Forensic Science International: Genetics 27 (2017) 82–91

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures

Michael A. Marciano[1,*], Jonathan D. Adelman[1]

Forensic & National Security Sciences Institute, College of Arts and Sciences, Syracuse University, 107 College Place 1-014 Center for Science and Technology, Syracuse, NY, 13244, USA

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 27 October 2016
Accepted 22 November 2016
Available online 24 November 2016

Keywords:
DNA mixtures
Number of contributors
Probabilistic
Machine learning

ABSTRACT

The deconvolution of DNA mixtures remains one of the most critical challenges in the field of forensic DNA analysis. In addition, of all the data features required to perform such deconvolution, the number of contributors in the sample is widely considered the most important, and, if incorrectly chosen, the most likely to negatively influence the mixture interpretation of a DNA profile. Unfortunately, most current approaches to mixture deconvolution require the assumption that the number of contributors is known by the analyst, an assumption that can prove to be especially faulty when faced with increasingly complex mixtures of 3 or more contributors. In this study, we propose a probabilistic approach for estimating the number of contributors in a DNA mixture that leverages the strengths of machine learning. To assess this approach, we compare classification performances of six machine learning algorithms and evaluate the model from the top-performing algorithm against the current state of the art in the field of contributor number classification. Overall results show over 98% accuracy in identifying the number of contributors in a DNA mixture of up to 4 contributors. Comparative results showed 3-person mixtures had a classification accuracy improvement of over 6% compared to the current best-in-field methodology, and that 4-person mixtures had a classification accuracy improvement of over 20%. The Probabilistic Assessment for Contributor Estimation (PACE) also accomplishes classification of mixtures of up to 4 contributors in less than 1 s using a standard laptop or desktop computer. Considering the high classification accuracy rates, as well as the significant time commitment required by the current state of the art model versus seconds required by a machine learning-derived model, the approach described herein provides a promising means of estimating the number of contributors and, subsequently, will lead to improved DNA mixture interpretation.

1. Introduction

DNA mixtures are defined as a mixture of the genomic DNA from 2 or more donors. The ability to separate or "deconvolute" the individual donors from a DNA mixture remains one of the most critical challenges in the field of forensic DNA analysis. Several metrics are required to accurately interpret and deconvolute DNA mixtures; a selection of the most critical include the number of contributors, the minimum expected heterozygote balance, the ratio of contributors, the amount of DNA template (per sample per contributor) and the probabilities of allele drop-out and drop-in. Specifically, the number of contributors is widely considered the most critical component in leading to an accurate DNA mixture deconvolution, in large part due to the deconvolution's sensitivity to whichever number of contributors is assumed. Likelihood-based deconvolution methods require the potentially erroneous assumption that the number of contributors is known by the analyst. Indeed, the assumption of the number of contributors can greatly affect the resulting conclusions [1]. Establishing the number of contributors permits the analyst to set a range of potential alleles at a particular locus within the sample and proceed with mixture deconvolution. However, the use of incorrect assumptions regarding the number of contributors can have effects on the resulting likelihood ratios [2,3] and, subsequently, the mixture interpretation as a whole. Therefore, making high-probability

[*] Corresponding author.
E-mail address: mamarcia@syr.edu (M.A. Marciano).

[1] Both authors contributed equally.

http://dx.doi.org/10.1016/j.fsigen.2016.11.006
1872-4973/© 2016 Elsevier Ireland Ltd. All rights reserved.

estimates of the number of contributors in a given mixture should be considered a vital component of DNA mixture deconvolution.

Several methods have been proposed to estimate the number of contributors by setting this value to the minimum number required to explain the observed profiles. The oldest and still most prevalent is a qualitative method known as the maximum allele count (MAC) method [4,5]. This method of allele counting requires the identification of the locus or loci with the maximum number of allelic peaks; the minimum number of contributors would then be calculated by counting the number of allelic peaks divided by 2 (to account for ploidy) and rounding up to the nearest whole number:

$$\textit{minimum number of contributors} = \frac{t_a}{\textit{ploidy}} \quad (1)$$

where t_a = total number of alleles at a DNA locus; and *ploidy* = the genomic copy number per cell. This method does not work well with complex mixtures in large part due to the potential for allele sharing between contributors; that is, MAC does not take into account the frequency of the alleles and the propensity for multiple contributors to possess common alleles. A quantitative assessment of the peak heights and areas can be combined with MAC and minimum heterozygote balance information to yield a minimum number of contributors while considering some level of allele sharing. However, several studies have shown that MAC tends to underestimate the number of contributors due to masking, that approximately 76% of 4 person mixtures actually look like 3 person mixtures if based solely on the number of peaks (the underestimates are based on 13 loci [6] and 15 loci [7]; dropout present in both data sets) [6,7] and that the method gives biased estimates that are likely to increase in case of population subdivision [6,8]. These observations have led to a consensus that the determination of the number of contributors must be based on more than merely the number of peaks observed [8,9]. More recently, Coble et al. demonstrated that, as expected, as the number of loci in amplification kits increases it become less likely that higher order mixtures of 5 and 6 contributors will be misclassified as samples with 1–3 contributors [10]. However, despite the increased number of loci included in the newer amplification kits, the problem simply shifts from potential misclassification of lower order mixtures to higher order mixtures, for example the probability of 5-contributor samples appearing as a 4-contributor sample.

Egeland et al. (2003) proposed estimating the true number of contributors using a maximum likelihood-based approach when diallelic markers (SNPs) are used [11]. This was later extended to include multiallelic markers (e.g. microsatellites) [9]. In addition to using the number of peaks observed, this likelihood-based method used the background frequencies of the alleles above a user-defined analytical threshold to calculate the likelihood a locus contains alleles from a certain number of contributors. The maximum likelihood estimation (MLE)-based approach was found to correctly estimate the number of contributors to 2- and 3-person mixtures more than 90% of the time, representing a modest improvement over MAC. For 4- and 5-person mixtures, the approach provided a correct estimation between 64%–79% of the time compared to approximately 30% of the time for MAC (15 loci, level of dropout not noted) [12]. Another study proposed an estimator that simply considered the number of allele peaks across an entire profile rather than looking at the maximum number of alleles at any 1 locus [7], and results were comparable to MLE-based approaches. It should be noted that these approaches all utilize qualitative data, but do not consider available quantitative data such as peak height.

More recently developed methods by Taylor et al. and NOCIt infer the number of contributors in a DNA sample by calculating the posterior probability via a Monte Carlo-based approach [13,14]. The NOCIt method utilizes both qualitative and quantitative data regarding the DNA sample for its inferences, and was shown to slightly outperform pre-existing methods. NOCIt was also shown to be insensitive to injection time, whereas MLE-based approaches suffered at low injection times. The overall accuracy across 5, 10 and 20 s injections using a "maximum probability" approach, in which the number of contributors with the highest probability was chosen by the software, was 83% using the authors' testing data set. Using a maximum probability method with NOCIt resulted in accuracies of approximately 94% for single source samples, 98% for 2 contributor samples, 87% for 3 contributor samples, and 63% for 4 contributor samples. Note that these approximate accuracies are based on interpretation of the histograms in Swaminathan et al. [14]. A drawback of NOCIt, however, is the required processing time; 5-person mixtures take up to 9 h to evaluate [14]. This time sink can be significantly prohibitive in a forensic lab requiring rapid analysis and with a growing backlog. Further, in spite of the clear advances beyond MAC shown in these more recent approaches, the correct estimation rate for the number of contributors in a DNA mixture arguably remains poor as the number of contributors increases.

Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience [15]. A machine learning algorithm can, after exposure to an initial set of data, be used to generalize; meaning that it can evaluate new, previously unseen examples and relate them to the initial "training" data. Machine learning is a widely-used approach with an incredibly diverse range of applications, including object recognition [16], natural language processing [17], and DNA sequence classification [18]. It is ideally suited for classification problems involving implicit patterns, and is most effective when used in conjunction with large amounts of data. Although machine learning has not previously been used within the domain of DNA mixture analysis, the problem area is well-suited to such an endeavor due to 2 key problem characteristics: there exists a large repository of human DNA mixture data in electronic format, and these data are high-dimensional and complex; patterns in such data are often non-obvious and beyond the effective reach of manual analysis but can be statistically evaluated using 1 or more machine learning algorithms.

In this study we describe a novel method to probabilistically infer the number of contributors in a mixed DNA sample using a machine learning approach. The conclusions generated are based on the use of both categorical (qualitative) data such as allele labels, dye channels and continuous and discrete (quantitative) data such as stutter rates, peak heights, heterozygote balance, and mixture ratios that describe the DNA sample. The method is computationally inexpensive, and results are obtained in a maximum of 10 s using a standard desktop or laptop computer with 6–8 GB RAM and an Intel i5 1.9 gHz processor.

2. Materials and methods

2.1. Data acquisition and exportation

The system was trained, tested and validated using electronic data (.fsa files) obtained from 1405 non-simulated DNA mixture samples comprised of 1–4 contributors and generated from a combination of 20 individuals. The set was obtained through publicly available data sets (http://www.bu.edu/dnamixtures/pages/help/introduction/) and validation data provided by collaborators. This set of 1405 samples included 35 different DNA template amounts from 0.0125 ng to 10 ng (0.0125, 0.025, 0.05, 0.0625, 0.075, 0.1, 0.125, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2, 1.3, 1.5, 1.6, 1.7, 2.0, 2.5, 3.0, 3.3, 3.5, 4.0, 5.0, 6.0, 7.0, 7.5, 8.0, 9.0, 10.0 ng) and 28 different ratios of contributors (1.6:3:1:1, 1.6:3:1:2, 1.6:3:2:1, 1.6:3:2:2, 1.6:6:1:1, 3:3:1:1, 3:6:1:1, 1.6:12:1, 1.5:3:1, 1.5:3:2, 1.5:6:1, 1.6:3:4, 12:1:1, 1:3:1/3:1:1/1:1:3, 3:1:2, 3:1:4, 3:2:1, 3:3:1, 3:4:1, 6:1:1, 6:3:1, 1:0, 1:1, 1:19/19:1, 1:2/2:1, 1:4/4:1, 1:9/9:1, 10:1). Further information regarding the samples can be found in Appendix A in Supplementary material, Tables 1 and 2. These samples were previously amplified using the AmpFLSTR® Identifiler® PCR Amplification Kit (ThermoFisher Scientific Inc.) (28 cycles) with subsequent detection performed on 5 different 3130/3100 Genetic Analyzers (ThermoFisher Scientific Inc.). Fragment analysis was performed using GeneMarkerHID® v2.8.2 (SoftGenetics® LLC) using a threshold of 10 relative fluorescent units (rfu) without stutter filters. All data were conservatively preprocessed by an analyst to remove artefacts such as dye blobs and electrical spikes. Data were exported from GeneMarkerHID® v2.8.2 (SoftGenetics® LLC) for further analysis using the PACE software. PACE-based preprocessing of the exported fragment data included the application of stutter filters (Appendix A in Supplementary material: Table 3) and an analytical threshold. The analytical threshold was dynamically applied based on the level of baseline noise encountered at each sample-locus combination. Rather than employing a traditional static threshold, this dynamic baseline that we term locus-sample-specific "thresholding" (LSST) considers the mean baseline noise in areas flanking the allele calling region of each locus. Regions with pull-up are preferentially avoided in the calculation of the analytical threshold. The LSST is a component built into the PACE software. Stochastic thresholds were not employed.

2.2. Data partitioning

A machine learning algorithm will "learn" a predictive model, and that model in turn, is potentially capable of classifying new, unfamiliar data. This ability to predict outcome values for previously unseen data is termed generalization. Merely providing training data to a learning algorithm is an insufficient generalization strategy; the algorithm may end up learning specific patterns only found in the training data by chance, and would then erroneously leverage those patterns to aid with classification. This is analogous to a curve-fitting problem in which a high-degree polynomial is used to fit generally linear data; such a choice may well result in a high correlation coefficient, but as soon as even 1 additional data point is added, the coefficient's value can plummet. Such a scenario highlights the problem of overfitting a model to data; such a model is not generalizable. To ensure generalization and avoid potential overfitting, the learning algorithm's model must instead be both trained and tested, and it is the resulting testing accuracy, not the training accuracy, that serves to validate the learned model. Such an approach requires that the initial data set, in this case, the library of DNA mixtures, must correspondingly be partitioned into completely separate training and testing subsets. For all modeling efforts herein, the training data set was created by randomly selecting 75% of the initial data, with the other 25% used for testing how generalizable the learned model is; this amounted to 1063 samples for training and 352 samples for testing. This set of 352 samples contained 753 sample-locus instances of allelic dropout over 159 total samples (allelic dropout identified using the PACE-LSST thresholding method). The computation time for model creation (training and testing) was approximately 90 min on a standard laptop computer (Intel Core i5–3230 M CPU @ 2.6 GHz, 8GB RAM and required less than 400MB of hard drive space).

A second data set, termed the truncated-degradation testing set, was compiled to examine the effects of fewer loci and simulated degradation on PACE. These data were initially identical to the 352-sample testing set previously described, but the large loci (CSF1PO, FGA, D18S51 and D2S1338) were removed from samples that had both a total template DNA amount of 0.25 ng or lower and a ratio of major to smallest minor no greater than 3:1 (this included 1:0, 1:1, 1:2/2:1, 1:4/4:1, 1:1:3/3:1:1/1:3:1, 3:3:1:1, 1.6:3:1:1). These requirements were imposed to ensure that the template amounts of each contributor would be reasonable considering the probability of locus dropout.

Machine learning algorithms contain hyperparameters which can be loosely thought of as knobs that tune an algorithm and thereby affect its behavior. Some hyperparameters can have a non-trivial impact on the resulting training time or even classification accuracy. Any attempt to tune these hyperparameters and thereby maximize an algorithm's classification accuracy are typically accompanied by a further partitioning of training data to ensure that data used for "tuned" algorithm validation are not also used to validate the final model. A viable alternative to data partitioning is the technique of *k*-fold cross-validation. In this technique the algorithm is trained a total of k times, with a fraction 1/*k* of training examples left out each time for validation purposes [19], leading to *k* distinct "folds". Each fold provides summary metrics that describe the algorithm's performance for that particular training, and the results from each of the folds are averaged to provide an overall assessment of model performance. All hyperparameter tuning in this study utilizes 5-fold cross-validation on the training data set. Hyperparameter optimization was performed using a limited grid search, in which hyperparameters related to cost or regularization were tuned to produce the highest possible accuracy so long as training and testing accuracies remained within 2% of one another. No other hyperparameters were optimized.

2.3. Feature scaling

Learning algorithms make use of data instances, each one of which has a corresponding feature vector. In this context, features can be considered data categories such as peak height, allele count per locus, etc. Most machine learning algorithms cannot appropriately utilize the raw features in this vector because feature scales can be wildly different from one another. The template DNA feature, for example, has mean and variance several orders of magnitude smaller than those of the maximum peak height feature. Distinct means and variances can lead to some features' importance being artificially inflated by learning algorithms, which are spending disproportionate amounts of time minimizing the larger errors produced by the features with the larger variances. While many researchers choose to resolve this concern by simply normalizing feature data via min-max scaling to a range of 0–1, some learning algorithms learn model weights more quickly and are more robust in the face of data outliers if features are instead standardized:

$$X_{std}^{(i)} = \frac{X^{(i)} - \mu_x}{\sigma_x} \tag{2}$$

Here $X^{(i)}$ is a given feature, μ_x. is the feature's mean, and σ_x is the corresponding standard deviation. All feature scaling in this study was performed using Eq. (2).

2.4. Feature selection

Bellman's "curse of dimensionality" [20] refers to the exponential rise in the time and space required to compute an approximate solution to a problem as the dimension increases. In other words, the continual addition of new features to a feature vector ultimately leads to decreased accuracy of the resulting classification model. Just as data are partitioned into a training set and a testing set to allow model validation and avoid data overfitting, so too must the feature vector be set to an optimal size – not too small, where vital information with predictive value is ignored, but not too large, where additional dimensions in the problem's feature space lead to a very high training accuracy and a very low testing accuracy: a hallmark of overfitting. It is therefore of great importance to only include features that strongly contribute to the subsequent classification problem, and to remove features that fail to contribute.

One metric that can estimate a feature's classification "importance" is the Kullback–Leibler divergence (Equation (3)), which is a measure of the reduction in entropy of the class variable (in this case, the true number of contributors) after the value for the feature is observed.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{3}$$

All candidate features are ranked by divergence, and any candidate feature with a divergence below 0.01 is removed prior to machine learning. Calculations of Kullback–Leibler divergence were performed using the Weka Knowledge Analysis Environment, version 3.8 [21].

2.5. Machine learning algorithms

No single machine learning algorithm is ideally suited for all classification problems [22]. The best-suited algorithm often depends on the size, quality, and characteristics of the associated training data. This study makes no presuppositions concerning which algorithm is optimal for the problem of classifying the number of contributors in a DNA mixture, and instead evaluates five candidate machine learning algorithms:

(1) The **k-Nearest Neighbors** (*k*-NN) algorithm determines an instance's class based on the most common class of its neighbors. Specifically, each object is classified by a majority vote of its nearest *k* neighbors, with the object then assigned to the most common class among those neighbors. This non-parametric algorithm is perhaps the simplest among those chosen as candidate learners.

(2) The **Classification and Regression Trees** (CART) algorithm is one of the most commonly used decision tree classifiers [23]. The decision tree itself is a foundational machine learning concept, in which feature space is continually subdivided into smaller regions of roughly uniform values, and in which the leaves of the tree represent the possible classes into which an object can be classified. Decision trees often provide high-accuracy models, but can also overfit data more frequently and are therefore potentially worse than the other candidate algorithms at generalizing to unseen data.

(3) The **Multinomial Logistic Regression** (Logit) algorithm is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. It is an extension of binary logistic regression, capable of generalizing to multiclass problems.

(4) The **Multilayer Perceptron** (MLP) algorithm is an artificial neural network in which backward propagation of errors (backpropagation) is used to train the network's weights and thresholds. In this study, a single hidden layer of neurons was used, and the 4 output nodes correspond to the 4 classes of number of contributors.

(5) The **Support Vector Machine** (SVM) algorithm used in this study attempts to optimize classification by maximizing the distance between the margins of classes. There are both linear and non-linear versions of this classifier, the latter of which is specifically designed for classes that cannot be linearly separated.

All machine learning algorithms were implemented using Python's Scikit-Learn library [24] with the exception of the multilayer perceptron, which was implemented in Weka using that software's default architecture [21].

2.6. Algorithm and model evaluation

A learning curve is a plot of training and cross-validation accuracy as a function of the number of training data used. That is, a learning curve measures how much better a classification model is at predicting as the number of instances used to train that model is increased. Learning curves can also be used to gauge how well a model might perform when faced with new, previously unseen data by comparing training and testing accuracies and noting the degree of convergence. A model is said to suffer from high bias – the persistent or systematic error that the learning algorithm is expected to make when trained on training sets – if its accuracy is low, and is said to have suffer from high variance – the expected value of the squared difference between any particular hypothesis and the averaged hypothesis, with respect to all training samples – if its training and testing accuracies are dissimilar. Bias is often associated with under-fitting a model, whereas variance is often associated with over-fitting. A high-variance model is very sensitive to the sample used to build that model; its error depends in large part on the training set used, and as that error is evaluated across distinct cross-validation folds, the resulting variance in error is high. Models in this study were assessed using both classification accuracy and the degree to which testing accuracy converged with training accuracy.

The accuracy of a model's classifications (e.g. 1, 2, 3 or 4 contributors) was determined through comparison of the results obtained from the classification of the testing data set with the known number of contributors for each of the testing data samples. Correct calls represent instances for which the known number of contributors is equal to the class with the highest associated probability. For example, consider a hypothetical sample *A* taken from the testing data set, where *A* is known to be a 3-contributor mixture. $\Pr(A = x)$ is the probability of A being a member of class *x*. If $\Pr(A{=}1) = 0.0001$, $\Pr(A{=}2) = 0.0005$, $\Pr(A = 3) = 0.9771$, and $\Pr(A = 4) = 0.0223$, the model finds the class associated with the highest probability (in this case, 3 contributors) and then classifies sample *A* as a mixture with 3 contributors – in this case, a correct classification. This approach to measuring model accuracy is the same as that used to determine the "max probability" accuracy in Swaminathan et al. [14]. Note that none of the classification algorithms described herein explicitly learn the underlying probability distribution over the set of classes (in this case, the various possible numbers of contributors) in order to make their predictions. These are deterministic learning methods, not probabilistic ones. However, even though the algorithms do not attempt to learn said probability distribution as part of their classification strategies, they can all still be placed within a probabilistic framework through the use of post-hoc approaches [25]. For example, Support Vector Machines' output can be expressed probabilistically via approaches such as Platt scaling [26] or isotonic regression [27]. All probabilities in this study were obtained using isotonic regression.

It should be noted that in any system's attempt to classify the number of contributors in a DNA mixture, such a system will necessarily overestimate its own accuracy for the highest-numbered class in its associated data set. The lowest-numbered class for DNA mixture interpretation is 1, but any chosen upper bound for a system's ability to classify the number of contributors

Table 1
Kullback–Leibler divergence of nine candidate features. The maximum number of contributors is the number of allelic peaks assuming that each peak represents 1 homozygous donor with no allele sharing. For example, if a locus had 2 peaks, in this context, the maximum number of contributors would be 2, 2 peaks contributed by 2 homozygotes with no allele sharing. In contrast, the minimum number of contributors refers to the minimum number of contributors given the number of peaks present (Eq. (1)). Using the previous example and applying Eq. (1), the minimum number of contributors for a locus with 2 allelic peaks is 1.

D_{KL}	features
1.638	sample-wide peak count
1.308	maximum number of contributors
1.060	minimum number of contributors
0.823	template DNA amplified
0.512	locus-specific peak count
0.358	min/max observed peak heights
0.309	probability of dropout
0.090	minimum observed peak height
0.038	maximum observed peak height
0	size of locus

is less than the number of contributors in some hypothetical mixture. PACE currently evaluates mixtures of 1–4 contributors, but 5- and 6-contributor mixtures are plausible samples. A modified system that evaluated mixtures of 1–6 contributors would still be unable to differentiate a 7-contributor sample, and so on. These mixtures with larger numbers of contributors would serve as a source of incorrect classification if they were present, especially when they contain confounding characteristics such as allele sharing. Their absence at the upper end of a system's classification limit therefore leads to overestimation of the system's accuracy at that upper limit. For example, this study's training data set contains mixtures of 1, 2, 3, and 4 contributors; if a significant number of 5- and 6-contributor samples were included, some of the samples currently (and correctly) classified as 4 number of contributors would likely be misclassified when the model is forced to consider a larger number of possible classes. Similarly, 5- and 6-contributor samples might be incorrectly classified as 4-contributor samples. One way to indirectly account for overestimation of 4-contributor classification accuracy is to observe that same overestimation in 3-contributor classification when all 4-contributor samples are removed. The reduction in accuracy for 3-person mixture classification can be considered a plausible lower bound for reduction in this study's 4-contributor classification accuracy.

3. Results

Kullback–Leibler divergence was calculated for ten candidate features (Table 1), and the base pair size of a locus was removed from the list of candidate features after achieving a divergence of 0. All other features were kept, and used in subsequent machine learning.

Summary metrics for all learning algorithms are found in Table 2. Sample sizes for each class in the testing data set are as follows: 94 samples with 1 contributor, 155 samples with 2 contributors, 74 samples with 3 contributors, and 29 samples with 4 contributors. The total testing set (352 samples) included 753 sample-locus instances in 159 samples where allele dropout occurred (using the PACE dynamic analytical threshold). A non-linear support vector machine using a radial basis function kernel produces a tightly converging model with high classification accuracy rates; it scores second-highest of all classifiers in both convergence and accuracy, and lacks the tendency to over fit associated with CART, making it the preferred candidate for subsequent analysis. The learning curve for the non-linear SVM model (Fig. 1) illustrates model convergence for the top-performing algorithm's model.

The performance of the SVM-derived classification model (PACE) was compared to the MAC method using the testing data set. Four analytical thresholds were compared; the dynamic threshold used by PACE as well as 50rfu, 100rfu and 150rfu (Fig. 2).

Table 2
Summary metrics for six machine learning algorithms' learned models for number of contributor classification. Training and testing accuracies are used to evaluate model convergence, and represent total accuracy across all 4 classes. Hyperparameter tuning was limited to hyperparameters impacting model variance, and the reported metrics describe optimized models.

Classifier	Number of Contributors	Precision	Recall	f1-score	Informedness	Training/testing Accuracy
k-NN	1	0.96	0.99	0.98	0.940	0.981/0.955
	2	0.98	0.97	0.97		
	3	0.98	0.87	0.92		
	4	0.79	0.98	0.88		
CART	1	0.97	1.00	0.98	0.965	0.974/0.975
	2	0.98	0.98	0.98		
	3	0.99	0.93	0.96		
	4	0.93	0.98	0.96		
Logistic regression	1	0.97	0.98	0.97	0.949	0.963/0.961
	2	0.97	0.98	0.98		
	3	1.00	0.89	0.94		
	4	0.83	1.00	0.90		
MLP	1	0.97	0.96	0.96	0.943	0.970/0.962
	2	0.96	0.97	0.96		
	3	0.96	0.95	0.96		
	4	0.95	1.00	0.97		
SVM (linear)	1	0.91	0.96	0.94	0.842	0.912/0.894
	2	0.89	0.90	0.89		
	3	0.89	0.77	0.82		
	4	0.88	0.96	0.92		
SVM (non-linear)	1	0.96	0.99	0.97	0.957	0.982/0.971
	2	0.98	0.97	0.97		
	3	1.00	0.93	0.96		
	4	0.93	1.00	0.96		

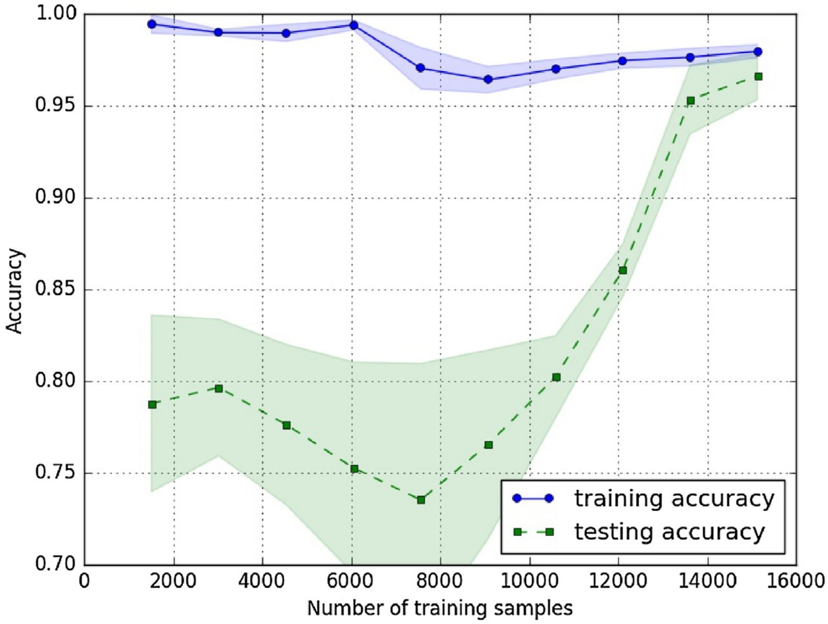

Fig. 1. Learning curve for a number of contributor estimation model derived from a support vector machine with a Gaussian kernel. The shaded area represents one standard deviation. Testing accuracy: 0.980. Note: the number of samples in this figure represent the number of sample-locus instances; therefore, because the Identifiler® amplification kit was used (15 STR loci per sample), 9000 sample-locus instances would be equivalent to 600 samples (9000/15 = 600).

The training data set (352 samples) displayed differing levels of dropout based on the analytical threshold being used. The application of the PACE-LSST threshold yielded 753 instances of allele dropout out across 159 samples, whereas 1739 alleles dropped out across 188 samples when using a 50rfu threshold, 3169 instances of dropout over 224 samples when using a 100rfu threshold, and 4329 dropped out across 251 samples when using a 150rfu threshold; see Table 1 in Appendix A in Supplementary material for sample-specific data. As expected the accuracy of the MAC method decreases with increasing numbers of contributors and thresholds. The dynamic threshold used with a trial of MAC and PACE are comparable at up to 2 contributors and have consistently higher accuracy rates than all other MAC trials. PACE results clearly differentiate themselves at 3 contributor estimation and above. Ultimately, the PACE SVM-derived model outperformed MAC in all cases. A direct comparison between PACE and NOCIt was untenable within the scope of this study, as the time and computational resources required to evaluate NOCIt using PACE's sizable testing set is significantly prohibitive.

Training and testing sets were compiled from the aforementioned samples using a proportionally stratified sampling of the overall data set (Table 3, Appendix A Table 4 in Supplementary material). The contributor classes (e.g. 1, 2, 3 or 4 contributors) were proportionally represented in the training and testing sets, with no overlap in the samples included in each set; classifications resulting from the samples in the testing set are therefore independent of the samples used to create the model. The overall model accuracy is 98.5%, meaning that 98.5% of the sample classifications (i.e. 1, 2, 3 or 4 contributors) were correct based on a comparison of the model's classifications with the known number of contributors. Classification of unknown single source and 4-contributor samples yielded 100% accuracy, with 94 and 29

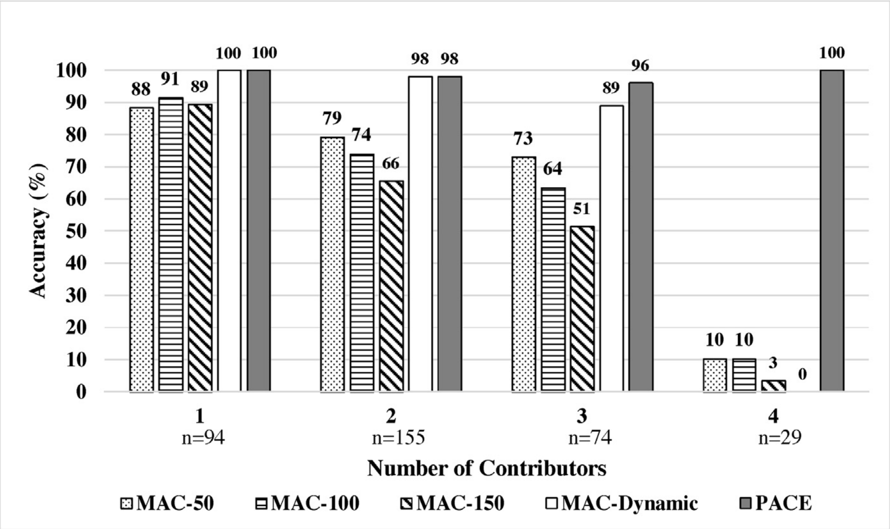

Fig. 2. Accuracy rates for several number of contributor estimation models, PACE (dynamic threshold), MAC at 50rfu, 100rfu, 150rfu and dynamic threshold.

Table 3
Sample sets used for the training and testing of number of contributor classification models created by machine learning algorithms. Stratified sampling was used to ensure a proportional representation of each contributor class in the 2 distinct sample sets.

Contributor Number	Training Set		Testing Set	
	Sample Count	Percentage	Sample Count	Percentage
1	290	27.3%	94	26.7%
2	457	43.0%	155	44.0%
3	202	19.0%	74	21.0%
4	113	10.6%	29	8.2%
Total	1062		352	

samples, respectively. The 2- and 3-contributor samples displayed 98.1% (152/155) and 95.9% (71/74) accuracy, respectively (Table 4).

All samples with incorrect classifications were misclassified by a maximum of ±1 contributor; for example, the 3 misclassifications in the 2-contributor group were misclassified as a single source, and the 3 samples misclassified in the 3-contributor group were classified as either 2-contributor or 4-contributor samples. Figs. 3 and 4 display the accuracy of the model across the DNA template (ng) amplified, with expanded data regarding misclassification shown in Table 5, with the ratio of contributors further explaining the misclassifications given the DNA template amount per contributor within a sample. (As anticipated, 3 of the 6 misclassifications occur at low DNA template amounts, at or below 0.25 ng of template DNA amplified, which is below the typical 1.0 ng target template amount for single source samples. The 2-contributor misclassified samples with 2.0 ng of template DNA amplified were both mixtures of 2 contributors at a 1–19 ratio, therefore the minor component is expected to contribute approximately 0.1 ng total to the mixture. Additionally, these two samples exhibit nearly equal probabilities in the 1 and 2 contributor classes (Table 5). Finally, one of the 3 contributor 7.0 ng samples was misclassified as a 4 contributor sample. As expected, this sample had exceptionally high baseline likely leading to the misclassification. It is unlikely a laboratory will encounter samples with greater than 2.0–3.0 ng of template DNA amplified.

Electropherograms of 5 of the 6 misclassifications can be found in Appendix A in Supplementary material, Figs. 1–5; one electropherogram was not included due to privacy requests. The samples misclassified by PACE were also assessed using MAC-based contributor estimations which yielded inconsistent results (Appendix A in Supplementary material, Table 5). Although in some cases MAC results identify the correct contributor number, PACE has higher overall accuracies and provides more consistent results.

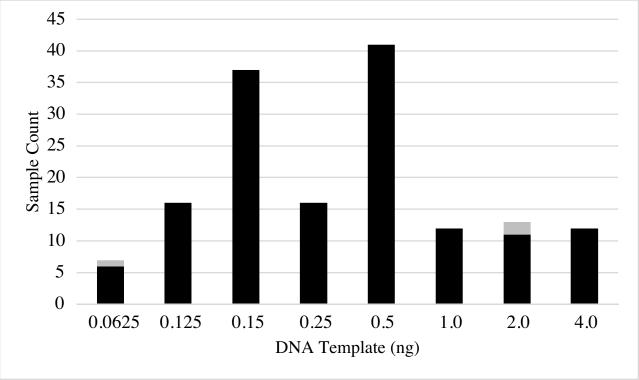

Fig. 3. Accuracy rates for number of contributor classification of 2-person mixtures at various template amounts. Correct (■); Incorrect (░).

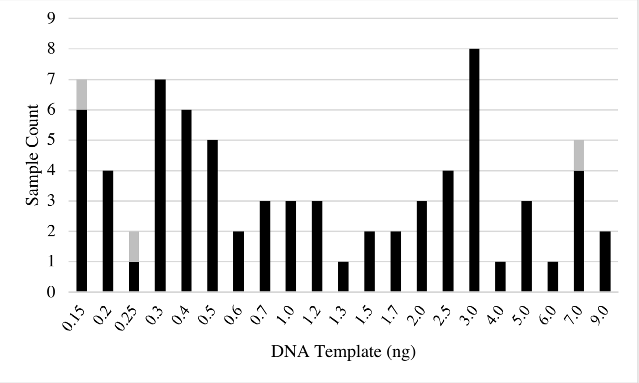

Fig. 4. Accuracy rates for number of contributor classification of 3-person mixtures at various template amounts. Correct (■); Incorrect (░).

A second model created from a machine learning algorithm exposed only to 1–3 contributor DNA samples was compared to the primary model described above, which was learned from an algorithm exposed to the entire data set (1–4 contributors), with resulting classification accuracy rates shown in Table 6. A “1 to 3 contributors” model should overestimate the accuracy of 3-person mixture classification, and a comparison of the 2 models’ accuracy rates for 3-person mixtures provides a potential lower bound of

Table 4
(A) – Number of contributor classification model accuracy rates when used to classify samples from the testing set. (B) – A confusion matrix representing the classifications of those data by PACE. Italicized values represent the number of correctly called samples.

(A)

Contributor #	% Correct	Incorrect Count	Correct Count	Over-estimate	Under-estimate
1	100%	0	94	0	0
2	98.1%	3	152	0	3
3	95.9%	3	71	1	2
4	100%	0	29	0	0

(B)

		Predicted Number of Contributors			
		1	2	3	4
Actual Number of Contributors	1	*94*	0	0	0
	2	3	*152*	0	0
	3	0	2	*71*	1
	4	0	0	0	*29*

Table 5
Summary of misclassifications by the PACE number of contributor classification model. The proportion of allele dropout was calculated by dividing the total number of alleles below the analytical threshold in a sample by the total number of alleles expected in the sample. Italicized values represent the maximum probabilities based on PACE output.

Contributor Number	Sample ID	DNA Template (ng)	Ratio of Contributors	Percentage of Dropout	PACE Model Estimate	PACE Contributor Probabilities			
						Pr(1)	Pr(2)	Pr(3)	Pr(4)
2	225	2.0	1–19	9.3	1	*0.53808*	0.46164	0.00028	0.00000
	226	2.0	1–19	14.0	1	*0.55271*	0.44729	0.00000	0.00000
	295	0.0625	1–19	34.9	1	*0.92698*	0.07302	0.00000	0.00000
3	317	7.0	3 to 1–1	0	4	0.00791	0.01679	0.01650	*0.95880*
	251	0.15	1 to 1–3	21.7	2	0.00865	*0.88060*	0.00208	0.00000
	262	0.25	1.5 to 3–1	22.4	2	0.00211	*0.99549*	0.00240	0.00000

Table 6
Number of contributor classification model accuracy rates for 2 separate models. The first model is trained using only 1–3 contributor samples, while the second model is also exposed to 4 contributor samples.

Contributor #	1–3 contributor accuracy	1–4 contributor accuracy
1	1.0	1.0
2	0.98	0.98
3	0.98	0.96
4	N/A	1.0

Table 7
A confusion matrix summarizing the accuracy of the PACE system given the truncated-degradation training data set with CSF1PO, FGA, D18S51 and D2S1338 removed from samples that both had a total template DNA amount of 0.25 ng or lower and a ratio of major to smallest minor no greater than 3:1 (this included 1:0, 1:1,1:2/2:1,1:4/4:1, 1:1:3/3:1:1/1:3:1,3:3:1:1, 1.6:3:1:1).

		Predicted Number of Contributors			
		1	2	3	4
Actual Number of Contributors	1	93	1	0	0
	2	32	123	0	0
	3	0	12	61	1
	4	0	4	0	25

overestimation for the highest-numbered class in a model. A reduction in classification accuracy for 3-person mixtures of 2% was observed, suggesting that the primary model's 4-person classification accuracy is likely overestimated by at least the same amount.

The truncated-degradation testing data set mirrored the previously described testing data however large loci (CSF1PO, FGA, D18S51 and D2S1338) were removed from samples that had both a total template DNA amount of 0.25ng or lower and a ratio of major to smallest minor no greater than 3:1 (this included 1:0, 1:1,1:2/2:1,1:4/4:1, 1:1:3/3:1:1/1:3:1,3:3:1:1, 1.6:3:1:1). This truncation was performed to both demonstrate the performance of PACE given fewer loci but also with simulated locus-dropout. This set is not meant to replace the need for empirically degraded samples in the testing data set but to demonstrate PACE's potential performance when samples such as these are encountered.

The confusion matrix yielded a significant increase in the number of misclassified 2 and 3 contributor samples (Table 7). This result is not unexpected due to the relative information content within the larger loci – the loci having the highest levels of allelic diversity. Despite this result, the maximum probability for the 50 misclassified samples ranged from 0.392–0.995, with 49 of the 50 samples having a maximum probability below 0.96 (48 of 50 below 84.3%). This is an encouraging result because the decrease in overall information content is captured in the PACE probabilities. A threshold of 0.99 could be applied, and, based on the truncated-degraded testing set, only 1 sample would be incorrectly classified; a 1.5:3:1 3-contributor sample (individuals BAC) with 0.25 ng of template DNA

4. Discussion

The proposed probabilistic method for estimating the number of contributors is a robust and reproducible method that was developed using an expansive data set comprised of samples amplified using the AmpFLSTR® Identifiler® PCR Amplification Kit (ThermoFisher Scientific Inc.). Our focus on the Identifiler® data set was due to current availability of samples; however, the method is applicable to any amplification system in use. Similar training data sets can be compiled from multiple laboratories' validation studies, with additional samples run as needed. A noteworthy aspect of this method is its independence based on instrument and injection parameters. The data presented were compiled from 5 different capillary electrophoresis instruments at 2 different laboratories, and had varied injection times (2–22 s) and kV used for injection (1–5 s). This is a significant advantage that would permit the model to be easily transferred between laboratories and would not require significant resources to perform internal validation.

A shortcoming of the data set is the lack of degraded and inhibited samples. The data set contained low DNA template samples (0.0125 ng to 0.0625 ng) that experienced a level of degradation common for such samples. We believe that the system would benefit from having degraded and inhibited samples included in the training set. The inclusion of the truncated-degradation training set and the use of low template samples that display typical degradation patterns indicate that generalizability and overall accuracy of the system may be significantly impacted during cases of extreme locus dropout, though less so if allele dropout is present in the absence of locus dropout. The decreased accuracy observed when locus dropout was simulated is not unexpected and will likely be mitigated through the adoption of amplification kits that include greater than 15 loci such as the PowerPlex® Fusion 6C System (Promega Corporation), GlobalFiler™ PCR Amplification Kit (ThermoFisher Scientific Inc.) and Qiagen Investigator® 24plex Kit (Qiagen N.V.). We anticipate that the effect will have minimal impact on the generalizability or overall classification accuracy. Finally, the data presented in this study focus on samples amplified using 28 cycles and the Identifiler® kit (ThermoFisher Scientific Inc.). While the PACE system operates independent of instrument, injection time and voltage, we anticipate that if different cycle numbers are used a stand-alone training set would need to be created or acquired. All other aspects of model learning via PACE would remain unchanged.

Injection and injection voltage were purposefully excluded from the feature vector, to test whether the predictive models could function independently of the variables. A predictive model that includes these features was created separately, however; the resulting model accuracy for predicting the number of contributors was slightly worse than what was reported in Table 2, suggesting that these features carry little predictive value. Therefore, excluding the injection time and injection voltage permits both higher accuracy, increased correct calls, and but also allows the system to be used independent of these features.

While eight of the nine initial candidate features were retained for machine learning, it should be noted that these are merely an initial set of candidates that impacted successful classification. Other candidates could increase (or decrease) classification accuracy if included. Of specific interest would be the inclusion of features more commonly associated with the process of mixture deconvolution, such as the ratio of contributors. PACE is specifically designed to be “feature agnostic”, and can construct classification models using feature vectors containing any type of data, whether they are numeric, nominal, binary, or even character-based. This feature of PACE will allow it to be seamlessly integrated into DNA sequence based forensic workflows.

Most of the candidate machine learning algorithms evaluated in this study produced similar accuracy rates and variances. A linear support vector machine performed noticeably worse, suggesting that the data are poorly separable using a linear decision boundary. While a non-linear SVM exhibited a much higher classification accuracy, a secondary benefit of using such a learning algorithm for model construction is its tendency to generalize well; in comparison, decision trees and *k*-NN classifiers are known for greater likelihood of learning models that overfit training data.

The overall accuracy of the model across all testing samples is over 98.5%, with only 6 misclassifications observed in the 2- and 3-contributor sample groups. The 100% accuracy experienced in the 4-contributor group is in part due to the lack of 5 (or greater) contributor samples for training; this group can be more accurately considered a “$\geq$4 contributor” group. Based on analysis of overestimation for 1–3 contributor classification, this model's accuracy would be expected to drop by at least 2% from the observed 100% classification accuracy if 5 contributor samples are included in the training data, for example. The practical utility of classifying a profile with having $\geq$4 contributors is significant, as many laboratories choose to not interpret DNA profiles with greater than 3 contributors. The system could be further strengthened through the inclusion of degraded and 5- contributor samples in future training and testing sets.

The PACE system is proposed as a valuable tool in the analyst assessment of the number of contributors. Of the 6 misclassifications (out of 352 total samples), 3 of the samples could be corrected if an analyst briefly reviewed the data, through the identification of artifactual peaks such as minus A and pull-up and indications of peaks below thresholds. The remaining 3 instances were due to allele dropout, allele sharing and high template effects such as elevated stutter, whereas an analyst or software did not have significant evidence to accurately predict the number of contributors. With analyst input the model has an accuracy rate of over 99.0% when evaluated using the testing data set. In addition, if imposing a threshold of a PACE probability of 99.0% only 1 sample would be incorrectly classified in both the full testing set and the truncated − degradation testing set.

Regardless, a formal evaluation of a machine learning-derived model alongside other top performers in a variety of laboratories operating under a variety of conditions may be of great interest to the community. While model validation via testing data performance was a vital component of this study, the potential for poor model generalization still exists if the underlying training data set poorly reflects the reality of DNA mixtures. And while the dataset itself is arguably massive enough to allay such concerns, any novel approach in a scientific discipline connected to law and court-based proceedings is likely to face increased levels of scrutiny and mistrust, and should be held to extremely high standards. Laboratory-based validation is therefore a necessary subsequent stage in this research.

4.1. Validation recommendations

Two plausible methods for implementation into standard sample analysis workflows are the following: (1) Use existing models to test laboratory-specific data sets to ensure the results that are obtained through PACE are within laboratory acceptable limits. This method mirrors common procedures for implementing new fragment analysis software. (2) Create new models using a laboratory-specific PACE training set that combines the existing PACE training data set with laboratory specific samples. Following model creation, one would test the system using samples that were not included within the training set to ensure the PACE output is within laboratory-specific limits.

5. Conclusion

We have presented a novel, highly accurate and rapid means of estimating the number of contributors in DNA mixtures. Achieving high classification accuracy, especially with complex mixtures containing many contributors, is a vital prerequisite to full mixture deconvolution, whether using basic manual mixture deconvolution or more advanced semi- and fully continuous probabilistic software suites. Apart from the high level of accuracy of this method, a key aspect is its lack of computational expense, regardless of the sample's complexity, rapid estimation can be completed using any standard laptop or desktop computer. The method achieves this through the use of machine learning, which leverages an initial training and testing data set to build the model; all of the computational “heavy lifting” is performed during data acquisition and model creation, as opposed to algorithms such as Markov Chain Monte Carlo methods that achieve their highest computational burden during classification This imparts both speed and reproducibility onto the end user, who will likely not be concerned with acquiring training data or evaluating candidate models. Although its purpose is not to fully automate the analytical process, the PACE system is proposed as a valuable tool in the assessment of the number of contributors. We believe that the field of forensic DNA analysis can greatly benefit from embracing machine learning as a key tool to combat complexity in analyses.

Conflicts of interest

None.

Acknowledgements

Data were provided by the Onondaga County Center for Forensic Sciences (Forensic Biology/DNA section), the Office of the Chief Medical Examiner New York City (Department of Forensic Biology), and Boston University. The authors would also like to thank the National Institute of Justice, Laura Haarer for invaluable assistance during software development, Kathleen Corrado Ph.D., Kathleen Hum M.Sc. and the Onondaga County CFS Forensic Biology/DNA section, Kevin Sweder Ph.D. and Chilukuri Mohan Ph.D., David Fisher M.S., Desarae Harmon M.S., Don Siegel Ph.D., and Timothy Kupferschmid MBA MFS.

This project was supported, in part, by Award No. 2014-DN-BX-K029, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2016.11.006.

References

[1] 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, Accessed: June 7, 2011.

[2] 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 (6) (2012) 697–707.

[3] 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.

[4] T.M. Clayton, J.P. Whitaker, R. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Sci. Int. 91 (1) (1998) 55–70.

[5] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Academic Press, 2014.

[6] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J. Forensic Sci. 50 (6) (2005) 1361.

[7] 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 (3) (2011) 314–326.

[8] 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 (1) (2007) 20–28.

[9] H. Haned, L. Pène, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count? J. Forensic Sci. 56 (1) (2011) 23–28.

[10] 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.

[11] T. Egeland, I. Dalen, P.F. Mostad, Estimating the number of contributors to a DNA profile, Int. J. Legal Med. 117 (5) (2003) 271–275.

[12] H. Haned, L. Pène, 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 (4) (2011) 281–284.

[13] D. Taylor, J.A. Bright, J.S. Buckleton, Interpreting forensic DNA profiling evidence without specifiying the number of contributors, Forensic Sci. Int. Genet. 13 (2014) 269–280.

[14] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt: a computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci. Int. Genet. 16 (2015) 172–180.

[15] P. Flach, Machine Learning: the Art and Science of Algorithms That Make Sense of Data, Cambridge University Press, 2012.

[16] P. Duygulu, K. Barnard, J.F. de Freitas, D.A. Forsyth, Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, Springer Berlin Heidelberg, 2002, pp. 97–112 Computer Vision ECCV.

[17] D. Jurafsky, J.H. Martin, Speech and Language Processing, Pearson, 2014.

[18] S.B. Cho, H.H. Won, Machine learning in DNA microarray analysis for cancer classification, Australian Computer Society, Inc, Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics, 192003, pp. 189–198.

[19] Y. Bengio, Y. Grandvalet, No unbiased estimator of the variance of k-fold cross-validation, The Journal of Machine Learning Research 5 (2004) 1089–1105.

[20] R.E. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 10–18.

[22] D.H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural Comput. 8 (7) (1996) 1341–1390.

[23] L. Breiman, J. Friedman, C. J.Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[25] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2002), 2002.

[26] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classifiers 10 (3) (1999) 61–74.

[27] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, Proceedings of the 22nd International Conference on Machine Learning, ACM (2005), 2005.