Forensic Science International: Genetics 43 (2019) 102150

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach

Corina C.G. Benschop[a,1,*], Jennifer van der Linden[a,1], Jerry Hoogenboom[a], Rolf Ypma[b], Hinda Haned[c,d]

[a] Netherlands Forensic Institute, Division of Biological Traces, Laan van Ypenburg 6, 2497GB, The Hague, the Netherlands
[b] Netherlands Forensic Institute, Division of Digital and Biometric Traces, Laan van Ypenburg 6, 2497GB, The Hague, the Netherlands
[c] University of Amsterdam, the Netherlands
[d] Ahold Delhaize, the Netherlands

ARTICLE INFO

Keywords:
DNA profile interpretation
DNA mixtures
Number of contributors
Machine learning

ABSTRACT

The number of contributors (NOC) to (complex) autosomal STR profiles cannot be determined with absolute certainty due to complicating factors such as allele sharing and allelic drop-out. The precision of NOC estimations can be improved by increasing the number of (highly polymorphic) markers, the use of massively parallel sequencing instead of capillary electrophoresis, and/or using more profile information than only the allele counts.

In this study, we focussed on machine learning approaches in order to make maximum use of the profile information. To this end, a set of 590 PowerPlex® Fusion 6C profiles with one up to five contributors were generated from a total of 1174 different donors. This set varied for the template amount of DNA, mixture proportion, levels of allele sharing, allelic drop-out and degradation. The dataset contained labels with known NOC and was split into a training, test and hold-out set. The training set was used to optimize ten different algorithms with selection of profile characteristics. Per profile, over 250 characteristics, denoted 'features', were calculated. These features were based on allele counts, peak heights and allele frequencies. The features that were most related to the NOC were selected based on partial correlation using the training set. Next, the performance of each model (=combination of features plus algorithm) was examined using the test set. A random forest classifier with 19 features, denoted the 'RFC19-model' showed best performance and was selected for further validation. Results showed improved accuracy compared to the conventional maximum allele count approach and an in-house nC-tool based on the total allele count. The method is extremely fast and regarded useful for application in forensic casework.

1. Introduction

Forensic DNA profiling and DNA profile interpretation have evolved tremendously over the years. Whilst previously only single source or unambiguous two-person DNA mixtures were regarded interpretable and suitable for weight of evidence evaluations [1,2], nowadays more and more complex DNA profiles can be analysed as an effect of the increased sensitivity, larger number of marker systems and advanced probabilistic genotyping software [3]. Despite these developments, uncertainty regarding the composition of DNA mixtures still exists. Probabilistic genotyping systems account for much of these uncertainties. In particular, uncertainty on the number of contributors (NOC) is taken care of through conditioning on the minimum NOC that maximises the probability of the prosecution (Hp) and the defence (Hd) and/or considering multiple sets of propositions regarding a different NOC [4]. One can report the minimum weight of evidence that was obtained [5] or report the LR results from all sets of propositions [6] which places the decision regarding which weight of evidence to rely on with the court. However, as stated by Taylor et al. [7], when it is not possible for a DNA expert to define the NOC it may be a lot to ask the court to do so. Approaches have been suggested that treat the NOC as a nuisance parameter that is integrated out [7,8]. However, most of the current probabilistic genotyping systems require that the user specifies the NOC [6]. In this study, we focus on these latter systems for which it

* Corresponding author.
E-mail addresses: c.benschop@nfi.nl (C.C.G. Benschop), j.v.d.linden@nfi.nl (J. van der Linden), j.hoogenboom@nfi.nl (J. Hoogenboom), r.ypma@nfi.nl (R. Ypma), h.haned@uva.nl (H. Haned).
[1] Both authors contributed equally.

https://doi.org/10.1016/j.fsigen.2019.102150
Received 25 July 2019; Received in revised form 15 August 2019; Accepted 19 August 2019
Available online 23 August 2019
1872-4973/

is important to have an estimate of the NOC that is as accurate as possible. This is particularly relevant if the difference between two contributor numbers means the difference between excluding a person as being a possible contributor, and producing a statistic that favours its inclusion.

The most common approach to the inference of the minimum NOC is the maximum allele count (MAC) method, which uses the locus exhibiting the largest number of alleles that is divided by two and rounded up to the nearest whole number [9–11]. This method is easy to use, though mischaracterizations are expected with high order mixtures (three or more contributors) [12,13]. Under- or over-estimating the NOC can affect the weight of evidence [14], particularly with qualitative models [15–17], but also when using a quantitative model [18–21]. Increasing the number of loci, using loci with a higher discriminatory power or using massively parallel sequencing data of short tandem repeat loci showed less misinterpretations of the NOC when using the MAC method [22–25]. Alternative methods using the total number of alleles (total allele count, TAC), the distribution of allele counts over the loci, the population's genotype frequencies, peak heights (PH), replicates, probability of allelic drop-out and stutter, or a Bayesian network approach have shown to yield improved NOC estimates [16,26–35]. Despite these efforts, estimating the NOC can be extremely complex with high order mixtures, high allele sharing, degraded and/or low-template DNA. Absolute certainty on the NOC is therefore regarded not possible. However, the fewer mischaracterizations the better. To that end, it is desired to make optimal use of the available profile information which can be achieved by using a machine learning approach [36]. An additional benefit of using a machine learning approach is that the estimation of the NOC can be performed in seconds, which is of importance in forensic casework requiring rapid analyses. Machine learning can be explained as a systematic study of algorithms and systems that improve their knowledge or performance with experience [37]. In machine learning, predicting the NOC can be regarded a classification problem with classes being the numbers of contributors. Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations with known categories. In model development, first, DNA profile characteristics, or features, that can be informative for the NOC are selected and engineered for a set of DNA profiles with variation as can be expected in forensic casework. Next, machine learning algorithms, or classifiers, are selected. The combination of features and an algorithm results in a model. By providing labelled training data (*i.e.* per DNA profile its features and known NOC) to a model, the algorithm acquires experience regarding the features that best fit the single donor profiles, the two-person mixtures, etcetera. The performance of the trained model is subsequently tested using another dataset, whose labels are not provided to the model. To our knowledge, one machine learning model for the estimation of the number of contributors in autosomal DNA profiles was published, *i.e.* the PACE software created by Marciano et al. [36]. This software relies on a Support Vector Machine-derived classification model and showed improved performance when compared to the MAC method [36]. The PACE software was trained on single donor up to four-person Identifiler® profiles and is not freely available. In this study, we used single donor up to five-person PowerPlex® Fusion 6C profiles, regarded 278 different profile characteristics, examined ten different algorithms, trained and tested various machine learning models, and selected the best performing one for further examination. This NOC machine learning model is based on a random forest classifier with 19 features and denoted the 'RFC19'. All Python code used is accessible *via*https://www.forensicinstitute.nl/research-and-innovation/european-projects/DNAxs. The RFC19 model can be examined using a laboratory's own data and will be made available through implementation in our DNA eXpert System, DNAxs [38].

Table 1
Overview of the numbers of DNA profiles per NOC in the train, test and hold-out dataset.

Number of contributors	Train	Test	Hold-out	Total
1	99	33	33	165
2	56	20	20	96
3	74	25	25	124
4	78	27	27	132
5	43	15	15	73
Total	350	120	120	590
Number of different donors	695	273	273	1174

2. Materials and methods

2.1. DNA profiles

A set of 590 single donor up to five-person (1p-5p) PowerPlex® Fusion 6C (PPF6C) profiles were generated from a total of 1174 different genotypes. These PPF6C profiles were used to train, test and validate the NOC machine learning model (Table 1). The DNA-profiles included:

A 165 single donor degraded or non-degraded profiles (48/165 had peak heights that decreased with 10 up to 62 RFU per base pair) generated from buccal swabs.
B 120 2p-5p mixed profiles selected from a previous study [21]. These profiles varied for the amount of DNA, level of drop-out, mixture proportion and level of allele sharing among the contributors [21].
C 15 mixed profiles that were artificially degraded following:

$$Degraded\ Peak\ Height = Initial\ Peak\ Height \times \beta^{\frac{fragment\ lengt\ of\ the\ allele-125}{100}}$$

(formula from [39], in which β is the degradation slope parameter per profile such that the peak height of the longest fragment is reduced to 40 RFU).

A 290 degraded profiles created by mixing DNA extracts that were randomly chosen from a set of DNA extracts from 2085 donors [40]. For these 290 profiles, degradation was created as DNA extracts were stored under non-optimal conditions prior to DNA profiling. The degraded DNA-profiles showed descending peak height patterns at the longer fragment lengths (0.08–109 RFU decrease in PH per base pair) and had up to six complete locus drop-outs.

These PPF6C DNA-profiles were chosen to cover the variation that can be encountered in forensic casework. The DNA-profiles were examined by an experienced reporting officer performing forensic casework on a daily basis to verify that this set is regarded representative for real casework.

In addition, a set of 35 extremely complex PPF6C profiles was selected to examine the limitations of the NOC machine learning model. These concerned six-person (6p) mixtures ($n = 16$), DNA mixtures generated from related individuals (brothers) with drop-out ($n = 10$) and extremely degraded mixed DNA profiles with a minimum of three locus drop-outs ($n = 9$) (Supplementary Table 1). Such DNA-profiles might be considered as not suitable for comparison or weight of evidence evaluation.

Replicates were not included in the train, test or hold-out dataset, but were used as a separate set to examine the effect of replicates. These concerned 114 2p-5p mixtures with three replicates each, resulting in a total of 342 DNA-profiles (Supplementary Table 2, replicate profiles from [21]).

All DNA profiles were analysed as described in [21], except for higher GeneMarker™ HID dye specific detection thresholds, namely

FL = 95, JOE = 140, TMR = 85, CXR = 135, TOM = 95 RFU, and a heterozygous peak imbalance percentage of 3%. Appendix 1 presents a representative selection of the electropherograms (EPGs).

2.2. Feature engineering and selection

A set of 11 locus features and 25 sample features were defined (Supplementary Tables 3 and 4). These features relate to aspects such as, the number of alleles, the peak heights, the length of the locus, or the allele frequencies. Features that make use of the allele frequencies were computed using a Dutch population database [40]. The features were engineered for each of the DNA profiles using the Python programming language (version 2.7.6) and were normalized using StandardScaler, from Scikit-Learn (version 0.19.1), by "removing the mean and scaling to unit variance" [41]. Amelogenin and the Y-chromosomal markers were excluded, resulting in a total of 278 features per profile (25 samples features + (11 locus features * 23 autosomal loci)).

The procedure of feature selection and ranking was as follows:

1) The features MAC and TAC were fixed as features 1 and 2 as these are informative on the NOC and are both used in our laboratory for estimating the NOC.
2) Iteratively, features were selected that had the highest partial correlation with the NOC, controlling for the effect of previously selected features. This partial correlation is a measure of how much two variables are correlated while removing the effect of other variables.
3) The procedure was repeated until 50 features were selected.

These 50 features were kept for further examination, in the order in which they were selected.

2.3. Machine learning models

A diversity of algorithms, *i.e.* classifiers, was tested, as there is not one algorithm that is specific for our classification problem. We used the Scikit-Learn (version 0.19.1) implementation in Python (version 2.7.6), a de facto standard in the field [41]. In total ten different supervised learning algorithms were tested (Table 2).

2.4. Model selection and validation

The procedure of model selection was as follows:

1) Each of the ten algorithms was trained and tested 50 times, where for *n*-th training features 1, 1 + 2, 1 + 2+3, *etc.*resulting from the partial correlation were used
2) Hyperparameter optimization, or tuning, was performed using 'GridSearch'. This procedure goes through all possible combination of the parameters of a given model and selects the combination that leads to the highest accuracy (number of correct predictions divided by total number of prediction) of the classifier on the train set. The parameters of the GridSearch are provided in Appendix 2. For each algorithm, the number of features with highest accuracy was stored and used in subsequent steps. In case of equal accuracies, models with small difference to the accuracy for the training set were selected, as a large difference is indicative of overfitting.
3) Per NOC and for each of the machine learning models we tested in this study, two performance metrics were computed: precision and recall defined as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Where the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) were recorded. For example, a 2p mixture classified as 2p mixture was defined as a TP. A 2p mixture classified as any other NOC as a FN. A predicted NOC of 2p while it is not a true 2p mixture was regarded a FP. Lastly, any non-2p mixture classified as a non-2p mixture was regarded a TN.

Precision results the percentage, or proportion, of correctly predicted NOC (the positive predictive value) and recall indicates the proportion of actual positives that was identified correctly.

1) The models resulting in the highest accuracy for the test set were selected for detailed examination with respect to:
2) the probabilities assigned by the classifier to each class of NOC; to assess whether these can be informative to the end user.
3) the samples with an incorrect estimate for the NOC; these samples (EPGs) were manually inspected to assess whether these mischaracterizations were expected.
4) The overall best performing model was selected for validation, *e.g.* testing on the hold-out set, and to define its applicable domain as described in the following section.

2.5. Validation and defining the applicable domain of the selected machine learning mode

The overall best performing model was further validated using the hold-out dataset (Table 1). In addition, its limitations were assessed by using the 'extremes' dataset (Supplementary). DNA profiling data from replicates were used to examine the performance when the outcomes of replicates are combined. With replicates, the NOC class that was obtained for at least two of three replicates was compared against the NOC prediction obtained for one replicate. When three replicates yielded three times a different NOC, the intermediate value was used.

2.6. Maximum allele count (MAC) and nC-tool

The outcomes of the NOC machine learning model were compared to the outcomes of the MAC method and nC-tool [51]. The MAC was calculated by taking the locus with the largest number of alleles, dividing this by two and rounding up to the nearest whole number. The nC-tool is an in-house excel spread sheet that uses the (5% corrected) TAC of the particular DNA profile and compares this to the TAC obtained using simulation data. The output is a probability per number of contributors (one up to five) per probability of allelic drop-out [51]. In this study the known allelic drop-out per sample was used and the NOC that yielded the largest probability was selected. Comparisons were performed for the same 120 samples as presented in [51] (which is a selection of 2p-5p PPF6C profiles from [21]), except that we used GeneMarker™ HID settings as described in section 2.1.

Table 2
Overview of the ten machine learning algorithms as used in this study.

	Model	Acronym	Reference
1.	Decision Tree Classifier	DTC	[42]
2.	Gaussian Naive Bayes	GaussianNB	[43]
3.	Gradient Boosting Classifier	GBC	[44]
4.	k-Nearest Neighbors Classifier	k-NN	[45]
5.	Linear Discriminant Analysis	LDA	[46]
6.	Linear Support Vector Classification	LSVC	[47]
7.	Logistic Regression Classifier	LRC	[48]
8.	Multi-layer Perception Classifier	MLPC	[49]
9.	Random Forest Classifier	RFC	[50]
10.	Support Vector Classification	SVC	[47]

3. Results and discussion

3.1. Initial features selection

For each of the DNA profiles, 278 features were engineered. A large number of features may result in over fitting, *i.e.* very good predictions for the training set but poor results when using the test set, while a small number may ignore vital information with predictive value [36]. To select those features that are most informative of the NOC, partial correlation calculations were performed. The features MAC and TAC were fixed in partial correlation as these are used in our laboratory and can be informative on the NOC [51]. For example, MAC and TAC had the highest correlation with NOC of all features (0.92 and 0.91). Supplementary Table 5 shows the top 50 features that included information regarding the number of alleles, allele frequencies and peak heights and were spread across the loci and dye channels. This top 50 did not include features regarding *e.g.* degradation slope and/or features counting loci with seven or more alleles. Apparently, these features did not add much to the information already obtained from the other features.

3.2. Accuracy of the machine learning models

The top 50 features that were obtained by partial correlation were used in the training and testing phase of each of the ten algorithms (see section 2.3). Supplementary Fig. 1 shows their accuracy as a function of the number of features. The number of features that resulted in the highest accuracy for the test set was selected and is presented in Table 3 (details on GridSearch parameters are presented in Appendix 2). Accuracy ranged between 0.792 and 0.833 which is lower than presented by Marciano et al., which ranged from 0.894 to 0.962 [36]. The lower accuracy in our study can be expected as our dataset included more complex profiles and more contributors (up to five instead of up to four contributors). However, our models are not strictly comparable as our dataset not only included more complex profiles, it also differed regarding the STR typing kit that was used and the features that were engineered.

For all models but one (LSVC4), all incorrect predictions on the test set were one lower or one higher than the true NOC (Supplementary Fig. 2). Two models showed best train and test accuracy, *i.e.* Random Forest Classifier with 19 features and Linear Discriminant Analysis with 40 features, denoted as RFC19 and LDA40, respectively. Both models yielded 83.3% correctly predicted NOC for the test set and comparable accuracy for the training set (Table 3 and Supplementary Fig. 1).

Besides accuracy, precision and recall are measures of relevance and can be used to compare the performance of the various models. When generating precision-recall plots and comparing the ten models it becomes clear that not one of the models outperformed all others as precision and recall differed per NOC (Supplementary Fig. 3). It was expected that precision and recall decreases with an increasing NOC except for the precision of 5p mixtures as these cannot be over-estimated using the machine learning models in this study. This trend was observed for most of the models, including LDA40 and RFC19 (Supplementary Fig. 3).

RFC19 and LDA40 were selected to further assess their performance.

Table 3
Train and test accuracy obtained per model (algorithm plus features) sorted from overall best to least performing model. For each algorithm, only the best performing number of features is shown.

Model	Number of features	Accuracy training set	Accuracy test set
LDA	40	0.888	0.833
RFC	19	0.874	0.833
LRC	8	0.823	0.825
GBC	11	0.886	0.817
SVC	5	0.811	0.808
LSVC	4	0.803	0.808
k-NN	3	0.834	0.800
DTC	14	0.843	0.800
MLPC	8	0.842	0.798
GaussianNB	4	0.794	0.792

3.3. Model selection

When comparing RFC and LDA, LDA is the least complex algorithm. However, this algorithm required more than double the number of features when compared to RFC (40 *vs.* 19, respectively) to obtain the same accuracy. The results of both models were examined in more detail to enable choosing between the two models. Both models have the same mischaracterization rate for the test set (20/120) and incorrect predictions in this set were always one contributor number lower or higher than the true NOC. Only minor differences were observed when examining the predictions per NOC (Fig. 1). Both models incorrectly classified 20 samples. Out of these, 13 were incorrectly classified with both models. The models predicted the same NOC for these 13 samples.

Both RFC and LDA present a probability for their predicted NOC. These probabilities are most useful if they are high for correct predictions and low for incorrect predictions. With RFC19 highest probabilities were obtained for single source profiles and probabilities tended to be lower with a higher NOC (Fig. 2A), which can be expected as higher order mixtures are more complex. With LDA40, the probabilities for 2p mixtures were often larger than those for single donor profiles. Furthermore, the predictions for 4p and 5p mixtures showed almost equal probabilities for being a 3p, 4p or 5p mixture (Fig. 2B). Only six out of the 42 4p/5p mixtures received a large probability of, for example, > 0.5 (Supplementary Fig. 4B). However, these six all resulted in an incorrect classification. For 18/20 and 17/20 incorrect predictions, the true NOC received the second largest probability when using RFC19 and LDA40, respectively.

Overall, RFC19 and LDA40 had comparable performances. There was a slight preference for RFC19, since this model required less features and the probabilities seemed more useful. Therefore, RFC19 was selected for further validation.

3.4. RFC with 19 features

RFC19 showed good overall performance regarding the test set (Table 3) and was selected for further evaluation and validation. Table 4 shows the 19 features used in this model which include information regarding allele counts, peak heights and allele frequencies. In total, eight out of the 25 sample features and five out of the 11 different locus features were included. The sample features regarding degradation slope were not included. Locus features in this model included the four types of data as listed in the candidate features (Supplementary Table 5); *i.e.* allele counts, minimum NOC, peak heights and allele frequencies. The locus features are located at various loci of various fragment length and in four of the five dye channels (Fig. 3). Interestingly, locus SE33 with high discriminatory power was not included, though various loci with lower discriminatory power (*e.g.* TPOX, Penta E, Penta D and TH01) were included. Solely, the loci with low discriminatory power were not very informative on the NOC: sorted on correlation to the NOC, they were listed position 188 or lower. The high ranking using the partial correlation approach shows the information that low discriminatory power loci have on the NOC is somehow independent of the information held in other features.

3.5. Validation of the RFC19 model using the hold-out dataset

The accuracy of the selected RFC19 model was further examined using the hold-out dataset (see Table 1). The percentage of correctly classified NOC for this dataset was similar to that of the test set (83.3%

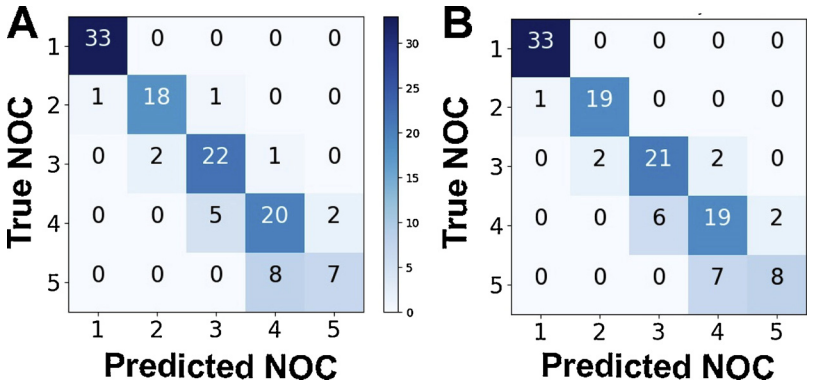

Fig. 1. Predicted *versus*true NOC for the test set when using RFC19 (A) or LDA40 (B). Accuracy, defined as the sum of the values on the diagonal divided by the total sum, is equal for both methods.

*versus*82.5% correctly classified, respectively), which gives confidence in the RFC19 model. Mischaracterizations for samples in the test or hold-out dataset (41/240) were one contributor number lower or one higher than the true NOC. Mischaracterizations occurred most often with the high order mixtures, in particular with the 5p mixtures that were predicted as 4p mixtures (Fig. 4). To further understand the behaviour of the RFC19 model, we performed detailed analyses of the 41 misclassified samples from the test (n = 20) and hold-out dataset (n = 21). For 18/41 (44%) samples the probability for the predicted NOC was smaller than 0.6 and for 13/41 (31.7%) the difference in

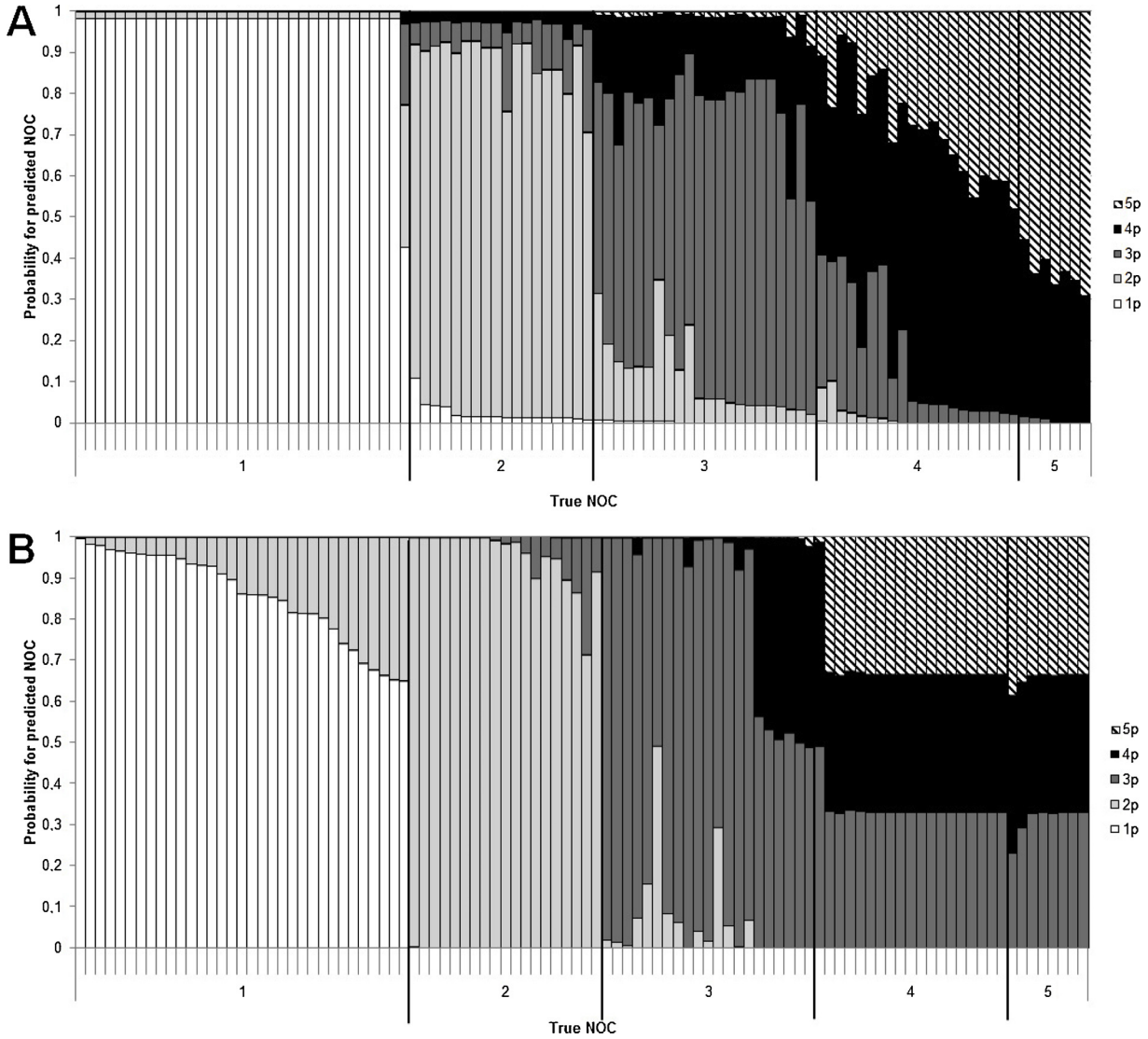

Fig. 2. Probabilities per NOC for DNA profiles in the test set that were correctly classified when using RFC19 (A) or LDA40 (B). The predicted NOC is presented as white, light grey, dark grey, black or white with black diagonal stripes for 1p, 2p, 3p, 4p or 5p, respectively.

Table 4
Overview of the 19 features as used in the selected RFC19 model.

Nr.	Feature	Sample/ locus feature	Details
1	MAC	Sample	Maximum allele count (MAC); Maximum number of alleles observed on a locus
2	TAC	Sample	Total allele count (TAC); Total number of alleles per profile
3	Standard deviation Allele Count	Sample	Standard deviation of the number of alleles per locus
4	Allele Count_D3S1358	Locus	Number of alleles at D3S1358
5	AC 5-6	Sample	Number of loci with an allele count of 5 or 6
6	Minimum NOC_Penta E	Locus	Allele count / 2, rounded up to 0 decimals at Penta E
7	Minimum NOC_Penta D	Locus	Allele count / 2, rounded up to 0 decimals at Penta D
8	AC 0	Sample	Number of loci with 0 alleles (locus drop-out)
9	Standard deviation PH_vWA	Locus	Standard deviation of the peak heights at locus vWA.
10	Match Probability	Sample	The probability of a random, unrelated person matching to this DNA profile. The probability is calculated using a Dutch frequency database [40].
11	Number of peaks below ST	Sample	Number of peaks with a peak height below the stochastic threshold of 800 RFU
12	Minimum NOC_TPOX	Locus	Allele count / 2, rounded up to 0 decimals at TPOX
13	Minimum NOC	Sample	Allele count of the locus with the largest number of alleles / 2, rounded up to 0 decimals
14	Minimum NOC_CSF1PO	Locus	Allele count / 2, rounded up to 0 decimals at CSF1PO
15	Minimum NOC_D16S539	Locus	Allele count / 2, rounded up to 0 decimals at D16S539
16	Sum AF_TH01	Locus	The sum of the allele frequencies of alleles at TH01
17	Allele Count_TPOX	Locus	Number of alleles at TPOX
18	Percentage of AF_D1S1656	Locus	For locus D1S1656, the percentage of alleles that are within the population database
19	Allele Count_D8S1179	Locus	Number of alleles at D8S1179

probability compared to the next highest probability was less than 0.2. Such small probabilities and differences compared to the next highest probability indicate uncertainty about these predictions and may give the end user a trigger that the prediction could be incorrect. Such results were, however, also obtained with a correctly predicted NOC, though to a lesser extent: 24% (48/199) had a class probability < 0.6 and 17% (33/199) yielded a probability that had a difference that was less than 0.2 compared to the next highest probability. Seventy-eight percent (32/41) of the incorrect predictions were obtained for samples that originated from the degraded mixed profiles set. Note that 52% of the complete dataset of 590 profiles (and 72% of the mixed profiles) contained profiles showing a degradation pattern (see section 2.1). These were included as forensic casework samples often exhibit degradation to some extent. Nine out of the 41 profiles resulted in an incorrect classification that was not expected based on manual inspection of the EPGs. These concerned samples in each of the NOC categories which yielded an under-estimated NOC. For each of these nine samples the true NOC received the second highest probability.

3.6. RFC19 model performance on the extremes dataset

Three types of extreme samples (Supplementary Table 1) were used to examine the limitations of the RFC19 model. The first type were 6p mixtures. As the NOC machine learning model was trained using 1p-5p mixtures, a NOC of six is not a possible outcome. These can thus only be under-estimated using the model and would at best be classified as 5p mixtures. The second type of extreme samples included low-template 2p, 3p and 4p mixtures generated from DNA of two or three brothers. As the model is not trained using samples with relatives, under-estimates were expected when using the NOC machine learning model. The third type of extreme samples included 3p, 4p and 5p mixtures that were severely degraded and showed seven up to 22 locus drop-outs per DNA-profile. These extremely complex samples yielded, as expected, mostly incorrect predictions using the NOC machine learning model (Table 5). On most (88%) of the 6p mixtures the RFC19 model predicted five contributors. One low-template 2p mixture with DNA from brothers yielded an over-estimated NOC (3p). Based on manual examination of the EPG, this profile could be well explained by two contributors; the profile had three loci with three or four alleles, 17 loci with one or two alleles and three complete locus drop-outs. The over-estimated NOC was thus unexpected for this sample.

Overall, the training, test and hold-out dataset in this study included complex low-template, degraded and low- and high-allele sharing DNA-profiles for which high accuracies were obtained (Fig. 4). However, the RFC19 machine learning model was not trained on DNA-profiles such as in the extremes dataset and thus is likely to yield mischaracterizations for such profiles.

3.7. Use of replicates

In some laboratories it is common practice to perform replicate analyses, *i.e.* generate multiple DNA-profiles using the same DNA extract. In our laboratory it is common to initially generate one replicate and in some cases three replicates. Replicate analysis is mainly

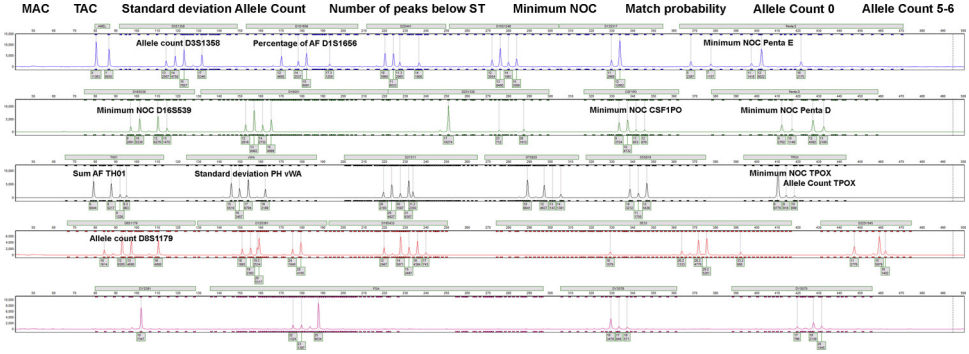

Fig. 3. Overview of the 19 features used in the RFC19 machine learning model shown in an exemplar 3p PPF6C profile. Sample features are indicated above the electropherogram (EPG) and locus features are shown at the particular loci within the EPG.

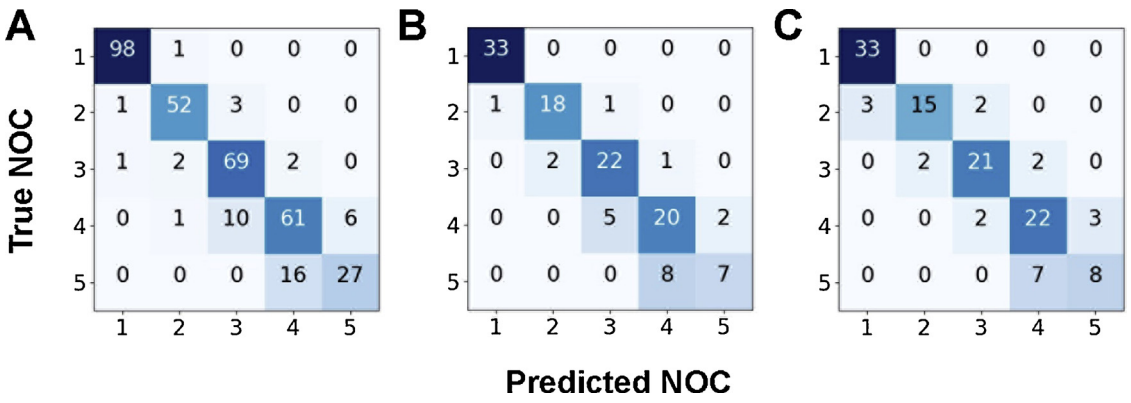

Fig. 4. Confusion matrix of the RFC19 machine learning model on the A) train, B) test and C) hold-out dataset.

performed in case more information is required from low-template contributors. The RFC19 machine learning model was trained on individual replicates and presented a prediction per replicate. However, multiple replicates can be used jointly in likelihood ratio calculations for which one NOC has to be defined for these replicates. In case the training dataset includes replicates, then models can be trained with features of the combined data and present one NOC per set of replicates. However, in our laboratory, replicates are not generated in all of the cases and such a model would narrow down its applicable domain. To that end, we selected the NOC that was predicted for the majority of replicates. There were no samples with a different NOC for each of the three replicates. The precision of the RFC19 model was higher for the replicates jointly than for the individual replicates (Table 6), indicating that this is a feasible approach for interpreting the outcomes of the RFC19 machine learning model when multiple replicates are available.

3.8. Comparison to alternative methods

The performance of the RFC19 NOC machine learning model was compared against the MAC approach and the nC-tool. The MAC gives an estimate for the minimum NOC, which we used as a set number for comparison. The nC-tool uses the TAC of the DNA-profile as input and outputs a probability per NOC for four different categories for allelic drop-out. In forensic casework, this drop-out category must be decided by the user. In this study, we used the true drop-out rate and noted the accompanying NOC for comparison. Table 7 shows that the higher the true NOC, the lower the accuracy of the nC-tool and that of the machine learning model. The MAC approach, however, showed a higher accuracy for the 3p mixtures than for the 2p mixtures which was the effect of elevated stutter peaks residing in these PPF6C profiles. In 2p mixtures these resulted in an over-estimated NOC (Supplementary Table 6). The nC-tool and machine learning model are less hindered by elevated stutter peaks and/or allelic drop-in as the overall profile is considered in the nC-tool and a variety of different features is regarded by the RFC19 machine learning model. The MAC approach yielded an equal percentage of correct classifications for 3p mixtures when compared to the RFC19 machine learning model (Table 7), though the RFC19 model yielded more correct classifications for the 2p, 4p and 5p mixtures.

Table 6
Percentages of correctly predicted NOC with the RFC19 model when using an individual replicate or the results of three replicates jointly.

	Individual replicates		Three replicates jointly	
True NOC	Number correct	Accuracy (%)	Number correct	Accuracy (%)
2	84 / 90	93.3%	29 / 30	96.7%
3	76 / 84	90.5%	27 / 28	96.4%
4	73 / 87	83.9%	26 / 29	89.7%
5	49 / 81	60.5%	17 / 27	63.0%
Total	282 / 342	82.5%	99 / 114	86.8%

Table 7
Percentage of correct classifications for 2p-5p PPF6C profiles when using the MAC approach, nC-tool or RFC19 machine learning model.

True NOC	*n*	Percentage of correct predictions		
		MAC	nC-Tool	RFC19 machine learning model
2	30	66.7%	100%	100%
3	30	96.7%	83.3%	96.7%
4	30	76.7%	70.0%	83.3%
5	30	36.7%	53.3%	60.0%
Total	120	69.2%	76.7%	85.0%

Overall, the machine learning model outperformed both MAC and the nC-tool: 85% correct classifications for the NOC machine learning model *versus*69.2% and 76.7% correct for the MAC and nC-tool, respectively (Table 7).

3.9. Future work

The RFC19 machine learning model showed high precision and runs within a second. We therefore regard this a useful tool for forensic scientists performing DNA casework analyses. Hence, this RFC19 model will be implemented in the DNA eXpert System, DNAxs [38].

The applicable domain of the model is 1p-5p PPF6C profiles

Table 5
Outcome of the RFC19 machine learning model for the extremely complex DNA-profiles. Grey cells indicate that the predicted NOC is equal to the true NOC.

Type of extreme samples	*n*	True NOC	Predicted NOC 1	2	3	4	5
Six-person mixtures	16	6	-	-	-	2	14
DNA mixtures from brothers with allelic drop-out	5	2	2	2	1	-	-
	3	3	-	3	-	-	-
	2	4	-	2	-	-	-
Extremely degraded mixed DNA profiles	3	3	3	-	-	-	-
	3	4	1	-	2	-	-
	3	5	1	1	1	-	-

generated using the PCR, CE and analyses settings as used in this study. This study can be used as an example to develop a machine learning model for DNA profiling data obtained using other STR typing kits and our code is made available through GitHub and *via* https://www.forensicinstitute.nl/research-and-innovation/european-projects/dnaxs.

The PPF6C profiling data includes 27 markers, although, in this study we used the 23 autosomal markers only. The amelogenine and Y-chromosomal markers were excluded from the analyses as our dataset consisted of male DNA profiles only. Extending the dataset with DNA profiles of females and male/female mixtures could further decrease the mischaracterisation rates and would enable expressing not only the total NOC but also the number of male/female contributors. Another way to increase the number of correct classifications could rely on using new features, features that combine various profile characteristics and/or using massively parallel sequencing data instead of data from capillary electrophoresis.

The RFC19 model was trained on individual replicates and improved accuracies were obtained when combining data from multiple replicates. In case training data includes replicates, then models can be trained with features of the combined data which may further improve accuracies for data from replicates.

As is common in machine learning approaches, we treated prediction of the NOC as a classification problem. However, as the target classes are in fact integers, a regression approach is also possible. This would have the benefit that the relationship between the classes would be clear to the algorithms. In other words, in a regression approach misclassifying a 4p mixture as a 2p mixture would yield a stronger penalty then misclassifying a 4p as a 3p mixture. In our classification approach, both receive the same penalty (although the former is rare, Fig. 4). Furthermore, a regression approach would yield real numbers rather than integers, where the distance to the integers could be interpreted as a measure of uncertainty, similarly to how we currently use the computed probabilities. Future work may study possible improvements in performance when switching from classification to regression.

Lastly, the RFC19 machine learning model presents a prediction for the NOC and a probability for this NOC. The higher the probability, the more certain the model is about the predicted NOC which can be useful to the end user. However, to the end user it is unknown which features made the model decide on the NOC that was presented which makes it a black box model. For example, for 9/41 misclassified samples, it was unsure why the error was made. Various methods exist that enable opening black boxes [52,53] by presenting the feature importance, thus the features that positively or negatively contributed to a given prediction [54]. This could help the end-user in understanding the outcome of the model, it would also help enhancing user trust in adopting a machine learning model for NOC estimation. Future research will focus on these methods to further improve our model and our understanding of it.

4. Conclusion

This study describes the development and validation of the RFC19 machine learning model that enables fast, automated and accurate classification of the NOC in 1p-5p autosomal PPF6C profiles. This model included various profile characteristics, such as allele counts, allele frequencies and peak heights. We demonstrate that feature engineering coupled with extensive model selection can produce high accuracy for classifying the NOC even for highly complex samples. Furthermore, our RFC19 machine learning model outperformed the MAC approach and nC-tool.

Acknowledgements

This study was partly funded by the European Union's Internal Security Fund — Police (Proposal Number: 820838, Proposal Acronym: DNAxs2.0).

We are thankful to Titia Sijen for useful discussions and suggestions, to Anouk Backx for creation of a portion of the DNA mixtures, to Margreet van den Berge for technical support in the laboratory and to Francisca Duijs for help with analysis of the DNA-profiles.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.fsigen.2019.102150.

References

[1] P.M. Schneider, R. Fimmers, W. Keil, G. Molsberger, D. Patzelt, W. Pflug, T. Rothamel, H. Schmitter, H. Schneider, B. Brinkmann, The German Stain Commission: recommendations for the Interpretation of mixed stains, Int. J. Legal Med. 123 (2009) 1–5.
[2] A.J. Meulenbroek, T. Sijen, C.C.G. Benschop, A.D. Kloosterman, A practical model to explain results of comparative DNA testing in court, Forensic Sci. Int. Genet. Supplement Series 3 (2011) e325–e326.
[3] P. Gill, H. Haned, Ø. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development, Forensic Sci. Int. Genet. 18 (2015) 100–117.
[4] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Sci. Int. 160 (2–3) (2006) 90–101.
[5] C.H. Brenner, R. Fimmers, M.P. Baur, Likelihood ratios for mixed stains when the number of donors cannot be agreed, Int. J. Legal Med. 109 (1996) 218–219.
[6] M.D. Coble, J.A. Bright, Probabilistic genotyping software: an overview, Forensic Sci. Int. Genet. 38 (2019) 219–224.
[7] 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.
[8] K. Slooten, A. Caliebe, Contributors are a nuisance (parameter) for DNA mixture evidence, Forensic Sci. Int. Genet. 37 (2018) 116–125.
[9] 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 (1998) 55–70.
[10] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Low-level DNA and Complex Mixtures, Academic Press, 2014 Chapter 7.
[11] SWGDAM interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories, Available from: http://www.fbi.gov/about-us/lab/codis/swgdam-interpretation-guidelines.
[12] D.R. Paoletti, T.E. Doom, C.M. Krane, D.E. Raymer, M.L. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (2005) 1361–1366.
[13] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci. Int. Genet. 1 (2007) 20–28.
[14] B.S. Weir, C.M. Triggs, L. Starling, L.I. Stowell, K.A.J. Walsh, J. Buckleton, Interpreting DNA mixtures, J. Forensic Sci. 42 (1997) 213–222.
[15] C.C. Benschop, H. Haned, T.J. de Blaeij, A.J. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: a descriptive study, Forensic Sci, Int. Genet. 6 (2012) 697–707.
[16] C.C.G. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci. Int. Genet. 19 (2015) 92–99.
[17] H. Haned, C.C.G. Benschop, P.D. Gill, T. Sijen, Complex DNA mixture analysis in a forensic context: evaluating the probative value using a likelihood ratio model, Forensic Sci, Int. Genet. 16 (2015) 17–25.
[18] J.A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci. Int. Genet. 12 (2014) 208–214.
[19] J.S. Buckleton, J.A. Bright, K. Cheng, Kelly H, D.A. Taylor, The effect of varying the number of contributors in the prosecution and alternate propositions, Forensic Sci. Int. Genet. 38 (2019) 225–231.
[20] T. Bille, S. Weitz, J.S. Buckleton, J.A. Bright, Interpreting a major component from a mixed DNA profile with an unknown number of minor contributors, Forensic Sci. Int. Genet. 40 (2019) 150–159.
[21] C.C.G. Benschop, A. Nijveld, F.E. Duijs, T. Sijen, An assessment of the performance of the probabilistic genotyping software EuroForMix: trends in likelihood ratios and analysis of Type I & II errors, Forensic Sci. Int. Genet. 42 (2019) 31–38.
[22] M.D. Coble, J.-A. Bright, J.S. Buckleton, J.M. Curran, Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci. Int. Genet. 19 (2015) 207–211.
[23] J. Curran, J. Buckleton, Uncertainty in the number of contributors for the European standard set of loci, Forensic Sci. Int. Genet. 11 (2014) 205–206.
[24] G.M. Dembinski, C. Sobieralski, C.J. Picard. Estimation of the number of contributors of theoretical mixture profiles based on allele counting: Does increasing the number of loci increase success rate of estimates? Forensic Sci. Int. Genet. 33 (2018) 24–32.
[25] B.A. Young, K. Butler Gettings, B. McCord, P.M. Vallone, Estimating number of contributors in massively parallel sequencing data of STR loci, Forensic Sci. Int.

Genet. 38 (2019) 15–22.
[26] H. Haned, L. Pene, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci. Int. Genet. 5 (2011) 281–284.
[27] 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 (2011) 23–28.
[28] A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Sci. Int. Genet. 6 (2012) 689–696.
[29] T. Tvedebrink, On the exact distribution of the numbers of alleles in DNA mixtures, Int. J. Legal Med. 128 (2014) 427–437.
[30] C.C.G. Benschop, H. Haned, T. Sijen, Consensus and pool profiles to assist in the analysis and interpretation of complex low template DNA mixtures, Int. J. Legal Med. 127 (2013) 11–23.
[31] D.R. Paoletti, D.E. Krane, T.E. Doom, M. Raymer, Inferring the number of contributors to mixed DNA profiles, IEEEACM Trans. Comput. Biol. Bioinform. 9 (2012) 113–122 (January–February (1)).
[32] J. Perez, A.A. Mitchell, N. Ducasse, J. Tamariz, T. Caragine, Estimating the number of contributors to two-, three-, and four-person mixtures containing DNA in high template and low template amounts, Croat. Med. J. 52 (2011) 314–326.
[33] C.C.G. Benschop, C.P. van der Beek, H.C. Meiland, A.G. van Gorp, A.A. Westen, T. Sijen, Low template STR typing: effect of replicate number and consensus method on genotyping reliability and DNA database search results, Forensic Sci. Int. Genet. 5 (2011) 316–328.
[34] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOClt: A computational method to infer the number of contributors to DNA samples analysed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180.
[35] L.E. Alfonse, M. Tejada, M.S. Swaminathan, D.S. Lun, C.M. Grgicak, Inferring the number of contributors to complex DNA mixtures using three methods: exploring the limits of low-template DNA interpretation, J. Forensic Sci. 62 (2017) 308–316.
[36] M. Marciano, J. 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.
[37] P.A. Flach, Machine Learning: the Art and Science of Algorithms That Make Sense of Data, Cambridge University Press, 2012 Prologue, Chapter 1, 2, 7, 9 & 10.
[38] C.C.G. Benschop, J. Hoogenboom, P. Hovers, M. Slagter, D. Kruise, R. Parag, K. Steensma, K. Slooten, J. de Jong, C. Creeten, T. Sijen, A.L.J. Kneppers, DNAxs/DNAStatistX: development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles, Forensic Sci. Int. Genet. 42 (2019) 81–89.
[39] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int. Genet. 21 (2016) 35–44.
[40] A.A. Westen, T. Kraaijenbrink, E.A. Robles de Medina, J. Harteveld, P. Willemse, S.B. Zuniga, K.J. van der Gaag, N.E.C. Weiler, J. Warnaar, M. Kayser, T. Sijen, P. de Knijff, Comparing six commercial autosomal STR kits in a large Dutch population sample, Forensic Sci. Int. Genet. 10 (2014) 55–63.
[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Bubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucker, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[42] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.
[43] SKLearn user manual section 1.9.1 Gaussian Naive Bayes: https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes.
[44] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (2001).
[45] SKLearn user manual section 1.6.2 Nearest Neighbors Classification: https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification.
[46] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Section 4.3 (2008), pp. 106–119.
[47] J.C. Platt, Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, Adv. Large Margin Classif. 10 (2000).
[48] SKLearn user manual section 1.1.11 Logistic regression: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.
[49] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[50] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[51] C. Benschop, A. Backx, T. Sijen, Automated estimation of the number of contributors in autosomal STR profiles, Forensic Sci. Int. Genet. Suppl. Ser. (2019) Manuscript in preparation.
[52] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti. A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018).
[53] https://christophm.github.io/interpretable-ml-book/, Accessed June 2019.
[54] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier. Proceeding KDD 2016, Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. August 13–17, 2016, pp. 1135–1144, , https://doi.org/10.1145/2939672.2939778.