Technical note: proposed method to objectively evaluate gunshot residue comparisons does not generalize to different-location settings

Bie, K. de; Vinkenoog, M.; Ariëns, S.; Dirkse, T.L.; Kukurin, C.; Ham, C.J.M. van der; ... ; Ypma, R.J.F.

Forensic Science International 369 (2025) 112414

Contents lists available at ScienceDirect

Forensic Science International

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

Technical note: Proposed method to objectively evaluate gunshot residue comparisons does not generalize to different-location settings

K. de Bie [a], M. Vinkenoog [a,b,*], S. Ariëns [a], T.L. Dirkse [a], C. Kukurin [a], C.J.M. van der Ham [a], W. Bosma [a], R.J.F. Ypma [a]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497 GB, the Netherlands
[b] Leiden University, Rapenburg 70, Leiden 2311 EZ, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Statistics
Likelihood ratios
Gunshot residue analysis
Forensic science

ABSTRACT

Previously, we proposed a likelihood ratio system for pairwise source comparison of gunshot residue (GSR) samples based on elemental composition. Only pairs of GSR samples from the same location type (e.g., hand-hand) were considered, as the data originates from casework and ground truth is not available for samples from different location types (e.g., hand-cartridge case). This lack of sample pairs taken from different locations is a limitation, as casework will usually require such sample pairs to be evaluated. In this study, we test the impact of this sampling location limitation by evaluating the model with an experimental dataset for which ground truth was available for same-location as well as different-location pairs. We find a sharp decline in the performance of the model, with C_{llr} deteriorating from 0.35 to 0.98. Additional exploration of various system extensions does not lead to significant improvements. We discuss the potential causes for this decline and conclude the system is not currently ready for application in forensic casework.

1. Introduction

The use of machine learning methods in the development of likelihood ratio (LR) systems is increasingly explored in forensic sciences. In 2022, Matzen et al. [1] published an LR system for comparative analysis of gunshot residue (GSR). GSR analysis involves detecting and analyzing particles formed during and after firing a shot with a firearm. These particles are expelled by the firearm and deposited on people and objects in the vicinity or may remain inside the firearm and the cartridge case. In forensic comparative GSR analysis, collections of particles found on suspects and/or objects are compared to link a person or an object to a crime scene. Scanning electron microscopy (SEM) coupled with energy dispersive X-ray (EDS) analysis is commonly used to analyze inorganic GSR particles [2]. The method proposed in [1] leverages the elemental composition of the particles as measured by SEM/EDS to construct an LR system.

The dataset used to train and evaluate the LR system described in [1] suffers from an important limitation. To train the model, pairs of GSR stubs are needed, for which it is known whether they originate from the same source or a different source. Here, by "source" we mean a shot or subsequent shots from the same firearm and with the same ammunition type. Because the dataset originates from casework, it is never certain if two stubs that are sampled at different location types (e.g., one is sampled from a suspect's hand, the other from the victim's clothes) originate from the same source. Therefore, for the same-source pairs, only stubs from the same location (e.g., two stubs from the same suspect's hand) are included. However, for different-source pairs, pairs from different location types could be constructed, as it is assumed that two stubs from different cases are not related. We refer to the constraint that the dataset does not include same-source, different-location comparisons as the *sampling location limitation*. As described in [1], this limitation means that the model performance may be overestimated: the LR system is only required to discriminate same-source, same-location pairs from different-source, (mostly) different-location pairs, which is an easier task than discriminating between same-source, different-location pairs and different-source, different-location pairs.

To be useful for casework, the LR system should be able to discriminate between same- or different-source pairs in a different-location setting. In this study, we evaluate the system performance on different-location pairs, using a dataset of GSR samples that were collected in an experimental setup. This way, same-source, different-location pairs can be constructed, as the origin of each shot is known. We

* Corresponding author at: Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague 2497 GB, the Netherlands.
E-mail address: m.vinkenoog@liacs.leidenuniv.nl (M. Vinkenoog).

https://doi.org/10.1016/j.forsciint.2025.112414
Received 4 July 2024; Received in revised form 6 January 2025; Accepted 20 February 2025
Available online 21 February 2025

find that the system's performance as reported in [1] does not hold for this dataset. Instead, performance declines to a level close to random. Additionally, we explore various extensions of the system and show that none of them lead to an acceptable improvement in performance. We explore several potential causes of this performance drop and conclude the system is not suitable for application in practical casework settings.

2. Datasets

We use the *mixed samples dataset* described in [1] to reproduce the LR system. This dataset consists of 1953 stubs from 210 criminal cases. As explained above, for same-source samples, only same-location stub pairs can be constructed. For further details regarding the mixed samples dataset, we refer to [1]. In addition, to test the sampling location limitation, we use an *experimental dataset* of GSR measurements for which the ground truth is known. It consists of 46 series of shooting experiments, where stubs were collected and measured at several locations, which allows for construction of different-location stub pairs. Location types include the cartridge cases (one or two stubs in each experiment), the bullet hole (one stub), and a proxy for the shooter's hand (one stub placed next to the weapon). The dataset contains a range of weapon and ammunition types. In total, 178 stubs were measured.

Within the experimental dataset, same-source as well as different-source pairs are constructed, where same-source pairs contain two stubs from one experiment with the same weapon and ammunition lot and different-source pairs contain stubs from two different experiments, similar to different shooting incidents in casework. In each case, the stubs in the pair can be from the same location type, referred to as same-location (i.e. cartridge case-cartridge case) or from two different location types, referred to as different-location (hand-cartridge case, hand-clothes or cartridge case-clothes). Table 1 shows how many pairs of each type can be constructed. Note that not all possible different-source pairs are used to train and evaluate the model in order to have balanced outcome classes, as described in [1].

For the experimental dataset, stubs were analyzed using a SEM/EDS according to the standard operating procedures based on ASTM 1588-20 [3]. The device used was a JEOL IT300-HR equipped with two Oxford X-max 50 detectors, used in backscatter mode. The automated particle detection and analysis was performed with the Aztec software (version 3.4, Oxford Instruments). The measurement files were exported using Aztec version 6.0 to the open format H5OINA to allow further analysis of the raw data.

3. Performed experiments

To create the LR system, we use the approach outlined in [1], with identical hypotheses, data preprocessing and modelling steps. In summary, for a given stub, only relevant GSR particles are filtered in order to get rid of background/environmental noise. For these particles, the weight percentages of 89 elements (see Appendix) are converted to a rank (i.e., their relative value across the entire dataset), after which their average is taken across all GSR particles on the stub. The feature vector for a pair of stubs is constructed by calculating the absolute difference between each element, resulting in a vector of length 89. In addition, a single Bhattacharyya coefficient is calculated to capture some of the variance in the sample (see [1] for details). An XGBoost model [4] with a logistic calibrator is trained to discriminate between same-source and different-source pairs. To evaluate the LR system, we use the C_{llr}, a common metric for the evaluation of LR systems, which takes into account both separation and calibration. Values closer to 0 indicate a better performance, and a C_{llr} of 1 equals a random performance [4]. Matzen et al. [1] report a C_{llr} of 0.35 on the mixed samples dataset using cross-validation, which means that parts of the same dataset were used for training, calibration and evaluation.

We aim to replicate the results from [1] on the experimental dataset, which unlike the mixed samples dataset contains same-source, different-location pairs and does not suffer from the sampling location limitation. In initial experiments, we used the mixed samples dataset to train the model, 50 % of the experimental dataset to calibrate the scores and the remaining 50 % to evaluate the results. First, we exclude the same-source, different-location pairs (as in the mixed samples dataset), to verify that the LR system performs well for the experimental dataset if these pairs are absent. Next, we use all stub pairs, including the same-source, different-location pairs, to assess the effect of the sampling location limitation. As this showed a poor performance (see Results section), various modifications and extensions of the method were conducted in an attempt to achieve better results. These experiments were performed using cross-validation on the experimental data only, rather than training on the mixed samples dataset and calibrating/evaluating on the experimental data. Moreover, to mitigate computational costs, the inclusion of the Bhattacharyya coefficient was

Table 1
The number of pairs that can be constructed from the 178 stubs measured within the 46 shooting series in the experimental dataset.

	Same-source	Different-source	**Total**
Same-location	45	5.905	**5.950**
Different-location	213	9.590	**9.803**
Total	**258**	**15.495**	**15.753**

Table 2
Modifications and extensions used in the additional experiments. The options corresponding to the method described in the original paper are shown in boldface.

Type of modification/ extension	Options	Additional information
Element selection	**89 elements** 63 elements 40 elements	GSR experts reduced the list of 89 elements in two steps, first removing all elements that do not contain GSR information (resulting in 63 elements), then removing all elements that rarely contain GSR information (resulting in 40 elements). See Appendix for the elements.
Element feature	**Percent rank** Weight percentage	Weight percentages are returned by the SEM/EDS software. These may be replaced by percent ranks, indicating the typicality of the sample.
Feature aggregation	**Mean** Weighted mean Normalized mean	The method used when aggregating over all particles on a stub; either the mean, weighted mean (weights based on the relevance level of the particle classes) or the normalized mean (for weight percentages: when fewer than 89 elements are used, the weight percentages are first re-calculated to sum to 1).
Adding standard deviation	**No** Yes	Instead of aggregating over all particles on a stub using only the mean, the standard deviation can also be added.
Adding sampling location	**No** Yes	Add a one-hot encoded vector of the sampling location of both stubs (either 'hand', 'clothes', or 'cartridge case') to the feature vector, increasing the vector length by 6.
Particle classes instead of elements	**No** Yes	Instead of using the mean percent rank (or weight percentage) of each element over the particles on a stub, each stub is described by the proportion of particles belonging to different particle classes.
Pairing of vectors	**Absolute difference** Concatenated	When pairing feature vectors of two stubs, the vectors can either be subtracted, yielding the absolute differences, or concatenated, resulting in a vector twice the length of a single stub vector.

omitted, given its minimal impact on the C_{llr}. Table 2 presents the different modifications and extensions that were attempted.

4. Results

Fig. 1 shows the distribution of the LRs calculated by the model, which was trained on the mixed samples dataset, then calibrated and evaluated on the experimental dataset, excluding same-source, different-location pairs as previously described. The C_{llr} of 0.19 confirms that the LR system performs well on the experimental dataset and verifies that we can reproduce the results reported in [1], given the sampling location limitation.

Fig. 2A shows the distribution of LRs for the system that is trained on the mixed samples dataset, then calibrated and evaluated on the full experimental dataset (with same-source, different-location pairs included). The distributions of the LRs for same- and different-shot stub pairs almost completely overlap, which indicates that the model is not able to discriminate between these pairs. The C_{llr} of 0.98 confirms that the LR system's performance is indeed close to random when the sampling location limitation is lifted. Fig. 2B and 2C show the LR distributions separately for same-location and different-location pairs respectively, using the same model. For the same-location pairs, as shown in Fig. 2B, the LR system is able to discriminate between same- and different-shot pairs reasonably well, which is in line with the results presented in [1]. However, for the different-location pairs (Fig. 2C), the distributions for the same- and different-shot pairs are nearly identical, showing that the system cannot discriminate between these pairs. Note that the system was calibrated only once (for all pairs in the calibration set), which explains why the LR distributions in Fig. 2B and Fig. 2C are not centered around a $\log_{10}$ LR of 0. For the same reason, we only calculate the C_{llr} for the full evaluation set.

In an attempt to improve performance on the experimental dataset, we perform additional experiments incorporating minor adjustments to the method proposed in [1]. We attempt all modifications outlined in Table 2. For these experiments, we train, calibrate and evaluate on the full experimental dataset (using cross-validation). None of these experiments lead to a significant improvement of the C_{llr} beyond the value of 0.98 reported for the initial settings, and we are unable to replicate the C_{llr} of 0.35 reported in [1] for the mixed samples dataset and 0.19 for the filtered experimental dataset at the beginning of this section.

4.1. Analysis of the feature vector using the sum of absolute differences

To explain the observed drop in system performance between the mixed samples dataset and the experimental dataset, an investigation of the feature vector was performed. We consider the four groups of pairs (same/different shot and same/different location) in the part of the experimental dataset used for evaluation of the LR system. Each pair of stubs is represented by a vector of length 89, corresponding to the 89 elements that were used in the analysis. For each element, the average weight percentage rank across all particles for each stub is calculated. Then, we compute the feature vector as the absolute difference between the average weight percentage ranks for the two stubs in the pair. If the two stubs in the pair are identical, all 89 absolute differences are equal to 0, obtaining a sum of absolute differences equal to 0. In contrast, for stubs that have very different weight percentage ranks for all elements, we expect very large absolute differences and therefore a very large sum of absolute differences. In this way, the sum of absolute differences allows us to quantify the degree to which the average weight percentage ranks differ across different subsets of stub pairs.

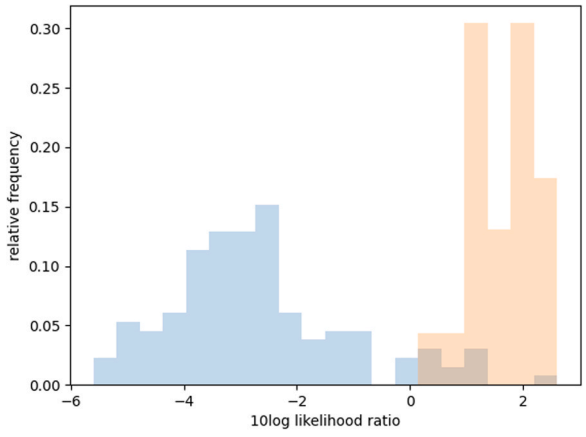

Fig. 1. Distribution of the $\log_{10}$ LRs after training the model on the mixed samples dataset, and calibrating and evaluating on the experimental dataset excluding same-source, different-location pairs. Ground truth is indicated by color: blue indicates different-source pairs and orange indicates same-source pairs.

The distributions of the sums are shown in Fig. 3. This analysis shows that same-shot, same-location pairs have very low sums, indicating that the feature vector is very similar for both stubs in the pair. Moreover, we observe a narrow distribution across these pairs, showing that they tend to have similar feature vectors across the dataset. In contrast, all different-location pairs as well as same-shot, different-location pairs have much larger sums, as well as a much wider distribution. This indicates that GSR samples from the same shot, sampled at the same location, are much more similar to each other than all other pairs. This implies that with the sampling location limitation (and in the absence of same-shot, different-location pairs), the LR system is essentially only required to recognize whether the absolute difference vectors of a stub pair contain small or large values in order to achieve a good performance.

5. Discussion

In this study, the LR system proposed in Matzen et al. [1] to differentiate same-shot from different-shot GSR samples is validated on an experimental dataset to test its main limitation: the absence of same-shot, different-location stub pairs in the mixed samples dataset used for developing the original LR system. We find that the LR system can discriminate between same-location pairs in the experimental dataset but not different-location pairs. C_{llr} values indicate near-random performance and did not improve sufficiently with several modifications and extensions. This is a crucial limitation for the LR system that prevents it from use in casework, as GSR comparisons in this setting are (nearly) always between different locations, such as a stub sampled on a suspect's hand and a stub sampled on the victim's clothes.

As we observe that the model fails specifically for different-location stub pairs, we believe that the drop in performance is due to the more complex nature of distinguishing different-location stub pairs as compared to same-location stub pairs. This hypothesis is further substantiated by the analysis of the feature vectors that are used by the LR system, which strikingly shows that feature vectors for same-shot, same-location stub pairs are much more similar to each other than feature vectors for all other stub pairs. There may be several alternative explanations for the poor performance of the model on the experimental dataset. For example, there may be something particularly challenging about the experimental dataset that does not occur in casework data, or the fact that the dataset was measured using a different electron microscope may have unforeseen effects. However, we believe that these explanations are not sufficient, as evaluation on the experimental data shows that the LR system performs well as long as the sampling location limitation applies to the evaluation dataset. Even when this is not the case, the system can discriminate reasonably well between same-shot and different-shot pairs for same-location stubs, which is in line with the reported results for the mixed samples dataset.

While the present study shows that the method proposed in [1] is not ready to be used in casework in its current form, we believe there are

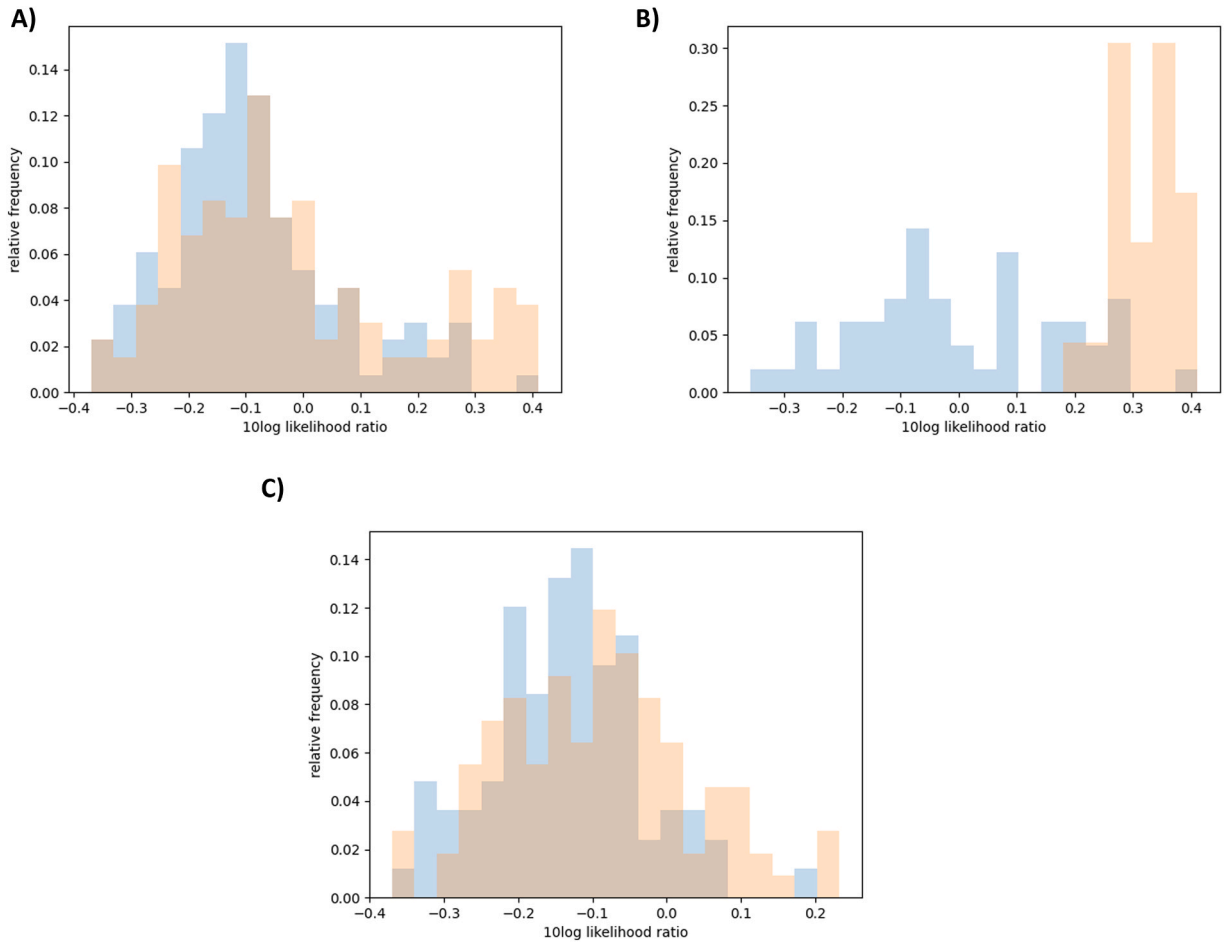

Fig. 2. Distribution of the $\log_{10}$ LRs after training the model on the mixed samples dataset, and calibrating and evaluating on the experimental dataset. Ground truth is indicated by color: blue indicates different-source pairs and orange indicates same-source pairs. A) shows the LR distribution for all evaluation pairs, B) for same-location pairs only and C) for different-location pairs only. Note that calibration was performed only once, for a mix of same- and different-shot samples, which explains why the $\log_{10}$ LRs are not centered around 0 for Fig. 2B and 2C.

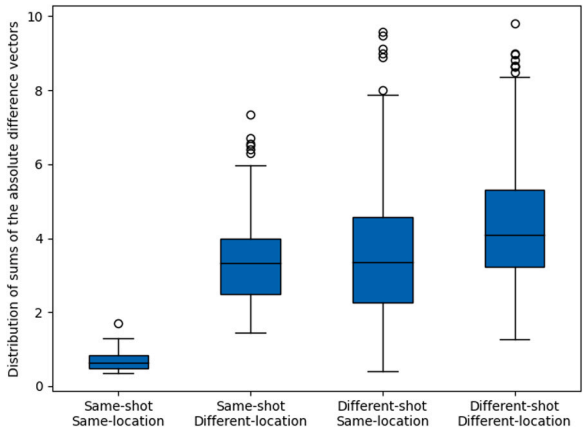

Fig. 3. The distribution of the sums of the absolute difference vectors as used by the LR system for the sample pairs in the evaluation dataset.

many ways forward. As noted in [1], the preprocessing method for constructing the feature vectors leads to a large information loss. In their casework, GSR experts perform an analysis that is much more complex than simply comparing the average weight percentages across all GSR particles in a stub, which is in essence the approach of the present LR system. Amongst others, GSR experts are interested in the variance of the characteristics of the particles on the stub, and this information is lost in the present approach. More sophisticated feature vectors and model architectures may be better equipped to capture the inherent complexity and nonlinear relationships within the data. As described in [1] however, increasing the number of parameters to be fit (either by increasing model or feature vector size) comes with increased requirements regarding the size of the training dataset. However, increasing the dataset size significantly is challenging, given the limited number of measurements available from casework, the absence of ground-truth labels for different-location pairs in casework samples, and the high cost and difficulty of obtaining representative samples through experiments. Alternatively, we suggest focusing on more sophisticated encoding of expert knowledge rather than aiming for automated pattern recognition from data. For example, more refined methods for preprocessing, such as excluding environmental debris, or mimicking the expert's thought process when comparing samples, may allow for a more cautious and thus efficacious method grounded in scientific knowledge.

Declaration of Competing Interest

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

Acknowledgements

We thank Ewout van Dijk, Teus Sloot and Vincent Wesselink for their

software contributions to this study.

Appendix

89 elements: Ac, Ag, Al, Ar, As, At, Au, B, Ba, Bi, Br, Ca, Cd, Ce, Cl, Co, Cr, Cs, Cu, Dy, Er, Eu, F, Fe, Fr, Ga, Gd, Ge, Hf, Hg, Ho, I, In, Ir, K, Kr, La, Lu, Mg, Mn, Mo, N, Na, Nb, Nd, Ne, Ni, Np, O, Os, P, Pa, Pb, Pd, Pm, Po, Pr, Pt, Pu, Ra, Rb, Re, Rh, Rn, Ru, S, Sb, Sc, Se, Si, Sm, Sn, Sr, Ta, Tb, Tc, Te, Th, Ti, Tl, Tm, U, V, W, Xe, Y, Yb, Zn, Zr

63 elements: Ag, Al, As, At, Au, Ba, Bi, Br, Ca, Cd, Ce, Cl, Co, Cr, Cs, Cu, Fe, Fr, Ga, Gd, Ge, Hf, Hg, I, In, Ir, K, La, Mg, Mn, Mo, Na, Nb, Ni, Os, P, Pb, Pd, Po, Pt, Ra, Rb, Re, Rh, Ru, S, Sb, Sc, Se, Si, Sm, Sn, Sr, Ta, Tc, Te, Ti, Tl, V, W, Y, Zn, Zr

40 elements: Ag, Al, As, Au, Ba, Bi, Br, Ca, Ce, Cl, Co, Cr, Cu, Fe, Ga, Gd, Ge, Hg, I, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Se, Si, Sm, Sn, Sr, Ti, V, W, Zn, Zr

References

[1] T. Matzen, C. Kukurin, J. van de Wetering, S. Ariëns, W. Bosma, A. Knijnenberg, A. Stamouli, R.J.F. Ypma, Objectifying evidence evaluation for gunshot residue comparisons using machine learning on criminal case data, Forensic Sci. Int. 335 (2022) 111293, https://doi.org/10.1016/j.forsciint.2022.111293.

[2] R.S. Nesbitt, J.E. Wessel, P.F. Jones, Detection of gunshot residue by use of the scanning electron microscope, J. Forensic Sci. 21 (1976) 595–610, https://doi.org/10.1520/JFS10532J.

[3] E30 Committee. Practice for gunshot residue analysis by scanning electron microscopy/energy dispersive X-ray spectrometry, ASTM International, West Conshohocken, PA (2020). https://doi.org/10.1520/e1588-20.

[4] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232. ⟨https://www.jstor.org/stable/2699986⟩.