Forensic Science International 317 (2020) 110502

Contents lists available at ScienceDirect

Forensic Science International

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

Evaluating Likelihood Ratio (LR) for firearm evidence identifications in forensic science based on the Congruent Matching Cells (CMC) method

John Song[a,*], Zhe Chen[a,b], Theodore V. Vorburger[a], Johannes A. Soons[a]

[a] Physical Measurement Laboratory (PML), National Institute of Standards and Technology (NIST), Gaithersburg, MD, 20899, USA
[b] School of Mechatronics Engineering, Harbin Institute of Technology (HIT), Harbin, 150001, China

ARTICLE INFO

Article history:
Received 19 February 2020
Received in revised form 6 July 2020
Accepted 2 September 2020
Available online 15 September 2020

Keywords:
Firearm identification
Congruent Matching Cells (CMC)
Forensics
Likelihood Ratio (LR)

ABSTRACT

Firearm evidence identification has been challenged by the 2008 and 2009 National Research Council (NRC) reports and by legal proceedings on its fundamental assumptions, its procedure involving subjective interpretations, and the lack of a statistical foundation for evaluation of error rates or other measures for the weight of evidence. To address these challenges, researchers of the National Institute of Standards and Technology (NIST) recently developed a Congruent Matching Cells (CMC) method for automatic and objective firearm evidence identification and quantitative error rate evaluation. Based on the CMC method, a likelihood ratio (LR) procedure is proposed in this paper aiming to provide a scientific basis for firearm evidence identification and a method for evaluation of the weight of evidence. The initial LR evaluations using two sets of 9 mm cartridge cases' breech face impression images with different sample sizes, imaging methods and ammunition showed that for all the declared identifications of the tested 2D and 3D image pairs, the evaluated LRs for the least favorable scenario were well above an order of 10^6, which provides Extremely Strong Support for a prosecution proposition (e.g. a same-source proposition) in a Bayesian frame. The LR evaluations also showed that for all the declared exclusions of the tested 3D image pairs, the evaluated LRs for the least favorable scenario were above an order of 10^2, which provides Moderately Strong Support for a defense proposition (e.g. a different-source proposition) in a Bayesian frame.

Published by Elsevier B.V.

1. Background

Bullets and cartridge cases fired or ejected from guns pick up characteristic surface topographies from the gun parts, resulting in toolmarks of a special kind, called "ballistics signatures", on the surface of the bullets and cartridge cases. Passage of the bullet through the gun barrel results in striation signatures. Impact of the cartridge case with the firing pin, breech face and ejector results in impression signatures. Both the striation and impression signatures are predominantly considered to be unique and reproducible to the firearm [1]. By analyzing these ballistics signatures, firearm examiners can connect a firearm to criminal acts [1].

Side-by-side tool mark comparison for firearm evidence identification has a history of more than a hundred years. However, the scientific foundation of firearm and tool mark identification has been called into question by several government funded reports in 2008, 2009 and 2016 [2–4] and by recent court decisions [2] on its fundamental assumptions, its procedure involving subjective interpretations, and the lack of a statistical foundation for evaluation of error rates or other measures for the weight of evidence.

Firearm evidence identification based on image comparison is fundamentally probabilistic in nature [2–5]. However, at present, most experts and institutes in the U.S. present the results of a bullet or cartridge case comparison in a "yes/no/inconclusive" range of conclusions [1] without a quantitative statement of error rate and likelihood ratio (LR). It ignores the probabilistic nature of firearm identification, and forces the expert to either defend a "yes" or "no" position based on more subjective grounds, or throw away valuable information by merely giving an "inconclusive" result [5].

There are several reasons underlying the existing practice [5]. The most important one might be the statement by the Committee for the Advancement of the Science of Firearm & Toolmark Identification that a condition of sufficient agreement between signatures implies that the likelihood that two different firearms generated the marks is so remote as to be considered a practical impossibility [1,5]. Another reason for the use of a yes/no/inconclusive format, instead of a probabilistic one, might be that

* Corresponding author.
E-mail address: song@nist.gov (J. Song).

http://dx.doi.org/10.1016/j.forsciint.2020.110502
0379-0738/Published by Elsevier B.V.

the current practice of bullet or cartridge case comparison does not result in objective numerical values for the agreement of the signatures [5].

Since the 1980's, estimates of likelihood ratio and coincidental match probability (CMP) have been used for specifying uncertainty of DNA identifications: "*The courts already have proven their ability to deal with some degree of uncertainty in individualizations, as demonstrated by the successful use of DNA analysis (with its small, but nonzero, error rate)*" [3]. It is therefore a fundamental challenge in forensic science to establish a scientific foundation and statistical method for probabilistic, quantitative expressions of the weight of evidence to support firearm evidence identifications, in the same way that reporting procedures have been established for forensic identification of DNA evidence [3,4]. Several experimental and theoretical efforts have been pursued along this line including the machine learning approach of Petraco et al. [6,7], the work on likelihood ratio by Riva et al. [8], the study of examiner error rates by Baldwin et al. [9] and Mattijssen et al. [10], the feature-based matching algorithm of Lilien et al. [11,12], the random forest approach of Hare et al. [13], and the work on image cross correlations and congruent matching cells (CMC) of Song et al. [14–23].

For firearm evidence identifications, in an "*ideal world*" – as stated by Kerkhoff et al. [5], "*the appropriate numbers for each individual case could be established with an objective, quantitative method.*" The expert would then be able to formulate an identification or exclusion conclusion associated with an error rate and likelihood statement, based on which the judge or jury can assess this information in combination with other aspects of the case and decide whether "*they are willing to assume*" that a bullet or cartridge case was indeed fired from a submitted gun [5]. In the ideal world, the expert provides, in a scientifically sound way, a quantitative measure for the weight of the evidence, and the judge or jury decides whether to accept the evidence and what weight to assign to it [5]. This raises an interesting question about the yes/no/inconclusive conclusion format: "*Why should the expert decide on the yes/no question in the absence of numbers, if it would be left to a judge or jury if numbers were available*? [5]"

In this paper, we aim towards the "*ideal world*" to provide "*the appropriate numbers for each individual case*" as mentioned above [5], i.e. to establish an automated and objective method for firearm evidence identification and quantitative LR evaluation procedure based on the congruent matching cells (CMC) method developed at NIST [16–23], and to provide a statistical basis for firearm evidence identifications in forensic science. The initial LR evaluation results using two sets of 9 mm cartridge cases' breech face impression images with different sample sizes, imaging methods and ammunitions show that, for all the declared identifications of the tested 2D and 3D image pairs, the evaluated LRs for the least favorable scenario were well above an order of 10^6. According to the verbal scale example of the 2010 Guideline of the European Network of Forensic Science Institutes (ENFSI) [24], those LRs provide an "Extremely Strong Support" to the forensic findings for the first proposition (same-source) compared to the alternative (different-source). The LR evaluation results also show that, for all the declared exclusions of the tested 3D image pairs, the evaluated LRs for the least favorable scenario are above an order of 10^2, which provides "Moderately Strong Support" of the forensic findings for the first proposition (different-source) compared to the alternative (same-source) [24].

In the following, we introduce the CMC method for automatic and objective firearm identification and error rate evaluation in Section 2, then discuss LR evaluation based on the CMC method in Section 3. In Section 4, we introduce initial LR evaluation results, and finally, we discuss our results and future work in Section 5.

2. CMC method for firearm evidence identifications and error rate evaluation

2.1. CMC method for firearm evidence identification

The congruent matching cells (CMC) method was developed for the objective correlation of impressed toolmarks [16,17]. This method is based on discretization – it divides a toolmark image into small correlation cells and uses pairwise cell correlations instead of correlation of the entire images. Multiple identification parameters are defined for quantifying the topography similarity of the correlated cell pairs, and the pattern congruency of the cell registration locations on both images [16,17]. The set of cell pairs that meet both the topography similarity and pattern congruency qualifications are defined as congruent matching cell pairs (CMCs). When the correlation result shows that the number of CMCs is equal or larger than an identification criterion say, CMC $\geq$ 6 [16,17] (or above a given LR threshold if the LR is used for reporting the weight of evidence), one might consider that the correlated images originate from the same firearm with a stated error rate (or LR).

The initial validation tests for the CMC method were conducted using a set of breech face impression images of cartridge cases, called the Fadul dataset consisting of 40 cartridge cases ejected from guns with 10 consecutively manufactured pistol slides [25]. The breech face impression topographies were captured by a disk scanning confocal microscope [18]. Fig. 1 shows the test results consisting of 717 known non-matching (KNM) and 63 known matching (KM) image pairs. The KM and KNM distributions show a significant separation. The number of congruent matching cell pairs (CMCs) for the 63 KM topography pairs ranges from 9 to 26; while the number of CMCs for the 717 KNM topography pairs ranges from 0 to 2. With a predetermined identification criterion for the number of CMC cells C $\geq$ 6 [16,17], there are no false positive or false negative results. The estimated KM and KNM distribution models enable choices for the C value based on the acceptable false positive/negative error rates. The estimated KM and KNM distribution models also enable LR evaluation for reporting the weight of evidence. Note that the identification criterion C is not

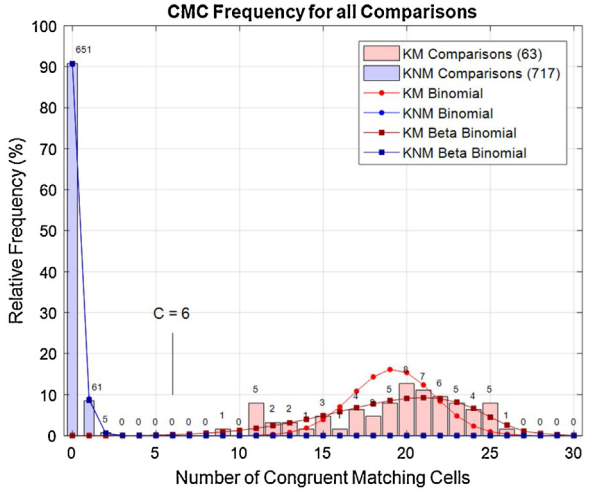

Fig. 1. Relative frequency distribution of image pairs vs. CMC number for 63 KM and 717 KNM image pairs. The KM and KNM distributions are each scaled to their sample size. The red and brown curves represent binomial and β-Binomial distribution models, respectively, for the KM data. The overlapping blue curves represent the two models for the KNM data. Note that the distribution models are discrete histograms, with the connecting lines drawn for visualization. The number of image pairs having a specific CMC value is shown just above each bar in the histograms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

requested when presenting the weight of evidence for a CMC correlation result using the LR approach.

2.2. Statistical data fitting based on the Binomial and β-Binomial models

The initial data fitting as shown in Fig. 1 was based on two statistical models: Binomial model and β-Binomial model with a fixed cell number N [23]. For the 717 KNM image pairs, both models showed very good fitting results: the overlapping blue curves represent the two model fittings for the KNM data. However, for the 63 KM image pairs, the red and brown curves representing Binomial and β-Binomial distribution models, respectively, which showed differences in the fitting results: the β-Binomial distribution model (brown) showed a closer fit to the KM data than the Binomial distribution model (to be discussed in Section 5).

It must be noted that, the actual cell number N for each cell pair's comparison is a variable. N is determined by the specific location of the cell grids on the reference images, and it will affect the shape of the β-Binomial distribution. In our initial fitting of the β-Binomial model, we started with a fixed cell number N. As a result, the fitted curve of the β-Binomial model was a smooth curve as shown in Fig. 1. [23]. Considering the actual cell numbers N varying between cell pair comparisons, we used the actual cell numbers N to fit the β-Binomial model. As a result, the fitting results showed many ups and downs [23]. Fig. 2 shows an additional test on the same set of Fadul breech face impressions using an improved version of the correlation software [22] and using the variable cell number N.

The experimental distributions for the KM and KNM image pairs in Fig. 2 have a larger separation than those shown in Fig. 1. For the 63 KM image pairs, the CMC values range from 15 to 30 and the number N of evaluated cell pairs ranges from 19 to 31. For the 717 KNM image pairs, the number of correlated CMCs ranges from 0 to 2 with N ranging from 16 to 32.

When we fit the histogram data of Fig. 2 with the variable cell numbers N [23], the red-cross curve in Fig. 2 represents a β-Binomial distribution model for the KM data, and the blue-dot curve represents a Binomial model for the KNM data. The green line shows the identification criterion C = 6 [16,17]. The brown line, C = 8, shows an alternative identification criterion located approximately midway between the KM and KNM correlation results [23].

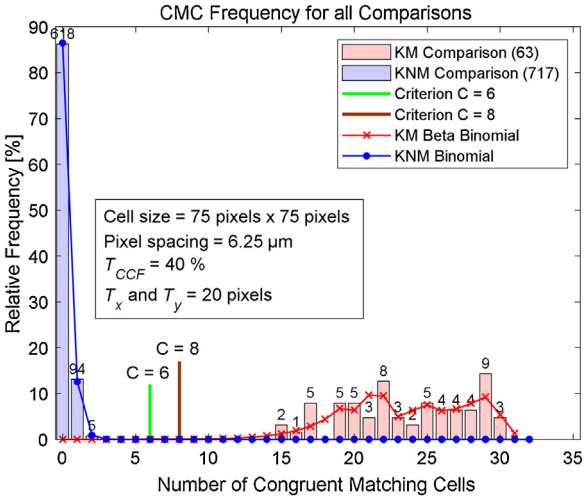

Fig. 2. Relative frequency distribution of image pairs vs. CMC number for 63 KM and 717 KNM image pairs of the Fadul dataset. The KM and KNM distributions are each scaled to their sample size.

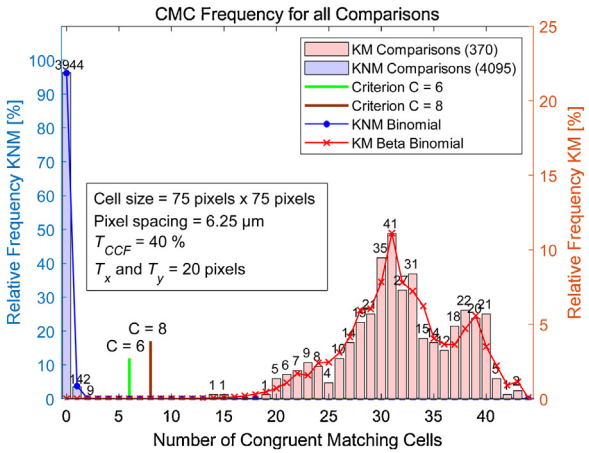

Fig. 3. Relative frequency distribution of CMC numbers for 370 KM and 4095 KNM image pairs of the Weller dataset [26].

The use of the variable cell number models in Fig. 2 does not change the fitted KNM distribution significantly but does produce a KM distribution with slope discontinuities. As a result, it fits the KNM histogram data significantly better than fitting either the Binomial of the β-Binomial model (Fig. 1). For the KNM data, we obtained a maximum likelihood estimator for the probability of a cell to be a CMC cell, $\hat{p}_{KNM}$ [23], of 0.0053. For the KM data, we obtained maximum likelihood estimators for the $\hat{\alpha}$ and $\hat{\beta}$ parameters of the β-Binomial model [23], of 9.22 and 0.87, respectively.

Fig. 3 shows another validation test using the Weller dataset consisting of 4465 topography image pairs of 95 cartridge cases ejected from guns with 11 pistol slides, 10 of which were consecutively manufactured [26]. The experimental distributions for the KM and KNM image pairs also show significant separation. For the 370 KM cartridge pairs, the number of CMCs ranges from 14 to 43, and N ranges from 22 to 44. For the 4095 KNM cartridge pairs, the number of CMCs again ranges from 0 to 2, and N ranges from 22 to 45. The different values of N are again taken into account when fitting the Binomial distribution model to the KNM data, yielding a value for $\hat{p}_{KNM}$ [23] of 0.0011, and when fitting the β-Binomial distribution to the KM data, yielding values for $\hat{\alpha}$ and $\hat{\beta}$ [23] of 10.61 and 0.81, respectively. The resulting modeled frequency distribution for the KM values of the dataset provides a better fit to the experimental data when using the variable cell numbers N, as shown in Figs. 2 and 3.

2.3. Error rate evaluation based on the CMC method

Error rates can be considered from two points of view [27]. The first point of view addresses the reliability of the identification system and procedure [27]. This reliability is specified and characterized by the cumulative false positive and false negative error rates for a given set of KM and KNM samples with an identification criterion C [17,23]. The cumulative false positive error rate represents the probability of obtaining erroneous result of identifications (declared matches) when comparing samples from different sources (KNM). The cumulative false negative error rate represents the probability of obtaining erroneous result of exclusions (declared non-matches) when comparing samples from the same source (KM). The cumulative false positive and false negative error rates can be used as a measure of the reliability of the identification system and procedure. In this paper and our previous publication [17,23], the cumulative false positive and false negative error rates for a given identification system are represented by E_1 and E_2.

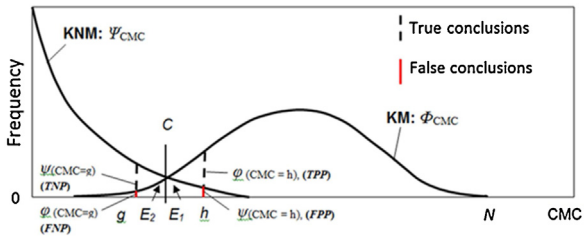

Fig. 4. Conceptual diagram of the CMC probability mass functions for KM and KNM comparisons, Φ_{CMC} and Ψ_{CMC} [17,23]. CMC is the number of congruent cells. The regions E_1 and E_2 under the curves represent cumulative false positive and false negative error rates. The lines at h and g represent probabilities of specific CMC scores representing an identification conclusion (CMC = h) with true positive and false positive probability (*TPP* and *FPP*); or an exclusion conclusion (CMC = g) with true negative and false negative probability (*TNP* and *FNP*) [17,23].

The second point of view addresses the probability of incorrect conclusion for an identification (declared match) or exclusion (declared non-match) from a case work [27]. This probability is of interest during legal proceedings. For example, when a firearms examiner concludes that the evidence and reference items are from the same source, an attorney may ask: "What is the probability that these two items are actually from different sources?" This probability is represented by an individual false positive error rate R_1 for a given evidence/reference image pair to be concluded erroneously as derived from the same source. The symbol R_2 represents the error rate for the opposite condition [23].

The large number of pairwise cell correlations using the CMC method can facilitate a statistical approach to modelling the distribution of correlation scores. These distributions can be used to evaluate cumulative false positive and false negative error rates and LRs. Fig. 4 shows a conceptual diagram of the CMC probability mass functions for KM and KNM correlations: Φ_{CMC} and Ψ_{CMC}. CMC is the number of congruent cells. To illustrate clearly the listed quantities, the schematic depicts the discrete probability distributions as continuous density functions that overlap much more than they would be expected to in practice. We use an identification criterion C to facilitate the description of the concepts. However, as we mentioned before, C is not a necessary factor for LR estimation. The LR value could be presented directly to the judge or jury without an identification criterion C and a declared match/non-match statement of conclusion, that would be left to a judge or jury when they assess the LR information in combination with other aspects of the case, and decide whether "*they are willing to assume*" that a bullet or cartridge case was indeed fired from a submitted gun [5].

The regions E_1 and E_2 under the curves represent the cumulative false positive and false negative error rates [17,23], which describe the reliability of the identification system and procedure as specified in Ref. 27. E_1 and E_2 can be calculated from the KM and KNM probability mass functions Φ_{CMC} and Ψ_{CMC} associated with the known number of correlated cells N and the identification criterion C [23]:

$$E_1 = \sum_{CMC=C}^{CMC=N} \Psi_{(CMC)} = \Psi_{(CMC=C)} + \Psi_{(CMC=C+1)} + \cdots + \Psi_{(CMC=N)} = 1 - \left(\Psi_{(CMC=0)} + \Psi_{(CMC=1)} + \cdots + \Psi_{(CMC=C-1)}\right) \quad (1)$$

$$E_2 = \sum_{CMC=0}^{CMC=C-1} \Phi_{(CMC)} = \Phi_{(CMC=0)} + \Phi_{(CMC=1)} + \cdots \Phi_{(CMC=C-1)}. \quad (2)$$

For the two sets of cartridge case correlations showing in Figs. 2 and 3, the cumulative false positive and false negative error rates E_1 and E_2 are calculated by Eqs. (1) and (2) and shown in Table 1. In our calculations, we use a Binomial distribution model for the KNM probability mass function Ψ_{CMC}, and a β-Binomial model for the KM probability mass function Φ_{CMC}, both with parameters estimated from the experimental data [23]. With the initial identification criterion C = 6 [16,17], the cumulative false positive error rates E_1 are 7.94×10^{-9} and 3.94×10^{-12}, and the cumulative false negative error rates E_2 are 3.86×10^{-5} and 1.05×10^{-6} for Fadul and Weller datasets, respectively. With the alternative identification criterion C = 8 [23], the cumulative false positive error rates E_1 are 2.07×10^{-12} and 8.91×10^{-17}, and the cumulative false negative error rates E_2 are 2.30×10^{-4} and 7.41×10^{-6} for Fadul and Weller dataset, respectively. It can be seen that the cumulative false negative error rates E_2 are significantly larger than the cumulative false positive error rates E_1, because of the larger dispersion of the CMC distributions of the KM image pairs (see Figs. 2 and 3). From a legal perspective, low false positive error rates E_1 are critical in avoiding wrongful convictions.

For each "declared match" CMC score h ($h \geq$ C), Fig. 4 illustrates the probabilities of achieving this score for both KM samples (black dashed bar which extend down to the x-axis) and KNM samples (red bar). For each "declared non-match" CMC score g ($g <$ C), it illustrates the probabilities of achieving this score for both KNM samples (black dashed bar) and KM samples (red bar). Note that in both cases, these probabilities would not represent the probability of an incorrect "identification" or "elimination" conclusion for a case work (it must be represented by the individual false positive and false negative error rate R_1 and R_2 [17,23]). In the case work of firearm identifications, there are two kinds of probabilities needed for evaluation: 1) probability for the "identification" hypothesis of the prosecution proposition *Hp* (i.e. a same-source proposition), and 2) probability for the "elimination" hypothesis of the defense proposition *Hd* (i.e. a different-source proposition). Both evaluations require knowledge about the prior odds on the reference image belonging to the dataset to be referred, i.e., a knowledge about the reference image being submitted for comparison and the reference dataset being referred for evaluation of the weight of evidence. These propositions are often referred to as hypotheses in Bayesian terms [28].

Based on the probabilities of achieving a specific CMC score (h or g) from both KM and KNM samples illustrated in Fig. 4, we can derive a likelihood ratio procedure for firearm evidence identifications.

3. Evaluation of likelihood ratio in firearm evidence identifications

3.1. The likelihood ratio in firearm evidence identifications

The Likelihood Ratio (LR) is a numerical value that expresses the weight of the forensic evidentiary findings *E*. It is obtained by the ratio of the probabilities (Pr) of the findings *E* (i.e. CMC correlation result h or g) under competing propositions from the prosecution (*Hp*) and the defense (*Hd*) [8]:

$$LR = \frac{\Pr(E \mid Hp, I)}{\Pr(E \mid Hd, I)} \quad (3)$$

In cases involving a suspect firearm and a questioned cartridge case (specific-source scenario), these propositions can be formulated in the following way [8]:

Hp: The questioned cartridge case was fired by the suspect firearm.

Table 1
The cumulative false positive and false negative error rates E_1 and E_2 for the two sets of cartridge case correlations shown in Figs. 2 and 3.

Sample set	Cartridge number	Image pairs KNM/KM	Statistical Models KNM/KM	ID Criteria	E_1	E_2	Note
Fadul	40	717/63	Bi./β-Bi.	C = 6	7.94×10^{-9}	3.86×10^{-5}	Fig. 2
			Bi./β-Bi.	C = 8	2.07×10^{-12}	2.30×10^{-4}	Fig. 2
Weller	95	4095/370	Bi./β-Bi.	C = 6	3.94×10^{-12}	1.05×10^{-6}	Fig. 3
			Bi./β-Bi.	C = 8	8.91×10^{-17}	7.41×10^{-6}	Fig. 3

Hd: The questioned cartridge case was not fired by the suspect firearm but by another unknown firearm.

The background information (I) represents case-specific information that may help to specify the nature of the relevant population considered under the proposition *Hd* [8].

3.2. Evaluating LR based on the CMC method

Different methods are used for evaluation of LR values, including feature-based and score-based methods [29–31]. Typical usage of the feature-based method is the LR evaluation of DNA tests, while the score-based method is typically used for pattern evidence, such as firearm toolmarks. In a score-based method, the LR value is calculated from a comparison or a correlation score that summarizes the degree of similarity between the relevant properties of two samples [23,31]. For a specific-source scenario in firearm identification, the reference sample is typically a test fire obtained from the suspect firearm, and the questioned (trace or evidence) sample is a sample found at the crime scene.

As a starting point of LR evaluation, we assume that the questioned and the reference toolmark images are random draws from the same statistical population whose KM and KNM score distributions have been previously characterized. As shown in Fig. 4 right, the black dashed bar (which extends down to the x-axis) represents the probability of obtaining a declared match score h ($h \geq$ C) assuming the proposition *Hp* that both samples are from the same firearm is true. This is the true positive probability *TPP*, or $\Pr(\mathrm{CMC} = h \mid Hp,\ I)$ [28]. The red bar on the right side of Fig. 4 represents the probability of a declared match score h ($h \geq$ C) assuming the proposition *Hd* that both samples are from different firearms is true. This is the false positive probability *FPP*, or $\Pr(\mathrm{CMC} = h \mid Hd,\ I)$ [28]. From Eq. (4) and Ref. [28], the ratio of *TPP* vs. *FPP*, or $\Pr(\mathrm{CMC} = h \mid Hp,\ I$ vs. $\Pr(\mathrm{CMC} = h \mid Hd,\ I)$ represents the true positive likelihood ratio (LR_1) for a declared identification resulting from a correlation result of CMC = h ($h \geq$ C, see Fig. 4, right):

$$LR_{1(CMC=h)} = \frac{\Pr(\mathrm{CMC} = h \mid Hp,\ I)}{\Pr(\mathrm{CMC} = h \mid Hd,\ I)} = \frac{\varphi_{(\mathrm{CMC}=h)}}{\psi_{(\mathrm{CMC}=h)}} = \frac{TPP_{(h)}}{FPP_{(h)}} \quad (h \geq C). \tag{4}$$

Similarly, for CMC = g ($g <$ C, see Fig. 4 left):

$$LR_{2(\mathrm{CMC}=g)} = \frac{\Pr(\mathrm{CMC} = g \mid Hd,\ I)}{\Pr(\mathrm{CMC} = g \mid Hp,\ I)} = \frac{\psi_{(\mathrm{CMC}=g)}}{\varphi_{(\mathrm{CMC}=g)}} = \frac{TNP_{(g)}}{FNP_{(g)}} \quad (g < C). \tag{5}$$

The weight of the forensic findings is essentially a relative and conditional measure that helps to progress a case in one direction or the other depending on the magnitude of the likelihood ratio [24]. As stated in the ENFSI Guidelines for Evaluative Reporting in Forensic Science: "*The conclusion shall be expressed either by a value of the likelihood ratio and/or using a verbal scale related to the value of the likelihood ratio. The verbal equivalents shall express a degree of support for one of the propositions relative to the alternative. The choice of the reported verbal equivalent is based on the likelihood ratio and not the reverse*" (see Table 2 below) [24].

It must be noted that both the LR value and the LR evaluation procedure are independent of any identification criterion C. However, for the practical usage of CMC method for evaluating of error rates [17,23], two identification criteria C_1 and C_2 for firearm evidence identification could be used: a high criterion C_1 for identification conclusions and a low criterion C_2 for exclusion conclusions. The gap between C_1 and C_2 would be an inconclusive region. The values of C_1 and C_2 can be determined by the proposed LR values (Table 2) [24].

4. Initial results of likelihood ratio evaluation

4.1. Influence quantities for likelihood ratio evaluation

Based on the KM and KNM distributions developed from the CMC method [17,23], and the relationship between LR and the true positive and false positive probabilities *TPP* and *FPP* (see Fig. 4 right and Eqs. (3) and (4)), we can evaluate the true positive likelihood ratio $LR_{1(\mathrm{CMC}=h)}$ for identification results concluded by CMC = h ($h \geq$ C). We can also evaluate the true negative likelihood ratio $LR_{2(\mathrm{CMC}=g)}$ based on the relationship between LR and the true negative and false negative probabilities *TNP* and *FNP* (see Fig. 4 left and Eqs. (3) and (5)) for elimination results concluded by CMC = g ($g <$ C). However, as mentioned above, from a legal perspective, the true positive likelihood ratios LR_1 are critical for firearm evidence identifications as they can be used to yield a probability of false positives (false identifications), which are to be avoided at almost any cost.

Table 2
The values of likelihood ratio and the verbal equivalents proposed as examples in the 2010 ENFSI Guideline for Evaluative Reporting in Forensic Science [24].

Values of likelihood ratio	Verbal scale
1	The forensic findings do not support one proposition over the other
2–10	Weak support of the forensic findings for the first proposition compared to the alternative
10–100	Moderate support . . . . .
100–1000	Moderately strong support . . .
1000–10,000	Strong support . . .
10,000–1,000,000	Very strong . . .
1,000,000 and above	Extremely strong . . .

Table 3
Likelihood ratio evaluation for two sets of samples.

Sample set	Cartridge number	Image pairs KNM/KM	Statistical Models KNM/KM	Number of cells N	CMC values for KNM	True negative $LR_{2(CMC=g_{max})}$	CMC values for KM	True positive $LR_{1(CMC=h_{min})}$	Note
Fadul	40	717/63	Bi./β-Bi.	19	0–2	1.13×10^3	15–30	2.14×10^{29}	Fig. 2
Weller	95	4095/370	Bi./β-Bi.	24	0–2	2.27×10^3	14–43	6.52×10^{32}	Fig. 3

In this context, it should be noted that conclusions of exclusion based on individual characteristics are somewhat controversial. Some labs only allow conclusions of exclusion when there are differences in class characteristics. There are many reasons for this restriction, for example, the firearm could have been tampered with or have excessive wear or damage since the incident, or the firing conditions and ammunition may have been very poor for the replication of individual characteristics. For matched class characteristics without support of matched individual characteristics, these labs may use different classes of inconclusive if agreement between individual characteristics is not sufficient for the declaration of a declared match.

There are several uncertainty sources for LR evaluation. As a starting point, we evaluate possible variation in LR values based on the first three of the following critical influence quantities:

- **Datasets:** The choice of the relevant population regarding firearm breech face manufacturing method or ammunition, as well as the datasets used to represent the population, can have a major influence on the correlation score distributions, in particular those for KM samples. In this study, two datasets were used: 1) the Fadul dataset with 40 cartridges ejected from 10 consecutively manufactured Ruger[1] P95PR15 slides [25] yielding correlation results shown in Fig. 2; and 2) the Weller dataset with 95 cartridges fired from 10 consecutively manufactured Ruger P95DC slides (plus one extra slide) [26] yielding correlation results shown in Fig. 3. The breech faces in both sets were finished using a sand/bead blasting process. The ammunition used in the Fadul dataset (Federal) is different from that used in the Weller dataset (Winchester).
- **2D vs. 3D imaging method:** Historically, comparing 2D reflectance images obtained with a comparison microscope had been the only method for firearm evidence identification. However, 2D images are significantly affected by optical conditions and surface properties that are not relevant to identification, such as lighting direction, image intensity, surface color, surface reflectivity etc., which may significantly affect the firearm evidence identification results [32,33]. In the past 20 years, 3D topography measurement instruments have demonstrated significant advantages. However, the optical comparison microscope is still the instrument used for the one-to-one sample comparisons required for court proceedings. It is therefore of interest to compare the LR ratio evaluations of 2D optical reflective images and 3D topography images using the same set of samples.
- **Statistical model and identification criteria:** The choice of the appropriate statistical model and identification criteria for firearm evidence identification is particularly important. For the KNM image pairs, both the statistical models of Binomial and β-Binomial distributions showed very close results (Figs. 2 and 3, also see [23]). The Binomial model is used here for describing the distribution of the KNM comparison scores. For the KM image pairs, however, the β-Binomial fitted the data more closely than the Binomial model (Figs. 2 and 3, also see [23]). Therefore, the β-Binomial model was used for describing the distribution of the KM comparison scores.
- **Measurement uncertainty:** There are many influence quantities for error rate and LR evaluations, each may have a large variation range that may significantly increase the error rates and decrease the LR values and may result in a large uncertainty range. It will be discussed in Section 5 – Future work.

For each set of LR estimations, we calculate LR values for the different choices of scenario described above and report only the LR values for the least favorable scenario in the histogram data. For example, for the identification results where CMC = h ($h \geq$ C, see Fig. 4, right), we report the minimum true positive likelihood ratio LR_1 obtained at the smallest observed CMC value, that is, $h_{\min} = 14$ ($N = 24$) for the Weller data (Fig. 3) and $h_{\min} = 15$ ($N = 19$) for the Fadul set (Fig. 2). For the exclusion results where CMC = g ($g <$ C, see Fig. 4, left), we report the minimum true negative likelihood ratio LR_2 obtained at the maximum observed CMC score, that is, $g_{\max} = 2$ for both datasets in Figs. 2 and 3. The estimated LR_1 and LR_2 are shown in Table 3.

As mentioned in Section 2.2, the actual cell number N for each cell pair's correlation is variable. In our initial fitting of the β-Binomial model, we started with a fixed cell number N, as a result, the fitted curve of the β-Binomial model was a smooth curve as shown in Fig. 1 [23]. Considering the actual cell numbers vary between cell pair comparisons, when using the actual cell numbers N to fit the β-Binomial model, the fitting results are with so many ups and downs as shown in Figs. 2 and 3 (also see Ref. 23). In the following discussions, we use the variable cell number N to fit the β-Binomial distribution, and use the averaged value at each CMC point of the KM distribution for evaluation of the LR.

4.2. Initial results of LR evaluation

Table 3 shows the initial results of LR evaluation for the two sample sets as shown in Figs. 2 and 3. The respective LR values are presented for two statistical models: Binomial (Bi) for the KNM distributions and β-Binomial (β-Bi) for the KM distributions. For the KM image pairs, the minimum true positive likelihood ratios $LR_{1(\mathrm{CMC}=h_{min})}$ are (2.14×10^{29}) and (6.52×10^{32}) for the two sets of samples, which are well above an order of 10^6 that provides an Extremely Strong Support for a same-source proposition, see Table 2 [24]. For the KNM image pairs, the minimum true negative likelihood ratios $LR_{2(\mathrm{CMC}=g_{max})}$ are (1.13×10^3) and (2.27×10^3), which are above an order of 10^3 that provides a Strong Support for a different-source proposition, see Table 2 [24].

Table 4 shows LR evaluations obtained from 2D optical reflective images and 3D topography images for the same Fadul dataset [25]. The 2D images were captured using a comparison microscope with a top-ring lighting source. Detailed measurement procedures and correlation results can be found in [20]. For both the 2D and 3D image data, the cell size was increased to 500 μ

[1] Certain commercial equipment, instruments, or materials are identified in this paper to specify adequately the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Table 4
Likelihood ratio estimations for the Fadul dataset using 2D reflective images vs. 3D topography images.

Sample set	Cartridge number	Image pairs KNM/ KM	Statistical Models KNM/KM	Number of cells N	CMC values for KNM	True negative $LR_{2(CMC=g_{max})}$	CMC values for KM	True positive $LR_{1(CMC=h_{min})}$	Note
Fadul (2D)	40	717/63	Bi./β-Bi.	21	0–3	0.16	9–28	7.98×10^{12}	Ref. [20]
Fadul (3D)	40	717/63	Bi./β-Bi.	16	0–2	587	11–26	6.72×10^{18}	Ref. [23]

m × 500 μm, yielding a lower number N for 3D image correlations and, on average, lower CMC scores. For the 63 KM image pairs, the minimum true positive likelihood ratio $LR_{1(\mathrm{CMC}=h_{min})}$ is (6.72×10^{18}) for the 3D method, and it decreases to (7.98×10^{12}) for the 2D method. But both are still above an order of 10^6 that provides an Extremely Strong Support for a same-source proposition, see Table 2 [24]. For the 717 KNM image pairs, the minimum true negative likelihood ratio $LR_{2(\mathrm{CMC}=g_{max})}$ is 587 for the 3D method, which provides a Moderately Strong Support for a different-source proposition, see Table 2 [24]. However, for the 717 KNM image pairs using the 2D method, the minimum true negative likelihood ratio $LR_{2(CMC=g_{max})}$ is significantly reduced to less than one (LR = 0.16). As mentioned before, the quality of 2D optical reflective images largely depends on the surface properties and lighting conditions. For example, when changing the lighting direction only a few degrees for one image of the compared image pair, a perfect matched sample pair may result in a declared non-match due to the change in reflections and shadows of the compared image [32,33]. The higher variability of the image data results in a higher probability of false negatives, which may be reflected by the lower value (less than 1) of the true negative likelihood ratio LR_2.

5. Conclusion, discussion and future work

Likelihood ratio evaluation aims to provide a quantitative measure for the weight of evidence in firearm identifications. It makes it possible for ballistics experts to formulate, in a scientifically sound way, an identification or elimination conclusion associated with a LR statement. Based on the LR values and the verbal equivalents, the judge or jury can assess this information in combination with other aspects of the case and decide whether "*they are willing to assume*" [5] that a bullet or cartridge case was indeed fired from a submitted gun and what weight to assign to it. The automatic, objective and quantitative LR evaluation can provide unbiased support to firearm identifications in forensic science, and provide a powerful tool for firearm examiners' case work and court testimonies.

The congruent matching cells (CMC) method was developed at NIST for automatic and objective firearm evidence identification and quantitative error rate and LR estimation. In this paper, a procedure is proposed for evaluation of the LR for firearm evidence identifications. Initial LR estimations using two sets of 9 mm cartridge cases with different sample sizes, imaging methods and ammunition showed that, for all the declared identifications of the tested 2D and 3D image pairs, the estimated LRs were well above the order of 10^6 that provides Extremely Strong Support for a prosecution proposition (e.g. a same-source proposition). The LR estimations also showed that, for all the declared exclusions of the tested 3D image pairs, the estimated LRs were above the order of 10^2, that provides Moderately Strong Support for a defense proposition (e.g. a different-source proposition).

The CMC method for LR evaluation can be applied to both 3D topography images and 2D optical reflective images. However, the 2D optical reflective images are significantly affected by optical conditions and surface properties that are not relevant to identification, which may significantly affect the firearm evidence examinations. As a result, the LR values for 2D image comparisons were significantly smaller than those for 3D image comparisons, especially for the true negative likelihood ratio LR_2.

LR estimations depend on the chosen statistical models and parameters, the relevant firearm/ammunition image datasets used to estimate distribution parameters, correlation programs and measurement procedures. In this paper, the initial LR estimations are limited to two small datasets of KM and KNM image pairs of breech face impressions with granular marks. Our next steps are to apply the procedure to larger datasets, selected from the NIST Ballistics Toolmark Research Database [34] with different breech face and firing pin impression marks, and test CMC correlation algorithms and error rate procedures accessible for evaluation of the LR. We expect that LR and error rate evaluation for case work will require procedures for the selection of the distributions of the reference datasets that match different types of evidence images, such as granular, striation and mixed type of images.

It must be noted that the extrapolation of parametric distributions at their tails of Binomial and β-Binomial model leads to great uncertainty. Perhaps other parametric approximations would be more appropriate at the tails of the graph. A NIST statistician is working to develop a new statistical model, and compare with the existing Binomial and β-Binomial model using the *qq*-plot sometimes used for comparison of two statistical models [35]. We are also working to characterize LR using the Tippett plots, as that used for report of forensic fingerprint evaluation method [36].

There is an interesting question regarding the Binomial and β-Binomial distribution: "Why the Binominal model fitted the KNM scores very well, while it didn't fit the KM scores well?" We believe the major reason is that the Binomial distribution is based on two key assumptions: 1) the comparisons between cell pairs are independent from each other, and 2) each cell pair comparison for the KNM images has the same probability $p = p_{\mathrm{KNM}}$ [23], and each cell pair comparison for the KM images has the same probability $p = p_{\mathrm{KM}}$ [23] to be qualified as a CMC. The resulting Binomial model fitted the KNM data quite well, because the KNM cell pair comparisons are independent from each other, and each cell pair comparison for the KNM images has the same probability $p = p_{\mathrm{KNM}}$ to be qualified as a CMC. As a result, the KNM cells appear to be qualified as a CMC cell pair is likely driven by random, non-selective factors, and systematic measurement errors are not significant factors in the evaluation.

On the other hand, both assumptions (independent cell comparisons and constant p value) are not fulfilled for KM cells. The cell pair comparisons may not be independent among the neighborhood cell pairs, and each cell pair comparison has different probability $p = p_{\mathrm{KM}}$ to be qualified as a CMC. Furthermore, systematic variations in firing conditions, firearm wear, contaminants and breech face impression area may cause variations in the size and quality of the common valid correlation areas of a KM image pair, which may cause additional variations in the probability p_{KM} of the cell pairs to be qualified as CMCs [23]. These effects can be improved using the β-Binomial distribution with variable p_{KM} values which can improve the KM distribution fitting [23].

One of our future works is to develop an uncertainty analysis procedure for both the error rate and LR evaluations. There are many influence quantities that may significantly increase error rates and decrease LR evaluations. However, the extremely large value of the true identification LR_1 estimated from different sets of 9 mm cartridge cases with different imaging methods (see Tables 3 and 4) suggest that it would be feasible to scale up the error rate and LR procedure for case work and database searching with large population size, and still arrive at reasonable and useful LR values to support firearm experts' case works and court testimonies.

CRediT authorship contribution statement

John Song: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Zhe Chen:** Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing - review & editing. **Theodore V. Vorburger:** Formal analysis, Investigation, Writing - review & editing. **Johannes A. Soons:** Formal analysis, Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Visualization, Writing - review & editing.

Acknowledgments

The research was supported by the Special Programs Office (SPO) of NIST. The authors are grateful to Dr. M. Tong and Mr. A. Zheng for providing the 2D and 3D images, and to Dr. N.F. Zhang, Dr. J. Yen, Dr. D.B. Ott and Dr. W. Chu for their contributions in the development of the CMC programs and error rate evaluation procedures.

References

[1] AFTE criteria for identification committee report: theory of identification, range of striae comparison reports and modified glossary definitions-an AFTE criteria for identification committee report, AFTE J. 24 (3) (1992) 336–340.
[2] The National Research Council, Ballistic Imaging, NRC, Washington, DC, 2008, pp. 82–84 3, 20.
[3] The National Research Council, Strengthening Forensic Science in the United States—A Path Forward, NRC, Washington, DC, 2009, pp. 153–154 184, 155.
[4] Executive Office of the President, Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, President's Council of Advisors on Science and Technology (PCAST), 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[5] W. Kerkhoff, R.D. Stoel, E.J.A.T. Mattijssen, R. Hermsen, The likelihood ratio approach in cartridge case and bullet comparison, AFTE J. 45 (3) (2013) 284–289.
[6] N.D.K. Petraco, et al., Application of Machine Learning to Tool Marks: Statistically Based Methods for Impression Pattern Comparisons, NIJ Report 239048, National Institute of Justice, Washington, DC, 2012.
[7] N.D.K. Petraco, L. Kuo, H. Chan, E. Phelps, C. Gambino, P. McLaughlin, F. Kammerman, P. Diaczuk, P. Shenkin, N. Petraco, J. Hamby, Estimates of striation pattern identification error rates by algorithmic methods, AFTE J. 45 (3) (2013) 235–244.
[8] 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, doi:http://dx.doi.org/10.1111/1556-4029.12382.
[9] D.P. Baldwin, S.J. Bajic, M. Morris, D. Zamzow, A Study of False-Positive and False-Negative Error Rates in Cartridge Case Comparisons, USDOE Technical Report #IS-5207, Defense Forensics Science Center, Forest Park, Georgia, 2014.
[10] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.N. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2019)110112.
[11] R. Lilien, Applied Research and Development of a Three-dimensional Topography System for Firearm Identification Using GelSight, NIJ Report 248639, National Institute of Justice, Washington, DC, 2016. https://www.ncjrs.gov/pdffiles1/nij/grants/248639.pdf.
[12] T. Weller, N. Brubaker, P. Duez, R. Lilien, Introduction and initial evaluation of a novel three-dimensional imaging and analysis system for firearms forensics, AFTE J. 47 (2015) 198–208.
[13] E. Hare, H. Hofmann, A. Carriquiry, Automatic matching of bullet land impressions, Ann. Appl. Stat. 11 (4) (2017) 2332–2356.
[14] J. Song, E. Whitenton, D. Kelley, R. Clary, L. Ma, S. Ballou, SRM 2460/ 2461 standard bullets and cartridge cases project, J. Res. Inst. Stand. Technol. 109 (6) (2004) 533–542.
[15] NIST SRM 2460/2461, Standard Bullet and Cartridge Cases Available at http://www.nist.gov/pml/div683/grp02/sbc.cfm (updated, May 25, 2017), (2017) .
[16] J. Song, Proposed NIST ballistics identification system (NBIS) using 3D topography measurements on correlation cells, AFTE J. 45 (2) (2013) 184–189.
[17] J. Song, Proposed "Congruent matching cells (CMC)" method for ballistic identification and error rate estimation, AFTE J. 47 (3) (2015) 177–185.
[18] T.V. Vorburger, J. Song, N. Petraco, Topography measurements and applications in ballistics and tool mark identifications, Surf. Topogr.: Metrol. Prop. 4 (2020), doi:http://dx.doi.org/10.1088/2051-672X/4/1/013002 2016-013002.
[19] W. Chu, M. Tong, J. Song, Validation tests for the congruent matching cells (CMC) method using cartridge cases fired with consecutively manufactured pistol slides, AFTE J. 45 (4) (2013) 361–366.
[20] M. Tong, J. Song, W. Chu, R.M. Thompson, Fired cartridge case identification using optical images and the congruent matching cells (CMC) method, J. Res. Inst. Stand. Technol. 119 (2014) 575–582, doi:http://dx.doi.org/10.6028/jres.119.023.
[21] H. Zhang, J. Song, M. Tong, W. Chu, Correlation of firing pin impressions based on the congruent matching cross-sections (CMX) method, Forensic Sci. Int. 263 (2016) 186–193.
[22] Z. Chen, J. Song, W. Chu, J.A. Soons, X. Zhao, A convergence algorithm for correlation of breech face images based on the congruent matching cells (CMC) method, Forensic Sci. Int. 280 (2017) 213–223.
[23] J. Song, T.V. Vorburger, W. Chu, J. Yen, J.A. Soons, D.B. Ott, N.F. Zhang, Estimating error rates for firearm evidence identifications in forensic science, Forensic Sci. Int. 284 (2018) 15–32.
[24] ENFSI Guidelines for Evaluative Reporting in Forensic Science, European Network of Forensic Science Institutes (ENFSI), 2010. http://enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science.
[25] T.G. Fadul Jr., G.A. Hernandez, S. Stoiloff, S. Gulati, An Empirical Study to Improve the Scientific Foundation of Forensic Firearm and Tool Mark Identification Utilizing 10 Consecutively Manufactured Slides. NIJ Report No. 237960, National Institute of Justice, 2012.
[26] T.J. Weller, A. Zheng, R. Thompson, F. Tulleners, Confocal microscopy analysis of breech face marks on fired cartridge cases from 10 consecutively manufactured pistol slides, J. Forensic Sci. 57 (4) (2012) 912–917, doi:http://dx.doi.org/10.1111/j.1556-029.2012.02072.x.
[27] C. Aitken, P. Roberts, G. Jackson, Fundamentals of probability and statistical evidence in criminal proceedings, R. Stat. Soc. (2010).
[28] F. Habibzadeh, P. Habibzadeh, The likelihood ratio and its graphical representation, Biochem. Med. (Zagreb) 29 (2) (2019)020101. https://www.biochemia-medica.com/en/journal/29/2/10.11613/BM.2019.020101.
[29] D. Meuwly, et al., A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, Forensic Sci. Int. (2016), doi:http://dx.doi.org/10.1016/j.forsciint.2016.03.048.
[30] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D.T. Toledano, J. Ortega-Garcia, Emulating DNA: rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Trans. Audio Speech Lang. Process. 15 (7) (2007) 2104–2115.
[31] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, Forensic Sci. Int. 219 (1–3) (2012) 129–140.
[32] J. Song, W. Chu, T.V. Vorburger, R.M. Thompson, J. Yen, T.B. Renegar, A. Zheng, R. M. Silver, Development of ballistics identification – from 2D image comparison to 3D topography measurement in surface metrology, Meas. Sci. Technol. 23 (4) (2012). http://stacks.iop.org/0957-0233/23/054010.
[33] J. Song, T.V. Vorburger, S. Ballou, R.M. Thompson, J. Yen, T.B. Renegar, A. Zheng, R.M. Silver, M. Ols, The national ballistics imaging comparison (NBIC) project, Forensic Sci. Int. 216 (2012) 168–182, doi:http://dx.doi.org/10.1016/j.forsciint.2011.09.016.
[34] NIST Ballistics Toolmark Research Database, (2020) . https://tsapps.nist.gov/NBTRD.
[35] Aris Spanos, Probability Theory and Statistical Inference—Economical Modelling with Statistical Data, Cambridge University Press, 1999 p. 243.
[36] D. Ramos, R. Haraksim, D. Meuwly, Likelihood ratio data to report the validation of a forensic fingerprint evaluation method, Data Brief 10 (2017) 75–92. https://www.journals.elsevier.com/data-in-brief.