familiar with specific tasks that arise when conducting forensic examinations.

This manuscript aims to briefly describe the key capabilities of ML and provide a critical review on its potential applications in forensic DNA profiling. Due to a relatively wide variety of such applications, this paper mainly focuses on the area of Human Identification (HID), briefly delving into additional forensic applications. Section 1 (subsection 1.1) introduces the topic of machine learning and its mathematical and computational foundations; subsection 1.2 presents a brief overview of the main machine learning methods; subsection 1.3 gives examples of problems that can be solved with ML approach, such as classification, clustering, regression analysis, dimensionality reduction and generative models, while in subsections 1.3.4 and 1.3.5, we discuss the topics of dimensionality reduction and generative models by using examples from forensic intelligence applications, such as prediction of externally visible characteristics (EVCs) and biogeographical ancestry (BGA) from DNA.

Section 2 focuses on ML applications for forensic STR analysis – the main focus of this article, and provides detailed insights into the challenges related to HID applications, emphasizing how the power of machine learning is harnessed to address these challenges. We explore the application of ML in STR genotyping using CE and MPS data, and its potential to enhance accuracy and efficiency. We begin by delving into the complexities arising from the use of MPS technology for STR genotyping, which generates vast genomic data and presents the need for more sophisticated bioinformatic tools. Furthermore, we explore the role of ML in DNA mixture interpretation, a particularly challenging aspect of forensic DNA analysis. As we delve into the applications of ML in STR genotyping, we see the logical progression of leveraging ML for improved forensic DNA analysis.

1.1. Machine learning approach

In the classical scientific approach to the study of a phenomenon or effect, scientists usually explore all the relationships between the elements of the system under study. Once identified, these relationships are analysed and used to create a comprehensive mechanistic model that describes the system. This is how most of the basic formulas used in physics, mechanics, thermodynamics, chemistry and other sciences were derived. In the engineering sciences, however, the analytical formulas developed for ideal models do not always accurately reflect the real world. In addition to the main parameters that determine the behaviour of the model, various random factors play their roles. In such cases, scientists would usually develop empirical formulas that include one or more numerical variables (coefficients) designed to factor in the influence of unknown environmental factors.

In forensic science, probabilistic genotyping algorithms use empirical formulas to provide the probability of obtaining the profile if a nominated individual is a DNA donor, compared to if they are not a DNA donor. For example, STRmix™ (one of the main probabilistic genotyping software) uses an algorithm that incorporates variables like observed and expected allele and stutter peak heights at specific DNA loci. To estimate the likelihood of the observed peak heights, the algorithm uses a log normal distribution in order to assign the probability of observed peak heights when the factors relating to various potential contributors are varied [5–8]. This empirical formula allows the software to predict peak heights in a probabilistic manner, taking into consideration the inherent variability and unknown influences in the real-world experimental data which are not able to be fully described by mechanistic models using analytical formulae. Additionally, the computational cost of using the empirical formula may be lower compared to more complex analytical formulas, making it more practical for implementation in software. So, the empirical formula provides a way to model the relationship between peak height and various unknown factors in a more comprehensive and flexible manner, taking into account the complexities of the real-world system being studied. This is especially useful when the relationships between the data are not fully known or when the amount of data or potential number of variables is large and requires statistical analysis techniques to derive meaningful insights.

The essential dependence relationship between the variables in the functions (for empirical formulas) can be obtained by correlating the results of experiments and observations. Based on the data obtained, it is possible to construct a function that will reflect the relationship between these data in a convenient way. The form of the obtained function is usually set by the researcher, the amount of data analysed is finite, and the relationship between variables is modelled. However, what if the relationships between variables are not known, and the amount of data as well as potential variables are exponentially large? In this case, statistical analysis such as ML, can help.

ML represents systems of statistical analysis that enables identification of dependencies in large volumes of data. In other words, ML algorithms can be described as a transformation (T) that predicts a vector (i.e. a set of values) of output variables ($\mathbf{Y}$) by learning from many examples of the input variables ($\mathbf{X}$). The function however, is usually "invisible" or cannot be easily articulated. In the most "non-forensic" applications this is not necessarily a problem, as the primary aim is to make accurate predictions. As such, the predictive model should be able to envisage the $\mathbf{Y}$ for every new $\mathbf{X}$ by mapping $\mathbf{Y}$ as a function of $\mathbf{X}$ for reference (training) data: $\mathbf{Y} = T(\mathbf{X})$. In the forensic arena however, there is a general expectation for full transparency and standardisation of the scientific methods used to analyse physical evidence in a criminal case. Given the intrinsic complexity of the ML algorithms, coupled with a relative scarcity of experts who can scrutinise these methods, forensic implementation of ML methods should follow relevant developmental and internal validation procedures, such as the SWGDAM Validation Guidelines for DNA Analysis Methods [9].

The first methods for ML began to be developed in the 1950 s [10, 11], although the first fundamental publication proposing the idea of neural networks (one of the dominant ML methods, discussed below) was published earlier in 1943 [12]. Since then, a large number of different ML methods have been developed, suited for a variety of data and types of tasks to be solved. Such methods include linear regression, linear discriminant analysis, k-nearest neighbours (k-NN) algorithms, naive Bayes algorithms, decision trees, random forest algorithms, various types of neural networks, and other methods.

When processing the initial data using ML methods, several approaches can be used. In general, the particular strategy will depend on the problem being solved and the form of data presentation. In the next paragraphs we will describe the main ML approaches and their potential applications.

1.2. Types of machine learning

Machine learning can be arbitrary categorized into four main types: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning [13]. Each of these categories is briefly discussed in the below paragraphs and summarized in Fig. 1.

1.2.1. Supervised learning

Supervised learning is the most common type of ML in general and in forensic DNA analysis in particular. In this type of learning, the predictive model is first trained with a large structured dataset - input variables and corresponding output variables [14]. The training data in this approach is given in the form of samples with labels. For example: sequences of DNA fragments annotated with labels corresponding to STR loci and flanking regions. The learning algorithm essentially learns the mapping function between the values based on a training dataset with labelled examples and subsequently assigns the corresponding label for each new example based on the established rules. Supervised learning is generally a very effective ML method, but requires

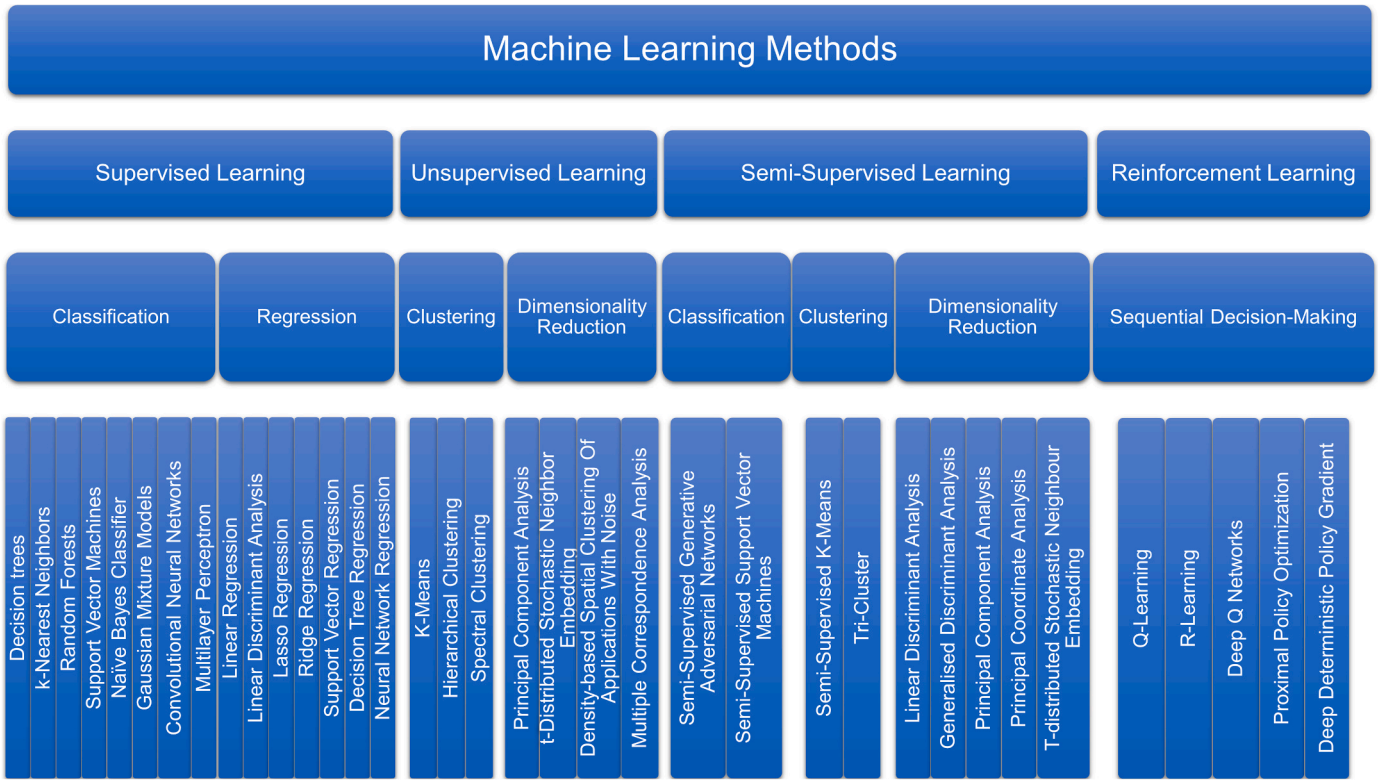


Fig. 1. A simplified diagram describing various ML algorithms. Please note that this diagram is not exhaustive and includes only some commonly used ML methods under the specified categories. The tasks such as classification, clustering, regression and dimensionality reduction can be solved with different ML methods, hence the overlap in the diagram.

high-quality, normalised, and thoroughly cleaned training data to reduce potential bias (e.g. overfitting) in the output. Supervised learning can be executed by two main approaches: classification and regression analyses (described below). However, it must be noted that each of these problems (i.e. classification, regression) can also be approached with other ML methods, categorized under unsupervised or semi-supervised learning categories (as outlined in Fig. 1).

1.2.2. Unsupervised learning

Unsupervised learning does not have an a priori-structured data-target format. In other words, it can also develop a function based on the input data (**X**), but does not require the corresponding output labels (**Y**) for the training data as in a supervised learning approach [15]. Given the lack of the predefined labels, the algorithms need to be fed with a very large comprehensive dataset in order to accommodate most of the different scenarios of the **X**-**Y** connections. This is essential, so the model can be efficiently trained to understand the properties of the data. In this approach, the task is to organise the data in a similar way as humans do but, as for humans, this depends on which features are presented to and extracted by the algorithm. For example, when presented with images of cats, dogs, tigers and wolves, a child might recognise cats and dogs as "pets" and recognise tigers and wolves as "wild animals". Alternatively, if cat-like and dog-like features were emphasised (extracted), the child might recognise tigers as a form of "cat" and wolves as a form of "dog". Unsupervised learning is particularly beneficial when there is a need to automatically organise terabytes of unlabelled data into similar clusters with minimal manual intervention, but classification depends on features of the data which may or may not be obvious.

1.2.3. Semi-supervised learning

Semi-supervised learning is essentially a blend between supervised and unsupervised learning approaches. This type of ML is usually used when there is a large amount of input data, while only a small portion of the data is labelled. The goal of semi-supervised learning is to leverage the additional information contained in the unlabelled data to improve the model's performance, especially when obtaining labelled data is costly, time-consuming, or limited. For example, a model might be supplied with a very large number of raw electropherograms to learn to distinguish an allele from background noise, while only a subset of electropherograms and/or alleles have been pre-labelled [16].

1.2.4. Reinforcement Learning

Reinforcement learning is a type of ML where an agent (e.g. computer program or an autonomous robot) learns to make decisions by interacting with an environment. The agent takes actions in the environment, receives feedback in the form of rewards or penalties, gathering experience, and uses this experience to improve its policy through trial and error. This type of ML is commonly used in tasks where an agent needs to make informed decisions by learning through trial and error, such as game playing, robotics, and autonomous systems.

1.3. Types of problems that can be solved with ML approach

There is a large variety of data processing tasks that are well suited for ML methods. Some of the problems that can be solved with ML include:

- Classification: assigning input data to predefined categories or classes;
- Regression: predicting continuous numerical values based on input data;
- Clustering: grouping similar data points together based on their characteristics;
- Dimensionality reduction: retaining the most meaningful features of the data while eliminating redundant or less informative ones;

- Image and Video Recognition: analysing and understanding visual data, such as object detection, facial recognition, identification and classification of bloodstains, sperm cells, and image captioning;
- Anomaly Detection: identifying unusual patterns or outliers in data, which can be valuable for fraud detection, fault diagnosis, or detecting anomalies in sensor data;
- Natural Language Processing: understanding and processing human language;
- Generative Models: creating new data that resembles the training data distribution, such as generating realistic images, music, or text;

The main types of these tasks are briefly discussed below.

1.3.1. Classification

Data classification is usually performed by a supervised learning method and requires an annotated dataset for efficient training (Fig. 2). The resulting model would be a function that determines which of the predefined groups the new data will be assigned to. In addition, the likelihood of belonging to each group can also be calculated. The classification models can be binary, where the data is divided into two groups, such as "true signal" vs "noise", "allele" vs "stutter"; "STR locus" vs "flanking sequence" and so on. In other cases, non-binary models are able to classify into multiple categories by assessing their best fit to one of the several groups. An example of such a classification approach is handwriting text recognition and interpretation (Fig. 3).

Classification tasks can be approached with many different ML methods, including linear discriminant analysis (LDA), logistic regression, naive Bayes, decision trees and random forest, k-nearest neighbours (k-NN), support vector machine (SVM) and neural networks (discussed below). A more detailed review of these classification techniques can be found in [14].

1.3.2. Clustering

The clustering problem is very similar to the classification problem with the main difference being the training approach, which uses unlabelled data (unsupervised learning). Clustering is used to find common aspects (patterns) within a dataset and distinguish between groups of data based on the features of the data. If the study can assume

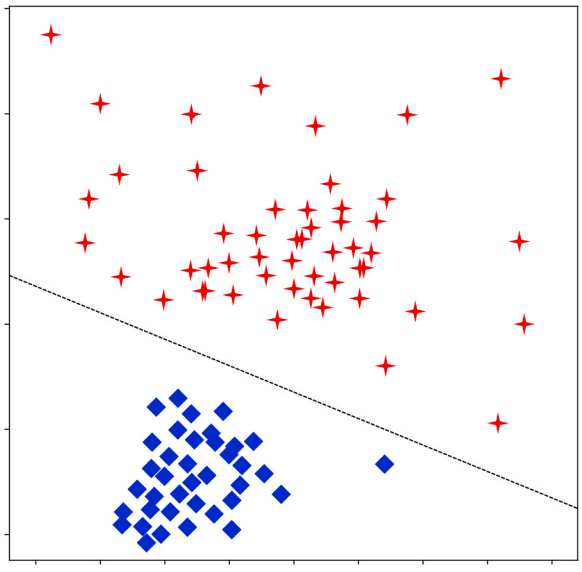


Fig. 2. An illustration of linear classification into two groups where predetermined annotation is represented by red star and blue diamond labelled data points which are in turn functions of two variables (one quantified on the horizontal axis and the other quantified on the vertical axis). The line represents a boundary that best separates the two groups based on the features (variables) presented to the model.

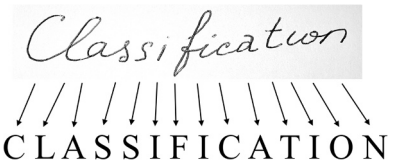


Fig. 3. An illustration of multi-label classification for text recognition where the dataset for classification includes examples of hand writing and the prediction categories are individual letters of the alphabet and/or words.

that the objects described by the data can be divided into a certain number of groups, then application of the clustering approach would result in segregating all objects into that number of groups. The clustering of the objects is performed according to the presence of the most similar characteristics (features) within each group, while being mostly different from objects in other groups (Fig. 4). The resulting model can be subsequently used to classify new data. The qualities of the objects that determine which clusters they belong to are not necessarily known but may correspond with other meta data associated with the objects.

Clustering can be solved with many different ML methods, such as: k-Means, density-based spatial clustering of applications with noise (DBSCAN), hierarchical clustering, fuzzy clustering and others [17,18].

Model-based likelihood estimation is a particular type of clustering algorithm that is used extensively in forensic DNA analysis for population assignment. The most popular form of this algorithm is represented by Structure [19] which is a Bayesian algorithm that uses a matrix (i.e. a rectangular arrangement of values or data representing a mathematical object or its particular property) of known genotypes (**G**) to estimate a matrix of (unknown) genotype frequencies in K ancestral populations (**P**) and a matrix of (unknown) genetic contributions (**Q**) to each ancestral population, according to:

$$\mathbf{G} = \mathbf{QP}$$

For a given value of K, Structure updates prior estimates of **P** and **Q** according to the posterior probability distribution:

$$P(\mathbf{Q},\mathbf{P}|\mathbf{G}) \propto P(\mathbf{Q})P(\mathbf{P})P(\mathbf{G}|\mathbf{Q},\mathbf{P})$$

Markov chain Monte Carlo (MCMC) simulations of P(**P**), P(**Q**) and P(**G**|**Q**,**P**) enable sampling from the posterior probability distribution. A log likelihood estimation is maximised until convergence. **P** and **Q** are initially modelled by Dirichlet distributions, which define variance in matrix elements but these are allowed to drift until they converge to stable values [20,21].

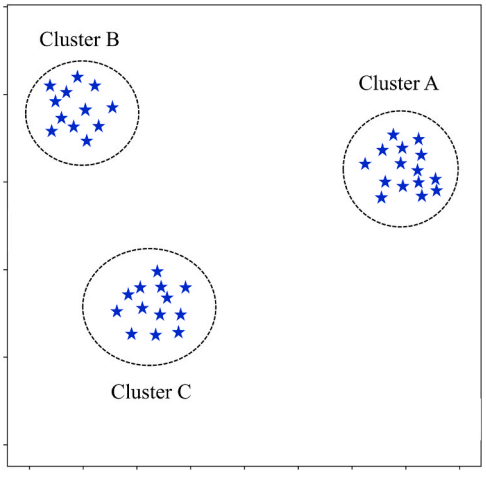


Fig. 4. An illustration of the data clustering approach where three clusters (A, B and C) are defined by their co-location in a co-ordinate system defined by two variables (one quantified on the horizontal axis and the other quantified on the vertical axis).

1.3.3. Regression analysis

When trying to solve a complex scientific problem, researchers perform a series of experiments or observations in which the behaviour of an object or a process itself is studied under different conditions. With this standard scientific method, the process or object under study is described only at points with given initial parameters, rather than the whole process. Nevertheless, with the collected data, one can produce a mathematical function that would describe the process with a given accuracy at other points for which experimental data was not collected. In cases where the number of variable parameters is relatively small, and the type of relationship between the parameters is known (e.g. various dependences: linear, quadratic, trigonometric, power, etc.), the construction of such a function is usually not a particularly difficult task.

However, if the studied object or phenomenon is described by a large number of parameters, and the relationship between these parameters does not lend itself to analytical description, then the methods of mathematical regression might better suit this task. As for the classification problem, constructing a regression function is an example of a supervised learning method. The only difference between the two is that when classifying, the studied data is predicted to belong to a particular category, while the result of regression is a number or a vector.

One of the most known (and arguably useful) applications of ML for constructing regressions is in predicting the results of stock trading. This method however, can be successfully applied in solving many other problems. In forensic DNA phenotyping for example, this method can be used for predicting externally visible characteristics, such as iris, skin and hair pigmentation from a DNA sample based on a large dataset of respective phenotypes with corresponding genotypes [22–24].

The problems requiring regression analysis can be approached with several ML methods including linear and polynomial regressions, logistic regression, decision trees and random forest, neural networks and others [3,25].

1.3.4. Dimensionality reduction

Quite often, a study involves collection of a large amount of different types of data related to an object or phenomenon under study. In such a case, it may turn out that some of the parameters that were measured are only generally related to the object of study and their change does not affect the change in the resulting values. In other cases, such a relationship may be present, but its effect is minimal compared to changes in other parameters and can therefore be disregarded. Consequently, it would be more productive (efficient) to discard such irrelevant parameters in order to facilitate the construction of a model describing the object/phenomenon under study. Another variation of this scenario is when there is a functional or statistical relationship between several parameters, which essentially means that a group of parameters can be expressed by fewer parameters. In such cases, it may be appropriate to reduce the number of represented variables through dimensionality reduction techniques. This approach transforms the data from a high-dimension space into a low-dimension space, while retaining the principal properties of the data. Such tasks can be approached via linear discriminant analysis (LDA), multiple correspondence analysis (MCA), principal component analysis (PCA), principal coordinate analysis (PCoA), generalised discriminant analysis (GDA), t-distributed stochastic neighbour embedding (t-SNE) and other methods [26,27].

In addition to simplifying the model and making it easier to analyse, dimension reduction is often used in the visualisation of results, when a function of n-variables is reduced to 2 or 3 variables, which makes it easier to display it graphically. In the area of forensic DNA analysis, dimensionality reduction is often applied in order to predict the BGA or EVCs of a person (such as facial appearance and other complex traits) from a DNA sample (also discussed below in subsection 1.3.5). Here, highly dimensional data exist in the form of multi-locus single nucleotide polymorphism (SNP) genotypes, where the dimension can vary from a few dozen SNPs (e.g the 34-plex autosomal SNP single base extension assay developed by Phillips et al. [28] to over a million SNPs (e.g. microarray genotypes). These data are reduced to the two or three dimensions (as principle components or principle coordinates) which explain most of the variance in the data and which can be easily visualised in two- or three-dimensional plots.

BGA inference by dimensionality reduction illustrates the issue of data features mentioned earlier. The 34-plex assay described above is comprised of SNPs that have been specifically chosen to differentiate between BGAs, that is, they have features that describe BGA. The 52-plex SNP assay developed by Phillips et al. (2006) [29] is comprised of SNPs that have been specifically chosen to differentiate between individuals, that is, they have features that describe individuals. When PCoA is used to, firstly, reduce dimensionality and, secondly, to cluster genotypes derived from three different populations, the 34-plex is much more effective than the 52-plex, even though it has fewer SNPs (Fig. 5). For the 34-plex, the first principle co-ordinate (PC) explains 41% of the variance in the data while the second PC explains 22%, making a total of 63% of the variance explained. Dimensionality reduction has resulted in the loss of just 37% of variance. For the 52-plex, the first PC explains 12% of the variance in the data while the second PC explains 7.5%, making a total of only 20% of the variance explained (80% unexplained). The 52-plex does not perform as well as the 34-plex because it consists of data with less relevant features for BGA (compare with the child who will only be able to classify tigers and wolves as cat-like and dog-like, respectively, if images of tigers and wolves emphasise cat-like and dog-like features). STRs are particularly unuseful for BGA inference because they have fewer BGA-like features (the higher mutation rates of STRs, relative to SNPs, means that allele frequency differences between populations are diminished). It should be noted that BGA estimation represents just an example for the dimensionality reduction approach and highlights the broader potential of genetic data analysis in forensic science, which is beyond the focus of the current review.

1.3.5. Generative models

By combining various methods of ML, it is possible to build more complex models which could be used to analyse a variety of incoming information and, on this basis, predict not only the class of an object or a specific value of a parameter, but create a comprehensive model of a system. This approach is generally known as 'deep learning'. It falls under the umbrella of machine learning and is inspired by the structure and functionality of the human brain with the aim of creating intelligent machines capable of making independent decisions [30]. Deep learning methods utilise variations of a hierarchical (layered) organisation of artificial neurons with connections to other neurons (similar to the brain structure). These neurons pass a signal to other neurons based on the received input – a process that can be repeated multiple times, which ultimately creates a complex network capable of intuitive learning: an artificial neural network – ANN [31,32].

A major public demonstration of deep learning occurred in 2016 when the AlphaGo computer program (developed by DeepMind Technologies and based on a deep neural network) defeated Lee Sedol in four games of Go (Lee won one game) [33]. Until this point, it was thought that AI would not be able to beat Go world champion because the landscape of potential moves was so great and because intuition was considered necessary to win. Previously, in 1997, the Deep Blue supercomputer (IBM) had beaten chess grandmaster Gary Kasparov using "brute force computation", but AlphaGo was an example of true deep machine learning [34].

In an artificial neural network, each layer of neurons performs transformations using weights, biases, and activation functions. This process continues until the data reaches the output layer, where final predictions or results are generated. ANNs are incredibly versatile and can be applied to various machine learning tasks, including classification, regression, image recognition, natural language processing, and more. What sets neural networks apart from other machine learning methods is their ability to identify and emphasize features that distinguish each class during the learning process. In classification tasks,

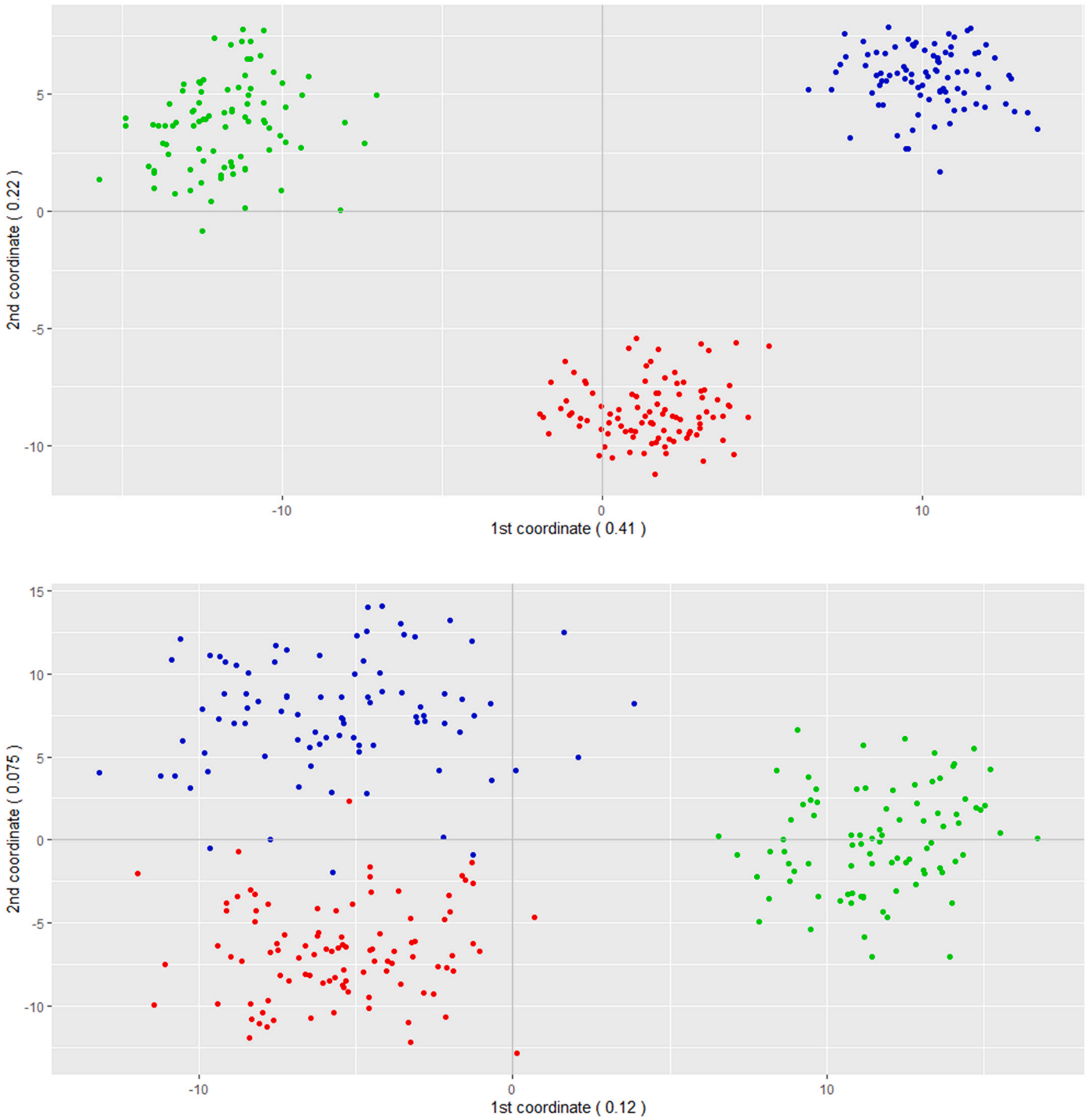


Fig. 5. PCoA applied to individuals genotyped with the 34-plex (above) and the 52-plex (below). The BGAs of the individuals are colour coded red, blue and green. The numbers in parentheses on the axis (co-ordinate) labels indicate the proportion of variance explained by each coordinate.

neural networks compare the features of an object with the features of each class, rather than comparing it with the features of each object in the training dataset [35].

Generative models, a subfield of deep learning, often incorporate layered networks that utilize multiple levels of nonlinear data processing to extract and transform features of interest. One popular example is generative adversarial networks (GANs). [36]. The key innovation behind GANs is their two-network architecture, which consists of a generator network and a discriminator network. The generator creates synthetic data, and the discriminator evaluates whether the data is real or fake, leading to a competitive training process where the generator becomes increasingly skilled at producing highly realistic data (as exemplified below). These models are highly versatile and can combine both supervised and unsupervised learning methods in a single model as reviewed elsewhere [37].

Deep learning and particularly GANs have proven to be a powerful approach producing reliable results for predictive tasks of diverse nature. One such task that is routinely deciphered in many fields including forensic science, is pattern recognition [38]. For instance, this approach can be used for artificial aging of an individual's photograph or biological age prediction from a facial image – another attempt to imitate a 'very ordinary' human capability [39,40]. Forensic molecular

phenotyping in general and the problem of predicting the externally visible appearance from a DNA sample in particular is another forensic application of the generative ML approach [41]. These characteristics can include physical traits such as eye colour, hair colour, skin colour, and even facial morphology. One of the most intriguing applications of generative machine learning in forensic science is the prediction of an individual's facial appearance from a DNA sample. This task is exceptionally challenging due to the complex relationship between genetic and epigenetic factors and facial traits [42–45]. Given the relatively limited knowledge of the underlying genetic architecture of craniofacial morphology, coupled with the complex task of phenotyping facial traits, generative machine learning models, particularly deep neural networks, may offer a promising solution [46]. This approach leads to more accurate predictions of complex variables, such as facial traits, by iteratively refining the generated output (as exemplified in Fig. 6).

Briefly, this complex process can be described as follows:

- Data Collection: To train a deep neural network for facial appearance prediction, a dataset is collected. This dataset includes both genetic data (DNA samples) and corresponding 2D/3D facial images of thousands of individuals. The genetic data provide the input to the model, and the images represent the target output – the individual's facial appearance.
- Feature Extraction: The deep neural network employs multiple layers of artificial neurons to extract relevant features from the genetic data. These features could be related to genetic or epigenetic variations, and other factors that contribute to facial traits.
- Training Process: The network is trained using a large dataset of genetic data and facial images. During training, the network learns to map genetic information to facial features. This process is guided by the images in the training dataset, allowing the network to make predictions about an individual's facial appearance based on their DNA.
- Discriminator Feedback: One of the key elements in generative models is the discriminator, which assesses the quality of the generated output. In this case, the discriminator acts as a classifier that evaluates how well the predicted facial features match the actual images in the dataset. The neural network iteratively refines its predictions by learning from the discriminator's feedback.
- Generating Facial Predictions: Once the network is trained, it can take a new 'questioned' DNA sample as input and generate predictions about the individual's facial appearance. These predictions can include the shape of the face, the size of facial features, and other visible traits.

The potential applications of this technology in forensic science are significant. For example, it could be used to create composite images of suspects based on DNA evidence left at crime scenes or reconstruct the facial appearance of a person based on partial skeletal remains, helping in the identification process [47,48]. While this field is still in its early stages, the power of generative machine learning offers exciting possibilities for forensic science. It highlights the potential of AI and deep learning to provide valuable insights and predictions from genetic data, ultimately aiding investigations and solving complex cases.

Similar to previous sections, this subsection employs examples of machine learning applications in forensic science to illustrate the principles of generative models, even though it is not the primary focus of this review.

1.4. *Benefits of machine learning methods*

One of the main advantages of ML methods is their capacity to

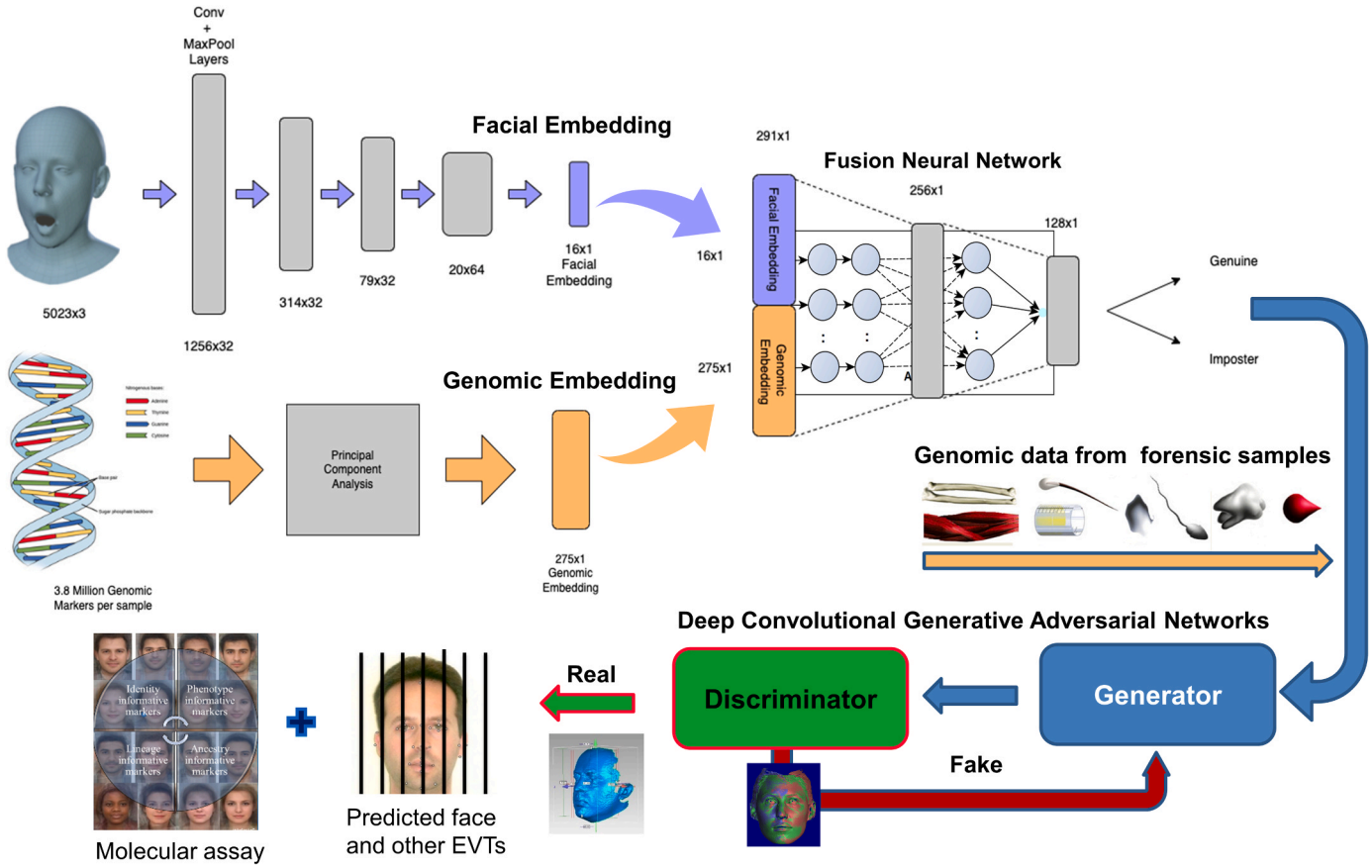


Fig. 6. Illustration of the proposed GANs approach for predicting complex phenotyping traits from genomic data (see [46] for more details).

streamline the processing of a large amount of 'big data' (and in many cases make their analysis possible in the first place), while such processing does not require deep specialised knowledge of statistics from the user [49]. This is possible due to a wide variety of ML algorithms, which have been developed for various applications. These algorithms have been incorporated into a large number of computer programs and libraries for different programming languages (such as Python, R, Matlab, etc.), which allows the use of ML to solve a research problem without extensive knowledge of the data processing algorithms themselves. The researcher only needs to formulate the research questions and the computational task, and then collect and prepare data (e.g. label, verify and clean the training dataset) prior to its input. Consequently, the automation capabilities of ML methods significantly reduce the burden of manual data analysis tasks, freeing up scientists to focus on higher-level problem-solving and creativity. Whether it's data pre-processing, feature extraction, or model optimization, ML algorithms can handle very large datasets through automating various processes, thereby streamlining research workflows and accelerating discoveries. Furthermore, machine learning algorithms can be applied in cases where some data is missing [50]. The algorithms are able to use partial data or will automatically fill the gaps (e.g. via interpolation, imputation, or Dirichlet process priors) in order to create a model and predict the final result [51–53].

The benefits of machine learning methods in data analysis are vast and continue to evolve as researchers explore new applications and develop more advanced algorithms. From aiding decision-making and forecasting to enabling personalized experiences and automation, ML stands as a transformative force in modern research and technological advancement. Embracing and harnessing the potential of machine learning will undoubtedly play a pivotal role in shaping the future of scientific inquiry and problem-solving across diverse domains.

1.5. *Weaknesses and pitfalls of machine learning methods*

The process of preparing and pre-processing data (e.g. cleaning, editing, etc.) is a crucial and time-consuming step in building ML models. This time is of course, spent in addition to the program code writing and algorithm training. Currently, this step requires mostly manual intervention. However, computational techniques can be leveraged to automate and streamline these tasks by automatically detecting, cleaning and removing inconsistencies, errors, outliers, anomalies and unusual patterns from the dataset. For example, techniques like clustering, classification, and regression can be employed to identify and handle missing values, correct data entry mistakes, and impute values based on patterns in the data (as discussed in Sections 1.3.1, 1.3.2 and 1.3.3). Nevertheless, human experts are currently still required to review and validate the results generated by these techniques to ensure the quality and accuracy of the prepared data before proceeding with training the ML algorithm.

The use of separate train/test/validation sets is an important aspect of ML which is also linked to its limitations. These sets are used to evaluate the performance of the ML algorithms and ensure that they are not overfitted to the training data. Overfitting occurs when the ML algorithm becomes too complex with more variables and and/or hidden variables than is justified by the data. In this case, the algorithm performs well on the training data but poorly on new data, reducing the reliability of the algorithm. Proper evaluation using separate and large train/test/validation sets can help mitigate this issue and increase the transparency and reliability of the results. While large datasets can improve the accuracy of the ML algorithms, they can also introduce bias and errors that can undermine the reliability and validity of the results produced by the algorithms. This phenomenon is often called the "curse of dimensionality" and refers to degrading performance of ML algorithms as the dataset's dimensionality (number of features or variables) increases [54,55]. To mitigate the curse of dimensionality, some techniques can be applied, including dimensionality reduction methods (e.g., principal component analysis), as discussed in Section 1.3.4. Either way, a dataset that is not representative of the population(s) being studied may result in inaccurate predictions and conclusions. Such an outcome may have significant implications for the criminal justice system.

Another problem is the over-reliance on ML algorithms as a substitute for human judgment [56–58]. While ML algorithms can provide valuable insights and support decision-making processes, they should not be used as a sole basis for legal decisions. Human experts should always be involved in interpreting and validating the results produced by the ML algorithms.

Moreover, the potential lack of transparency can undermine the validity and acceptability of the results produced by the ML algorithms, particularly in legal contexts, where forensic evidence is often used to support court decisions. This problem is known as the "black box" issue. In this context, explainable AI (XAI) is a critical area of research that aims to develop ML algorithms that are transparent and interpretable [59,60]. XAI refers to the ability of an AI system to provide explanations for its decisions, so that human experts can understand the underlying reasoning and have confidence in the accuracy and reliability of the results. Several techniques have been proposed to improve the transparency of ML algorithms, including decision trees, rule-based models, and Bayesian networks [61–63]. These models are transparent and interpretable, allowing human experts to understand the decision-making process and validate the results produced by the algorithms.

As ML continues to permeate various fields, it is essential to recognize that these methods are not infallible and may produce erroneous results if not appropriately handled. The inherent weaknesses and pitfalls of ML can significantly impact decision-making and the reliability of outcomes, making it imperative for researchers to navigate these limitations cautiously. By addressing these challenges, the scientific community can work towards enhancing the transparency and robustness of ML techniques, ensuring their responsible and effective deployment.

2. Machine learning applications in forensic DNA analysis

Application of ML methods can be especially beneficial for the field of DNA analysis. Continuous advancements in high-throughput genomic technologies such as massively parallel DNA sequencing coupled with improved algorithms for bioinformatic data analysis lead to accumulation of complex genetic data. These data require extensive processing and annotation – a natural application for ML. In fact, ML approaches are used to solve numerous problems in genetics and genomics, such as: generating genomic annotations, predicting functional genomic elements, understanding mechanisms of gene expression and other problems requiring complex analysis of large and often unstructured datasets (for a comprehensive review see [64]).

Forensic science is a multidisciplinary field that analyses and draws conclusions from numerous sources of highly variable multidimensional data. Forensic analysts are expected to base their data interpretation on extensive professional knowledge, experience and rigid standards with zero-bias involved in their decisions. This task however is challenging, as biological data can be extremely complex and diverse. It includes human and non-human DNA markers (e.g. STRs, SNPs and microhaplotypes) used for identity and investigative purposes; RNA markers (e.g. mRNA and microRNA) used for confirmation of the biological tissue of origin; epigenetic markers (e.g. DNA methylation) used for biological age estimation and biological tissue of origin, and even protein markers which can be used as a proxy for genetic identification [65–68]. Each of these markers include numerous inherent (and often "hidden") variables, patterns and artefacts, requiring powerful software and thoroughly trained specialists for data analysis and interpretation. In many cases, the analysis of generated data requires complex mathematics and/or a large number of calculations, which are not feasible or

even possible without computational support. The computational methods allow inclusion of an increased number of both qualitative and quantitative factors, leading to creation of increasingly intricate models that often stray beyond the practical reach of analysts. The data generated in such diverse applications must meet several conditions prior to incorporation into a learning computational pipeline. One of the essential conditions for efficient learning is the size of a training set, which is expected to be significantly larger than the number of specific features (e.g. allele calls vs artefacts) to be extracted. This is necessary to mitigate against the risk of overfitting [69,70]. Sometimes however, it might be necessary to reduce the dimensionality of the data by extracting the most informative features, which allows more efficient use of computer algorithms. In most cases, the data requires extensive processing (e.g. labelling) for a more efficient classification via supervised learning.

Current applications of ML in forensic biology can be arbitrarily divided into:

1) Applications related to human identification, such as efficient designation of STR alleles from either CE- or NGS-generated data, while filtering for potential artefacts and assisting with DNA mixture interpretation, such as estimating the number of contributors (NoC) to a DNA mixture.
2) Applications related to forensic intelligence, such as forensic molecular phenotyping and biogeographic ancestry prediction.
3) Applications related to increasing the evidential value of DNA evidence, such as forensic microbiology (e.g. post-mortem interval estimation), biological tissue source confirmation and activity level reporting.

Considering the relatively large variety of ML applications in forensic DNA analysis, this literature review (and specifically Section 2) focuses on applications for HID only.

2.1. STR allele designation from CE- and MPS- generated data

The process of STR genotyping and data interpretation consists of numerous steps, including separation of signals from different colour channels, identification of all peaks that are present in an electropherogram (EPG), sizing DNA fragments, designating allele calls at each locus, removing artefacts and subsequently designating a DNA profile for further interpretation. The complexity of separating the alleles from background noise and artefacts is the basic problem in forensic DNA analysis. Currently, the process of STR genotyping is carried out in a semi-automatic manner, using dedicated expert software such as Genemapper™ (Thermo Fisher Scientific). This software separates allelic peaks from noise and artefacts with the help of built-in algorithms and utilises a number of validated thresholds. The process subsequently results in a designated DNA profile, although manual confirmation by an analyst is required. The resulting DNA profile can be used for comparison with a reference profile or for further analysis with probabilistic genotyping (PG) software.

The majority of published studies in this niche undertook a similar approach. Briefly, the proposed ML methods make no a priori assumptions and rely almost entirely on the raw EPG data to generate a model, while eliminating the data analysis thresholds which are used in contemporary STR profiling. Instead of using static analytical and stochastic thresholds (ST), they utilise either a dynamic threshold [71] or refrain from applying any thresholds altogether [72–74]. The proponents of this approach presume that application of a static threshold (ST) is too conservative and may lead to significant loss of valuable information (i.e. allelic peaks), which could be reduced if using a dynamic threshold (DT) or no thresholds (NT) at all. In other words, the platform-agnostic algorithms are designed to utilise all the available information (i.e. raw data) to learn and make informed predictions, thus maximising the information obtained from DNA evidence.

Traditionally, the decision on setting interpretation thresholds is based on extensive validation using a variety of samples and a particular piece of analytical equipment (i.e. genetic analyser). In the traditional approach, the signals below such thresholds (e.g. below AT) are considered unreliable (due to the inherent limitations of any analytical system) and discarded. This approach helps to minimise a situation when artefacts are labelled as allele peaks and/or true peaks are discarded. Most forensic DNA laboratories use a series of validated static thresholds, which aim to avoid Type I errors (e.g. designating an artefactual peak as an allele) and minimise Type II errors (e.g. incorrectly classifying true peaks as artefacts), as recommended by the SWGDAM [75]. Traditionally, the possibility of detecting false positives (Type I error) is considered a more consequential error in forensic science, as it may lead to false inclusion and wrongful conviction [76,77]. The application of a more conservative ST and especially AT minimises occurrence of Type I error at the expense of Type II error. On the other hand, the use of a more variation-sensitive DT reduces the possibility of losing essential allelic information. Considering the enhanced sensitivity of modern STR kits and genotyping technologies, it is important to rigorously validate proposed dynamic threshold methods to ensure at least equal performance (including concordance) to existing methods, especially with trace DNA samples.

Consequently, it is important that ML models which do not incorporate any thresholds should be validated to demonstrate concordance with the traditional threshold-based methods. In the following paragraphs we give a more detailed overview of the studies that have utilised either the DT, ST or NT (or their combination) approach to improve contemporary HID methods.

2.2. STR genotyping using CE - generated data

Pull-ups are among the most common technical artefacts of the CE process. These artefactual peaks can be observed when there is a partial overlap between the wavelength windows of two or more fluorescent dyes used to label and detect amplified STR fragments. The majority of these artefacts are zeroed by a spectral calibration mechanism during pre-run calibration of the CE system. Some, however, may remain and challenge DNA profile interpretation, especially mixed samples with minor contributors. Currently, the routine approach to detect and remove pull-ups is by manual inspection of the EPG raw data. This method however is time-consuming and not always accurate.

A possible solution to this problem has been offered in a recent study by Adelman et. al [78]. The authors describe a method for automatic detection and removal of fluorescence pull-ups that does not require prior removal of the technical or biological artefacts. The proposed computational pipeline utilises a symbolic regression model that is applied to the raw EPG data. In order to identify the candidate variables for building a model, the data were first filtered with a static AT of 10 rfu, followed by either a dynamic, locus– or sample-specific threshold. Subsequently, three more quantitative filters were applied, resulting in removal of all expected allelic and stutter peaks, leaving only the artefactual peaks caused by pull-ups.

Quite expectedly, peak height was found to be the most significant variable influencing pull-up frequency, while both the dye and the type of CE instrument were also found to be insignificant features. Accordingly, a direct relationship between the height of a pull-up peak and its removal accuracy was observed: the higher the peak, the more accurately it was trimmed by an algorithm. Notably, the dynamic locus-specific AT together with ML algorithms demonstrated better performance compared to static ATs of 50 rfu and 150 rfu. On the other hand, applying a dynamic AT alone left a larger number of pull-up artefacts than when using STs. The developed symbolic regression models have been optimised for EPG data generated with commonly used 3100 and 3500 genetic analysers and demonstrated 96% predictive accuracy when applied in combination with a DT.

Fluorescence pull-ups however are not the only artefactual signals

commonly observed in EPGs. Other common CE artefacts include electrical spikes, reverse and forward stutters, and background noise. A recent publication by Marciano et.al [71] proposes a sophisticated ML approach for EPG analysis intended to identify and remove such artefacts (except pull-ups), and subsequently yield a DNA profile [71]. The authors of the study developed an "intelligent Locus-Sample-Specific Threshold and Noise Reducer" (iLSST-NR) system, containing four computational modules. The first module utilises a locus- and sample-specific dynamic analytical threshold, which is calculated as four standard deviations above the mean of the background noise in the flanking regions of each locus, while a peak detection algorithm subsequently detects and removes artefacts such as pull-ups and spikes. The second module employs algorithms that detect forward and reverse stutters, based on user-defined thresholds. The third module incorporates trimming algorithms, designed to remove any additional background noise artefacts, not detected by previous modules, while retaining potential low-level alleles. The fourth module incorporates an SVM predictive model and is the actual ML component of the system, designed to classify the data, remove any remaining artefacts and extract the final DNA profile in each locus. The iLSST-NR system has been trained with EPG data that has been generated from 960 single source and mixed samples produced by three different CE platforms. The model performance has been tested with 341 samples generated in the same manner as the training data set. Overall, the iLSST-NR system demonstrated better performance compared to traditional ST methods with an overall accuracy of 97.2% for detecting true alleles and a false positive rate of 0.8% for detecting artefacts. Conversely, 11% of artefactual peaks were designated as true alleles, which represented 0.079% of the total alleles detected.

The main problem of this approach, according to the authors, is that the application of the iLSST-NR system successfully increased the number of interpretable alleles at the cost of calling an increased number of additional incorrect alleles compared to static thresholding methods. This problem seems to be mainly associated with mixed samples and specifically with low-template contributors at the background noise level. Nevertheless, the proposed ML approach in conjunction with DT has outperformed the traditional ST method in terms of maintaining the balance between retrieving the maximum information from EPG data, while minimising the number of artefactual peaks that were erroneously called alleles.

Based on outcomes of the aforementioned studies, the problem of accurate classification of such a diverse range of EPG signals presented in very large data sets required for model training, might be better addressed by a more sophisticated ML approach – deep learning. It is expected that deep learning algorithms with hundreds of artificial neurons in conjunction with model-free reinforced learning, are expected to produce a better classification outcome with CE- and especially MPS -generated data. Deep neural networks usually have more computational layers than traditional ML algorithms, thus offer an ability to learn more efficiently from large and diverse datasets. If the aim is not only to separate allelic signals from noise and artefacts, but provide a comprehensive genotyping solution for different types of forensic DNA samples, the DL models might be a preferable approach.

A number of recent studies have demonstrated successful use of ANNs to classify raw EPG signals as either allelic or artefactual [73,74]. In these studies, the researchers decided to eliminate all the genotyping thresholds completely. The studies followed a previous small-scale attempt by the same research group to use ANNs to categorise EPG data [72], which in-turn is based on a number of previously published studies (outside of the forensic science area) to analyse CE data with the ANN approach [79–81].

In this initial proof-of-concept, the authors built a neural network with 1206 input neurons (according to 201 scan points per EPG x 6 fluorescent dyes), one hidden layer with 100 neurons and an output layer with 5 neurons, (according to the number of classification categories) [72]. The ANN model was trained on a very limited dataset of only two single-source electropherograms (epg) decoded into a set of "scans", covering approximately 8 base pairs in both directions and resulting in 12,000 data sets of 1206 inputs. Each of those training sets went through 100 iterations (’epochs’ in the computer science lexicon) in order to recognise various features such as: baseline, alleles, pull-ups and stutters (both forward and reverse). The final prediction model demonstrated approximately 93% overall precision with a test dataset consisting of only one electropherogram, represented in the same manner as the training data set. Specifically, the ANN model demonstrated good performance for baseline and allele calling features classification (e.g. 96% precision) and inferior performance for artefacts prediction (e.g. 68% precision for the n-4 stutter prediction).

In their subsequent work, the researchers have substantially increased the amount of training data and number of computational networks [74]. The training dataset was represented by a total of 206 single – source and mixed DNA profiles of various template concentrations generated on two types of genetic analysers (3130xl and 3500xl). The ANN architecture was the same as in the preliminary model [72,73]. The authors compared performance of various ANN models (e.g. incorporating reference profiles, mixtures or both types of data generated on either a single CE platform or both) and found that an ANN model trained on mixed data demonstrated very similar or slightly improved performance compared to a model trained on a single data type, while the precision metrics ranged between 90% and 96% across all datasets.

The authors state that the developed ANN model can be potentially incorporated as a step before data analysis in the existing expert systems such as Genemapper™ to streamline the interpretation by saving substantial time required for manual analysis. In a more recent study, Lin et. al describe the developmental validation of FaSTR™ DNA analysis software following the FBI Quality Assurance Standards for Forensic DNA Testing Laboratories 2020 [82]. The developed software was compared with the GeneMapper™ ID-X software on a large dataset of 3403 single-source and mixed DNA profiles generated using seven different STR profiling kits. The study demonstrated almost 100% concordance in peak designations including an accurate designation of stutter peaks. In a more recent publication, the proposed ANN model has been improved, which together with the expanded training data, allowed it to classify an entire DNA profile, rather than individual profile features [83]. Furthermore, the FaSTR™ DNA software and its improved ANN model was recently validated by Forensic Science South Australia to determine whether one of the human DNA analysts could be replaced by an “ANN analyst” [84]. The validation demonstrated remarkable FaSTR™ DNA accuracy of 99.7% in detecting allele peaks in the reference profiles, with additional research ongoing [85,86].

2.3. STR genotyping using massively parallel sequencing data

Rapid progress in massively parallel DNA sequencing (MPS) and its superior capabilities have triggered the implementation of this technology for HID applications. Some operational laboratories are already transitioning from traditional CE fragment separation, especially for SNP genotyping. The computational methods designed for DNA sequencing data analysis face similar problems as those for CE-generated data, although MPS data brings additional challenges. For example, the MPS platforms generate a significantly larger amount of genomic data than CE genotyping methods. In addition, sequencing retrieves more comprehensive genomic information (e.g. the actual DNA sequence of STR loci including flanking regions vs just the number of tandem repeats), hence requiring extensive processing and interpretation. As a result, even more sophisticated bioinformatic tools are required to extract a DNA profile and resolve biological variants, intra-locus polymorphisms, sequencing errors and platform-specific artefacts, before the profile interpretation can be conducted. Given the complexity of MPS data analysis, the ML approach looks like a natural fit for solving this problem. From the ML point of view, this question fits into a pattern

analysis paradigm and can be approached with clustering or classification algorithms.

An attempt to develop a bioinformatic ML tool for extracting STR sequences from MPS raw data has been recently presented by Liu et al. [87]. Their *Fragsifier* software incorporates a random forest prediction model to designate STRs by locating both the repeat sequences and respective flanking sequences. This comprehensive strategy proved advantageous over the other previously published tools, which rely on either repeat or flanking region detection. *Fragsifier* is a platform-agnostic software tool that analyses sequencing reads in the FASTQ format and detects possible STR loci by the presence of consecutive k-mers of 3–6 nucleotides repeated more than two times, followed by alignment of STR flanking sequences to a reference genome. The prediction model has been established through supervised training with 7-mers of various STR loci, extracted from 40 sequenced samples and based on the reference dataset of approximately 2500 sequences extracted from STRait Razor v2s dataset [88]. In addition to 'positive feature training', the random forest prediction model was trained with 'negative feature training' examples that included trailing sequences used in the i7 sequencing adaptors to recognise 'what is not an STR'. The evaluation of the learning step produced a 30-tree forest model, which showed the best balance between prediction accuracy (99%) and low computational cost.

One of the limitations of the developed software is that it cannot correct sequencing errors or any sequence-specific artefacts such as homopolymer-read errors, which may affect its performance. Despite these limitations, the ML software produced generally concordant genotypes with CE and outperformed other previously published bioinformatic solutions for STR sequence extraction from MPS data. Another recent study by Yang et al. describes the use of linear regression and neural networks on MPS data of SNPs for DNA mixture interpretation [89]. The authors utilized a dataset of 10 single-source samples and 66 mixtures of various natures sequenced for 960 autosomal SNPs. The experiments demonstrated that both the linear regression and neural network models out-performed EuroForMix (probabilistic genotyping software in the R-coding language) in determining the minor contributor to DNA mixtures, including challenging profiles generated from highly degraded samples and closely related donors.

Another new and promising approach for mixture interpretation using ML was recently published by Crysup et.al [90]. This study explores the application of the previously described barcoding method using unique molecular identifiers (UMIs) [91] coupled with a machine learning bioinformatic pipeline to enhance the accuracy of allele calling in low-template DNA mixtures. Low-template samples and particularly low-template mixtures pose challenges in forensic DNA analysis due to PCR artefacts and the difficulty of distinguishing minor contributors from noise products. The authors developed a bioinformatic pipeline that utilizes UMIs attached to DNA sequences and employs ML techniques to filter out noise products. They conducted experiments with varying DNA input amounts and mixture ratios and found that using UMIs reduced noise, and machine learning further improved performance. The study demonstrates that UMIs coupled with a ML pipeline can significantly enhance allele calling accuracy by filtering out the noise products, especially in low-template mixtures, and offers potential benefits for forensic applications.

Despite a relatively small number of publications in this particular field, it is expected to increase steadily, in line with the growing number of forensic laboratories implementing MPS technology. The application of ML to mixture deconvolution is further discussed in the following subsection.

2.4. DNA mixture interpretation

Forensic DNA mixture deconvolution is one of the most challenging and controversial problems in forensic DNA analysis. Most challenges arise with interpreting complex DNA mixtures of three and more donors, especially if one or more of these contributors is represented by a very low-level portion of the mixture and/or the DNA molecules are degraded to varying degrees. In such a common situation, a true allele can be confused with various artefacts, including stutters, pull-ups, drop–in and drop-out. All these variables may influence one of the key parameters in DNA mixture interpretation – estimating the number of contributors (NoC). This assessment is fundamental for generating two competing propositions, usually defined as Hp (for the prosecution) and Hd (for the defence). Subsequently, these competing propositions are incorporated into Bayes theorem to evaluate the odds of different scenarios by calculating respective likelihood ratios. Due to the complexity of such calculations, it is usually completed using probabilistic genotyping software (PGS).

Most of the current PG systems rely on the manual assignment of the minimum NoC for the calculation of likelihood ratios. There are different methods to infer the NoC, such as the maximum allele count (MAC) [92], total allele count (TAC) [93] and the maximum likelihood estimation (MLE) [94]. The MLE approach is probabilistic by its nature and generally considered the most comprehensive: it can incorporate peak height information, allele sharing and probabilities of drop – in, drop – out, stutter and other artefacts. Conversely, the MAC method requires removal of all artefacts (especially stutters), which otherwise can lead to overestimating the NoC. Even after artefact removal, the MAC method is more likely to underestimate the NoC, due to the possibility of allele sharing. The uncertainty of mixture deconvolution by current PGSs increases proportionally with the number of contributors and/or when there is significant allele sharing between contributors [95].

One of the first attempts to efficiently decipher the NoC using an ML approach was described by Swaminathan et. al [96]. The authors developed a MCMC continuous probabilistic method - NOCIt - that utilises information about peak height, degradation, forward and reverse stutter, noise and allelic drop-out and other parameters to infer the number of contributors to a profile. In a follow up study, the model was calibrated with 100 single-source ground truth profiles, while its performance was consequently evaluated with 815 DNA profiles of varying quality, number of contributors, and mixture proportions [97]. Notably, this process included a manual step of artefact removal prior to incorporation into the computational pipeline. In a more recent publication by the same research group, the authors developed another continuous model, incorporating ANN, then performed benchmark comparison between the two methods and the commonly-used MAC approach [98]. Each method was tested with a standard set of 214 PROVEDIt mixtures [99] and focused on accuracy and precision of the true NoC estimation (based on the highest LR) among other parameters. As expected, the ANN required a much longer training period than the NOCIt (approximately 24 h vs several minutes). However, after the training, it ran faster and demonstrated higher precision metrics than the NOCIt, which required tens of minutes to run and was less repeatable. The authors also note that ML models in this particular study and in general, have certain limitations such as higher variability of results (i.e. precision, accuracy, robustness) compared to probabilistic models.

Another significant contribution to the field of DNA mixture analysis using computational methods was made in 2017 by the introduction of PACE: probabilistic assessment for contributor estimation software [100]. This study evaluated five different ML algorithms: k-NN, CART, MLP, SVM and MNLR for NoC prediction accuracy and artefact removal efficiency. The model was initially trained with STR data generated using the Identifiler™ amplification kit (Thermo Fisher Scientific). Evaluation of all five models demonstrated relatively similar performance. However, the non-linear SVM algorithm showed the best performance such as higher classification accuracy of all the tested algorithms, while the linear SVM algorithm demonstrated the worst metrics. Furthermore, the non-linear SVM model yielded NoC prediction accuracies of 100%, 98.1%, 95.9% and 100% for single – source samples and mixtures with two-, three- and four – contributors, respectively. It

should be noted that due to training dataset limitations, the accuracy for four donor mixtures is an estimation and is actually lower by approximately 2 per cent.

The subsequent development and validation of the latest PACE™ v1.3.7 software utilised a continuous probabilistic model in conjunction with the SVM classification approach [101]. The model has been extended to incorporate data from additional STR kits and trained on both single– person DNA profiles and up to four-person mixtures generated with the Globalfiler™ (Thermo Fisher Scientific) and PowerPlex Fusion 6c® (Promega) amplification kits. It also includes modules that permit automated artefact identification such as n-1 and n + 1 stutters, pull-up and background noise, using the iLSST-NR algorithms previously developed by the authors. The developmental validation of PACE demonstrated over 93.5% accuracy in automated artefact identification for both amplification kits and greater than 90% accuracy in predicting the number of contributors (up to four donors). According to the authors, the relatively high number of misclassified samples in validation experiments (10%) was due to samples with high levels of allele sharing, low DNA template and allelic dropout in degraded DNA samples. Notably, most of the misclassified samples were underestimated by one contributor only.

In the follow-up study, the model was updated to predict up to 5 contributors and demonstrated even higher accuracy in both artefact identification and NoC estimation (95% and 92%, respectively) [102]. The validation included complex DNA mixtures with various contributor ratios, samples with high levels of dropout due to degradation and low template samples. According to the authors, an additional benefit of the PACE software is its efficient usage of computational resources: the results can be obtained in seconds using a standard desktop or laptop computer (compared to hours with alternative methods implementing MCMC algorithms). Another benefit is the flexibility of the PACE model, which potentially can be trained to produce classifications from data generated with other amplification kits, such as MPS-based kits. Pending extensive validation, the current version of the software can be potentially incorporated into operational workflow *prior* to a PG software pipeline, to improve the accuracy and increase the speed of the NoC estimation, and subsequently, generate more accurate propositions for likelihood ratio calculations.

Another notable attempt to assist with the correct estimation of NoC has been published by Benschop et al. [103]. Their ML software is based on an RF classifier approach and named "RFC19" (due to 19 features it seeks to classify). These 19 features comprise information about various aspects of allele count, peak heights and allele frequencies (refer to Fig. 3 in the original article for more details). The final model has been selected upon evaluation of ten ML algorithms based on precision and recall metrics. Two models (RFC19 and LDA40) demonstrated comparable performance, which was superior to the remaining models. Considering that RFC19 required fewer features for classification (19 vs 40), it has been selected for further validation.

The final model was trained on single-donor samples and up to 5-donor mixtures of various quantity and quality from 1174 unique donors, which were used to construct 590 DNA profiles. The NoC classification accuracy of the final model was 83.3%, with most mischaracterisations occurring in challenging samples represented by complex mixtures of three to five donors with degraded DNA template. The comparison of RFC19 to other non-ML NoC estimation methods such as MAC and nC-tool (an in-house TAC-based approach), demonstrated its superior performance in accurate classification of 2- to 5-person mixtures (85% vs 69.2% vs 76.7%, respectively). One of the limitations of the current RFC19 model is the absence of female DNA profiles (both single and mixtures) in the training dataset. Incorporation of such data is essential and would affect (and potentially improve) the model's prediction accuracy. Overall, the performance of the RFC19 model seems to be slightly inferior to that of PACE™ (as discussed above). This comparison however, might not be entirely objective considering intrinsic differences in numerous parameters, such as training datasets (as discussed below). It is worth noting that a recent follow-up study describes another method for NoC estimation based on a less computationally intensive and 'simple' decision tree model (only two decision nodes) [104]. The proposed decision tree model was compared to RFC19, NOCIt and MAC approaches, demonstrating 77% accuracy in NoC estimation.

The automatic ML methods for NoC estimation can be used as standalone software or as an integral component of the PGS. To this point, McGovern et. al describe a developmental validation of another probabilistic algorithm termed the 'variable number of contributors' (VarNoC), integrated into the STRMixTM software [105], which is based on a previously proposed MCMC method by Weinberg [106]. The validation demonstrated a relatively good performance of the VarNoC module (i.e. higher LRs for the correct NoC) with a sample set of ten DNA mixtures with varying mixture proportions. As expected, with more challenging samples (i.e. more complex mixtures), the accuracy of NoC estimations and the reproducibility of the LRs were less accurate. Also, the VarNoC can be considered a semi-automatic approach as it requires manual assignment for the anticipated range of NoC (i.e. Nn to Nn+1) [107].

Based on the current literature, the ML approach is very promising and may provide a better answer for the accurate estimation of NoC (and mixture interpretation in general) over the existing methods. However, the implementation of this tool into forensic casework would require extensive testing to identify its capabilities and limitations as with any new tool in the forensic arsenal. On this point, it is important to emphasize a common problem of validation studies, which was also highlighted by the authors of the NOCIt study [98]. This problem is the lack of standardization between different studies. Significant variability in data sets and system parameters, differences in data pre-processing (e.g., stutter pre-filtering) and methods trained with different goals in mind (e.g., different maximum NoC to be predicted) limit the true face-to-face comparison between the NoC estimation methods. However, this problem is not unique and applies to other computational methods and forensic disciplines, as accentuated by the PCAST report and recent publications [108,109]. Therefore, a proper comparison of this kind would require a large standardised dataset of a wide range of sample types tested under variable analytical conditions, according to the SWGDAM validation guidelines [9110]. This requirement is an essential prerequisite for conducting proper validation studies and implementing a new method into operational practice.

Another significant problem that applies not only to this specific field, but forensic science in general is the "opacity" of the ML algorithms. In other words, the software user usually sees only the output (i.e. prediction of the NoC) without any reasoning for that conclusion. Considering the legal context of forensic evidence analysis, this situation has significant judicial implications. Recently, a few studies have attempted to produce more transparent models using XAI approaches (as introduced in Section 1.5). For example, a Realistic Counterfactuals (ReCo) method has been proposed to generate realistic counterfactual explanations for correlated data [63]. It uses a decision tree to provide realistic and interpretable explanations for the NoC predictions made by the model. The effectiveness of this approach was demonstrated on both simulated and real-world DNA datasets and can be subsequently used for any type of ML model. It can also be extended to other areas of forensic science, such as fingerprint analysis and ballistics, where XAI can play a similar role in improving the interpretability and transparency of predictions.

2.5. *Additional applications of machine learning in forensic DNA analysis*

While the focus to this point has been on ML to solve various problems with STR data analysis in the area of HID, ML methods have also been successfully applied to a number of other aspects of forensic DNA analysis. For example, a recent publication describes an application of an ML predictive model to infer the Y-SNP haplogroup from a Y-STR haplotype [111]. The authors evaluated several ML algorithms,

including kNN, NB, Logistic Regression, SVM, DeT, and RF. All models except NB, showed high prediction accuracy (above 95%) in the lowest haplogroup resolution, while for the highest haplogroup resolution, the RF model demonstrated a superior outcome compared to other models (77% accuracy).

Another recently published study describes ML models utilising SVM and RF classifiers for haplogroup prediction based on Y-STR data [112]. The study compared performance of the three ML models: SVM, RF and K-NN to build an effective function for accurately predicting a haplogroup. The developed training model was fed with data from 32 Y-STRs genotyped in 377 samples from 19 populations, while an additional 126 samples were used to test the model performance. The final model has been integrated into a free software "*PredYMaLe*" which was tested on additional published datasets and showed very high prediction accuracy (over 97%). According to the authors, the software is not limited to only the markers used for its training, and can be fed with any set of Y-STRs.

Another relevant HID application of ML has been demonstrated in a recent study by Woerner et. al who has investigated human individualisation with bioinformatic data generated from sequencing 286 genomic markers in 22 microbial species, sampled from three body sites in 51 individuals [113]. Specifically, the study has compared microbial genome composition and phylogenetic analysis incorporated into two ML classifiers: nearest neighbour and reverse nearest neighbour. The classifiers demonstrated a remarkable classification accuracy of 100% while using the maximum nearest neighbour conditioned approach. The study also suggested that microbial strain composition is more individualising than a phylogenetic approach.

Additional potential applications of ML methods in the forensic DNA field beyond the focus of this review include forensic intelligence and inference, such as prediction of the body fluid origin [114,115], visual appearance traits [116], biogeographical ancestry [117–119], biological age [120–122], and post-mortem interval estimation [123,124]. Most of these applications require analysis of complex MPS data, which can be streamlined using ML solutions, as discussed above.

3. Conclusion

ML algorithms are considered a subfield of artificial intelligence. They are designed to train a computer program how to learn from experience with respect to some tasks and its performance measure(s) [125]. These algorithms have a wide range of potential applications in forensic DNA analysis and beyond, as illustrated by a growing body of literature. The use of ML algorithms in forensic science in general and forensic DNA analysis in particular can provide valuable insights by supporting decision-making processes and minimizing human subjectivity. ML algorithms offer significant advantages over the current methods of DNA data analysis and interpretation and are gradually becoming incorporated into routine casework. However, as with any novel forensic tool, ML methods must be extensively tested and validated prior to operational implementation. Consequently, the properly validated computational algorithms can improve the throughput and reliability of evidence analysis and reduce the implicit subjectivity of human interpretation.

Some of the major limitations preventing rapid implementation of the ML methods and especially deep learning methods include the requirements for very large (and relevant) training sets representing the widest spectrum of possible cases, significant computational resources to allow superior processing capacity and training of highly-skilled personnel with the knowledge that is not readily available in most operational laboratories. Most importantly, the upcoming implementation of ML platforms in forensic DNA analysis (and other forensic disciplines) must remain as transparent as possible rather than" a process specified only in terms of the relationship between inputs and outputs that uses information to produce a particular set of results" [126], also known as a 'black box' approach. A lack of transparency can undermine the reliability and validity of the results produced by these algorithms. XAI techniques, such as decision trees, rule-based models, Bayesian networks, and the use of separate train/test/validation sets, can help improve the transparency and reliability of the ML algorithms used in forensic science. By developing more transparent and interpretable (i.e. "white-box") algorithms, we can increase confidence in the results produced by ML algorithms and ensure their acceptability in legal contexts.

The integration of ML methods in forensic science represents an objective and irreversible progression. In a similar manner, novel technologies like digital photography were gradually incorporated into expert practice. Initially, there was scepticism within the expert community concerning digital photography (authors' personal communication and [127]). Nevertheless, as image quality improved and digital technology advanced, its seamless integration into expert practice occurred without altering the fundamental principles of forensic photography. In the present day, it is nearly impossible to find documentation of a crime scene or an expert report without digital photographs. In a similar manner, the implementation of ML algorithms is expected to be a valuable asset for analysing and interpreting challenging forensic evidence. Their primary advantage lies in streamlining analysis, reducing workloads, and minimizing the subjectivity of the decision-making process. However, considering the technology's relatively nascent state, it may take a significant amount of time before machines can completely replace human experts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.M. Butler, Fundamentals of forensic DNA typing, Academic press, 2009.
[2] E. Pilli, A. Berti, Forensic DNA analysis: technological development and innovative applications, CRC Press, 2021.
[3] B. Mahesh, Machine learning algorithms - a review, Int. J. Sci. Res. 9 (2020) 381–386.
[4] P. Scaruffi, Intelligence is not artificial: A history of artificial intelligence and why the singularity is not coming any time soon, Creat. Indep. Publ. Platf. (2018).
[5] D. McNevin, K. Wright, M. Barash, S. Gomes, A. Jamieson, J. Chaseling, Proposed framework for comparison of continuous probabilistic genotyping systems amongst different laboratories, Forensic Sci. 1 (1) (2021) 33–45.
[6] D. Taylor, J.A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Sci. Int Genet 7 (5) (2013) 516–528.
[7] J.S. Buckleton, J.-A. Bright, D. Taylor, The continuous model. Forensic DNA Evidence Interpretation, CRC press, 2018.
[8] D. Taylor, J. Buckleton, J.A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Sci. Int Genet 21 (2016) 126–133.
[9] Scientific Working Group on DNA Analysis Methods (SWGDAM): Validation Guidelines for DNA Analysis Methods. <⟨https://www.swgdam.org/_files/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf⟩>, 2016 (accessed December 12,. 2021).
[10] F. Rosenblatt, The perceptron, a perceiving and recognizing automaton Project Para. <⟨https://blogs.umass.edu/brain-wars/files/2016/03/rosenblatt-1957.pdf⟩>, 1957).
[11] A.L. Samuel, Some studies in machine learning using the game of checkers, IBM J. Res. Dev. 3 (3) (1959) 210–229.
[12] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (4) (1943) 115–133.
[13] I.H. Sarker, Machine learning: algorithms, real-world applications and research directions, SN Comput. Sci. 2 (160) (2021).
[14] S.B. Kotsiantis, I. Zaharakis, P. Pintelas, Supervised machine learning: a review of classification techniques, Emerg. Artif. Intell. Appl. Comput. Eng. 160 (1) (2007) 3–24.
[15] Y. Bengio, A.C. Courville, P. Vincent, Unsupervised feature learning and deep learning: A review and new perspectives, CoRR, abs/1206.5538 1 (2012).
[16] O. Chapelle, B. Schölkopf, A. Zien, Learning, Semi-Supervised, MIT Press, 2006.
[17] M. Mittal, L.M. Goyal, D.J. Hemanth, J.K. Sethi, Clustering approaches for high-dimensional databases: a review, Wiley Interdiscip. Rev. -Data Min. Knowl. Discov. 9 (3) (2019), e1300.
[18] S.K. Popat, M. Emmanuel, Review and comparative study of clustering techniques, Int. J. Comput. Sci. Inf. Technol. 5 (1) (2014) 805–812.
[19] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2) (2000) 945–959.

[20] D. Falush, M. Stephens, J.K. Pritchard, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, Genetics 164 (4) (2003) 1567–1587.

[21] M.J. Hubisz, D. Falush, M. Stephens, J.K. Pritchard, Inferring weak population structure with the assistance of sample group information, Mol. Ecol. Resour. 9 (5) (2009) 1322–1332.

[22] F. Liu, K. van Duijn, J.R. Vingerling, A. Hofman, A.G. Uitterlinden, A.C. Janssens, M. Kayser, Eye color and the prediction of complex phenotypes from genotypes, Curr. Biol. 19 (5) (2009) R192–R193.

[23] S. Walsh, F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA, Forensic Sci. Int Genet 7 (1) (2013) 98–115.

[24] K. Zaorska, P. Zawierucha, M. Nowicki, Prediction of skin color, tanning and freckling from DNA in Polish population: linear regression, random forest and neural network approaches, Hum. Genet 138 (6) (2019) 635–647.

[25] A. Dasgupta, Y.V. Sun, I.R. Konig, J.E. Bailey-Wilson, J.D. Malley, Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience, Suppl 1, Genet Epidemiol. 35 (Suppl 1) (2011) S5–S11.

[26] X.Z. Xu, T.M. Liang, J. Zhu, D. Zheng, T.F. Sun, Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing 328 (2019) 5–15.

[27] L.C. Lee, A.A. Jemain, On overview of PCA application strategy in processing high dimensionality forensic data, Microchem. J. 169 (2021), 106608.

[28] C. Phillips, M. Fondevila, M.V. Lareau, A 34-plex Autosomal SNP Single Base Extension Assay for Ancestry Investigations, in: A. Alonso (Ed.), DNA Electrophoresis Protocols for Forensic Genetics, Humana Press, Totowa, NJ, 2012, pp. 109–126.

[29] J.J. Sanchez, C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C. D. Harrison, E. Musgrave-Brown, A. Salas, D. Syndercombe-Court, P. M. Schneider, A. Carracedo, N. Morling, A multiplex assay with 52 single nucleotide polymorphisms for human identification, Electrophoresis 27 (9) (2006) 1713–1724.

[30] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016.

[31] A.K. Jain, J.C. Mao, K.M. Mohiuddin, Artificial neural networks: a tutorial, Computer 29 (3) (1996) 31–44.

[32] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, A.M. Umar, O.U. Linus, H. Arshad, A.A. Kazaure, U. Gana, M.U. Kiru, Comprehensive review of artificial neural network applications to pattern recognition, IEEE Access 7 (2019) 158820–158846.

[33] B. News, Artificial Intelligence: Google's AlphaGo Beats Go Master Lee Se-Dol. <⟨https://www.bbc.com/news/technology-35785875⟩>, 2016 (accessed October 30.2023).

[34] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.

[35] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, Neural Comput. 29 (9) (2017) 2352–2449.

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.

[37] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3313–3332.

[38] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 770–778.

[39] G. Antipov, M. Baccouche, J. Dugelay, Face aging with conditional generative adversarial networks, 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 2089–2093.

[40] A. Anand, R.D. Labati, A. Genovese, E. Munoz, V. Piuri, F. Scotti, Age estimation based on face images and pre-trained convolutional neural networks, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017, pp. 1–7.

[41] M. Kayser, Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes, Forensic Sci. Int Genet 18 (2015) 33–48.

[42] M. Barash, P.E. Bayer, A. van Daal, Identification of the single nucleotide polymorphisms affecting normal phenotypic variability in human craniofacial morphology using candidate gene approach, J. Genet. Genome Res. 5 (1) (2018), 060814.

[43] A.T. Andriani, P.K. Zahra, E.I. Auerkari, Genetic contributions to craniofacial growth: a review, J. Phys.: Conf. Ser. 1943 (1) (2021), 012095.

[44] D. Sero, A. Zaidi, J. Li, J.D. White, T.B.G. Zarzar, M.L. Marazita, S.M. Weinberg, P. Suetens, D. Vandermeulen, J.K. Wagner, M.D. Shriver, P. Claes, Facial recognition from DNA using face-to-DNA classifiers, Nat. Commun. 10 (1) (2019) 2557.

[45] S. Naqvi, H. Hoskens, F. Wilke, S.M. Weinberg, J.R. Shaffer, S. Walsh, M. D. Shriver, J. Wysocka, P. Claes, Decoding the human face: progress and challenges in understanding the genetics of craniofacial morphology, Annu Rev. Genom. Hum. Genet 23 (1) (2022) 383–412.

[46] N. Pandkar, T.-S. Moh, M. Barash, 2022. 3D Facial Biometric Verification Using a DNA Sample for Law Enforcement Applications, 21st IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'22), IOS Press.

[47] J.D. White, K. Indencleef, S. Naqvi, R.J. Eller, H. Hoskens, J. Roosenboom, M. K. Lee, J. Li, J. Mohammed, S. Richmond, E.E. Quillen, H.L. Norton, E. Feingold, T. Swigut, M.L. Marazita, H. Peeters, G. Hens, J.R. Shaffer, J. Wysocka, S. Walsh, S.M. Weinberg, M.D. Shriver, P. Claes, Insights into the genetic architecture of the human face, Nat. Genet 53 (1) (2021) 45–53.

[48] P. Dabas, S. Jain, H. Khajuria, B.P. Nayak, Forensic DNA phenotyping: inferring phenotypic traits from crime scene DNA, J. Forensic Leg. Med 88 (2022), 102351.

[49] J. Qiu, Q. Wu, G. Ding, Y. Xu, S. Feng, A survey of machine learning for big data processing, 2016, EURASIP J. Adv. Signal Process. 1 (2016) 67.

[50] C. Wang, X. Liao, L. Carin, D.B. Dunson, Classification with incomplete data using Dirichlet process priors, J. Mach. Learn Res. 11 (12) (2010) 3269–3311.

[51] S. Saini, I. Mitra, N. Mousavi, S.F. Fotsing, M. Gymrek, A reference haplotype panel for genome-wide imputation of short tandem repeats, Nat. Commun. 9 (1) (2018) 4397.

[52] J. Chen, J. Yang, K. Li, Q. Ji, X. Kong, S. Xie, W. Zhan, J. Wu, S. Huang, H. Huang, R. Li, Z. Zhang, Y. Cao, Y. Yu, Z. Mao, Y. Yu, H. Lv, Y. Pu, F. Chen, P. Chen, Evaluation of a SNP-STR haplotype panel for forensic genotype imputation, Forensic Sci. Int Genet 62 (2023), 102801.

[53] J. Kim, N.A. Rosenberg, Record-matching of STR profiles with fragmentary genomic SNP data, Eur. J. Hum. Genet 31 (11) (2023) 1283–1290.

[54] V. Berisha, C. Krantsevich, P.R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, J. Liss, Digital medicine and the curse of dimensionality, NPJ Digit Med 4 (1) (2021) 153.

[55] E. Debie, K. Shafi, Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses, Pattern Anal. Appl. 22 (2) (2019) 519–536.

[56] C. Coglianese, D. Lehr, Regulating by robot: administrative decision making in the machine-learning era, Geo. LJ 105 (2016) 1147.

[57] M. Busuioc, Accountable artificial intelligence: holding algorithms to account, Public Adm. Rev. 81 (5) (2021) 825–836.

[58] S.K. Katyal, Private ACcountability in the Age of Artificial Intelligence, Ucla Law Rev. 66 (1) (2019) 54–141.

[59] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 1–42.

[60] P.J. Phillips, C.A. Hahn, P.C. Fontana, D.A. Broniatowski, M.A. Przybocki, 2020. Four principles of explainable artificial intelligence, Gaithersburg, Maryland. p.18.

[61] A.A. Solanke, Explainable digital forensics AI: towards mitigating distrust in AI-based digital forensics analysis using interpretable models, Forensic Sci. Int. Digit. Investig. 42 (2022), 301403.

[62] S.W. Hall, A. Sakzad, K.K.R. Choo, 2022. Explainable artificial intelligence for digital forensics, Wiley Interdisciplinary Reviews: Forensic Science. 4(2): e1434.

[63] M.S. Veldhuis, S. Ariens, R.J.F. Ypma, T. Abeel, C.C.G. Benschop, Explainable artificial intelligence in forensics: Realistic explanations for number of contributor predictions of DNA profiles, Forensic Sci. Int Genet 56 (2022), 102632.

[64] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet 16 (6) (2015) 321–332.

[65] D. Jordan, D. Mills, Past, present, and future of DNA typing for analyzing human and non-human forensic samples, Front. Ecol. Evol. 9 (172) (2021).

[66] B.R. McCord, Q. Gauthier, S. Cho, M.N. Roig, G.C. Gibson-Daw, B. Young, F. Taglia, S.C. Zapico, R.F. Mariot, S.B. Lee, G. Duncan, Forensic DNA analysis, Anal. Chem. 91 (1) (2019) 673–688.

[67] P.M. Schneider, B. Prainsack, M. Kayser, The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry, Dtsch Arztebl Int. 51–52 (51–52) (2019) 873–880.

[68] L.A. Marano, C. Fridman, DNA phenotyping: current application in forensic science, Res. Rep. Forensic Med. Sci. 9 (2019) 1–8.

[69] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49 (12) (1996) 1373–1379.

[70] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, John Wiley & Sons, 2013.

[71] M.A. Marciano, V.R. Williamson, J.D. Adelman, A hybrid approach to increase the informedness of CE-based data using locus-specific thresholding and machine learning, Forensic Sci. Int Genet 35 (2018) 26–37.

[72] D. Taylor, D. Powers, Teaching artificial intelligence to read electropherograms, Forensic Sci. Int Genet 25 (2016) 10–18.

[73] D. Taylor, A. Harrison, D. Powers, An artificial neural network system to identify alleles in reference electropherograms, Forensic Sci. Int Genet 30 (2017) 114–126.

[74] D. Taylor, M. Kitselaar, D. Powers, The generalisability of artificial neural networks used to classify electrophoretic data produced under different conditions, Forensic Sci. Int Genet 38 (2019) 181–184.

[75] Scientific Working Group on DNA Analysis Methods (SWGDAM): Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. <⟨https://www.swgdam.org/_files/ugd/4344b0_3f94c9a6286048c3924c58e2c230e74e.pdf⟩>, 2021 (accessed June 25.2022).

[76] W.C. Thompson, F. Taroni, C.G. Aitken, How the probability of a false positive affects the value of DNA evidence, J. Forensic Sci. 48 (1) (2003) 47–54.

[77] A. Kloosterman, M. Sjerps, A. Quak, Error rates in forensic DNA analysis: definition, numbers, impact and communication, Forensic Sci. Int Genet 12 (2014) 77–85.

[78] J.D. Adelman, A. Zhao, D.S. Eberst, M.A. Marciano, Automated detection and removal of capillary electrophoresis artifacts due to spectral overlap, Electrophoresis 40 (14) (2019) 1753–1761.
[79] B. Pokrić, N.M. Allinson, E.T. Bergström, D.M. Goodall, Dynamic analysis of capillary electrophoresis data using real-time neural networks, J. Chromatogr. A 833 (2) (1999) 231–244.
[80] G. Bocaz-Beneventi, R. Latorre, M. Farková, J. Havel, Artificial neural networks for quantification in unresolved capillary electrophoresis peaks, Anal. Chim. Acta 452 (1) (2002) 47–63.
[81] O.G. Mohammed, K.T. Assaleh, G.A. Husseini, A.F. Majdalawieh, S.R. Woodward, Novel algorithms for accurate DNA base-calling, J. Biomed. Sci. Eng. 6 (02) (2013) 165.
[82] M.-H. Lin, S.-I. Lee, X. Zhang, L. Russell, H. Kelly, K. Cheng, S. Cooper, R. Wivell, Z. Kerr, J. Morawitz, Developmental validation of FaSTR™ DNA: software for the analysis of forensic DNA profiles, Forensic Sci. Int.: Rep. 3 (2021), 100217.
[83] D. Taylor, Using a multi-head, convolutional neural network with data augmentation to improve electropherogram classification performance, Forensic Sci. Int Genet. 56 (2022), 102605.
[84] L. Volgin, D. Taylor, J.A. Bright, M.H. Lin, Validation of a neural network approach for STR typing to replace human reading, Forensic Sci. Int Genet. 55 (2021), 102591.
[85] D. Taylor, J. Buckleton, Combining artificial neural network classification with fully continuous probabilistic genotyping to remove the need for an analytical threshold and electropherogram reading, Forensic Sci. Int Genet 62 (2023), 102787.
[86] D. Taylor, D. Abarno, A lights-out forensic DNA analysis workflow for no-suspect crime, Forensic Sci. Int Genet 66 (2023), 102907.
[87] Y.Y. Liu, D. Welch, R. England, J. Stacey, S. Harbison, Forensic STR allele extraction using a machine learning paradigm, Forensic Sci. Int Genet 44 (2020), 102194.
[88] J.L. King, F.R. Wendt, J. Sun, B. Budowle, STRait Razor v2s: advancing sequence-based STR allele reporting and beyond to other marker systems, Forensic Sci. Int Genet 29 (2017) 21–28.
[89] T.-W. Yang, Y.-H. Li, C.-F. Chou, F.-P. Lai, Y.-H. Chien, H.-I. Yin, T.-T. Lee, H.-L. Hwa, DNA mixture interpretation using linear regression and neural networks on massively parallel sequencing data of single nucleotide polymorphisms, Australian, J. Forensic Sci. 54 (2) (2022) 150–162.
[90] B. Crysup, S. Mandape, J.L. King, M. Muenzler, K.B. Kapema, A.E. Woerner, Using unique molecular identifiers to improve allele calling in low-template mixtures, Forensic Sci. Int Genet 63 (2023), 102807.
[91] A.E. Woerner, S. Mandape, J.L. King, M. Muenzler, B. Crysup, B. Budowle, Reducing noise and stutter in short tandem repeat loci with unique molecular identifiers, Forensic Sci. Int Genet 51 (2021), 102459.
[92] T.M. Clayton, J.P. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int 91 (1) (1998) 55–70.
[93] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, Int J. Leg. Med 128 (3) (2014) 427–437.
[94] H. Haned, L. Pene, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (1) (2011) 23–28.
[95] J.M. Butler, H. Iyer, R. Press, M.K. Taylor, P.M. Vallone, S. Willis, DNA Mixture Interpretation: A NIST Scientific Foundation Review (draft), 2021.
[96] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int Genet 16 (2015) 172–180.
[97] C.M. Grgicak, S. Karkar, X. Yearwood-Garcia, L.E. Alfonse, K.R. Duffy, D.S. Lun, A large-scale validation of NOCIt's a posteriori probability of the number of contributors and its integration into forensic interpretation pipelines, Forensic Sci. Int Genet 47 (2020), 102296.
[98] J. Valtl, U.J. Monich, D.S. Lun, J. Kelley, C.M. Grgicak, A series of developmental validation tests for Number of Contributors platforms: Exemplars using NOCIt and a neural network, Forensic Sci. Int Genet 54 (2021), 102556.
[99] L.E. Alfonse, A.D. Garrett, D.S. Lun, K.R. Duffy, C.M. Grgicak, A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEDIt, Forensic Sci. Int Genet 32 (2018) 62–70.
[100] M.A. Marciano, J.D. Adelman, PACE: probabilistic assessment for contributor estimation- a machine learning-based assessment of the number of contributors in DNA mixtures, Forensic Sci. Int Genet 27 (2017) 82–91.
[101] M.A. Marciano, J.D. Adelman, Developmental validation of PACE: Automated artifact identification and contributor estimation for use with GlobalFiler and PowerPlex(R) fusion 6c generated data, Forensic Sci. Int Genet 43 (2019), 102140.
[102] M.A. Marciano, J. Adelman, L. Armogida, PACE™: Rapid and automated artifact identification and number of contributor prediction (Webinar). <⟨https://learning.forensicac.org/course/view.php?id=406⟩>, 2020 (accessed 07/28/2020.).
[103] C.C.G. Benschop, J. van der Linden, J. Hoogenboom, R. Ypma, H. Haned, Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach, Forensic Sci. Int Genet 43 (2019), 102150.
[104] M. Kruijver, H. Kelly, K. Cheng, M.H. Lin, J. Morawitz, L. Russell, J. Buckleton, J. A. Bright, Estimating the number of contributors to a DNA profile using decision trees, Forensic Sci. Int Genet 50 (2021), 102407.
[105] D. Taylor, J.-A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci. Int.: Genet. 13 (2014) 269–280.
[106] M.D. Weinberg, Computing the Bayes factor from a Markov chain Monte CArlo simulation of the posterior distribution, Bayesian Anal. 7 (3) (2012) 737–769.
[107] C. McGovern, K. Cheng, H. Kelly, A. Ciecko, D. Taylor, J.S. Buckleton, J.A. Bright, Performance of a method for weighting a range in the number of contributors in probabilistic genotyping, Forensic Sci. Int Genet 48 (2020), 102352.
[108] E.S. Lander, P.W. Group, Forensic science in criminal courts: ensuring scientific validity of feature-comparison methods, (2016).
[109] D. McNevin, K. Wright, J. Chaseling, M. Barash, Commentary on: Bright, et al., Internal validation of STRmix - a multi laboratory response to PCAST, Forensic Sci. Int Genet. 41 (2018) e14–e17, 2019.
[110] M.D. Coble, J. Buckleton, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmao, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S. T. Sherry, S. Willuweit, M. Prinz, DNA Commission of the International Society for Forensic Genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int Genet 25 (2016) 191–197.
[111] M.Y. Song, C.X. Zhao, Z. Wang, Y.P. Hou, Applying machine learning algorithms to a real forensic case to predict Y-SNP haplogroup based on Y-STR haplotype, Forens. Sci. Int-Gen. S 7 (1) (2019) 637–638.
[112] C. Bouakaze, F. Delehelle, N. Saenz-Oyhereguy, A. Moreira, S. Schiavinato, M. Croze, S. Delon, C. Fortes-Lima, M. Gibert, L. Bujan, E. Huyghe, G. Bellis, R. Calderon, C.L. Hernandez, E. Avendano-Tamayo, G. Bedoya, A. Salas, S. Mazieres, J. Charioni, F. Migot-Nabias, A. Ruiz-Linares, J.M. Dugoujon, C. Theves, C. Mollereau-Manaute, C. Nous, N. Poulet, T. King, M.E. D'Amato, P. Balaresque, Predicting haplogroups using a versatile machine learning program (PredYMaLe) on a new mutationally balanced 32 Y-STR multiplex (CombYplex): unlocking the full potential of the human STR mutation rate spectrum to estimate forensic parameters, Forensic Sci. Int Genet 48 (2020), 102342.
[113] A.E. Woerner, N.M.M. Novroski, F.R. Wendt, A. Ambers, R. Wiley, S.E. Schmedes, B. Budowle, Forensic human identification with targeted microbiome markers using nearest neighbor classification, Forensic Sci. Int Genet 38 (2019) 130–139.
[114] D. Iacob, A. Fürst, T. Hadrys, A machine learning model to predict the origin of forensically relevant body fluids, Forens. Sci. Int-Gen. S 7 (1) (2019) 392–394.
[115] R.J.F. Ypma, P.A. Maaskant-van Wijk, R. Gill, M. Sjerps, M. van den Berge, Calculating LRs for presence of body fluids from mRNA assay data in mixtures, Forensic Sci. Int Genet 52 (2021), 102455.
[116] M.A. Katsara, W. Branicki, S. Walsh, M. Kayser, M. Nothnagel, V. Consortium, Evaluation of supervised machine-learning methods for predicting appearance traits from DNA, Forensic Sci. Int Genet 53 (2021), 102507.
[117] K. Sun, Y. Yao, L. Yun, C. Zhang, J. Xie, X. Qian, Q. Tang, L. Sun, Application of machine learning for ancestry inference using multi-InDel markers, Forensic Sci. Int Genet 59 (2022), 102702.
[118] M. Hajiloo, Y. Sapkota, J.R. Mackey, P. Robson, R. Greiner, S. Damaraju, ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction, BMC Bioinforma. 14 (1) (2013) 61.
[119] E. Alladio, B. Poggiali, G. Cosenza, E. Pilli, Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field, Sci. Rep. 12 (1) (2022) 8974.
[120] A. Montesanto, P. D'Aquila, V. Lagani, E. Paparazzo, S. Geracitano, L. Formentini, R. Giacconi, M. Cardelli, M. Provinciali, D. Bellizzi, G. Passarino, A new robust epigenetic model for forensic age prediction, J. Forensic Sci. 65 (5) (2020) 1424–1431.
[121] A. Vidaki, D. Ballard, A. Aliferi, T.H. Miller, L.P. Barron, D. Syndercombe Court, DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, Forensic Sci. Int Genet. 28 (2017) 225–236.
[122] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. Henneman, A.D. Kloosterman, P.J. Verschure, Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression, Forensic Sci. Int Genet 31 (2017) 19–28.
[123] R. Liu, Y. Gu, M. Shen, H. Li, K. Zhang, Q. Wang, X. Wei, H. Zhang, D. Wu, K. Yu, W. Cai, G. Wang, S. Zhang, Q. Sun, P. Huang, Z. Wang, Predicting postmortem interval based on microbial community sequences and machine learning algorithms, Environ. Microbiol 22 (6) (2020) 2273–2291.
[124] H.R. Johnson, D.D. Trinidad, S. Guzman, Z. Khan, J.V. Parziale, J.M. DeBruyn, N. H. Lents, A machine learning approach for using the postmortem skin microbiome to estimate the postmortem interval, PLoS One 11 (12) (2016), e0167370.
[125] M.I. Jordan, T.M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.
[126] P. Procter, Cambridge international dictionary of English, (1995).
[127] T. Gloe, M. Kirchner, A. Winkler, R. Böhme, Can we trust digital image forensics?, Proceedings of the 15th ACM international conference on Multimedia, 2007. pp. 78–86.