**Realistic visual explanations for individual predictions of the number of contributors made by any machine learning model**

**Abstract**

Using machine learning to determine the number of contributors in short tandem repeat profiles has been shown to obtain good accuracy. However, these predictions are not understandable to the biologist user who would normally determine the number of donors themselves. Therefore, we have created a visual aid that incorporates explanations from SHAP values and counterfactual examples for each prediction. Existing methods for generating counterfactuals have not attempted to handle correlated features, causing those methods to find examples that are impossible given the feature combinations. Since the features from STR data are highly corelated, we have implemented a new method that generates realistic counterfactuals on highly correlated data, and also cuts the number of feature changes in half as compared to presenting the closest counterfactual training data point.

1. **Introduction**

Deriving the number of contributors from Short Tandem Repeat (STR) profiles is a challenging task due to occluding factors such as allele sharing between donors, or allelic drop out. This becomes increasingly difficult when the number of contributors rises. Most software that is used for DNA interpretation requires the NOC to be entered by the user [1]. This could make the difference of including or excluding a person of interest, which is relevant for court.

There have been a multitude of different methods to estimate the NOC. From methods that use only allele counts in a qualitative approach such as the MAC [2], Maximum likelihood approach [3],

Probabilistic mixture algorithm [4]

NOCit tool [5] which is a likelihood-based approach taking into account both qualitative and quantitative data.

1-3 person mixtures. Synthetically created from 58 known donors. Modelling the types of peaks, peak ehights, drop out and stutter rates. posterior probability via a Monte Carlo-based approach. maximum probability method with NOCIt resulted in accuracies of approximately 83%. Computationally slow! Up to 9h for a 5 person mixture.

True allele up to ten contributors? probability model

nC-tool [6] uses both the MAC and Total Allele Count (TAC), as well as categories of drop-out to obtaining better results.

quantitative continuous model peak heights in the DNA profile and considers the

effect of artifacts and allelic drop-out. By using this software, the likelihoods of 1–4 persons’

contributions are calculated, and the most optimal number of contributors is automatically

determined; Kongoh was validated using 27 two-person mixtures, 27 three-person mixtures, and 18 fourperson mixtures. These mixtures were experimentally prepared using non-degraded DNA

from pristine blood samples.

Machine learning approach

PACE random forest [7]: 969 non-simulated DNA samples of 1 to 5 contributors generated from a combination of 120 individuals. They achieved about 90% accuracy on identifying 1, 2, 3, 4+ profiles.

Bayesian probability framework TrueAllele [8]. Closed-source. Minutes to longer for more complex

Correct estimation of the NOC is important.

The results that machine learning methods can obtain has been demonstrated to outperform standard methods [9]. This raised the question however of making these predictions understandable for experts. Without explanations, the experts cannot determine if they should trust the prediction or not. It is also difficult to defend why the expert picked one NOC over the other, if they only relied on the output of a machine learning model.

A decision tree was presented in a paper that made an attempt to make predictions more transparent [10]. However, this led to less accurate predictions (77%) and relied heavily on the filtering of artefacts, which are inherently part of STR data and the NOC estimation process. Moreover, from the explainable AI community, decision trees are not considered an explanation [source]. You are also forced to then use a decision tree, while more interesting predictors are being modelled that may perform even better in the future. The advantage of using an explanation per prediction, is that they are more accurate, and there is no need to filter through an entire decision tree.

A counterfactual is defined as

The **contribution** of this paper is as follows:

* Introduce the field of eXplainable Artificial Intelligence (XAI) to the field of forensic science by demonstrating its value on a practical issue.
* Generate explanations for individual predictions of the NOC by any machine learning model.
* Present the explanations consisting of SHAP values and counterfactual examples in a visualization.
* Implement a new method for finding realistic counterfactuals (ReCo) by deriving them from the training data. This causes the number of feature changes to be cut in half compared to the training data, while maintaining high plausibility. To the best of our knowledge, this is the first method to automatically handle correlated features well.
* Create a simple realism metric that can score how plausible the proposed counterfactual is in terms of its feature combinations.

1. **Materials and methods**
   1. *Machine Learning model*

Originally, the estimation of the NOC was treated as a classification problem [1]. However, since the outputs of the model are ordinal, the problem could benefit from being tackled with a regression model. After a short benchmarking study (see Appendix 1), we concluded that a regression model can achieve better results. In this study, we used a Random Forest Regressor with default parameters.

* 1. *Data analysis and sampling*

The dataset initially consisted of 590 PowerPlex® Fusion 6C (PPF6C) profiles, either from a single donor, or a mixture up to 5 donors [10]. This NOC was based on ground-truth information about the profile. Each profile was then represented by 19 features such as allele counts, allele frequencies and peak heights. These features are almost all very highly correlated.

Tabular, numeric data. For the future, the features might change, but we expect them to remain numeric.

We noticed that this high-dimensional dataset was too sparse for generating counterfactual explanations using the training data; the most similar profile was still very different.

In sampling-based explanation approaches, a dense neighbourhood is created around the current profile to be explained. However, none of these methods can handle correlated features. Therefore, strange

* 1. ***Desiderata*** *explanations*

From meetings with the end users, we determined that there were two main questions of interest:

1. *What were the main reasons for the model to reach the current prediction?*
2. *How could we have arrived at a different prediction?*

To answer question 1, we determined that the use of SHAP values would be sufficient to give an impression of feature importance. For correlated features, SHAP values will turn out lower for each individual feature

It is an issue that feature importance methods split the impact on the model over correlated features. The result of this issue is that the importance value for correlated contributing features is underestimated, in contrast to if their importance was left undivided. However, since main goal of these values is to give an impression of the contributing factors to a prediction, the exact values are not important. This part of the explanation is to observe a general sense of which features contributed to the prediction in which direction. For this purpose, we deem SHAP adequate.

To map out what the counterfactual explanations must accommodate, a list of desiderata was determined.

* Model-agnostic
* Interactive (target can be chosen by user)
* Valid (has desired output)
* Sparse (has not too many feature changes)
* Proximal (is close enough)
* *Robust (the same every time for the same profile)*
* Realistic (makes combinations of feature values that make sense)

Since determining the NOC with machine learning is still a novel approach, there is no consensus about which type of model is most fit. The NFI is looking to improve the model and used features in the future. Therefore, a model-agnostic method is preferable.

The level of interaction we determined was most valuable, was letting the user input the target prediction. Most existing methods assume a binary case, and thus do not have to concern themselves with which target to pick other than the opposite. In this problem, the range of possible values is 1-5. It is not always straightforward to pick the next-best option; different users determine different ranges of possibilities.

The next three desiderata are straightforward as they have been mentioned by most literature on counterfactual explanations. We need counterfactuals that have the desired output (are valid) [11, 12], have the fewest number of feature changes in comparison to the profile we want to explain (are sparse) [11, 12], and must not modify each feature value too much (are proximal) [12, 13].

Though diversity is often encouraged [14], we do not see its value for this problem currently. This is mainly because the explanations are new to the users and they do not want to be confused by seeing multiple, possibly contradicting examples.

The above constraints have mostly been covered quite well in the literature.

However, **realism** is often overlooked or handled quite poorly. One way that realism is considered is the counterfactual point’s distance to the closest training data point [12]. Though several studies have brought up the issue there should be a way to handle **correlated** features [15], none have tackled this issue into their method. To the best of our knowledge, we are the first to develop a method that is suitable for datasets with correlated features.

We present a novel realism score which is calculated as follows

1. When the dataset is loaded, ReCo calculates each feature’s top correlated other features. . With being the feature of interest, and all other features.
2. Once a counterfactual is found, we define the differences in feature value between the counterfactual and original instance as:
3. Then for each element in , we perform the following steps:
   1. Look up its top correlated feature, and take that feature’s value
   2. Look up the combination of with of with the feature value in step a in the training data.
   3. If the combination exists, the realism score is incremented by 1, otherwise by 0.
   4. If the feature in step a was part of , we repeat step a with the next highly correlated feature.

To present the information to the user, a visual approach was used. We incorporated information that answered both questions into one **visualization**.

For tabular data, there have been several approaches to present the information. For example, by a conversational statement [16],

Random **sampling**-based approaches [11, 17-20], sampling through a genetic algorithm [12, 21]. Using the gradient of the loss with respect to the input [22], a method that is based on the data, not on the classifier like SHAP.

A similar piece of work uses **SHAP** values for the current instance to be explained, for both the predicted class A, and target class B [19]. Specific counterfactual instances are then generated by sampling nearest neighbours, changing only the features from the original instance that have negative SHAP values for class B. This approach suffers from the fact that only changing features with negative SHAP values, they limit the range of possible feature changes and therefore produce counterfactuals that are generally further away.

Similarly, a paper discusses using LIME and SHAP to generate counterfactuals from [11]. However, their method is based on highly-dimensional (1000+), behavioural or textual data. They produce counterfactuals by iteratively setting the top contributing features to 0, until the target class is reached. This is not viable in our dataset, since setting a value to 0 does not usually correspond to a realistic feature value.

Unlike studies about loan applications and similar situations, actionability is not a goal of this study. The DNA profiles cannot and will not be altered in the future.

The notion of Multi-Objective Counterfactuals was first proposed with four objectives, solved by a genetic algorithm [12]. Besides distance between x and x’, and the number of feature differences, they also consider the distance to the target outcome and plausibility based on the training data.

To measure the distance between two data instances, we implemented an L1 norm function as shown in

This seems to be the measure of choice by the literature as it does not blow outlier distances out of proportion [12].

|  |  |  |
| --- | --- | --- |
|  |  | () |

Where represents the range of the -th feature, the number of features, and

. We scale with each feature’s range to minimize the influence of different scales, variations, and distributions. This is quite robust for unscaled and unnormalized features with lots of outliers, which is the case in this dataset.

Desiderata for explaining decision trees [13].

1. **Results and discussion**
   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 [52]. 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.

* 1. *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.

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 |

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.

* 1. 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.

G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\Fig_test RFC19 vs LDA40 copy.tif

Figure 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.

G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\NOC probabilities RFC LDA correct copy.tif

Figure 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.

* 1. *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.

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 |

G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\Fig 19 features copy.tif

Figure 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.

* 1. 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% 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 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.

G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\TrainTestValidation_RFCGrid19 copy.tif

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

* 1. 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.

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 | - | - |

* 1. 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 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.

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% |

* 1. 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. 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).

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% |

* 1. 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 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 [53,54] by presenting the feature importance, thus the features that positively or negatively contributed to a given prediction [55]. 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.

1. **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.

**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, H, Kelly, 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 Science International: Genetics 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, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (January–February (1)) (2012) 113–122.
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. 2012, Prologue, Chapter 1, 2, 7, 9 & 10: Cambridge University Press.
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. Machine Learning Res. 12 (2011) 2825-2830.
42. F. Pedregosa-Izquierdo, Partial Correlation in Python (clone of Matlab's partialcorr). 2014 cited 2018; Available from: https://gist.github.com/fabianp/9396204419c7b638d38f.
43. L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
44. SKLearn user manual section 1.9.1 Gaussian Naive Bayes: https://scikit-learn.org/stable/modules/naive\_bayes.html#gaussian-naive-bayes
45. J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29 (2001).
46. SKLearn user manual section 1.6.2 Nearest Neighbors Classification: https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification
47. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Section 4.3, p.106-119, 2008.
48. J.C. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. Adv. Large Margin Classif. 10 (2000).
49. SKLearn user manual section 1.1.11 Logistic regression: https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression
50. D.E. Rumelhart, G.E. Hinton, R.J. Williams. Learning representations by back-propagating errors. Nature 323 (1986) 533-536.
51. L. Breiman, Random Forests, Machine Learning, 45 (2001) 5-32.
52. 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.
53. R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti. A survey of methods for explaining black box models. ACM Computing Surveys 51 (2018).
54. https://christophm.github.io/interpretable-ml-book/, Accessed June 2019.
55. 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. Pages 1135-1144. San Francisco, California, USA. August 13-17 2016. DOI: http://dx.doi.org/10.1145/2939672.2939778.

**Supplementary Material**

**G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\Train test accuracies 10 algorithms copy.tif**

Supplementary Figure 1. Train and test accuracies per algorithm and per number of features. A) LDA, B) RFC, C) LR, D) GBC, E) SVC, F) LSVC, G) k-NN, H) DTC, I) MLPC, J) GaussianNB. The selected number of features per algorithm is presented in the green circle.

**G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\Train test per NOC per model copy.tif**

Supplementary Figure 2. Confusion matrices for the train and test set per model. A) LDA40, B) RFC19, C) LR8, D) GBC11, E) SVC5, F) LSVC4, G) k-NN3, H) DTC14, I) MLPC8, J) GaussianNB4.



Supplementary Figure 3. Precision-recall plots for the test set per true NOC. A) LDA40, B) RFC19, C) LR8, D) GBC11, E) SVC5, F) LSVC4, G) k-NN3, H) DTC14, I) MLPC8, J) GaussianNB4.

G:\BI\BI_ALG\R&D\Research\Corina\Tables & figures\NOC tool\NOC probabilities RFC LDA incorrect sample comparison copy.tif

Supplementary Figure 4. Probabilities per NOC for DNA profiles in the test set that received an incorrect NOC prediction when using RFC19 (A) or LDA40 (B). Empty bars indicate that the NOC for this sample was correctly predicted by the particular model.

Supplementary Table 1. Overview of PowerPlex® Fusion 6C DNA profiles in the ‘extremes’ dataset.

|  |  |  |
| --- | --- | --- |
| **Extreme because:** | **Number of contributors** | **Number of samples** |
| More donors than the model was trained with | 6 | 16 |
| Relatives (brothers) within the mixtures in combination with a large degree of allelic drop-out | 2 | 5 |
| 3 | 3 |
| 4 | 2 |
| Extreme level of degradation (at least three locus drop-outs) | 3 | 3 |
| 4 | 3 |
| 5 | 3 |

Supplementary Table 2. Overview of the numbers of DNA-profiles used to examine the effect of replicates.

|  |  |  |
| --- | --- | --- |
| **Number of Contributors** | **Number of different DNA extracts** | **Total number of DNA-profiles**  **(including replicates)** |
| 2 | 30 | 90 |
| 3 | 28 | 84 |
| 4 | 29 | 87 |
| 5 | 27 | 81 |
| Total | 114 | 342 |

Supplementary Table 3. Overview of the 25 sample features. Sample features take account of all 23 autosomal loci within the PowerPlex® Fusion 6C profiles. Amelogenin and the Y-chromosomal markers were excluded from the analyses.

|  |  |  |
| --- | --- | --- |
| **Number** | **Feature** | **Details** |
| 1 | MAC | Maximum allele count (MAC); Maximum number of alleles observed on a locus |
| 2 | TAC | Total allele count (TAC); Total number of alleles per profile |
| 3 | Mean Allele count | Mean, median or standard deviation of the number of alleles per locus |
| 4 | Median Allele Count |
| 5 | Standard Deviation Allele Count |
| 6 | Minimum Allele Count | Minimum number of alleles observed per locus |
| 7 | Minimum NOC | Minimum number of contributors (NOC); Maximum Allele Count / 2, rounded up to 0 decimals |
| 8 | AC 0 | Number of loci with an allele count of 0 (*i.e.* empty loci/ locus drop-outs), 1 or 2, 3 or 4, 5 or 6, 7 or 8, or 9 alleles or more. |
| 9 | AC 1-2 |
| 10 | AC 3-4 |
| 11 | AC 5-6 |
| 12 | AC 7-8 |
| 13 | AC ≥9 |
| 14 | Maximum PH | Largest, lowest, mean, median or standard deviation of the peak height (PH) of an allele (RFUs) observed across the profile |
| 15 | Minimum PH |
| 16 | Mean PH |
| 17 | Median PH |
| 18 | Standard deviation PH |
| 19 | Degradation Slope1 | Slope of line where  Y= sum of peak heights per locus and X= average fragment length per locus |
| 20 | Degradation Slope2 | Slope of line where  Y= average of peak heights per locus and X= average fragment length per locus |
| 21 | Number of Peaks above ST | Number of peaks above or below, or the ratio between the number of peaks above/below or below/above the stochastic threshold (ST) of 800 RFUs |
| 22 | Number of Peaks below ST |
| 23 | Ratio peaks below/above ST |
| 24 | Ratio peaks above/below ST |
| 25 | Match probability | The probability of a random, unrelated person matching to this DNA profile. The probability is calculated using the allele frequencies of 2085 male Dutch individuals database [40]. |

Supplementary Table 4. Overview of the 11 locus features which were calculated for each of the 23 autosomal loci within the PowerPlex® Fusion 6C profiles.

|  |  |  |
| --- | --- | --- |
| **Number** | **Feature** | **Details** |
| 1 | Allele count | Number of alleles |
| 2 | Minimum NOC | Allele count / 2, rounded up to 0 decimals |
| 3 | Maximum PH | The largest, smallest, mean, median or standard deviation of the peak height (PH, in RFUs) of alleles at the particular locus |
| 4 | Minimum PH |
| 5 | Mean PH |
| 6 | Median PH |
| 7 | Standard deviation PH |
| 8 | Minimum AF | The lowest or highest allele frequency (AF) of an allele, or the sum of the allele frequencies of the alleles, or the percentage of alleles that are within the population database. *I.e.*: |
| 9 | Maximum AF |
| 10 | Sum AF |
| 11 | Percentage of AF |

Supplementary Table 5. Top 50 ranked features and their partial correlation (correlation for TAC and MAC). Features are ordered based on the sequential algorithm described in section 2.2. Further details on these features are presented in Supplementary Tables 3 and 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **Number** | **Feature** | **Sample/ locus feature** | **Partial Correlation** |
| 1 | MAC | Sample | 0.9209 |
| 2 | TAC | Sample | 0.9133 |
| 3 | Standard Deviation Allele Count | Sample | 0.4214 |
| 4 | Allele count\_D3S1358 | Locus | 0.2904 |
| 5 | AC 5-6 | Sample | 0.2577 |
| 6 | Minimum NOC\_Penta E | Locus | 0.2100 |
| 7 | Minimum NOC\_Penta D | Locus | 0.2123 |
| 8 | AC 0 | Sample | 0.1962 |
| 9 | Standard deviation PH\_vWA | Locus | 0.1911 |
| 10 | Match Probability | Sample | 0.1972 |
| 11 | Number of Peaks below ST | Sample | 0.1621 |
| 12 | Minimum NOC\_TPOX | Locus | 0.1547 |
| 13 | Minimum NOC | Sample | 0.1811 |
| 14 | Minimum NOC\_CSF1PO | Locus | 0.1435 |
| 15 | Minimum NOC\_D16S539 | Locus | 0.1518 |
| 16 | Sum AF\_TH01 | Locus | 0.1188 |
| 17 | Allele count\_TPOX | Locus | 0.1242 |
| 18 | Percentage of AF\_D1S1656 | Locus | 0.1290 |
| 19 | Allele count\_D8S1179 | Locus | 0.1174 |
| 20 | AC 3-4 | Sample | 0.1264 |
| 21 | Standard deviation PH\_D2S441 | Locus | 0.1330 |
| 22 | Minimum AF\_D22S1045 | Locus | 0.1369 |
| 23 | Sum AF\_SE33 | Locus | 0.1266 |
| 24 | Minimum AF\_Penta E | Locus | 0.1224 |
| 25 | Allele count\_D18S51 | Locus | 0.1264 |
| 26 | Minimum AF\_D5S818 | Locus | 0.0978 |
| 27 | Minimum AF\_D8S1179 | Locus | 0.0933 |
| 28 | Ratio peaks above/below ST | Sample | 0.0978 |
| 29 | Median PH\_D19S433 | Locus | 0.1038 |
| 30 | Minimum NOC\_vWA | Locus | 0.0951 |
| 31 | Median PH\_D7S820 | Locus | 0.0847 |
| 32 | Median PH\_Penta D | Locus | 0.1021 |
| 33 | Sum AF\_FGA | Locus | 0.1022 |
| 34 | Median PH\_TPOX | Locus | 0.0942 |
| 35 | Median PH\_D2S441 | Locus | 0.1190 |
| 36 | Maximum PH\_Penta D | Locus | 0.0964 |
| 37 | Maximum PH\_Penta E | Locus | 0.1266 |
| 38 | Minimum NOC\_D3S1358 | Locus | 0.0928 |
| 39 | Mean PH\_D19S433 | Locus | 0.0909 |
| 40 | Minimum AF\_CSF1PO | Locus | 0.0887 |
| 41 | Minimum NOC\_FGA | Locus | 0.0898 |
| 42 | Minimum AF\_D2S441 | Locus | 0.0898 |
| 43 | Standard deviation PH\_TH01 | Locus | 0.0917 |
| 44 | Minimum NOC\_D12S391 | Locus | 0.0806 |
| 45 | Percentage of AF\_D7S820 | Locus | 0.0837 |
| 46 | Minimum NOC\_D7S820 | Locus | 0.1172 |
| 47 | Sum AF\_D10S1248 | Locus | 0.0980 |
| 48 | Minimum PH\_Penta E | Locus | 0.0840 |
| 49 | Median PH\_Penta E | Locus | 0.0954 |
| 50 | Mean PH\_Penta E | Locus | 0.1616 |

Supplementary Table 6. Comparison of the performance of the MAC method, nC-tool and RFC19 machine learning model for 2p-5p PPF6C profiles.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MAC** | | | **TAC nC-tool** | | | **RFC19 machine learning** | | |
| **True NOC (*n* per method)** | **Under-estimated** | **Correctly estimated** | **Over-estimated** | **Under-estimated** | **Correctly estimated** | **Over-estimated** | **Under-estimated** | **Correctly estimated** | **Over-estimated** |
| 2 (*n*=90) | 0% | 67% | 33% | 0% | 100% | 0% | 0% | 100% | 0% |
| 3 (*n*=88) | 0% | 97% | 3% | 17% | 83% | 0% | 3% | 97% | 0% |
| 4 (*n* =89) | 20% | 77% | 3% | 27% | 70% | 3% | 17% | 83% | 0% |
| 5 (n =87) | 63% | 37% | 0% | 47% | 53% | not applicable | 40% | 60% | not applicable |
| Total (*n* =354) | 21% | 69% | 10% | 23% | 77% | 1% | 15% | 85% | 0% |

1. Taylor, D., J.-A. Bright, and J. Buckleton, *Interpreting forensic DNA profiling evidence without specifying the number of contributors.* Forensic Science International: Genetics, 2014. **13**: p. 269-280.

2. Clayton, T.M., et al., *Analysis and interpretation of mixed forensic stains using DNA STR profiling.* Forensic Science International, 1998. **91**(1): p. 55-70.

3. Haned, H., et al., *Estimating the Number of Contributors to Forensic DNA Mixtures: Does Maximum Likelihood Perform Better Than Maximum Allele Count?* Journal of Forensic Sciences, 2011. **56**(1): p. 23-28.

4. Paoletti, D.R., et al., *Inferring the Number of Contributors to Mixed DNA Profiles.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012. **9**(1): p. 113-122.

5. Swaminathan, H., et al., *NOCIt: A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping.* Forensic Science International: Genetics, 2015. **16**: p. 172-180.

6. Benschop, C., A. Backx, and T. Sijen, *Automated estimation of the number of contributors in autosomal STR profiles.* Forensic Science International: Genetics Supplement Series, 2019. **7**.

7. Marciano, M.A. and J.D. Adelman, *Developmental validation of PACE™: Automated artifact identification and contributor estimation for use with GlobalFiler™ and PowerPlex® fusion 6c generated data.* Forensic Science International: Genetics, 2019. **43**.

8. Bauer, D.W., et al., *Validating TrueAllele® Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors.* Journal of Forensic Sciences, 2020. **65**(2): p. 380-398.

9. Benschop, C.C.G., et al., *Automated estimation of the number of contributors in autosomal short tandem repeat profiles using a machine learning approach.* Forensic Science International: Genetics, 2019. **43**: p. 102150.

10. Kruijver, M., et al., *Estimating the number of contributors to a DNA profile using decision trees.* Forensic Science International: Genetics.

11. Ramon, Y., et al., *A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C.* Advances in Data Analysis and Classification, 2020. **14**(4): p. 801-819.

12. Dandl, S., et al. *Multi-Objective Counterfactual Explanations*. in *Parallel Problem Solving from Nature – PPSN XVI*. 2020. Cham: Springer International Publishing.

13. Sokol, K. and P. Flach. *Desiderata for interpretability: Explaining decision tree predictions with counterfactuals*. 2019.

14. Russell, C. *Efficient search for diverse coherent explanations*. 2019.

15. Barocas, S., A.D. Selbst, and M. Raghavan. *The hidden assumptions behind counterfactual explanations and principal reasons*. 2020.

16. Sokol, K. and P. Flach. *Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements*. 2018.

17. Wachter, S., B. Mittelstadt, and C. Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.* Harvard journal of law & technology, 2018. **31**: p. 841-887.

18. Grath, R.M., et al., *Interpretable Credit Application Predictions With Counterfactual Explanations.* ArXiv, 2018. **abs/1811.05245**.

19. Rathi, S., *Generating Counterfactual and Contrastive Explanations using SHAP*. 2019.

20. White, A. and A. Garcez. *Measurable Counterfactual Local Explanations for Any Classifier*. in *ECAI*. 2020.

21. Guidotti, R., et al., *Factual and Counterfactual Explanations for Black Box Decision Making.* IEEE Intelligent Systems, 2019. **34**(6): p. 14-23.

22. Moore, J., N. Hammerla, and C. Watkins, *Explaining deep learning models with constrained adversarial examples*. 2019. p. 43-56.