Forensic Science International 353 (2023) 111858

Contents lists available at ScienceDirect

Forensic Science International

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

Automated interpretation of comparison scores for firearm toolmarks on cartridge case primers

Martin Baiker-Sørensen [a], Ivo Alberink [a], Laura B. Granell [c], Leen van der Ham [a], Erwin J.A. T. Mattijssen [a,*], Erich D. Smith [c], Johannes Soons [b], Peter Vergeer [a], Xiaoyu A. Zheng [b]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, 2497GB Den Haag, the Netherlands
[b] National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[c] Federal Bureau of Investigation, 2500 Investigation Parkway, Quantico, VA 22134, USA

ARTICLE INFO

Keywords:
Objective comparison
Firearm toolmark
Computer aided system
Interpretation
Likelihood ratio
Weight of evidence
3D topography

ABSTRACT

An automated approach for evaluating the strength of the evidence of firearm toolmark comparison results is presented for *a* common source scenario. First, comparison scores are derived describing the similarity of marks typically encountered on the primer of fired cartridge cases: aperture shear striations as well as breechface and firing pin impressions. Subsequently, these scores are interpreted using reference distributions of comparison scores obtained for representative known matching (KM) and known non-matching (KNM) ballistic samples in a common source, score-based likelihood ratio (LR) system. We study various alternatives to set up such an LR system and compare them using qualitative and quantitative criteria known from the literature. As an example, results are applied to establish a system suitable for a firearm-ammunition combination often encountered in casework: Glock firearms with Fiocchi nickel primer ammunition. The system outputs an LR and a measure of LR uncertainty. The range of possible LR-values is limited to a minimum and maximum value in areas of the score domain with little reference data. Finally, the feasibility of combining LRs of different mark types into one LR for the entire primer is assessed. For the distribution models considered in this paper, different modeling approaches are optimal for different types of similarity scores. For the chosen firearm-ammunition combination, non-parametric Kernel Density Estimation (KDE) models perform best for similarity scores based on the correlation coefficient, whereas parametric models perform best for the Congruent Matching Cells (CMC) scores, assuming binomial and beta-binomial models for KM and KNM score distributions respectively. Finally, it is demonstrated that individual LRs of different mark types can be combined into one LR, to interpret a set of different marks on the primer as a whole.

1. Introduction

To evaluate whether two ballistic samples were fired from the same firearm, forensic examiners typically use a comparison microscope to manually compare firearm toolmarks on the samples. Both the evaluation of firearm toolmark similarity and its significance are based on subjective judgements and rely on the skill, training, and experience of the examiner. Several studies recommend a more objective approach for comparing toolmarks and estimating the strength of the evidence, based on quantitative measures of toolmark similarity, variability, and repeatability [1,2].

In recent years, more objective approaches for evaluating the similarity of firearm toolmarks and interpretation of the strength of the evidence were proposed for a variety of marks encountered in casework: breechface impressions [3–11], firing pin impressions [3–5,12], aperture shear striation marks [13–15] and bullet LEA (land engraved area) striation marks [13,16–21].

Most of these approaches rely on algorithms to calculate a similarity score of two toolmarks. While similarity scores do not include information about typicality of features, which might lead to a loss of information when determining toolmark similarity [22], it has been shown for a variety of mark types that similarity scores do contain enough information to be discriminative. This means that distributions of scores of known-matches (within-source distributions) are significantly different from distributions of scores of known non-matches (between-source distributions).

* Corresponding author.
E-mail address: e.mattijssen@nfi.nl (E.J.A.T. Mattijssen).

https://doi.org/10.1016/j.forsciint.2023.111858
Received 13 March 2023; Received in revised form 4 September 2023; Accepted 10 October 2023
Available online 12 October 2023

Examples of discriminatory scores are the profile correlation coefficient (CCF_{max}) for striated marks, such as aperture shear on the primer [13,14] and the land engraved areas (LEAs) on a bullet [16,20,23]; as well as the areal correlation coefficient ($ACCF_{max}$) [24,25], and the Congruent Matching Cells (CMC) score [5,6,26] for breechface and firing pin impressions. In addition, approaches based on multiple scores were proposed, combining scores of breechface and firing pin impressions [4,27] and different scores derived from the same LEA profile [28].

For interpreting the results of a score-based comparison algorithm, basically two approaches are discussed in the literature: 1) making an identification decision based on the score(s) observed and evaluating the associated error rates [8], or 2) assessing the evidential strength of a comparison result using score-based likelihood ratios [29]. In the current paper, we focus on likelihood ratios (LR).

To obtain an LR by automated means requires representative toolmark population databases of known matching (KM) and known non-matching (KNM) similarity scores (please refer to Section 4.3 for an in-depth discussion of representative databases). The respective score distributions are used to transform a comparison score to an LR. In the literature on automated LR-methods, this is achieved in several ways. First, score distributions can be generalized with non-parametric models [3,14] or parametric models [8,13]. Second, binary classifiers from machine learning can be used to transform the score to an LR. The soft (probability) output of the binary classifier is used for this purpose [30]. In this paper we focus on the first approach.

Based on the current literature, it is not obvious which distribution modeling approach is preferable for a particular mark type. The key challenge is that for a discriminatory score, the LR calculation requires evaluation of the tail of at least one of the score distributions, where typically little experimental data is available. Using parametric models has the advantage that it requires relatively less data to obtain well defined score distributions. However, unless the choice of the parametric model has a strong (physics-based) justification, evaluation of the tails of the model where little experimental data is available can lead to unsupported (extreme) LR values. In addition, parametric models may be too restrictive leading to a poor fit of the data. The non-parametric KDE (kernel density estimates) approaches do not require choosing a specific model. However, they require more data to generalize well, and the evaluation of a distribution tail is affected by the choice of the KDE kernel size parameters.

2. Goals and research questions

In this paper, we compare several automated LR approaches for the interpretation of comparison scores obtained for a set of toolmarks on the cartridge primer: the breechface (BF) impression, the firing pin (FP) impression and the aperture shear (AS) striations (Fig. 1, left). In addition, we study whether the interpretation results for the individual toolmarks can be combined into a single result.

Briefly, we aim at answering the following research questions:

- Among the approaches studied, which is optimal for the score-based interpretation of comparison results for three primer mark types (BF, FP and AS), using LRs?
- How can we interpret the combined comparison results for the various mark types on the primer?

3. Materials and methods

3.1. Firearms and ammunition

For this study, two hundred 9 mm caliber Glock firearms of types 17, 17 C, 19, 19 C, 26 and 34 of generation 2–4 were used [14]. These firearms typically leave similar types of marks with similar properties. Two test fires were obtained for each firearm, using ammunition with nickel plated primer (Fiocchi, 9 mm LUGER, Full Metal Jacket, FMJ). An example of a fired cartridge case is shown in Fig. 1, left.

3.2. Firearm toolmark acquisition

The three-dimensional (3D) surface topography of the three toolmark types evaluated in this study were acquired with Alicona Infinite Focus Microscopes (IFM), types G4 and SL [31]. These are optical microscopes that employ white light focus variation to measure sample topography [32]. Advantages of this measurement approach are that it does not require mechanical probes, is fast, has high vertical and horizontal resolution, and enables measurement of areas with relatively high slopes. While the various areas of the primer containing the toolmarks can all be measured with the IFM, this direct approach is complicated by the presence of deep firing pin impressions, strong reflections and variations in surface reflectivity and color. Therefore, the acquisitions were performed on a cast of the base of the cartridge cases. The casts were made using gray forensic silicone [33], which yields a detailed negative copy of the cartridge case base with homogeneous reflective properties. The measured 3D image was then mirrored in software to obtain the toolmark surface topography. Fig. 1 (Right) shows an example of the surface topography image obtained.

For this study, the three toolmark types were measured separately,

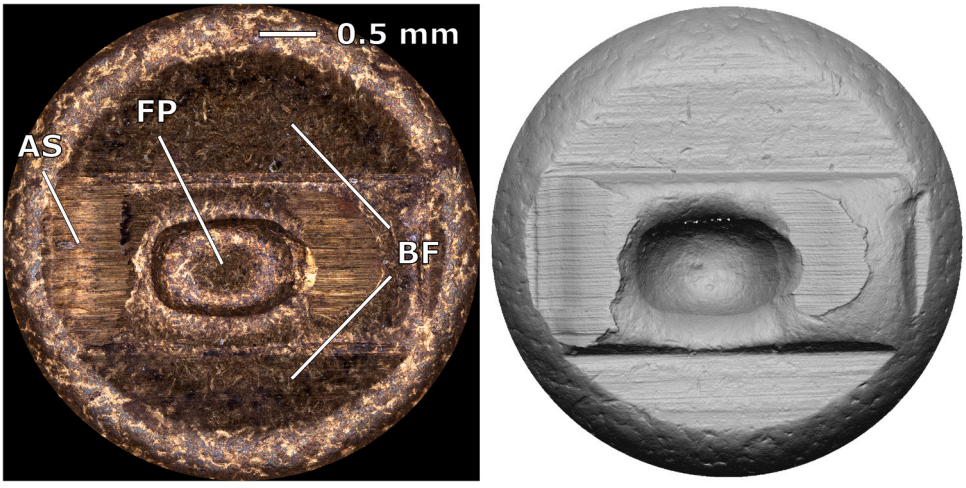

Fig. 1. (Left) Light microscopy image of cartridge case base after firing with a Glock firearm. Typical mark types are the firing pin (FP) impression, the breechface (BF) impression, and the aperture shear (AS) striations. (Right) Example of a 3D surface topography of a cast made from the cartridge case primer area.

with the following settings: objective with 20 x magnification and a numerical aperture (NA = 0.4), 200 nm vertical resolution and approximately 1 μm pixel spacing. For the firing pin impression, the acquisition was limited to the bottom of the impression. The acquisition time varied from approximately 30 s for the firing pin impression (FP) to approximately 10 min for both, breechface impressions (BF) and aperture shear (AS) striations.

3.3. Toolmark type definition and pre-processing

The different regions of interest (ROIs), each containing one toolmark type, were defined manually, using the in-house developed software package Scratch [34]. The software allows for the segmentation of the acquired image into ROIs using geometrical primitives such as rectangles, polygons, circles and ellipses, or combinations of those. All of these can be rotated. Each shape can be labeled foreground (i.e., its enclosed area contains the mark) or background (i.e., its enclosed area is not a part of the mark). For the different toolmark types, the following ROI shapes were chosen:

- BF:Circular foreground ROI and (rotated) rectangular background ROI
- FP: Elliptical foreground ROI
- AS:(Rotated) Rectangular foreground ROI

After ROI definition, the AS data and the FP data were resampled to a pixel separation of 1.5 μm, and the BF data to a pixel separation of 3.5 μm. During this process, noise and surface form components were attenuated using Gaussian regression filters [35] with a low-pass cutoff length of 5 μm and a high-pass cutoff length of 250 μm. Examples of cropped and filtered toolmarks are shown in Fig. 2.

3.4. Comparison algorithms for firearm toolmarks

3.4.1. Comparison of striation marks

Striation toolmarks typically only contain relevant information in a cross-section, orthogonal to the direction of the striations. Therefore, after cropping and filtering, a striation profile was generated by averaging the surface heights along the striations. The resulting averaged profile heights are more repeatable than the individual pixel height data [13]. Automated comparison of two aperture shear profiles was then conducted using a previously published approach [13,14]. First, two toolmark profiles are aligned (registered) using a multi-scale, Gaussian filter-based registration strategy [36,37] with two degrees of freedom, lateral scale and translation (shift). This means that during alignment, one profile is shifted and scaled with respect to the other profile such that the two profiles are most similar. The respective similarity metric, and resulting similarity score, is the Pearson Correlation Coefficient [20]. In the context of toolmark comparisons, this score is often referred to as the normalized Cross Correlation Function maximum CCF_{max}. The maximum allowed scale factor difference and translation were set to +/- 3% and +/- 100 μm respectively. The similarity score ranges between −1 and + 1, corresponding to full negative or positive linear dependency respectively, with 0 corresponding to no linear dependency. The correlation score is not sensitive to differences in the vertical scale of the compared profiles, which facilitates the comparison of weak and strong striations. Most studies [16–21] do not adjust comparison scores for differences in the lateral scale of the profiles. Here, the adjustment was applied to account for differences between the direction of the tool (firing pin aperture) motion and tool orientation.

3.4.2. Comparison of impression marks

Impression toolmarks were compared in two ways. The first approach evaluated the areal similarity of the entire surface topography of the impressed marks. Two images were aligned (registered) using a multi-scale, Gaussian filter-based registration strategy [36,37] with three degrees of freedom: rotation around the vertical axis and translation in the lateral X and Y directions. The respective similarity metric, and resulting similarity score, is again the Pearson Correlation Coefficient. In the context of tool mark comparisons, this score is often referred to as the normalized areal cross correlation function maximum $ACCF_{max}$ [24,25]. The second approach was the Congruent Matching Cells (CMC) method [5,6]. Briefly, the image domain of one image is divided into small image patches or cells, which are subsequently registered independently to the other full image with the same image registration strategy as for the whole marks described above. The similarity score is the number of congruent matching cells, i.e., the subset of cell pairs that meet or exceed a similarity tolerance value, as determined by their correlation coefficient, and that have an approximately congruent registration location (similar difference in position and orientation). The resulting CMC score is an integer value between 0 (no congruent matching cell pairs) and the number of available reference cells into which the image domain was split up (all cell pairs are congruent matching cells).

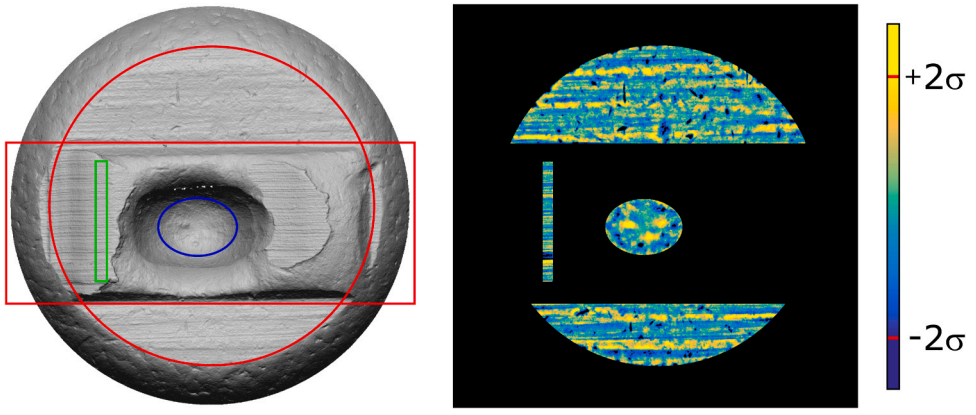

Fig. 2. Demonstration of the manual ROI determination step in Scratch. The red circle and rectangle were used to select the breechface impression, the blue ellipse was used to select the bottom of the firing pin impression and the green rectangle was used to select the aperture shear striations (left). Note that all elements can be rotated if required and that the rectangle of the breechface impression was set to 'background'. After ROI determination, noise and form components in the data were attenuated, yielding datasets similar to the false-color image shown on the right. The surface height is represented by colors, with the highest structures shown in yellow and the lowest structures shown in blue. To ensure a proper height-to-color mapping also in the presence of outliers, surface heights of more than two standard deviations above or below the average surface height are cut off and represented as yellow and blue respectively.

3.5. Interpretation of comparison results for individual marks

Following the strategies described in the previous subsections, distributions of known matching (KM) and known non-matching (KNM) scores were determined by comparing marks on test fires from the same firearm and different firearms respectively. For each pair of test fires, one AS, two BF, and two FP similarity scores were calculated. The reference similarity score distributions were subsequently modeled to interpret a comparison result quantitatively by estimating a score-based likelihood ratio (LR).

3.5.1. Approaches for estimating a likelihood ratio

In the context of this study, a score-based likelihood ratio is the ratio between the likelihood of obtaining the comparison score, given that the prosecutor hypothesis H1 is true (the samples were fired from the same firearm), and the likelihood of obtaining the comparison score, given that the defense hypothesis H2 is true (the samples were fired from different firearms):

$$LR(s) = \frac{p(s|H1)}{p(s|H2)}$$

Here, p(s|H1) means the probability of observing score (s) given H1. In this study, we compared two methods to estimate this measure of evidentiary strength. The methods require samples of two probability density or probability mass functions. The first one is a dataset of H1-scores, and the second one a dataset of H2-scores. H1-scores result from comparisons between marks that are known to have been made by the same firearm. These 'Known Match' (KM) scores correspond to the prosecutor hypothesis (H1). The H2-scores originate from comparisons between marks that are known to have been made by different firearms. These 'Known Non-Match' (KNM) scores correspond to the defense hypothesis (H2).

The two studied methods differ in the statistical models that are used to describe the observed distribution samples. The first method uses non-parametric distribution models, based on Kernel Density Estimation (KDE) and the second method Bayesian statistics and parametric distribution models, such as Gaussian, binomial, or beta-binomial distributions. The implementation details of these two methods are described in more detail in Sections 3.5.4 and 3.5.5 respectively.

Next to this direct approach to estimate an LR via probability densities, also an indirect approach via 'posterior odds' might be used. In the indirect approach, the sets of KM and KNM similarity scores are used to predict the probability of a KM. Because there are only two possible mutually exclusive comparison categories, either KM or KNM, this also fixes the probability of a KNM. By dividing their ratio (also known as the posterior odds) with the prior odds (typically determined by the ratio p(H1)/p(H2) of the dataset sizes for KM and KNM), an estimate for the LR can be obtained.

$$LR(s) = \frac{p(H1|s)}{1 - p(H1|s)} \frac{p(H2)}{p(H1)}$$

During preliminary studies, two regression models were used for the probability of a KM: logistic and isotonic regression. Compared to the direct 'probability density' models, these two indirect 'posterior odds' models performed similar or worse for all the criteria discussed in Section 3.5.3. Therefore, they were not used in further analyses.

3.5.2. General LR-method choices

Some of the LR-method choices that have been made are more general in nature and are applicable to both studied LR-methods. They are discussed first in this section.

The LR-methods in this study are based on a 'common source' framework. This means that the prosecutors hypothesis states that the two cartridge cases that are compared were fired with the same firearm. The defense hypothesis states that the two cartridge cases were fired with two different firearms. This framework corresponds to the forensic scenario where the firearm that may have fired questioned cartridges found at one or more crime scenes is not available. In a 'specific source' framework, on the other hand, the prosecutor's hypothesis states that a questioned cartridge case was fired by a particular firearm, and the defense hypothesis states that the cartridge case was not fired by that firearm. Both scenarios differ in the nature of the likelihoods in the LR and the respective reference score distributions. In a specific source scenario, the KM distribution could be based on a collection of test fires obtained from the firearm identified in the hypotheses, to accommodate differences observed in these distributions for different firearms of the same manufacturing model [3,8]. Capturing specific firearm characteristics is not possible for a common source scenario, where, instead, we model the KM distribution based on test fires from firearms that yield toolmark class characteristics similar to those observed on the compared cartridge cases. For both scenarios, it is important to match the characteristics of the test fire ammunition used to establish a distribution with that of the questioned cartridge case. See [38,39] for more background and discussion on this and other frameworks. The choice to use a common source framework is based on practical arguments: although initially more data is needed to develop the LR-method, requiring a large collection of pieces of evidence fired with representative firearms and ammunitions, no additional data is needed when interpreting comparison results in practice. The specific source approach would require building a firearm specific reference database for each case, including various ammunition types, which is very time-consuming and therefore less feasible in practice.

For the set of 200 firearms, a total of 19900 KNM firearm pairs can be identified. The set of comparison results for these pairs most likely contain some degree of dependency, as the same firearm is used in more than one comparison, which can lead to clustering. In this study, we used all available firearm pairs for the development of the LR-methods. For each KNM pair, multiple cartridge case comparisons can be performed, because two cartridges of the same ammunition type were fired for each firearm. To prevent introducing more dependency in the data, only one, randomly selected, similarity score was used for each of the firearm pairs. For each firearm and ammunition type, there was only one KM comparison available.

Both the KM and KNM datasets are samples of much bigger populations, as it is practically unfeasible to include all firearms of a particular brand that are in usage worldwide or in a specific country. The datasets are based on a realistic subset of the Glock firearm population of the Netherlands however, as they were randomly selected from cases involving shooting incidences. In this study, the resulting model-uncertainty that is associated with the use of a subset of the whole population was quantified by an interval around the predicted best estimate for the LR [40] (what is considered as the best estimate will be explained in the separate sections).

We note that there is controversy on the issue of interval estimation for LRs. It has been argued that using confidence/credible intervals in LR values does not fit in the Bayesian probability framework. On the other hand, it has been argued that information as to the preciseness of LR results is relevant information for the trier of fact. Besides, intervals may be seen as useful when evaluating performance, as it happens in this manuscript. More general considerations and opinions on the use of intervals for LRs are described in [41]. In the current study we use LR systems that result in intervals.

Before being used as input for an LR-method, the scores can be transformed, for example to obtain a domain from $-\infty$ to $+\infty$, to obtain a probability density function that better matches with a specific distribution, or to reduce the dimensionality. We used two transformations in this study. The cross-correlation scores were always transformed from a -1 to $+1$ domain to a $-\infty$ to $+\infty$ domain: first a transformation to a 0 to 1 probability domain takes place (add 1, divide by 2), followed by a log-odds transformation ($\log10(score/(1-score))$). When using CMC scores in a non-parametric model, a dimensionality-

reduction transformation was used: the number of congruent matching cells for each comparison was divided by the number of available reference cells.

When using an LR-method, it must furthermore be considered whether limitations on the attainable values of the LR should be enforced. For a score that has high sensitivity and specificity, the LR calculation requires evaluation of the tail of at least one of the score distributions where typically little experimental data is available. In many situations, the most extreme LR values supported by the available data are related to the sizes of the KM (smallest LR) and KNM (largest LR) datasets. Extrapolation beyond the experimental range of the respective score values should be done with caution. Different approaches to limit the LR are described in literature; see for example [22]. In the current study, we used the Empirical Lower and Upper Bound (ELUB) approach to truncate LR-values, as described by [42].

3.5.3. Comparing LR-methods

To compare different LR-methods, we considered two qualitative criteria and two more quantitative ones:

- Qualitative 1: Explainability
- Qualitative 2: Availability of a parametric distribution model
- Quantitative 1: Validity of predicted LRs
- Quantitative 2: Final-system characteristics

The first qualitative aspect is related to how easy it is to explain the LR-method and how intuitive it is to understand. In this study, we considered the non-parametric distribution model implementation (Section 3.5.4) to be easier to explain than the parametric model implementation (Section 3.5.5). The kernel density estimations are a rather intuitive way of modeling the observed data since they directly follow the data. Parametric models need to have a 'fit' with the data which involves more study and explanations. Next to this, in general, we considered the direct 'probability density' approaches to be easier to explain than indirect 'posterior odds' approaches.

The second qualitative aspect is related to the availability of commonly applied parametric statistical models. For the CMC-scores, the binomial (KNM) and beta-binomial (KM) models [8] are used. In this modeling approach, a comparison is viewed as a series of Bernoulli trials, each describing whether a reference cell is successful in becoming a CMC cell. The underlying assumptions are that the cell trials are independent and that each reference cell has the same success probability, which differs for KM or KNM comparisons. For the beta-binomial distribution, this assumption is relaxed in that the cell success probabilities are allowed to vary from comparison to comparison according to a beta distribution. This modification seeks to account for differences in mark reproducibility for different samples. For the cross-correlations scores, a one-dimensional Gaussian model was used, though its application to cross-correlation scores has not been studied as thoroughly as the modeling of CMC scores.

The first quantitative aspect is related to the validity of the estimated LRs. To assess this validity, both the KM and KNM datasets were randomly split into three equal-sized subsets, to be used in a three-fold cross-validation scheme. Iteratively, one part is used to construct an LR-system, while the other parts are used to assess to what extent the predicted LRs match with the known categories of KM or KNM. Each of these three subsets was used once as validation data, using the other remaining subsets as training data. The three subsets of predicted LRs were recombined into a single dataset, which had the same size as the original dataset. The predicted LRs in this recombined dataset were finally limited using the ELUB-method.

As a first investigative check, a quantity called devPAV has been calculated, based on the predicted LRs [43]. This quantity aims to assess to what extent the predicted LR values disagree with the corresponding LRs based on observed relative frequencies of the KM and KNM LR-data. Since this is an iterative procedure (we calculated LRs for a set of LRs), ideally an identity relation should be observed. A systematic deviation from identity indicates non-valid LRs. devPAV measures how much the predicted LRs deviate from the newly calculated LRs, on average. This deviation is measured on a 10 logarithmic scale. As a rule of thumb, we used a devPAV value above 0.5 as criterion for 'the disagreement needs further investigation', which corresponds to an average absolute deviation of a factor of 3 ($10^{0.5} \approx 3$) on the LR-scale. Next, the properties of the LR system were visualized using histograms of the predicted KM and KNM LRs as well as using an Empirical Cross Entropy (ECE) plot [44]. The ECE is a commonly used cost-function that mainly penalizes misleading evidence (KM data for which an LR below 1 is predicted, and KNM data for which an LR above 1 is predicted). Generally, a lower ECE is better. The ECE of an LR system should always be below the one of a 'neutral' system (a 'neutral' system always returns the uninformative result LR = 1), otherwise it would be better not to use the LR system at all.

The second quantitative aspect is related to the characterization of the final LR-system. This final LR-system used all available data as training data; no validation data was used. Because a larger training dataset was used compared to cross-validation, it was assumed that the validity of the final LR-system is at least as good as the cross-validation results. The best estimate for the LR was calculated for all possible score values, yielding a score-to-LR transformation function. The sampling uncertainty associated with the predicted LRs was quantified using an uncertainty interval. A 90% interval was calculated by generating distributions of the predicted LRs and using the 0.05 and 0.95 percentiles. The LR-distributions were obtained by developing LR-systems that were based on different samples of either the model parameter(s) or the score data, see Sections 3.5.4 and 3.5.5 for more details. The ELUB-method was applied to each of these samples separately, before the percentiles were calculated. The resulting interval is visualized by plotting the LR-percentiles as a function of the best estimate for the LR that belongs to the corresponding score.

3.5.4. Method 1 details: Non-parametric models and bootstrap samples

For the non-parametric distribution modeling approach, a Kernel Density Estimation (KDE) was used [45] based on Gaussian-shaped kernels. KDE requires specifications of the kernel bandwidth, which corresponds to the standard deviation of the Gaussian. In our experiments, the 'optimal' bandwidth was different for each dataset. Usually, it is a trade-off between capturing the essential properties of the distribution of the dataset, and overfitting of this distribution. Unfortunately, there is no single method to determine the optimal bandwidth for all possible datasets.

In this study, a two-step approach was used to select the bandwidth for each of the ten evaluated distributions, corresponding to the five different scores for both the KM and KNM datasets. First, a bandwidth was determined for all ten datasets individually, based on the number of peaks in the distribution of the dataset. The number of expected peaks was chosen visually using a histogram of the dataset. The smallest bandwidth that yielded that number of peaks in the resulting probability density function was used as bandwidth. Only one significant digit was used when determining this bandwidth, except between 1 and 2, in which case two significant digits were used.

Next, a single multiplication factor was chosen, which is applied to the bandwidths of all ten datasets. When choosing this factor, the two quantitative comparison aspects of LR-systems (as described in Section 3.5.3) are taken into account, in addition to the matches between the histograms of the datasets and their probability density functions. In this study, a multiplication factor of 1.5 was chosen. The resulting bandwidths were used for the original datasets, and also for the dataset splits that were used during cross-validation; they are listed in the Appendix.

To obtain a sampling uncertainty interval for the predicted LRs, bootstrapping was used. For each bootstrap sample, new sets of both KM and KNM scores were randomly drawn with replacement from the original KM and KNM datasets. A total number of 400 bootstrap samples

was used. The kernel bandwidth chosen for the original dataset was used for all bootstrap samples. Each of these samples resulted in slightly different probability density functions for both KM and KNM data, and therefore in slightly different LRs at a specific score. The LR calculated based on the original data sets was used as the best estimate for the LR.

3.5.5. Method 2 details: Parametric models and MCMC-simulations

For parametric score distribution models, there are different ways to calculate an LR that includes an uncertainty interval. In this study, we used the so-called 'Posterior-LR' approach [46,47]. It is based on Markov Chain Monte Carlo (MCMC) simulations to obtain samples from the posterior distributions of the score distribution model parameters. The KM and KNM score probabilities, obtained when evaluating the models at a specific score value, are slightly different for each sample of these distribution model parameters. When dividing the probabilities calculated with the KM model by those of the KNM model, a distribution of LRs is obtained for that specific score value. The median of this distribution was used as the best estimate for the LR at that score.

The MCMC-simulations were performed in Stan [48], using the default 'no-U-turn sampler' (NUTS) algorithm. After a warm-up or burn-in period of 10,000 samples, 100 samples were drawn from four chains, resulting in a total of 400 samples. Experiments conducted using a higher number of samples after burn-in did not yield significant changes.

3.6. Combining comparison results of different toolmark types

In the previous sections an approach was proposed for independent interpretation of comparison results, obtained for each type of toolmark. In casework however, it is desirable to additionally provide a combined interpretation result for all evaluated toolmarks for a given cartridge case. LR-based systems can do that by multiplying the LR results obtained from the comparisons of different toolmarks, provided that the comparison results are independent (given the hypothesis) [29]. The three toolmark types on the primer considered in this work are generated by different parts of the firearm. The BF toolmark by the slide's breechface, the FP mark by the firing pin and the AS mark by the edge of the firing pin aperture. They are therefore in general considered independent by forensic toolmark examiners in practice.

To study quantitatively whether the similarity scores ($CCF_{\max}$ for striation marks and $ACCF_{\max}$ as well as CMC for impression marks) are conditionally (in)dependent of each other, score combinations were analyzed in three different manners, qualitatively using scatterplots, and quantitatively using the distance correlation [49] and the Bergsma-Dassios sign covariance test of independence [50]. The distance correlation falls within a range of 0 and 1, with 0 meaning that there is no dependency. The Bergsma-Dassios tests of independence were done using the R-implementation in the TauStar-package [51], with $p = .05$ as significance level to reject the null hypothesis (two similarity scores are independent). Please note that we limited the analysis to testing for independence to between the continuous variables AS $CCF_{\max}$, BF $ACCF_{\max}$ and FP $ACCF_{\max}$. The CMC score includes the local $ACCF_{\max}$ for each patch that is compared as a crucial part of the decision whether a cell is congruent or not, and we therefore consider testing for independence between the FP $ACCF_{\max}$ and BF $ACCF_{\max}$ scores sufficient.

4. Results and discussion

4.1. Comparison results interpretation for separate marks

The results using both the non-parametric (KDE, M1) and parametric (MCMC, M2) distribution models are shown and compared separately, for each of the five combinations of toolmark and score type. For the first combination of firing pin impression and $ACCF_{\max}$ score, six figures are included in the main text: two figures with the observed transformed score distribution and the respective estimated probability density function, two figures with LR cross-validation results, and two figures with details of the final LR system characteristics. A single figure with the distributions of the observed untransformed scores of both the KM and KNM datasets is included in the Appendix. For the other four combinations, only the most relevant figures are included in the main text; the other figures are included in the Appendix.

For the two-dimensional parametric models of the CMC score, a different approach is used to present the score distributions. The combined overview of both datasets is given using a contour plot, based on grouped numbers of CMCs. The comparisons between the observed distributions and their respective distribution models are given using histograms, based on the largest group of grouped CMCs.

For these two-dimensional models, the characterization of the final system is shown as a function of the first score variable (number of matching cells) for both the minimum and maximum of the second score variable (number of reference cells) that was observed in the available dataset.

4.1.1. Firing pin impression – $ACCF_{max}$

Normalized histograms of the untransformed scores are shown in the Appendix (Fig. 14). Fig. 3 shows a comparison between the probability density functions of the two modeling methods with a histogram of the transformed scores, for both the KM and KNM datasets. For the KNM data, both methods yield comparable probability densities that visually correspond well with the histogram. For the KM data, the difference between the two methods is larger, but both still correspond well with the histogram.

Fig. 4 shows the results of the cross-validation calculations. The left graph shows the relative frequencies of the predicted log10-LRs (LLRs) for both the KNM and KM datasets, for both methods. For all datasets, the majority of the predicted LLRs are at the most extreme values, defined by the ELUBs. The results are very comparable for both methods. Comparing the Empirical Cross Entropy (ECE) plots of the two methods, the right graph, reveals that they are very similar as well, and both methods are clearly better than a neutral system (refer to Section 3.5.3). The calculated devPAV values are 0.21 and 0.45.

Fig. 5 shows comparisons between the characterization of the two final LR-systems. The left graph shows the score-to-LLR transformation functions. M2 reaches the highest LLR values at relatively lower scores. In addition, the lower bound on the LLR is slightly lower for M2. Fig. 5 (right) shows a comparison between the uncertainty intervals around the best estimates for the LLRs. At low LLRs, the two methods are comparable. At LLRs above zero, the interval of M2 is smaller than that of M1 due to higher 95%-percentile values of M1.

In summary, the cross-validation results of the two methods are comparable, while the final-system characteristics of M2 are a bit better than those of M1. In addition, the explainability of M1 is better than that of M2. Overall, since we prefer a well explainable model over a bit better performance, we chose to use M1 for this dataset, although M2 is also applicable.

4.1.2. Firing pin impression – CMC

Three figures related to the score distributions are included in the Appendix (Fig. 15 to Fig. 17); they show distributions combined for KM and KNM, and comparisons between the observed and modeled distributions for KM and KNM and for M1 and M2. The relative frequencies of the predicted LLRs are comparable for the two methods, as shown in the Appendix (Fig. 18). Fig. 6 shows that the Empirical Cross Entropy of both methods are very low compared to a neutral system (LR = 1, see Section 3.5.3), with M1 performing better than M2. The calculated devPAV values are 0.86 and 0.92, both above 0.5. These relatively high values are mainly explained by the limited amount of KM data in the area where the KM and KNM distributions overlap and a somewhat poor model fit for either the KM or KNM distributions for CMCs between 0 and 4. Overall, we consider the predictive performance of both

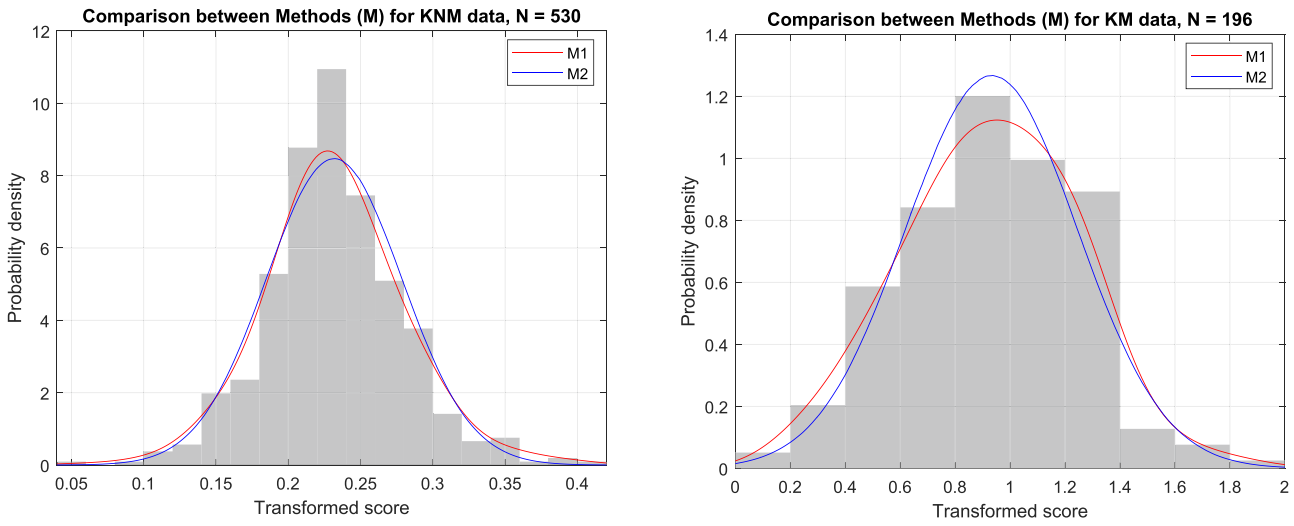

Fig. 3. Histogram of the KNM (left) and KM (right) transformed $ACCF_{max}$ scores of the firing pin impression marks and the estimated probability densities for the non-parametric (M1; red) and parametric (M2; blue) models.

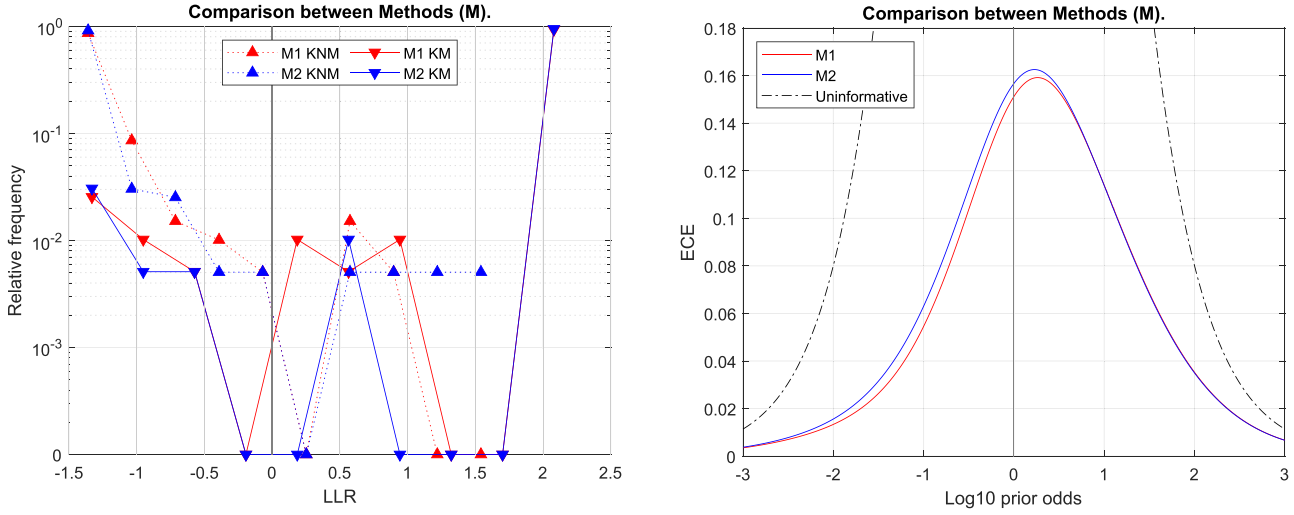

Fig. 4. Relative frequencies of the predicted LLRs for both the KNM and KM datasets, using the non-parametric (M1; red) and parametric (M2; blue) models, for the $ACCF_{max}$ scores of the firing pin impression marks (left). Empirical Cross Entropy (ECE) of the non-parametric (M1; red) and parametric (M2; blue) models, for the $ACCF_{max}$ scores of the firing pin impression marks (right).

methods sufficient. The performance could be further improved by acquiring more data in the overlap area and refitting the models.

Fig. 7 (left) shows the characteristics of the final systems, where the lower bound on the LLR is slightly lower for M2, but the upper bounds are similar. M1 seems to be more sensitive to the number of reference cells (*N*) than M2. At the minimum value of *N* (24), the difference between the methods is a shift of the transformation function. However, at the maximum value of *N* (48), M2 reaches the most extreme LLR values at less extreme scores.

The plot at the right of Fig. 7 shows that the two methods behave comparable at LLRs below zero, and that the interval of M2 is smaller than that of M1 above an LLR of zero, mainly due to a higher 95th-percentile. For M2, the interval appears to depend to a limited extent on the number of reference cells (N).

Summarizing, the cross-validation results of M1 are slightly better than those of M2, while the final-system characteristics for M2 are better than that for M1. The explainability of M1 is better than that of M2. However, suitable statistical models are available for CMC scores, which favors M2. Taking all aspects into consideration, we recommend using M2 for this dataset, but M1 is also applicable.

4.1.3. *Breechface impression – $ACCF_{max}$*

The probability densities of both untransformed score distributions, of the transformed KM data, and the ECE-plot are included in the Appendix (Fig. 19, Fig. 20 and Fig. 21); they show comparable results for the two methods. The calculated devPAV values are 0.25 and 0.30. On the other hand, the methods show relevant differences for the KNM data. Fig. 8 shows the probability density functions of the KNM data. Although there is a visible difference in the height of the modus of the distributions, the most significant difference is located at transformed scores between approximately 0.3 and 0.5. The M1 distribution reflects the presence of a few data points in this area. The Gaussian model of M2 is incapable of modeling two distribution peaks, and instead widens the distribution function.

Fig. 9 shows the relative frequencies of the predicted LLRs for the KM and KNM data. Although most of the data points are located at the ELUBs, a significant portion are not, especially for the KM data. M2 correctly predicts most KNM data points at the lower bound, but it also incorrectly predicts more KM data points at this lower bound than M1 does.

The characteristics of the final systems are shown in Fig. 10. Most

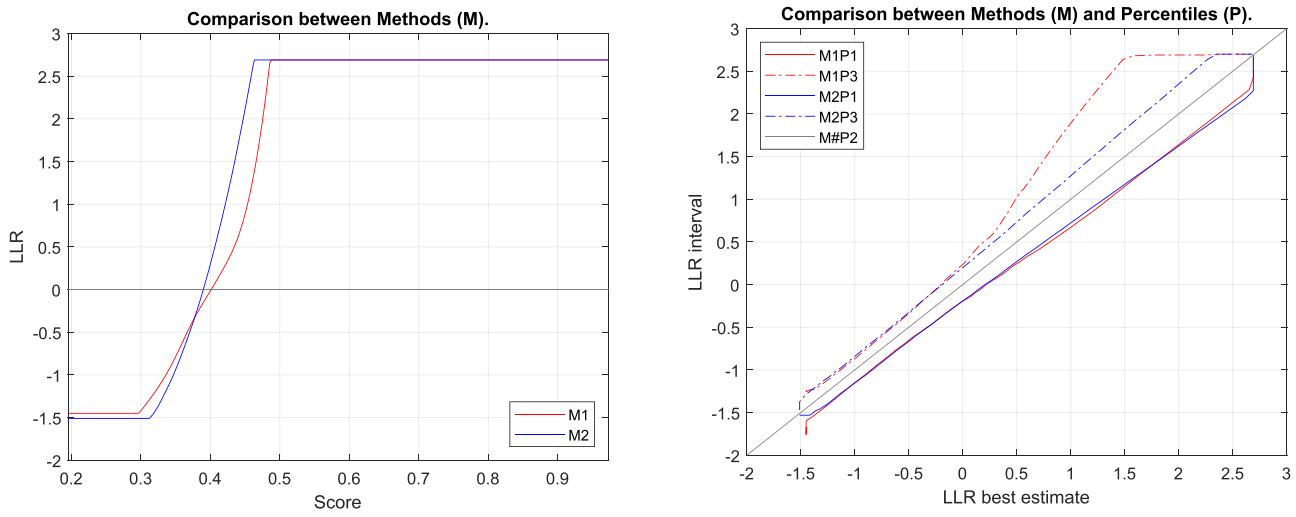

Fig. 5. Score-to-LLR transformation functions of the non-parametric (M1) and parametric (M2) models, for the $ACCF_{max}$ scores of the firing pin impression marks (left). Uncertainty intervals (90%; P1 = 5%, P2 = 50%, P3 = 95%) around the best estimates for the LLRs of the non-parametric (M1) and parametric (M2) models, for the $ACCF_{max}$ scores of the firing pin impression marks (right).

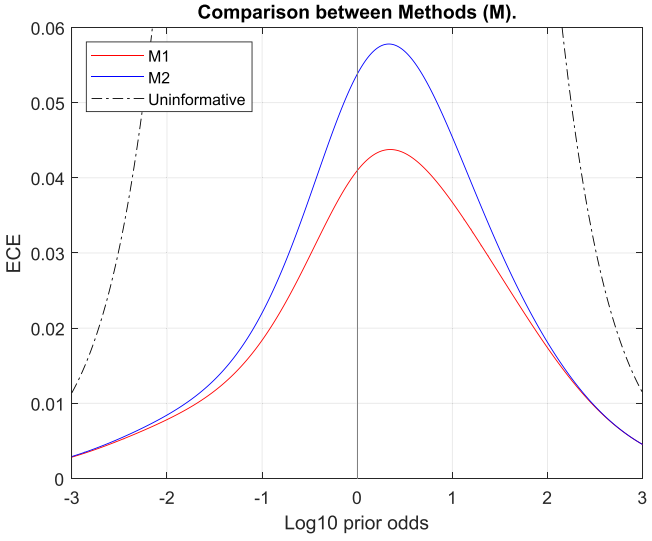

Fig. 6. The Empirical Cross Entropy (ECE) of the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the firing pin impression marks.

striking is the behavior of M1 between a score of 0.4 and 0.55: it increases up to its upper bound, decreases, and increases again to its upper bound. This effect is related to a small number of KNM data points with relatively high scores. The chosen parametric model for M2 is not able to capture this behavior. More experiments would be required to evaluate whether the rare high-score events occur in a wider score domain, which would smoothen the M1 LLR curve. The upper bounds of the two methods are comparable, while M1 has a smaller lower bound. M2 reaches its most extreme LLR values at less extreme scores. The uncertainty intervals of both methods are very comparable, except for the very high LLR values where M1 shows the effects of the rare KNM data points with high scores.

Summarizing, the cross-validation results of the two methods are comparable, while the final-system behavior for M1 is slightly better than that for M2. In addition, the explainability of M1 is better than that of M2. Overall, we choose to use M1 for this dataset, although M2 is applicable as well.

4.1.4. *Breech face impression – CMC*

Qualitatively, the results for CMC scores of this impression toolmark are comparable to those of the FPs mark (see Section 4.1.2). Fig. 25 and Fig. 26 (in the Appendix) show that the cross-validation results of the two models are very comparable. The calculated devPAV values are 0.30 and 0.21. Results for the characteristics of the final system shown in Fig. 11 are similar to those in Fig. 7 of the firing pin impression marks in that M2 is behaving better than M1. Overall, because of the better explainability, we choose to use M2 for this dataset, although M1 is also applicable.

Also included in the Appendix (Fig. 22 to Fig. 24) are the three figures related to the score distributions. They show distributions combined for KM and KNM, and comparisons between the observed and modeled distributions for KM and KNM and for M1 and M2.

4.1.5. *Aperture shear – CCF_{max}*

The probability density functions (Fig. 28 and Fig. 29, in the Appendix) and the cross-validation results (Fig. 30 and Fig. 31, also in the Appendix) are very comparable for M1 and M2, with M2 performing slightly better during cross-validation. The calculated devPAV values are 0.26 and 0.56. The latter is slightly higher than 0.5, but we consider the predictive performance of both methods sufficient. Also included in the Appendix are the normalized histograms of the untransformed scores of both the KNM and KM datasets (Fig. 27).

Fig. 12 shows the final-system characteristics of the two methods. M2 has a smaller lower bound, while M1 has a higher upper bound. M2 reaches its most extreme LLR values at relatively lower scores. At low LLRs, the uncertainty intervals of the two methods behave comparable. However, at LLRs above two, the interval of M1 becomes larger than that of M2.

Summarizing, the cross-validation results of the two methods are comparable, while the final-system characteristics for M2 are a bit better than that for M1. In addition, the explainability of M1 is better than that of M2. Due to the better explainability, we use M1 for this dataset, although M2 would be applicable as well.

4.2. *Combining comparison results of different toolmark types*

For qualitative assessment of the dependency between marks, scatterplots are presented in Fig. 13. Quantitative results are given in Table 1. No obvious dependencies could be found between the similarity scores of the different mark types, either for KM or KNM comparisons. Almost all distance correlations are close to 0 and for all combinations to

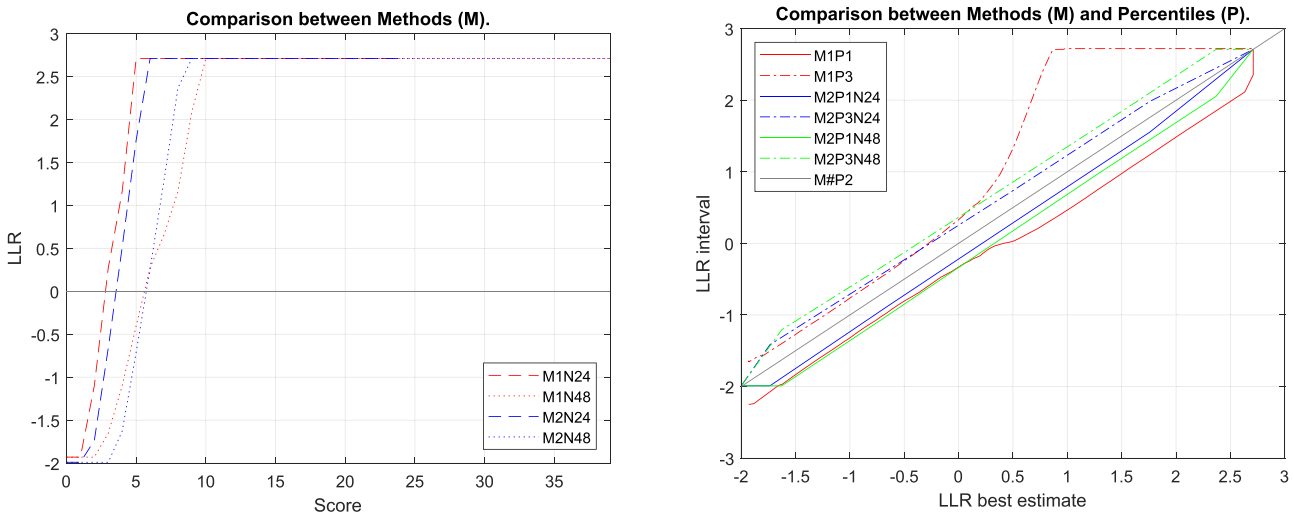

Fig. 7. Score-to-LLR transformation functions of the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the firing pin impression marks (left). Uncertainty intervals (90%; P1 = 5%, P2 = 50%, P3 = 95%) around the best estimates for the LLRs of the non-parametric (M1) and parametric (M2) models, for the CMC-scores of the firing pin impression marks (right). For the two-dimensional parametric model, intervals for the minimum (N = 24) and maximum (N = 48) total number of cells are shown.

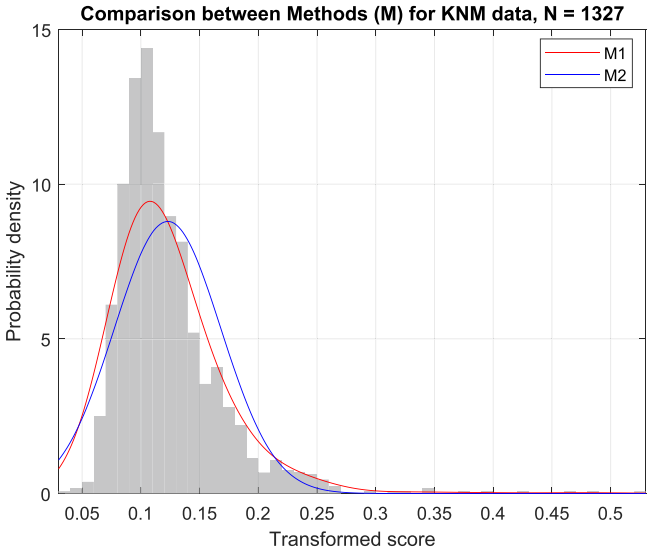

Fig. 8. Histogram of the original KNM data and the resulting probability densities of the non-parametric (M1; red) and parametric (M2; blue) models, for the transformed $ACCF_{max}$ scores of the breechface impression marks.

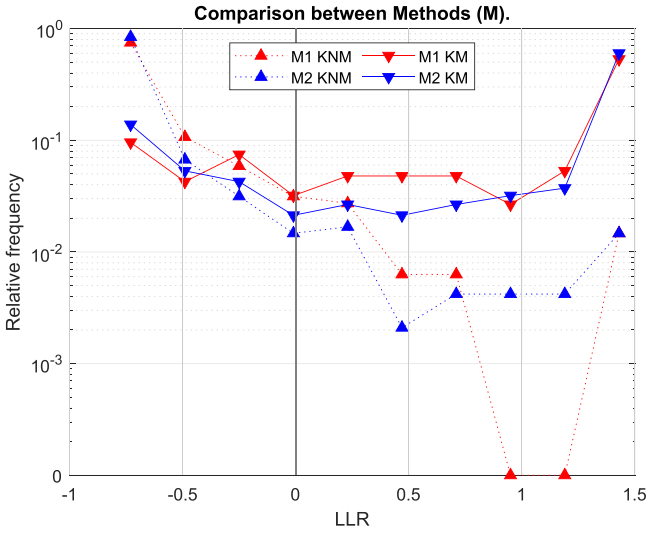

Fig. 9. Relative frequencies of the predicted LLRs for both the KNM and KM datasets, using the non-parametric (M1; red) and parametric (M2; blue) models, for the $ACCF_{max}$ scores of the breechface impression marks.

which the Bergsma-Dassios test was applied, the null hypothesis of similarity score independence could not be rejected. Exceptions are the combination of $ACCF_{max}$ and CMC, based on the *same* mark, in the KM case. Both for BF (Distance corr.: 0.56) and FP (Distance corr.: 0.62) there is a strong dependence between the similarity scores, but as it concerns the exact same mark, this is to be expected. Consequently, LRs calculated using these two similarity scores should not be multiplied. Rather one of the two should be chosen. Alternatively, a multi-dimensional or combined similarity score, combining CMC and $ACCF_{max}$ scores, may yield improved separation between KM and KNM.

A scenario in which the distance correlation is slightly elevated is for the combination KM BF CMC and FP CMC (Distance corr.: 0.24), although no obvious dependency can be seen in the scatter plot (Fig. 13). Note that this elevated dependency does not show for the KM BF $ACCF_{max}$ and FP $ACCF_{max}$ pair (Distance corr.: 0.15). Although all cartridges were of the same brand and had the same primer material, properties like primer material hardness might vary to some extent [52], which in turn may influence how well marks are transferred from the firearm to the primer, particularly for marks of the same category, impression or striation like BF and FP. That is not necessarily the case for different mark categories.

In conclusion, none of the results point towards any (strong) dependency between the similarity scores used for calculating the LRs for the different marks. This does not irrefutably prove independence, but provides enough support for the validity of multiplying LRs of different mark types in practice.

Consequently, we consider it reasonable for our studied dataset to calculate a combined LR for all studied marks on the primer by multiplying the individual LRs of the different mark types. Various examples of individual as well as combined LRs for both, known matching as well as known non-matching exhibits, are presented in Table 2. Note that the calculated LR intervals cannot be multiplied in a straightforward manner. Future research will have to be conducted to find the optimum strategy for determination of a combined LR including uncertainty intervals.

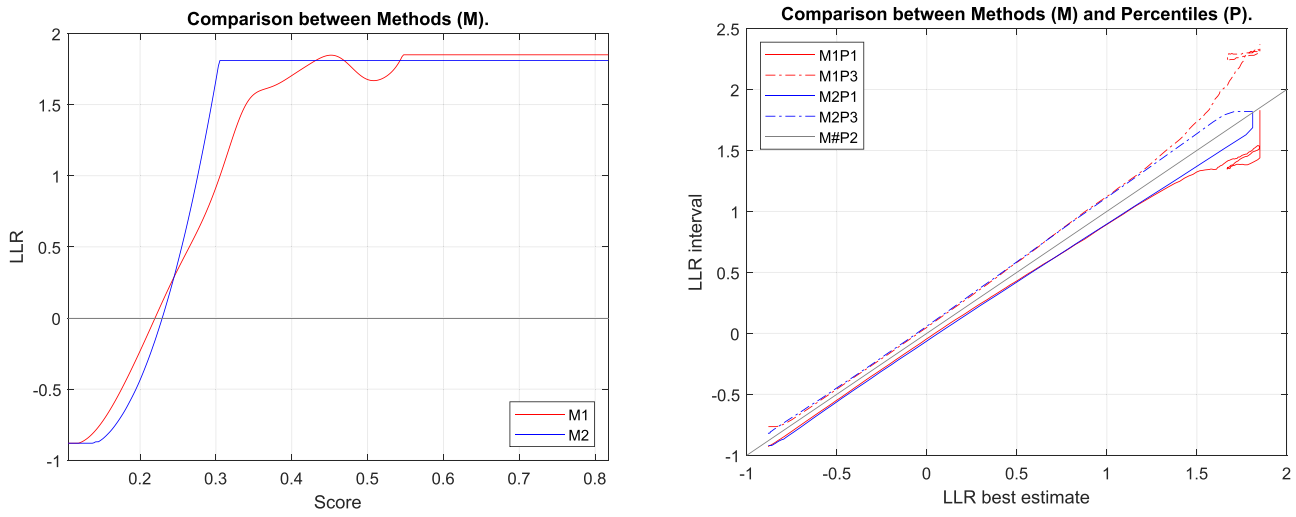

Fig. 10. Score-to-LLR transformation functions of the non-parametric (M1; red) and parametric (M2; blue) models, for the $ACCF_{max}$-scores of the breechface impression marks (left). Uncertainty intervals (90%; P1 = 5%, P2 = 50%, P3 = 95%) around the best estimates for the LLRs of the non-parametric (M1) and parametric (M2) models, for the $ACCF_{max}$ scores of the breechface impression marks (right).

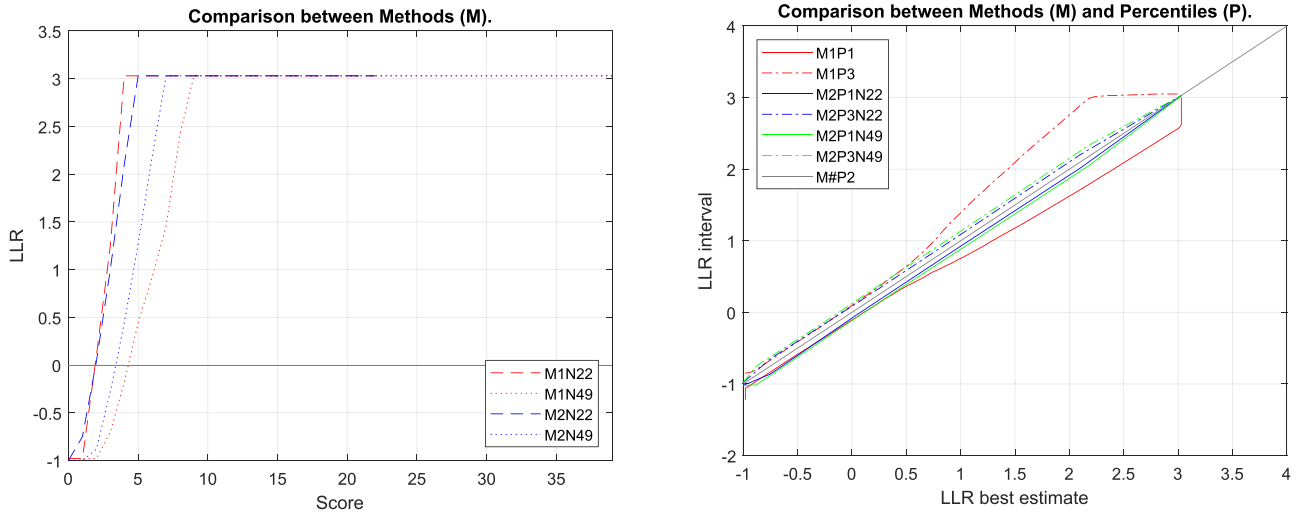

Fig. 11. Score-to-LLR transformation functions of the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the breechface impression marks (left). Uncertainty intervals (90%; P1 = 5%, P2 = 50%, P3 = 95%) around the best estimates for the LLRs of the non-parametric (M1) and parametric (M2) models, for the CMC-scores of the breechface impression marks (right). For the two-dimensional parametric model, intervals for the minimum (N = 22) and maximum (N = 49) number of reference cells are shown.

If dependencies would be present, combining LRs of different mark comparisons may still be possible by including the dependencies between scores in the statistical models. LRs would have to be estimated using procedures that sample the multivariate probability distribution of the various scores, similar to the case when more than one score is considered for each mark.

Note that material properties of the primer can influence all interpretation results. If the primer is very hard, especially impression marks may not transfer very well, causing the various mark comparisons to yield relatively lower similarity scores. This in turn will lead to LRs of the individual mark comparisons to be relatively lower as well. On the other hand, if the primer material is very soft, marks typically transfer better, leading to higher similarity scores and LRs. Thus, depending on material properties, all LRs can be relatively lower or higher, but can still be considered conditionally independent with respect to the firearm part creating the mark. In the current study we set up the reference database with samples of the same ammunition brand that have similar primer material properties. However, if in the future reference databases including samples of various ammunition brands, and thus primer material properties, should be combined, e.g. to obtain a larger database, LR multiplication might not be possible as a result of the primer material dependency of the various scores. Further research is required to study the best approach under such circumstances.

4.3. *Choosing representative background population samples*

The score-based LRs in this study are calculated using population samples of known matching and known non-matching comparison scores. The population from which the cartridge case samples are drawn is assumed to be representative of the case based on the forensic information available to the examiner. There are two major criteria that affect this determination. The first is the comparison scenario: same source vs specific source. Significant differences have been observed in the KM score distribution of firings obtained from similar firearms, even from (consecutively manufactured) firearms of the same manufacturing model [3,8]. Thus, the KM distribution for a specific source scenario may

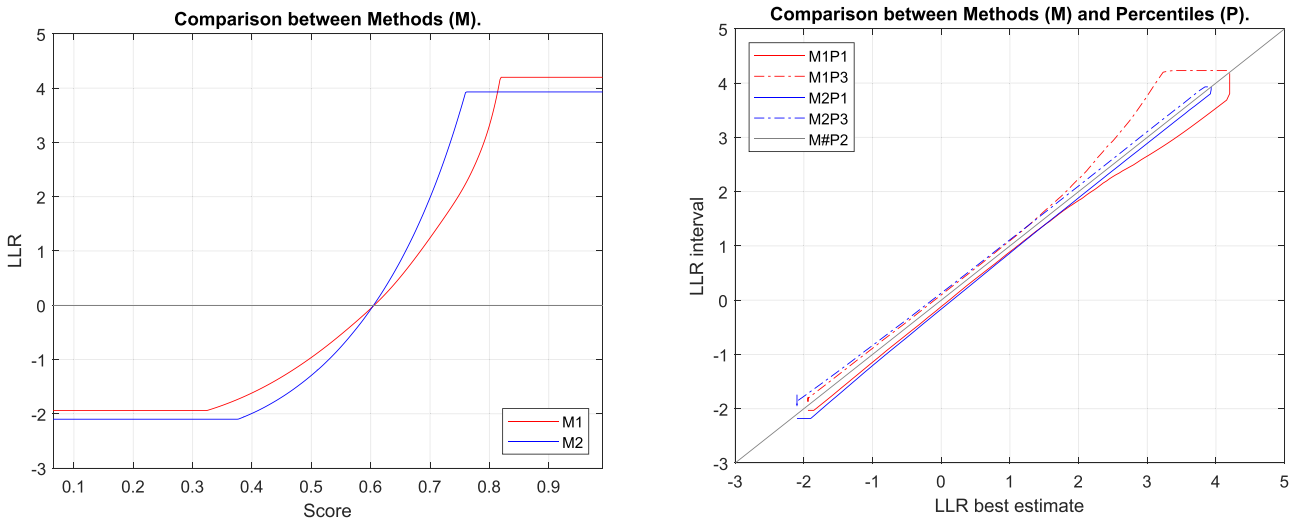

Fig. 12. Score-to-LLR transformation functions of the non-parametric (M1; red) and parametric (M2; blue) models, for the CCF_{max}-scores of the aperture shear marks (left). Uncertainty intervals (90%; P1 = 5%, P2 = 50%, P3 = 95%) around the best estimates for the LLRs of the non-parametric (M1) and parametric (M2) models, for the CCF_{max}-scores of the aperture shear marks (right).

have to be based on test fires obtained from the case firearm. For a common source scenario, the firearm is not available, and our only choice is to base the KM distribution on firearms that yield class characteristics similar to those observed on the questioned cartridge cases. Similarly, for both scenarios, one could argue that the KNM distributions should be based on comparisons that involve the questioned cartridge case(s). This could be achieved by (automatically) comparing the questioned cartridge case with cartridge cases sampled from a representative population. On the other hand, preliminary results obtained at NIST and NFI seem to indicate that KNM distributions are less sensitive to variations in firearms or ammunition (results in preparation for publication). The second aspect concerns the influence factors, and their values, which must be considered in selecting the representative population. All the LRs in this study were determined with a database of samples that have the same characteristics as the samples that are compared. They were all from the same firearms manufacturer and the ammunition brand was identical as well. In casework, this may not be feasible. Reference databases should therefore be set up with different firearm and ammunition brands with varying properties like primer material (e.g. nickel, brass and steel). Research is required to evaluate the extent to which these properties may vary from the case scenario and the respective effect on LRs. Currently, we consider a sampled population representative for the comparison at hand, if influence factors such as firearm manufacturing method/brand and ammunition brand are either identical to the considered case, or as long as statistical data on the expected variation of the LR, as a result of using samples from a wider population, is available.

5. Summary and conclusion

In this study we compared the performance of several common source LR systems for automated interpretation of score-based comparison results of firearm toolmarks on cartridge case primers. Performance was evaluated using four qualitative and quantitative system comparison criteria known from the literature. Combinations of various toolmark types and similarity scores were studied, leading to the following statistical model recommendations: Non-parametric Kernel Density Estimation models for the correlation metrics CCF_{max} (aperture shear striation mark) and $ACCF_{max}$ (firing pin and breechface impression toolmarks) and parametric models for the CMC metric (firing pin and breechface impression marks). The parametric models for CMC are a beta-binomial model for KM and a binomial model for KNM.

The proposed LR-method was applied to and evaluated for the combination Glock firearms and Nickel primer ammunition (Fiocchi). The system can thus be employed in casework, given that our framework is acceptable for the judicial system of the country in which it is used. Besides the specific combination of firearm and ammunition evaluated in this work, we expect that the framework is generic enough to be applied to other firearm/ammunition combinations as well, given that model parameters are optimized with that particular combination.

It was observed that most combinations for different toolmark types revealed no dependency between similarity scores, suggesting that LRs for separate toolmark types can reasonably be multiplied to yield an interpretation result for all considered toolmark types on a primer. A weak dependency was found for the KM CMC results of the breechface and firing pin impressions that requires more research.

In combination with automated toolmark comparison methods for all relevant toolmark types published previously, the presented interpretation framework provides an objective evaluation of the strength of the evidence in casework.

Disclaimer

Certain commercial equipment or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the NFI, NIST, or FBI, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

CRediT authorship contribution statement

Martin Baiker-Sørensen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Ivo Alberink**: Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Laura B. Granell**: Investigation, Data Curation, Writing – review & editing. **Leen van der Ham**: Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Erwin J.A.T. Mattijssen**: Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – review & editing, Supervision. **Erich D. Smith**: Conceptualization, Resources, Writing – review & editing. **Johannes Soons**: Conceptualization, Methodology, Investigation, Writing –

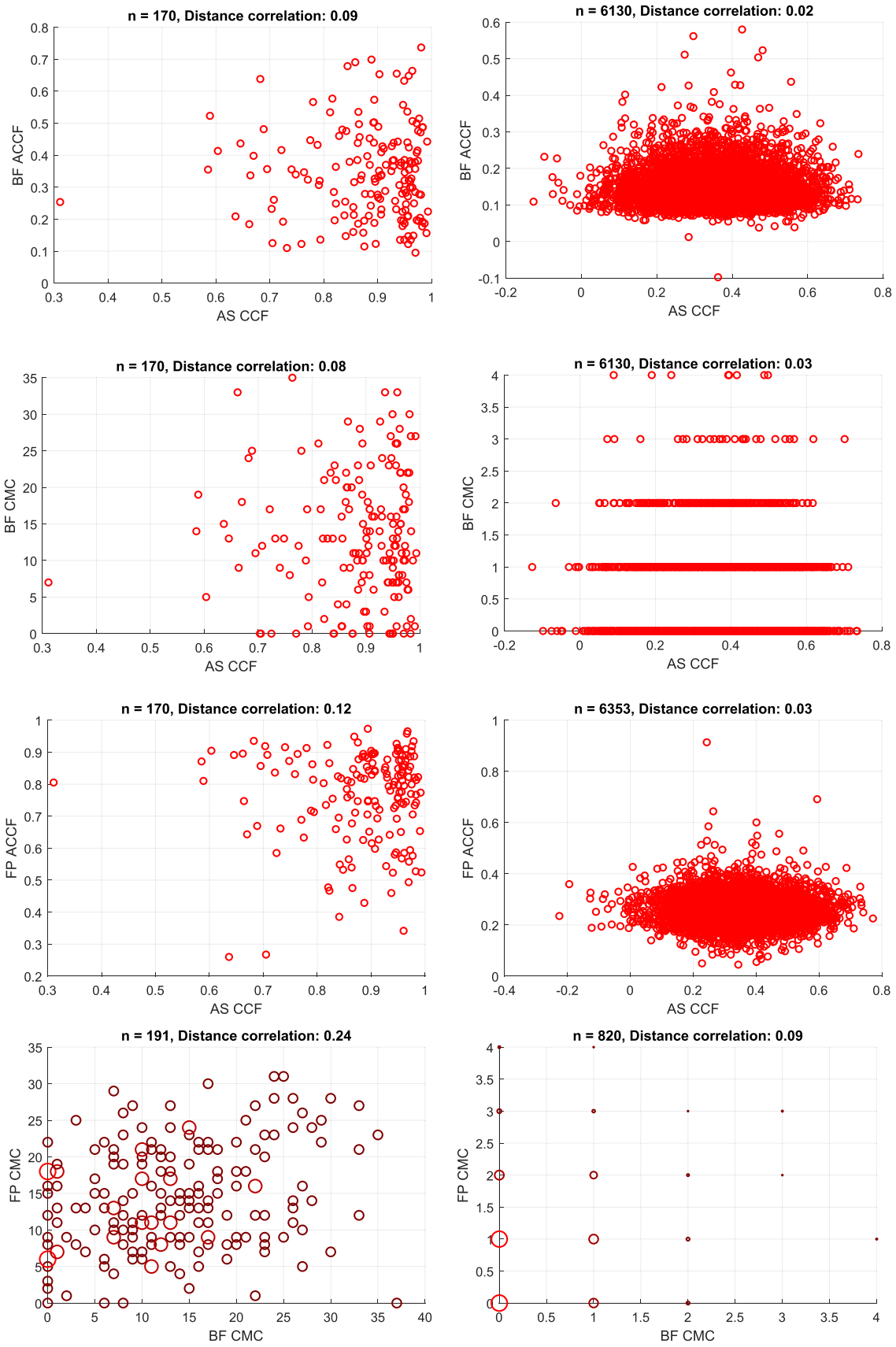

Fig. 13. Scatterplots for qualitative assessment of the similarity score distributions for various mark type combinations. For categorical vs. categorical score scatterplots (BF CMC vs. FP CMC), the size and color of the markers indicates the number of scores. Large markers with bright red color indicate a large number of scores, small markers with dark red color indicate a small number of scores at a given location.

Table 1
Distance correlations between similarity scores of toolmarks on the same primer (CCF_{max}, $ACCF_{max}$, CMC) as well as similarity score independence test results (p = . 05 significance level).

				BF		FP	
				ACCF	CMC	ACCF	CMC
KM	AS	CCF (n = 170)	Distance correlation	0.09	0.08	0.12	0.11
			Independence test (p-value)	0.94	-	0.78	-
	BF	ACCF (n = 191)	Distance correlation	-	0.56	0.13	0.15
			Independence test (p-value)	-	-	0.41	-
		CMC (n = 191)	Distance correlation	-	-	0.16	0.24
	FP	ACCF (n = 191)	Distance correlation	-	-	-	0.62
KNM	AS	CCF (n = 6130)	Distance correlation	0.02	0.03	0.03	0.02
			Independence test (p-value)	0.23	-	0.11	-
	BF	ACCF (n = 820)	Distance correlation	-	0.13	0.07	0.06
			Independence test (p-value)	-	-	0.39	-
		CMC (n = 820)	Distance correlation	-	-	0.08	0.09
	FP	ACCF (n = 820)	Distance correlation	-	-	-	0.06

Table 2
Examples of individual and combined LLRs for three known matching (KM 1–3) and three known non-matching (KNM 1–3) exhibits. The combined LLR is given for a combination of $ACCF_{max}$ LLRs (BF and FP) and the CCF_{max} LLR (AS) as well as for a combination of CMC LLRs (BF and FP) and the CCF_{max} LLR (AS). The combined LLRs are presented without uncertainty intervals, as more research is required to determine the optimum intervals for a combined LLR. Note that the set of KNM reference scores used in this example is larger than the one used during the evaluation. While for the evaluation only one measurement per firearm was included in the determination of KNM scores, for this example both measurements were included.

	Breech face KM: n = 188 KNM: n = 56918				Firing pin KM: n = 187 KNM: n = 42001				Aperture shear KM: n = 196 KNM: n = 76440		Combined LLR	
	ACCF	LLR (5%, 95%)	CMC	LLR (5%, 95%)	ACCF	LLR (5%, 95%)	CMC	LLR (5%, 95%)	CCF	LLR (5%, 95%)	ACCF ACCF CCF	CMC CMC CCF
KM 1	0.66	3.78 (3.47,3.91)	14/24	4.65 (4.65,4.65)	0.91	4.50 (4.47,4.54)	28/36	4.53 (4.53,4.53)	0.96	4.78 (4.75,4.81)	13.06	13.96
KM 2	0.57	3.78 (3.47,3.91)	8/32	4.65 (4.65,4.65)	0.97	4.50 (4.47,4.54)	32/42	4.53 (4.53,4.53)	0.89	4.78 (4.75,4.81)	13.06	13.96
KM 3	0.56	3.78 (3.47,3.91)	20/24	4.65 (4.65,4.65)	0.93	4.50 (4.47,4.54)	35/36	4.53 (4.53,4.53)	0.95	4.78 (4.75,4.81)	13.06	13.96
KNM 1	0.22	-0.10 (−0.13,−0.08)	0/24	-0.86 (−0.94,−0.84)	0.27	-1.55 (−1.90,−1.33)	0/36	-1.34 (−1.38,−1.32)	0.22	-1.91 (−2.01,−1.82)	-3.56	-4.11
KNM 2	0.12	-0.90 (−0.94,−0.81)	0/24	-0.86 (−0.94,−0.84)	0.27	-1.55 (−1.92,−1.33)	2/36	-1.14 (−1.36,−0.97)	0.33	-1.91 (−2.01,−1.77)	-4.36	-3.91
KNM 3	0.16	-0.71 (−0.74,−0.68)	2/32	-0.21 (−0.30,−0.13)	0.26	-1.55 (−1.96,−1.33)	0/42	-1.34 (−1.38,−1.32)	0.45	-1.39 (−1.55,−1.27)	-3.65	-2.94

review & editing. **Peter Vergeer**: Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Xiaoyu A. Zheng**: Conceptualization, Methodology, Investigation, Writing – review & editing.

Declaration of Competing Interest

None.

Appendix

KDE bandwidths

Table 3
Selected KDE bandwidths for the KM and KNM datasets of all five toolmark types and score type combinations.

Mark type	Score type	KDE bandwidth	
		KM	KNM
Firing pin impression	*$ACCF_{max}$*	*0.14*	*0.021*
Firing pin impression	*CMC*	*0.09*	*0.020*
Breechface impression	*$ACCF_{max}$*	*0.17*	*0.027*
Breechface impression	*CMC*	*0.14*	*0.017*
Aperture shear	*CCF_{max}*	*0.23*	*0.05*

Firing pin impression – $ACCF_{max}$

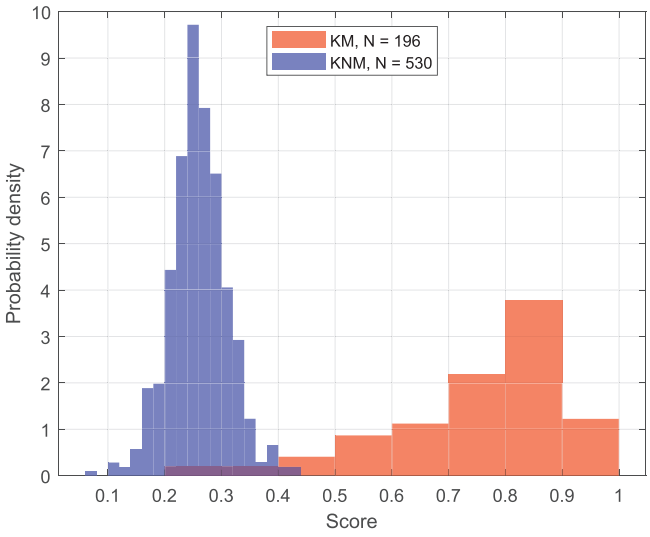

Fig. 14. Normalized histograms of the untransformed scores for both the KNM (blue) and KM (red) datasets of the $ACCF_{max}$ of the firing pin impression marks.

Firing pin impression – CMC

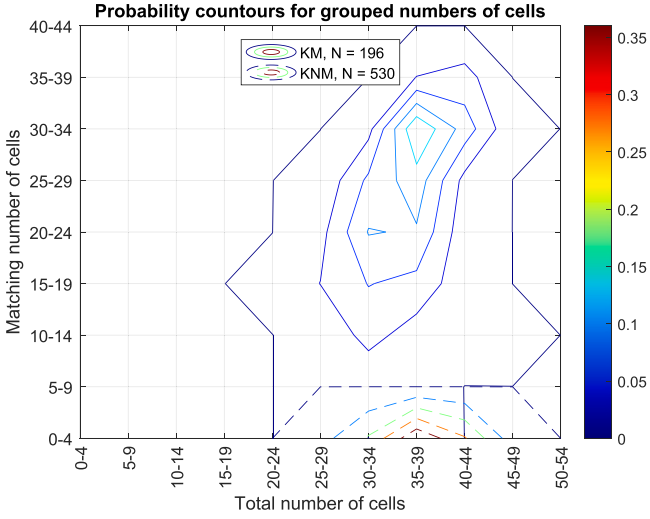

Fig. 15. Contour plot of the observed distributions of the grouped KM (solid lines) and KNM (dashed lines) CMC data of the firing pin impression marks. Group sizes of 5 were used.

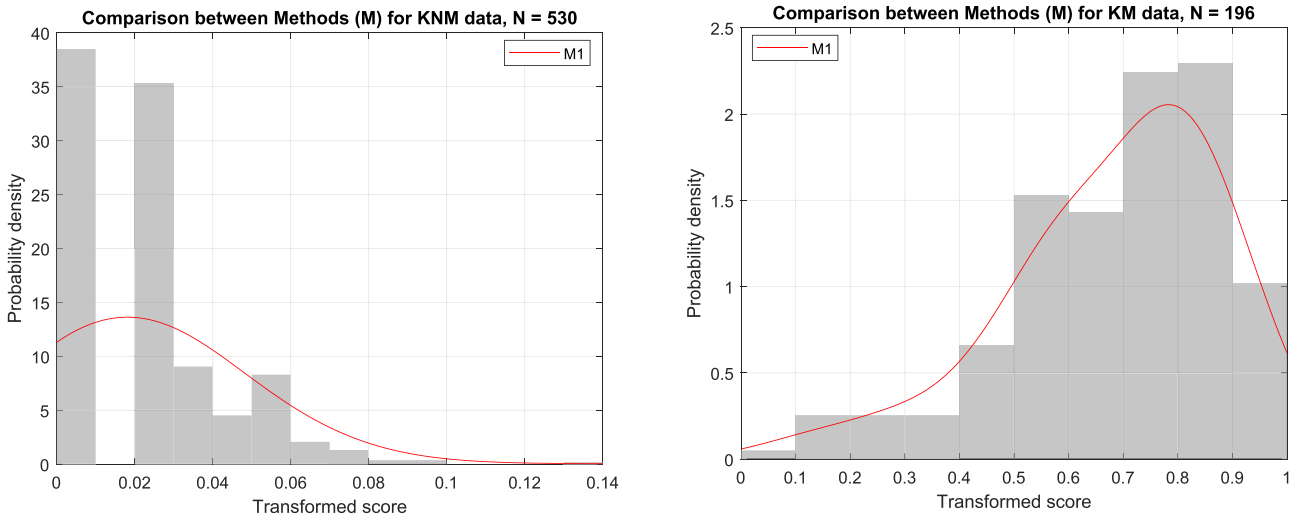

Fig. 16. Histogram of the KNM (left) and KM (right) transformed CMC scores of the firing pin impression marks and the estimated probability densities for the non-parametric (M1) model.

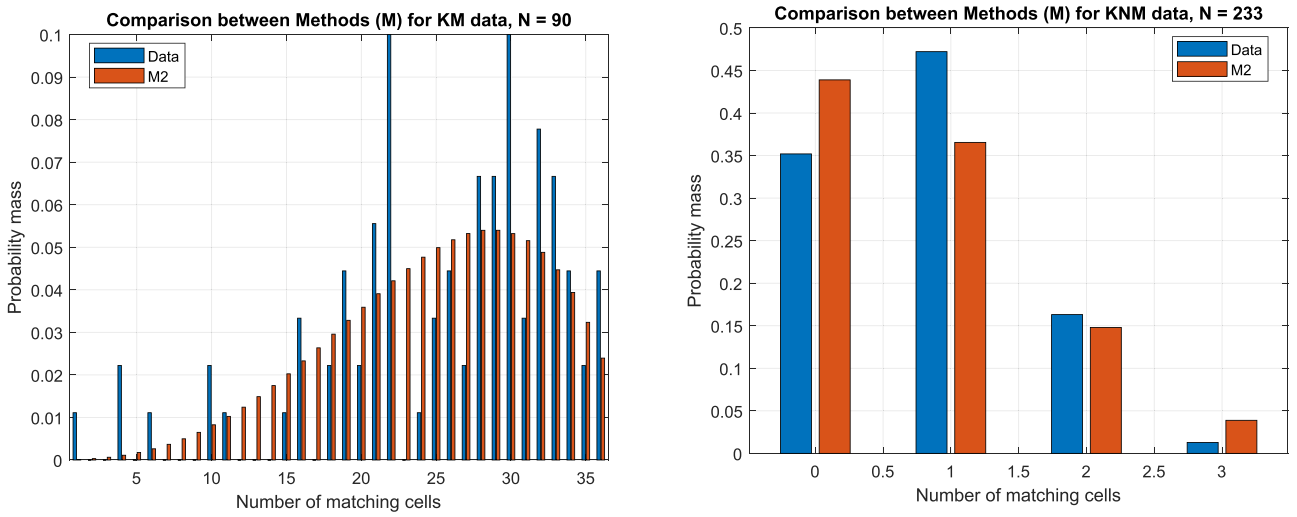

Fig. 17. Histogram of the grouped numbers of KNM (left) and KM (right) CMCs of the firing pin impression marks and the estimated probability densities for the parametric (M2) model.

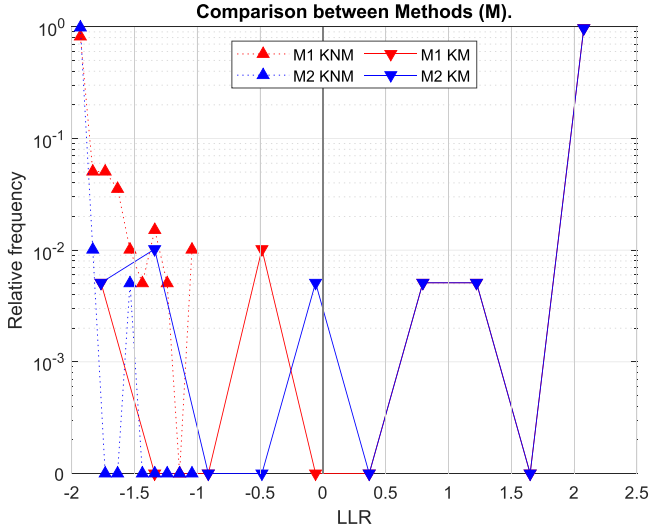

Fig. 18. Relative frequencies of the predicted LLRs for both the KNM and KM datasets, using the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the firing pin impression marks. Only towards the lowest LLRs of the KNM data there is a difference: M2 predicts almost all data at the LLR extreme, while M1 also predicts some data points in the adjacent bins.

Only towards the lowest LLRs of the KNM data there is a difference: M2 predicts almost all data at the LLR extreme, while M1 also predicts some

data points in the adjacent bins.

Breechface impression – $ACCF_{max}$

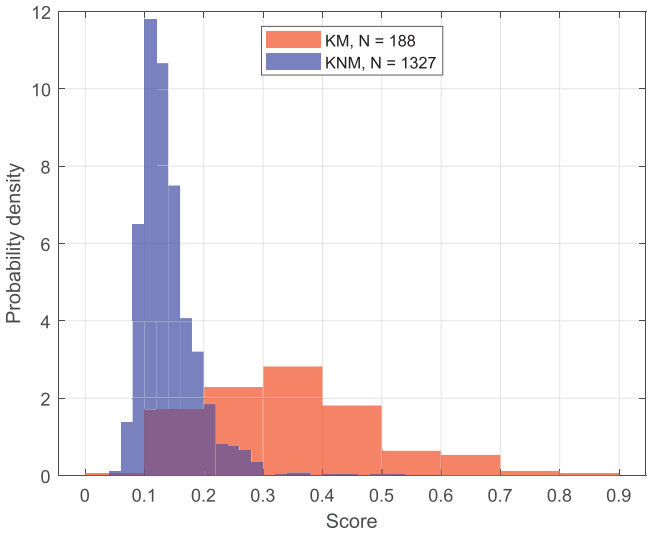

Fig. 19. Normalized histograms of the untransformed scores for both the KNM (blue) and KM (red) datasets of the untransformed $ACCF_{max}$ scores of the breechface impression marks.

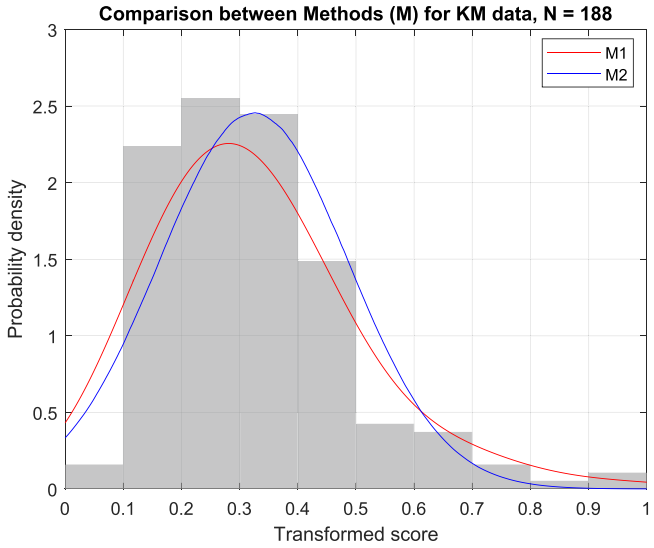

Fig. 20. Histogram of the original KM data and the resulting probability densities of the non-parametric (M1; red) and parametric (M2; blue) models, for the transformed $ACCF_{max}$ scores of the breechface impression marks.

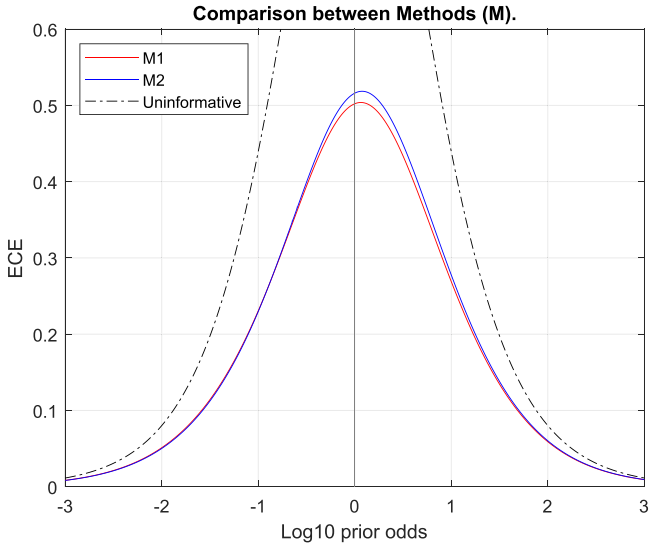

Fig. 21. Empirical Cross Entropy (ECE) of the non-parametric (M1; red) and parametric (M2; blue) models, for the $ACCF_{max}$ scores of the breechface impression marks.

Breechface impression – CMC

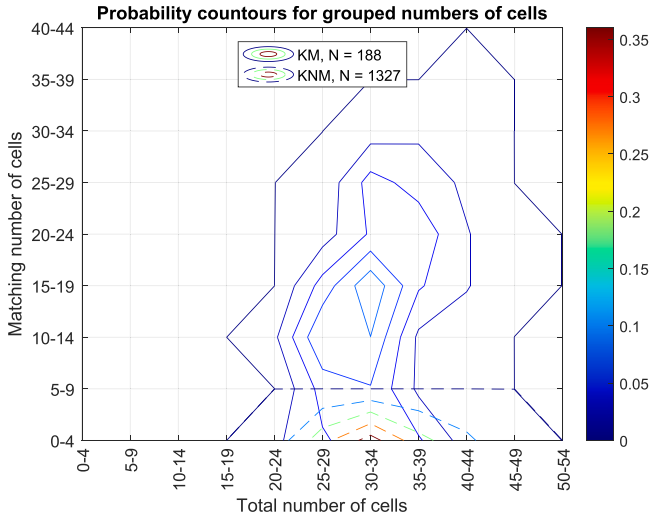

Fig. 22. Contour plot of the observed distributions of the grouped KM (solid lines) and KNM (dashed lines) CMC data of the breechface impression marks. Group sizes of 5 were used.

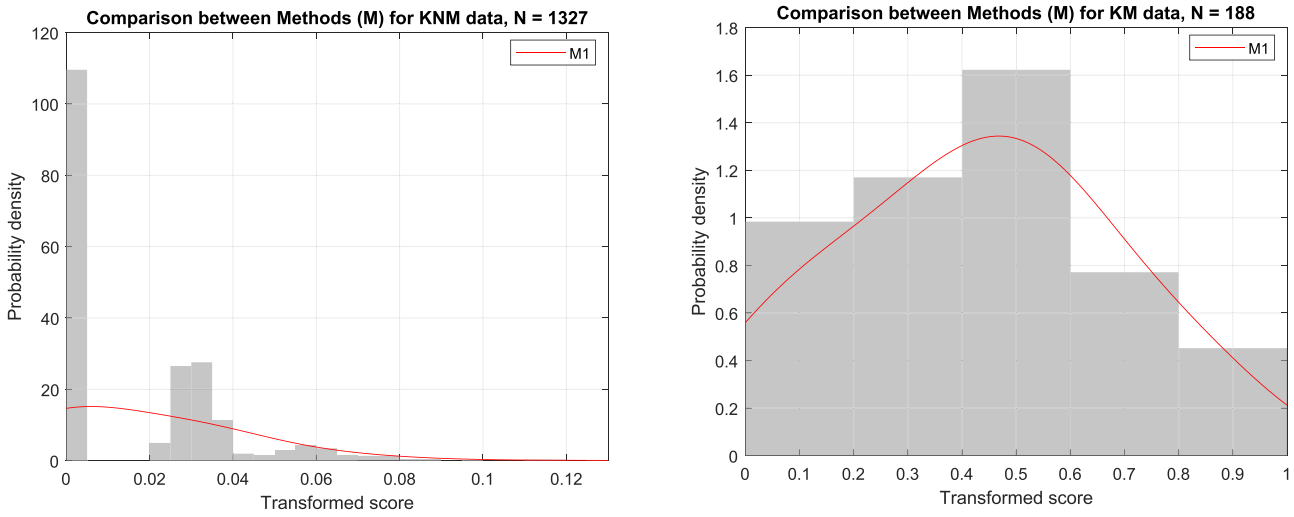

Fig. 23. Histogram of the KNM (left) and KM (right) transformed CMC scores of the breechface impression marks and the estimated probability densities for the non-parametric (M1) model.

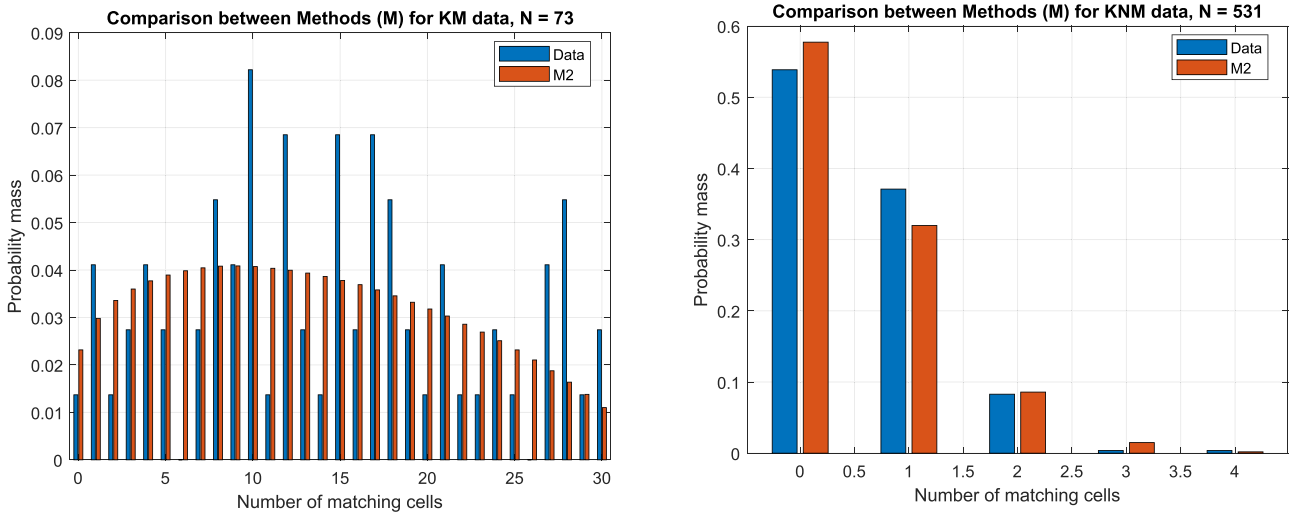

Fig. 24. Histogram of the grouped numbers of KNM (left) and KM (right) CMCs of the breechface impression marks and the estimated probability densities for the parametric (M2) model.

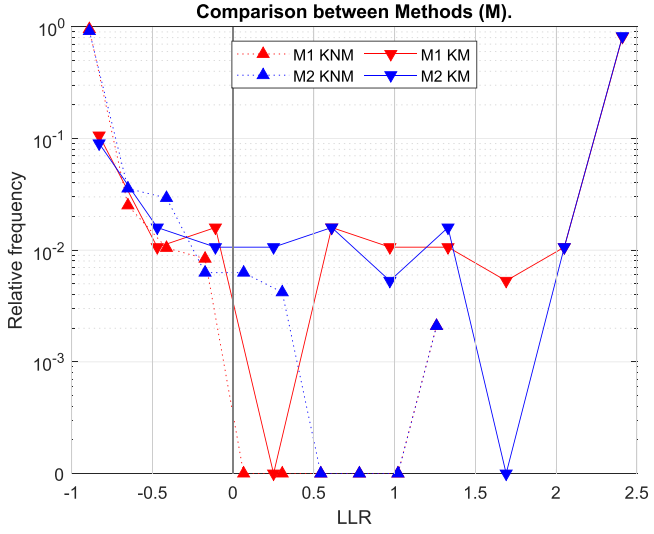

Fig. 25. Relative frequencies of the predicted LLRs for both the KNM and KM datasets, using the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the breechface impression marks.

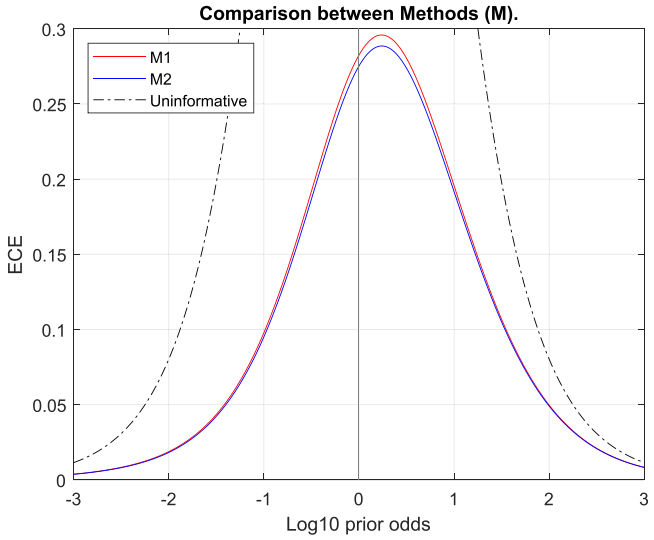

Fig. 26. Empirical Cross Entropy (ECE) of the non-parametric (M1; red) and parametric (M2; blue) models, for the CMC-scores of the breechface impression marks.

Aperture shear – CCF_{max}

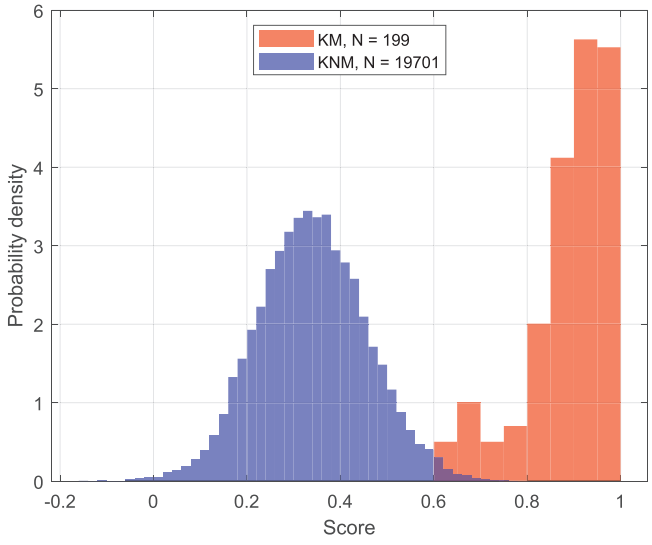

Fig. 27. Normalized histograms of the untransformed scores for both the KNM (blue) and KM (red) datasets of the CCF_{max}- scores of the aperture shear marks.

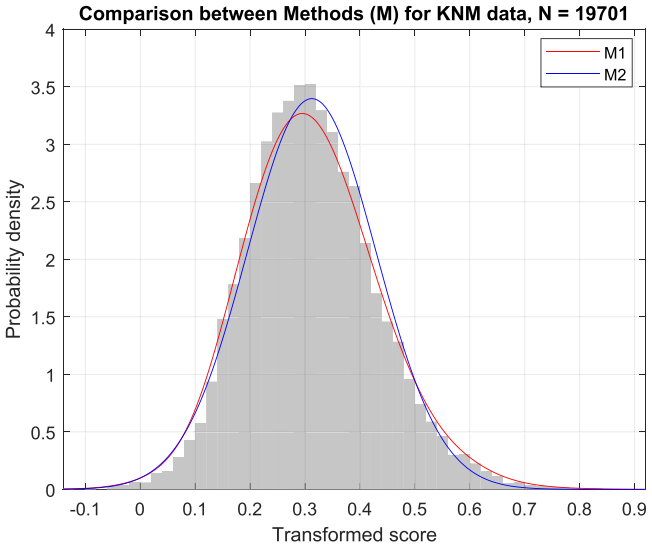

Fig. 28. Histogram of the original KNM data and the resulting probability densities of the non-parametric (M1; red) and parametric (M2; blue) models, for the transformed CCF_{max}- scores of the aperture shear marks.

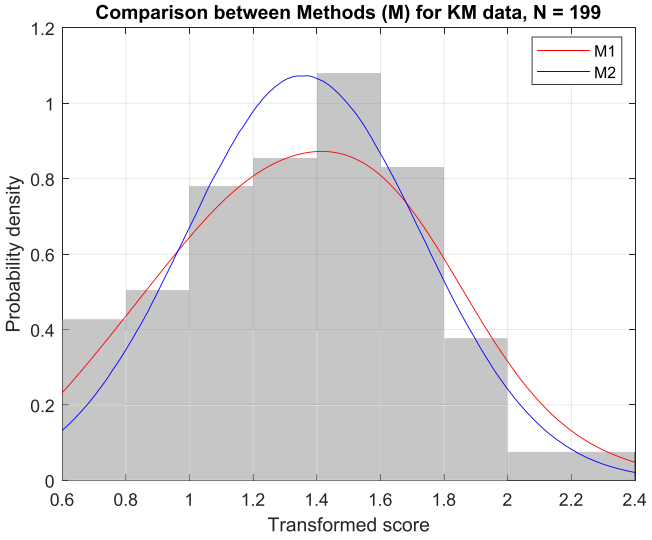

Fig. 29. Histogram of the original KM data and the resulting probability densities of the non-parametric (M1; red) and parametric (M2; blue) models, for the transformed CCF_{max}- scores of the aperture shear marks.

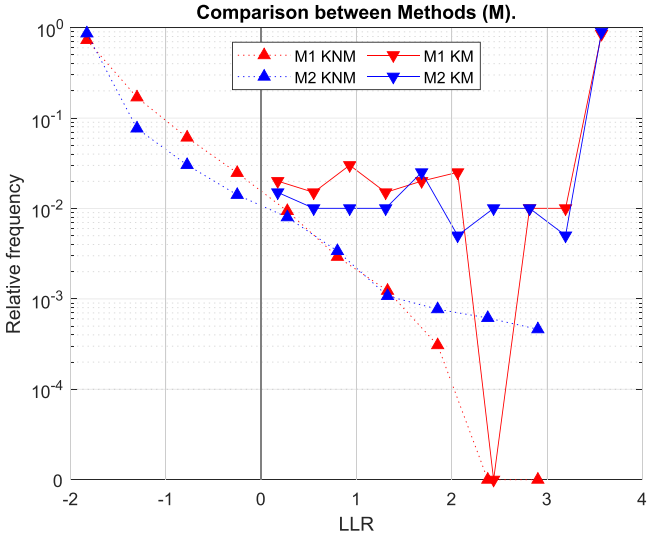

Fig. 30. Relative frequencies of the predicted LLRs for both the KNM and KM datasets, using the non-parametric (M1; red) and parametric (M2; blue) models, for the CCF_{max}- scores of the aperture shear marks.

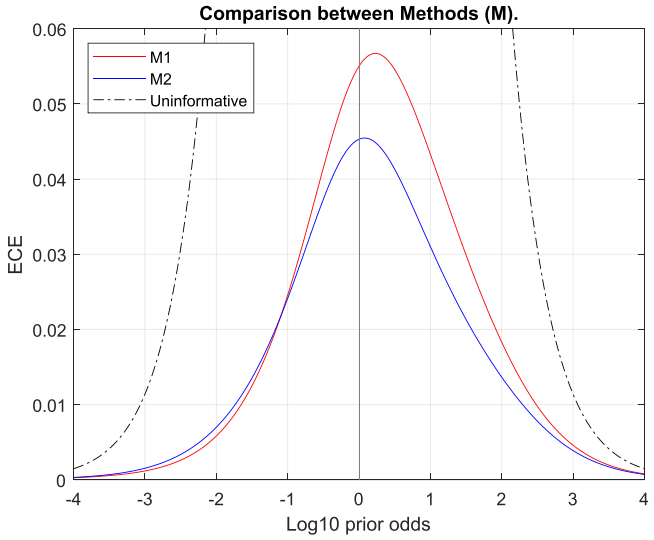

Fig. 31. Empirical Cross Entropy (ECE) of the non-parametric (M1; red) and parametric (M2; blue) models, for the CCF_{max}- scores of the aperture shear marks.

References

[1] Committee on Identifying the Needs of the Forensic Sciences Community: National Research Council, Strengthening Forensic Science in the United States: A Path Forward 2009 The National Academies Press, Washington, DC, USA.

[2] Lander, E.S., U.S. President's Council of Advisors on Science and Technology (PCAST), et al., Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016. ISBN-13: 9781539172000.

[3] F. Riva, C. Champod, Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases, J. Forensic Sci. 59 (3) (2014) 637–647.

[4] F. Riva, et al., Comparison and interpretation of impressed marks left by a firearm on cartridge cases – Towards an operational implementation of a likelihood ratio based technique, Forensic. Sci. Int. 313 (2020), 110363.

[5] J. Song, Proposed "NIST Ballistics Identification System (NBIS)" based on 3D topography measurements on correlation cells, AFTE J. 45 (2) (2013) 184–194.

[6] J. Song, Proposed "Congruent Matching Cells (CMC)" method for ballistic identification and error rate estimation, AFTE J. 47 (3) (2015) 177–185.

[7] J. Song, et al., 3D topography measurements on correlation cells—a new approach to forensic ballistics identifications, Meas. Sci. Technol. 25 (6) (2014), 064005.

[8] J. Song, et al., Estimating error rates for firearm evidence identifications in forensic science, Forensic Sci. Int 284 (2018) 15–32.

[9] X.H. Tai, W.F. Eddy, A fully automatic method for comparing cartridge case images, J. Forensic Sci. 63 (2) (2018) 440–448.

[10] S. Yammen, P. Muneesawang, Cartridge case image matching using effective correlation area based method, Forensic. Sci. Int. 229 (1–3) (2013) 27–42.

[11] H. Zhang, et al., Pilot study of feature-based algorithm for breech face comparison, Forensic. Sci. Int. 286 (2018) 148–154.

[12] H. Zhang, et al., Correlation of firing pin impressions based on congruent matching cross-sections (CMX) method, Forensic. Sci. Int. 263 (2016) 186–193.

[13] M. Baiker, et al., Quantitative comparison of striated toolmarks, Forensic. Sci. Int. 242 (2014) 186–199.

[14] E.J.A.T. Mattijssen, et al., Validity and reliability of forensic firearm examiners, Forensic. Sci. Int. 307 (2020), 110112.

[15] T.B. Weller, M. Duez, P. Lilien, Introduction and initial evaluation of a novel three dimensional imaging and analysis system for firearm forensics, AFTE J. 47 (4 - Fall) (2015) 11.

[16] B. Bachrach, Development of a 3D-based automated firearms evidence comparison system, J. Forensic. Sci. 47 (2002) 6.

[17] Z. Chen, et al., Fired bullet signature correlation using the Congruent Matching Profile Segments (CMPS) method, Forensic. Sci. Int. 305 (2019), 109964.

[18] Z. Chen, et al., Pilot study on deformed bullet correlation, Forensic. Sci. Int 306 (2020), 110098.

[19] L.S. Chumbley, et al., Validation of tool mark comparisons obtained using a quantitative comparative statistical algorithm, J. Forensic. Sci. 55 (4) (2010) 953–961.

[20] P. De Smet, et al., Experimental evaluation of the impact of seating depth variations on observed marks on primers, Forensic. Sci. Int. 179 (2–3) (2008) 163–171.

[21] C. Gambino, et al., Forensic surface metrology: tool mark evidence, Scanning 33 (5) (2011) 272–278.

[22] G.S. Morrison, E. Enzinger, Score based procedures for the calculation of forensic likelihood ratios - Scores should take account of both similarity and typicality, Sci. Justice 58 (1) (2018) 47–58.

[23] W. Chu, et al., Striation density for predicting the identifiability of fired bullets with automated inspection systems, J. Forensic Sci. 55 (5) (2010) 1222–1226.

[24] D. Ott, et al., Identifying persistent and characteristic features in firearm tool marks on cartridge cases, Surf. Topogr.: Metrol. Prop. (2017) **5**.

[25] D. Ott, R. Thompson, J. Song, Applying 3D measurements and computer matching algorithms to two firearm examination proficiency tests, Forensic. Sci. Int. 271 (2017) 98–106.

[26] Z. Chen, et al., A convergence algorithm for correlation of breech face images based on the congruent matching cells (CMC) method, Forensic. Sci. Int. (2017) 213–223.

[27] E.J.A.T. Mattijssen, et al., Firearm examination: examiner judgments and computer-based comparisons, J. Forensic. Sci. 66 (1) (2021) 96–111.

[28] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, Ann. Appl. Stat. 11 (4) (2017) 2332–2356.

[29] C. Aitken, F. Taroni, S. Bozza, Statistics and the evaluation of evidence for forensic scientists, Wiley,, 2020. ISBN-13: 9781119245223.

[30] G.S. Morrison, Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio, Aust. J. Forensic Sci. 45 (2) (2013) 173–197.

[31] Alicona Imaging GmbH, G, Austria. Infinite Focus SL. Accessed: 16 June 2022; Available from: https://www.alicona.com/products/infinitefocuss⟨l⟩.

[32] F. Helmli, Focus variation instruments, Book.: Opt. Meas. Surf. Topogr. (2011) 131–166.

[33] Loci Forensics B.V., N.-V., The Netherlands. ForensicSil. Accessed: 12 July 2022; Available from: http://www.lociforensics.nl/forensic-sil/forensic-sil.

[34] X.A. Zheng, et al., Reference Population Database of Firearm Toolmarks (RPDFT). in 53rd Annual, Training Seminar of the Association of Firearm and Tool Mark Examiners,, Atlanta, USA, 2022.

[35] Standardization, I.Of ISO 16610–71:2014, Geometrical product specifications (GPS) — Filtration — Part 71: Robust areal filters: Gaussian regression filters. [Accessed: 20 September 2022]; Available from: ⟨https://www.iso.org/standard/60159.html⟩.

[36] J.B.A. Maintz, M.A. Viergever, A Survey of medical image registration, Med. Image Anal. 2 (1) (1998) 1–36.

[37] B. Zitová, J. Flusser, Image registration methods: a survey, Image Vis. Comput. 21 (11) (2003) 977–1000.

[38] Ommen, D.M., Approximate Statistical Solutions to the Forensic Identification of Source Problem. 2017, PhD thesis, South Dakota State University.

[39] I.N. Van Dorp, et al., Value of evidence in the rare type match problem: common source versus specific source, Law Probab. Risk 19 (1) (2020) 85–98.

[40] A.J. Leegwater, et al., Performance Study of a Score-based Likelihood Ratio System for Forensic Fingermark Comparison, J. Forensic Sci. 62 (3) (2017) 626–640.

[41] G.S. Morrison, Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate, Sci. Justice 56 (5) (2016) 371–373.

[42] P. Vergeer, et al., Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating, Sci. Justice 56 (6) (2016) 482–491.

[43] P. Vergeer, et al., Measuring calibration of likelihood-ratio systems: a comparison of four metrics including a new metric devPAV, Forensic. Sci. Int. 321 (2021), 110722.

[44] D. Ramos, et al., Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci. 58 (6) (2013) 1503–1518.

[45] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC,, 1986. ISBN-13: 978-0412246203.

[46] A. van den Hout, I. Alberink, Posterior distributions for likelihood ratios in forensic science, Sci. Justice 56 (5) (2016) 397–401.
[47] I. Alberink, R.S. Bolton-King, S. Menges, Posterior likelihood ratios for evaluation of forensic trace evidence given a two-level model on the data, J. Appl. Stat. 40 (12) (2013) 2579–2600.
[48] S.D. Team Stan Model. Lang. Users Guide Ref. Man. 2 2022 30.
[49] G.J. Székely, et al., Measuring and testing dependence by correlation of distances, Ann. Stat. 35 (6) (2007) 2769–2794.
[50] W. Bergsma, A. Dassios, A consistent test of independence based on a sign covariance related to Kendall's tau, Bernoulli 20 (2) (2014) 1006–1028.
[51] TauStar: Efficient Computation and Testing of the Bergsma-Dassios Sign Covariance R package v.1.1.4 [Accessed: 28 August 2023]; Available from: ⟨https://cran.r-project.org/web/packages/TauStar/index.html⟩.
[52] K. Addinall, et al., The effect of primer cap material on ballistic toolmark evidence, Forensic. Sci. Int. 298 (2019) 149–156.