Forensic Science International 364 (2024) 112239

Contents lists available at ScienceDirect

Forensic Science International

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

Automated comparison and evaluation of striated cutting plier toolmarks on metal wires

Jean-Alexandre Patteet *, Christophe Champod

School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, 1015, Switzerland

ARTICLE INFO

Keywords:
Toolmarks
3D
Machine learning
LR

ABSTRACT

Toolmarks examination validity and subjectivity have come under scrutiny. This research focuses on the case of cutting plier marks. This paper presents an automatic comparison method and assesses its performance. It is designed to assign a weight to the forensic evidence (i.e, a comparison between toolmarks) with a likelihood ratio (LR). 3D topographies are acquired and treated to be compared using a set of correlation metrics. A machine learning algorithm combines comparison metrics and enables LR computation. Pliers of various brands and models were used to study the variability both within and between tools. We explained why the specific zone (area along the blade) has to be chosen to build the within-source variability and how the between-source variability can be built in different scenarios. Misleading evidence rates between 0 % and 4 % have been measured and it demonstrates the accuracy of the method when applied on the pliers used.

1. Research purpose

Forensic toolmark examination is a feature comparison method that ultimately assesses observations made during the comparative stage between a recovered mark and prints made by a tool of interest (TOI). The comparative stage involves observing and comparing features (e.g., two striated patterns side by side) and assessing these observations to help determine whether the TOI is or not at the origin of the mark. Both past and current practices depend on the examiner's skill in using a comparison microscope to identify and evaluate similarities and differences [32,8]. Patteet and Champod [34] did a review of the evolution of these practices with regards to striated toolmarks. Features have traditionally been separated into three categories, namely "class characteristics", "sub-class characteristics" and "individual characteristics", following the Association of Firearm and Tool Mark Examiners (AFTE) guidelines [1,2]. The method proposed in this paper deals with "individual characteristics" (striations) that will later be called 'accidental features'. "Sub-class characteristics" have also been observed and are described hereinafter as 'topographical texture'.

The field of toolmark analysis, like the rest of forensic feature-based comparison disciplines, is under scrutiny (mainly in the US), and the validity of its subjective approach has been questioned over the years [29,35]. To overcome this subjectivity (or this dependance on the operator), automatic comparison systems may offer a solution, coming in support to the examiner and not as a replacement.

The Genrich case is a good example of the difficulties courts may face having to rely solely on expert testimony (People v. Genrich, Case No. 1019COA132, 2019) [36]. In this case, James Genrich was convicted for first degree murder from a series of pipe bombs detonated in Grand Junction, Colorado, in 1991. Wire cutters were found at James Genrich's apartment and the examiner concluded that these tools were at the origin of the marks left on parts of the bombs. The expert claimed an identification "to the exclusion of all other tools", because "striated toolmarks are unique". Mr. Genrich's defense team relied on the 2009 NRC report to question the practice and its validity.

This case served as a starting point of this research. Indeed the scientific literature regarding cutting plier marks is limited [19,21,3,44,45] compared to other tools such as screwdrivers. We realised that the knowledge regarding marks produced by cutting pliers have yet to be studied.

The PCAST report highlighted the lack of objective methods in instances where marks are compared to reference material obtained from tools of interest. Often, the final decisions rest with the examiners, and there is a limited body of peer-reviewed research available, if not nonexistent. Consequently, the PCAST committee strongly recommend that the field of toolmark analysis adopt automated, more objective methods. Such methods should disclose error rates to demonstrate their validity and repeatability. The strategy proposed in this paper leverages

* Corresponding author.
E-mail addresses: jean-alexandre.patteet@unil.ch (J.-A. Patteet), christophe.champod@unil.ch (C. Champod).

https://doi.org/10.1016/j.forsciint.2024.112239
Received 23 March 2024; Received in revised form 16 September 2024; Accepted 27 September 2024
Available online 30 September 2024

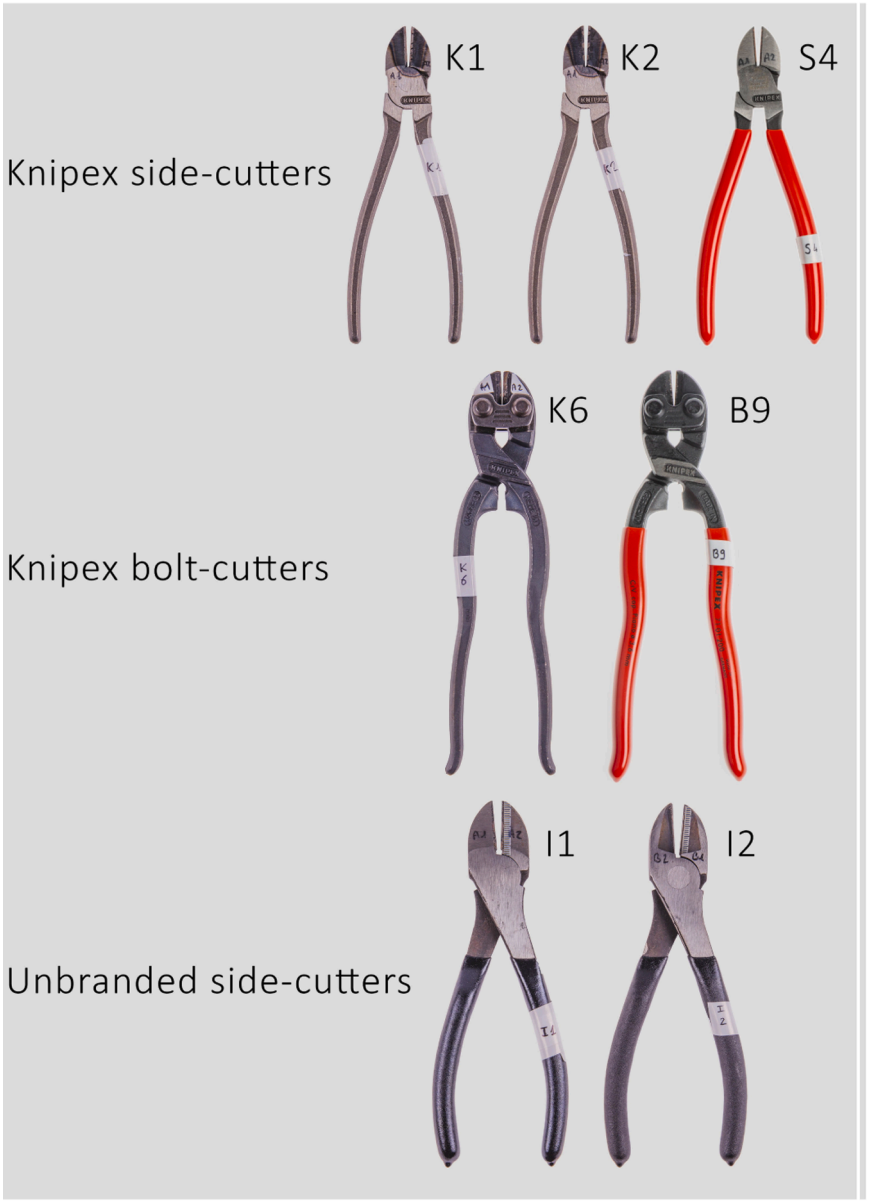


Fig. 1. Pliers used in this study, first row are Knipex side-cutters, middle row Knipex bolt-cutters and bottom row are two store bought pliers with no brand indication. The two-character code on the top right of each plier is its name for this research.

3D topographies to facilitate comparison outcomes between a mark and a set of reference impressions. This is achieved through the application of likelihood ratios (LRs), harnessing the power of machine learning (ML) algorithms.

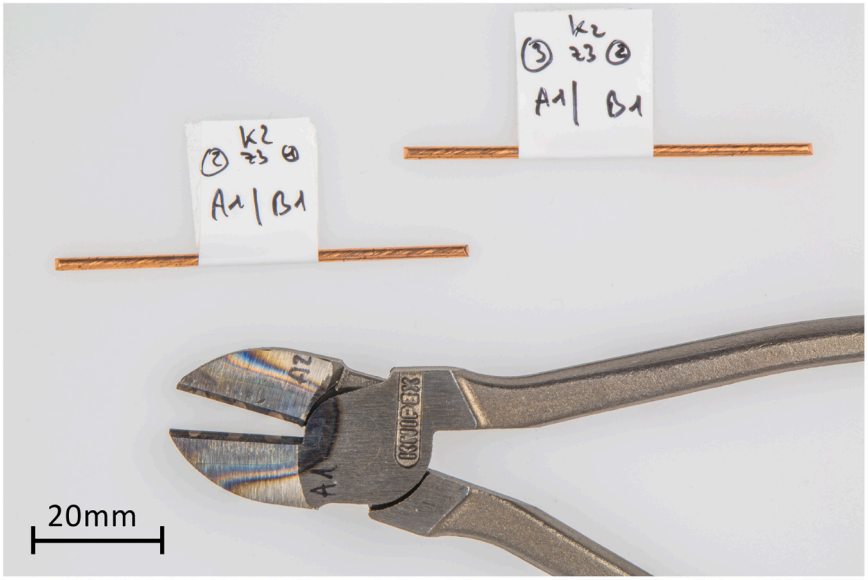
To mitigate the variability introduced by factors such as fluctuating illumination conditions, which are a concern when using optical comparison microscopes, we employ 3D confocal and focus variation microscopy as detailed in various studies [13,16,26,33,37,5,43,46,6,9]. This approach facilitates automatic data processing, extracting the necessary data for comparison. Comparison scores are calculated using various metrics, with the resultant data being refined through machine learning techniques.

LRs are obtained through the construction of within-source and between-source variability distributions [38,44,4,7]. The within-source variability distribution is dependent on the 'repeatability' of marks in a specific zone of the cutting plier captures how repeatable they are on cut wires when creating references sequentially using the same zone of the blades. The between-source variability distribution depends on the 'selectivity' of features of marks from different zones of a single tool or from different tools making the relevant population. In this work, two different scenarios are investigated to appreciate the selectivity and compute likelihood ratios. They will later be presented in the Section 2.5 'Sampling and datasets'. A detailed explanation on repeatability and selectivity is given in Champod et al. [12]. The likelihood-based interpretation framework that will take advantage of these distributions can support an examiner in the evaluation of his/her observations.

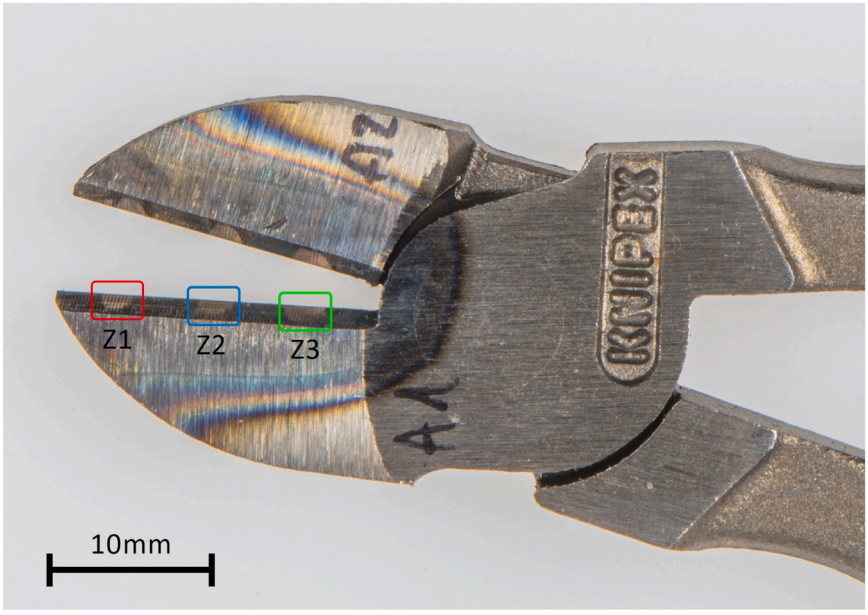
The question of source regarding cutting pliers is not necessarily the tool in its entirety but rather the specific zone of the cutting blade used to cut the wire. Indeed a single cutting plier can leave marks produced by different areas of its blades. These areas are called *zones* in this study. This aspect will be discussed later in this article.

The aim of this study is to develop a method for comparing and assessing marks made by cutting pliers. This method encompasses everything from data acquisition to LR computation. It will allow to answer questions such as:

How to conduct automatically comparisons of striated marks based on 3D data?, *How to build the underpinning distributions?*, *How to compute LRs?*, *What is the influence of data processing on the LRs?*, *Does the type of pliers*



(a) Plier K1 and wire samples.



(b) Plier K1 Zones.

Fig. 2. Recovered Knipex side cutter from the factory with annotated copper samples (a) and selected Zones (b). 1 in red, 2 in blue and 3 in green. A1 and A2 are the blades numbering. On the other side, there is B1 and B2.

(make and model) impact the results?, What are the limits of the computed LRs and how will it be used ?

The system is engineered to closely mimic real-world casework scenarios. Moreover, this study allows us to understand phenomena specific to cutting plier marks that have not been addressed in previous studies:

- The dynamics of cutting influence on characteristics variability of different areas of a single mark later called regions of interest (ROIs),
- The differences and similarities of marks made by a single plier depending on the blade, edge of the blade and location along the blade,
- The links between the two or four different marks created by a single cut.

The next sections will present the material, the developed method and the results.

2. Material and method

2.1. Pliers and wires

Ten pliers of two models were bought from professional construction store. Pliers from the same model were manufactured the same year (information retrieved from a code stamped on the handles). Pliers labeled with the letter S (S1-S10) are Knipex side-cutter model 70 01 160. Pliers labeled with the letter B (B1-B10) are Knipex bolt-cutter model 71 01 200. The same models were also obtained following a visit at the Knipex® factory in Germany, 4 years earlier. Tools K1, K2 and K3 are the same model as the S tools and tools K6 and K7 are the same model as B tools (Fig. 1). Other pliers from different brands were bought in DIY stores (Tools I1 and I2).

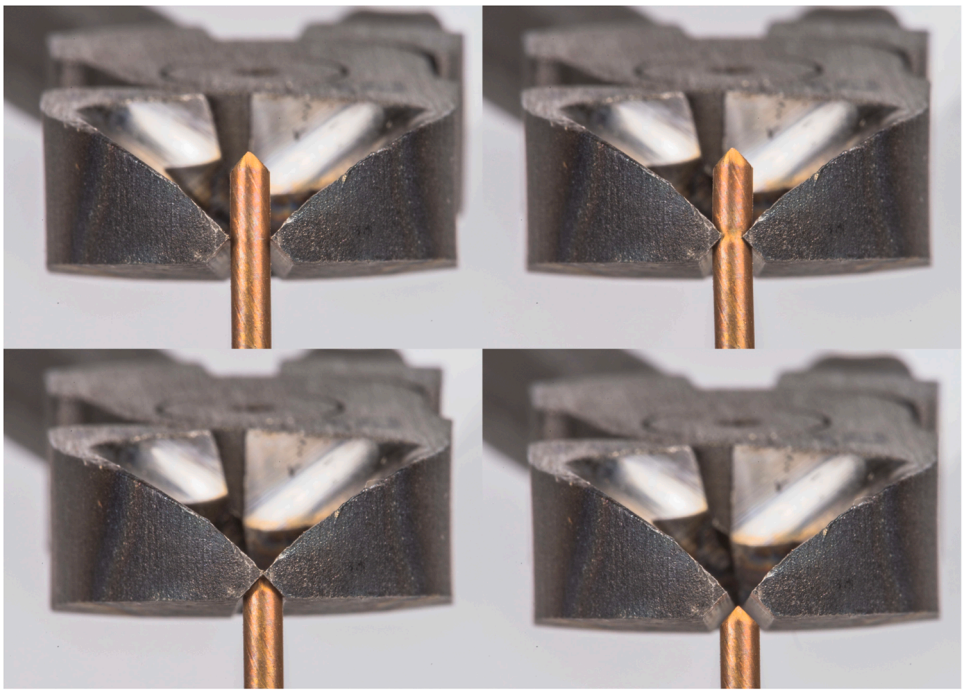


Fig. 3. Four stages of cutting over time, illustrating both the shape of the cut wire and the areas of the tool that come into contact with the sample.

Table 1
Table summarizing the tools that were used, their model and the zones where marks were made on the blades.

Tool	Model	Zone
S4	1	2,3
S9	1	3
B8	2	3
B9	2	3
BX	2	3

A visit to the Knipex factory allowed us to gain some insight into the production of these tools. The manufacturing process for these pliers follows a sequence of steps: molding, milling, and quenching. For the Knipex pliers selected for the study, there's a final manual sharpening step, which modifies the cutting surface of each blade. Moreover, it was not feasible to obtain pliers made sequentially as both parts are ultimately assembled after being picked from boxes holding hundreds of parts.

Research indicates that in the majority of cases, examiners can correctly associate a mark to its corresponding tool, even if the tools are produced consecutively [17,48,49,11,30,14]. However, in the infrequent cases where the manufacturing process is basic and only involves molding, significant similarities might be observed between marks from different tools [31]. Such scenarios do not pertain to the tools used in our study as many steps follow molding such as milling the edges of the blades, reheating and quenching to harden the surface and even a final sharpening step.

We use copper wires, approximately 2 mm in diameter, for our cuts. Copper was chosen due to its common usage, and the specific diameter ensured the selected pliers could easily cut through it.

Wires are cut by hand and each segment is labelled indicating the tool (K1, K2.), the blade (A1, A2…), the zone (1,2 or 3) and the number of the cut (1−10) as shown in Fig. 2a and Fig. 2b. The following Table 1 summarizes which pliers and associated zones were used to capture within-source variabilities. Table 1 shows that at least two tools from the same model and two tools from different models are selected to compare results between tools of a same model and tools from different models. Zone 3 is favored because it is the area on which pliers are usually used (closest to the pivot point of the blades).

2.2. Instrument and acquisition

After cutting the wires, their topographies are captured using a 3D microscope. The Sensofar Sneox is used with its confocal or focus variation technologies depending on the roughness of the sample. Topographies are acquired using stitched images with a 20x magnification lens resulting in a lateral resolution of 0.69 μm/pixel [10].

The instrument is fully calibrated (stage, lenses, noise, measurement capabilities) with the given Sensofar standard materials and additional measures of a selected toolmark are made prior to each acquisition session to ensure the reproducibility of the measurements over time.

The toolmarks produced have a V-shape because both blades work simultaneously, creating two sides (e.g., A1 and A2) as shown in Fig. 4a. Each side is acquired sequentially with the 3D microscope. Using a 3D microscope with a multi-axis system designed to hold thin cables is repeatable as it enables rotation and elevation [10]. Light and resolution remained constant for each topography, unless excessive missing points are detected. Results are 3D topographies made of [x,y,z] data points.

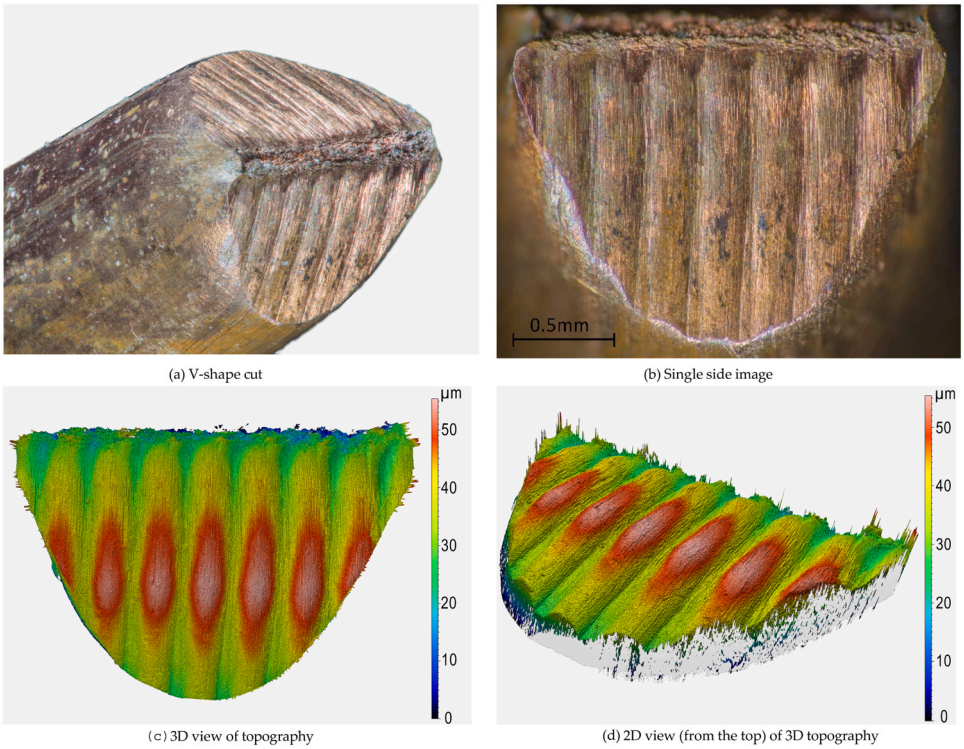
Fig. 4d illustrates a single side of a cut wire in 3D in the MountainsMap® 8 software (SensoMAP Standard V8) from DigitalSurf [15].

2.3. Data processing

MountainsMap® is used with the addition of the "Advanced Profile" module to treat topographies and extract profiles of all sides of marks.

To ease the development of this comparison methodology, one side of mark (half moon) is compared to another without combining the results of both sides of the mark. The combination of both sides will be presented in the last part of the Results section. All results before that last section are solely based on one side of marks.

We developed a dedicated data treatment script in MountainsMap®. Its steps are summarized in Fig. 5. Using this script ensures reproducible treatment across the entire dataset. The software provides access to the results of each step of the treatment, allowing for manual adjustments if errors arise.



(a) V-shape cut (b) Single side image

(c) 3D view of topography (d) 2D view (from the top) of 3D topography

Fig. 4. The original V-shape mark is shown in (a). (b) and (c) shows a single side of the V-shape mark in 2D (view from the top) and (d) in 3D.

The different steps of the processing are as follows:

A half circle area extracts the inside portion of the toolmark, the outside being prone to deformation, artefacts and sample manipulation marks. Fig. 6a shows a top view image of the mark without its edges.

An automatic leveling step corrects any acquisition sample tilt (Fig. 6b). It is also called normalization by Bachrach et al. [4].

Artefacts are removed automatically with the inbuilt function. They can appear if the illumination is too bright resulting in high peaks or because of dust particles that will scatter the light. Points creating a slope over 80° are discarded in this step as it is not expected to have such high slopes on the striated marks

A spline filter is then applied to reduce the gross waviness seen on the toolmark, as illustrated in Fig. 7a. When applying the spline filter in MountainsMap®, there are two outputs: a "waviness surface" and a "roughness surface." In this instance, the "roughness surface" is selected and the "waviness surface" is discarded.

This waviness often stems from an inherent topographical texture or a manufacturing pattern that recurs along the tool's blade. It's anticipated that tools produced in the same batch exhibit similar waviness

patterns. Notably, this surface characteristic can vary between blades of the same tool, requiring the examiner to select an appropriate cut-off filter value for each tool and blade. By eliminating this waviness, more distinct and finer striations come to the forefront, which are crucial for subsequent comparisons. We experimented with various combinations of spline and Gauss filter cut-offs, finding that the spline filter yielded the best results in our study.

Once the waviness is removed, the topography is rotated to align with the angle (or "direction") of the striation pattern, as shown in Fig. 7b. In MountainsMap®, it consists of selecting "align texture with X axis" in the rotation operator. The direction of any waviness pattern, if present, doesn't always align with that of the striation pattern, necessitating its prior filtering [40]. This automatic rotation ensures that profiles are taken perpendicular to the striation pattern. This perpendicular orientation is crucial, as non-perpendicular profiles can exhibit lag and variations in striae widths and depths.

We extract three rectangular regions of interest (ROIs) from the topography. These areas are positioned at the top (in reference to the orientation in Fig. 7b), where the end of the cut is located, at the middle, and at the bottom, where the blades initiate the cut. We found that, for most pliers, the striation pattern wasn't continuous from the bottom to the top of the toolmark. This inconsistency largely stems from the dynamics of the cutting process and the varying parts of the blade surface that make contact with the wire as it cuts through (Fig. 3). A similar phenomenon is observed on the aperture shear mark left by a Glock pistol on a fired cartridge case. Striae tend to be discontinuous along the mark, which led us to choose three distinct ROIs instead of an average profile encompassing the entire mark. Such an average profile wouldn't accurately represent the observable realities. While our comparisons heavily rely on automated features, it's crucial that our compared profiles genuinely reflect the toolmarks. Therefore, we opted for smaller ROIs to ensure the averaged profile aligns with visible striations on the 3D image. Moreover, we avoided focusing solely on one ROI to retain as much comparative information as possible.

From each of these three ROIs, we extract a mean profile. For instance, Fig. 8 showcases a profile derived from ROI1. Consequently, every toolmark's data is distilled into three distinct profiles. In the comparison process, profiles from ROI1 of two toolmarks are matched against each other, as are the profiles from ROI2, and similarly for ROI3. Hence, a single comparison between two toolmarks essentially comprises three individual profile comparisons.

Prior to the extraction of profiles, the last step consists of applying another spline filter with a 15 µm cut- off (keeping the waviness information this time) in order to remove residual noise (Fig. 8). After being cut to equal length, which will ease comparison in the next step, profiles coordinates are extracted from MountainsMap® as text files ([x,y] coordinates) to be imported into RStudio [42]. We did not encounter any excessive processing time but subsampling the data to reduce the amount of pixels could be done if needed.

2.4. Algorithm for comparison and evaluation

The algorithm methodology developed is described below. It is

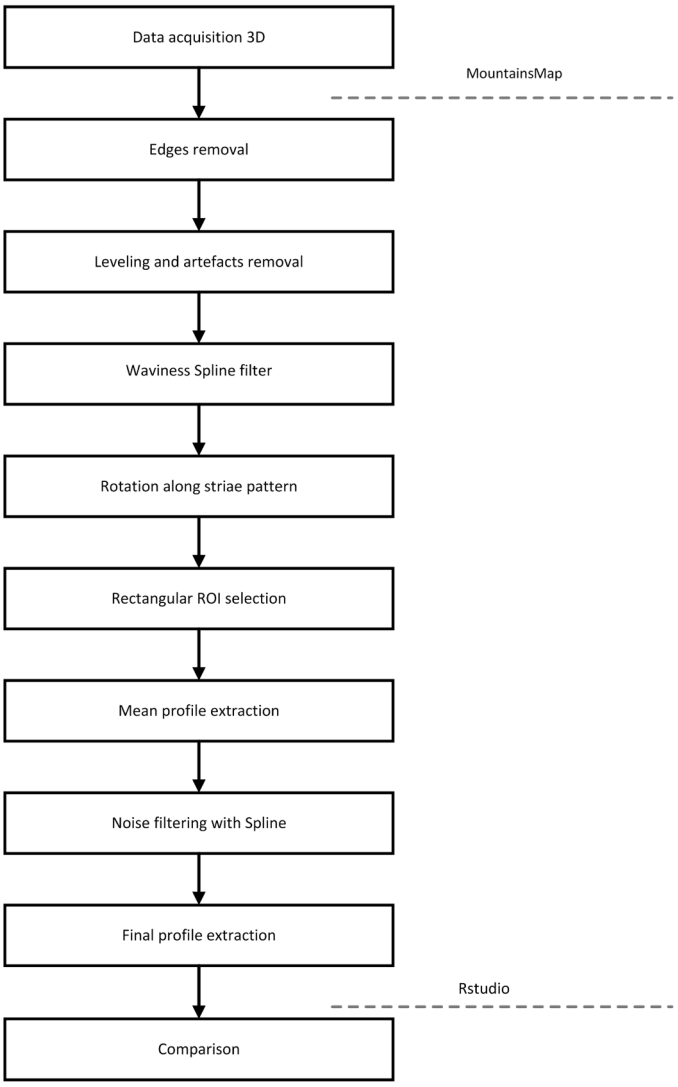


Fig. 5. Flowchart describing the data treatment from 3D acquisition to the profile extraction for comparison.

written in Rstudio [42] and mainly uses the "caret" and "tidyverse" packages [25,50]. The rest of the packages can be found on github (https://github.com/jpatteet/Thesis_Toolmarks), in the "Packages_and_Functions.R" file.

For each comparison, two profiles originating from the same ROI are taken. To ensure they are properly aligned for comparison, a cross-correlation function (CCF) is employed. This alignment step is crucial because even when profiles come from marks made by identical zones of a tool, there can be variations due to manual cutting and the automated positioning of the ROIs by MountainsMap®. These variations introduce a lag in the profile positions, even when they are supposed to represent the exact same zone. To compensate for this lag and achieve accurate alignment, the CCF proves to be an effective method. Heizmann [23,22,37,51]. After alignment, the similarity between both profiles is computed using several metrics (scores) such as Bachrach's relative distance or adapted versions of Chumbley's method which separates the profiles in different windows [13,19,44,4]. Each comparison is thus described by several similarity scores. Four are used for the results presented here:

- The CCF score,
- The Bachrach relative distance,
- The normalized Congruent Matching Profile Segments (CMPS) score [47],
- A custom score which computes CCF between 50 points windows of profiles. A threshold value of 0.7 was chosen and CCF scores are adjusted based on the distance with that threshold. The new scores are all averaged to get a final score between −1 and 1. The threshold and window values were selected after testing a range of parameters and running the machine learning algorithm for each pair. The pair of parameters (window=50and threshold=0.7) with the highest accuracy was ultimately chosen.

Each comparison is assigned a class value: either 'W' or 'B'. 'W' stands for 'within' and is used when the compared profiles come from the same tool and identical zone. Conversely, 'B' stands for 'between' and indicates that the profiles either originate from different tools or the same tool but different zones. In forensic cases, the 'W' class corresponds to the prosecutor hypothesis (*Hp*), and the 'B' class corresponds to the defense hypothesis (*Hd*).

A dataset comprising comparisons from both classes is compiled. All these comparisons are consolidated into a single dataframe. For each

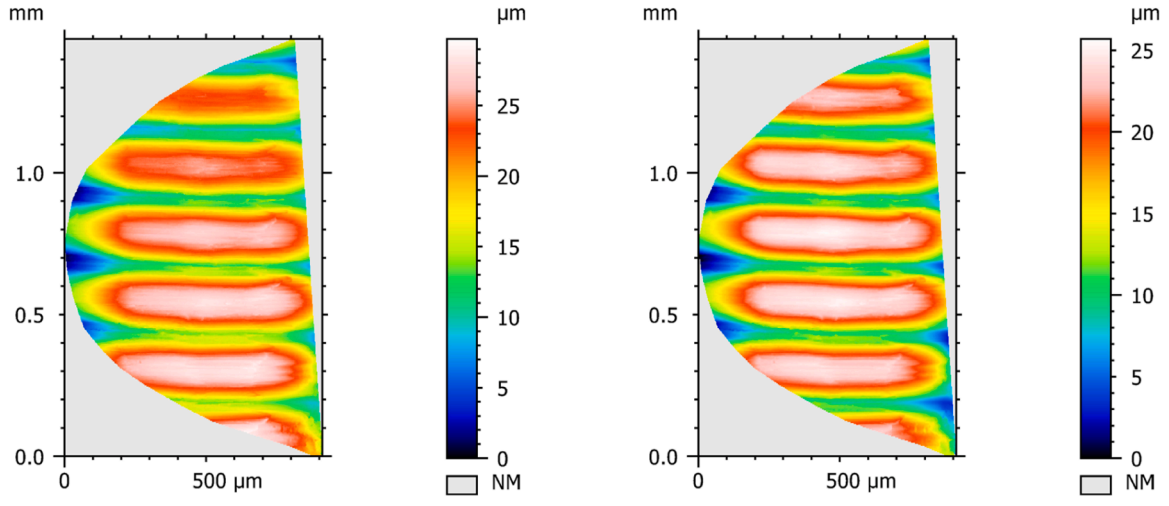
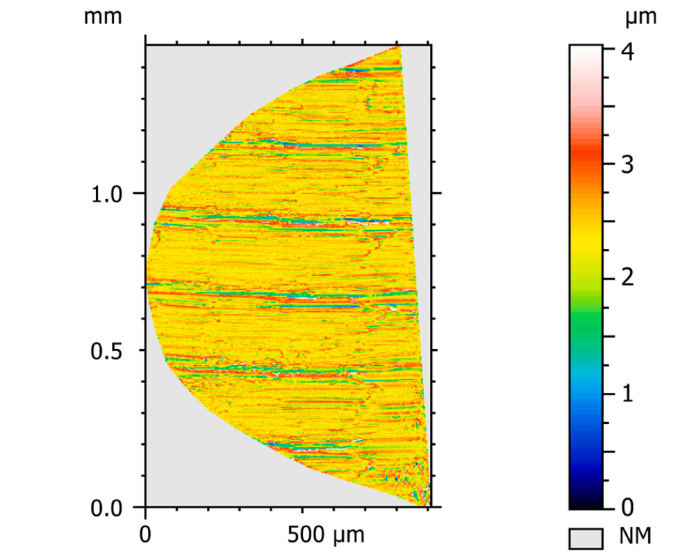
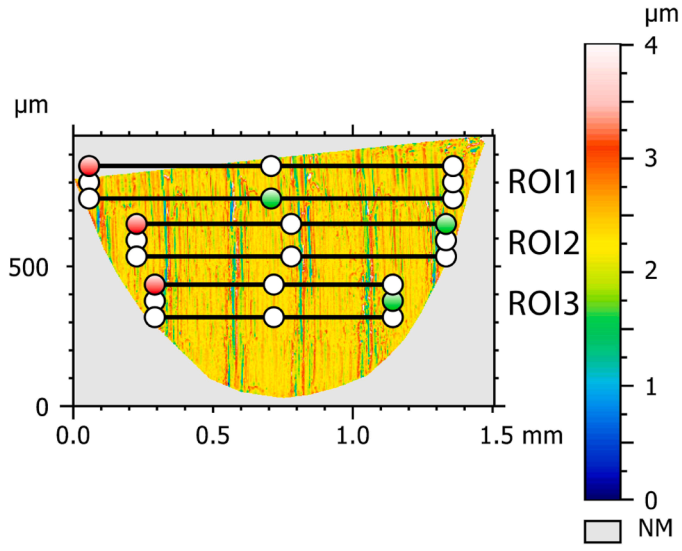


Fig. 6. Flat representation of the topography with color depth scale before treatment (a) and after leveling (b).



(a) Spline filtering result.



(b) ROI selection.

Fig. 7. Topography after spline filtering (keeping the roughness) with a cut-off value of c=40 μm (a). (b) shows the rotated image along the striation pattern and the three ROI where profiles are going to be averaged and extracted.

comparison, the associated row in the dataframe captures the 4 similarity scores, the lag between the two profiles, and the class designation (either 'W' or 'B', as shown in Table 2).

As a comparison between two marks is made of three separate comparisons (one for each ROI), 12 scores (four metrics times 3 comparisons) are used as predictors for the machine learning step. We did not carry out any reduction of the number of predictors nor did we attempt to reduce dimensionality by PCA for example. All predictors were used as such. We employ machine learning (ML) techniques to determine the optimal model and parameters for distinguishing between the two classes ('W' or 'B'), using the score values as the predictors. The lag value is not used as a predictor. To facilitate this learning task, we randomly partition the data into a training set and a testing set. The training set is used to train the algorithm, while the testing set evaluates independently its accuracy. We utilize the train() function from the caret package in RStudio with a 10-fold cross-validation, allocating 50 % of the data for testing. We explored various classification methods — including randomforest, lda, svmRadial, glm,glmnet and gbm — to gauge their accuracy. These methods were chosen to encompass a broad spectrum of complexity, allowing us to optimize increasing numbers of hyper parameters.

In our analysis, the 'glmnet' method yielded the lowest error rates, as gauged by accuracy, sensitivity, and specificity values. Although all tested classifiers we tried delivered generally accurate results, glmnet stood out for its conservativeness in assigning probabilities to its predictions due to its penalization parameter [20]. The glmnet classifier is an extension of generalized linear models that can attribute penalties to the coefficients of the predictors. When comparing marks from a dataset, the training was re-ran after the first ML, this time swapping the testing and training sets. This allow us to double our results without introducing any duplicates.

The following list summarizes the specifications of the classification:

- 12 scores are used as predictors (4 scores for each profile comparison — ROI1vsROI1, ROI2vsROI2 and ROI3vsROI3),
- The training and testing set are created randomly allocating 50 % of the data in each one,
- a 10-fold cross validation procedure is used on the training set to optimize the hyperparameters of the model,
- The glmnet classifier is used,
- The output used is the predicted probability of each observation of the testing set.

The classifier predicted probability on the target class W for a given

Table 2
Summary of the variabilities when comparing marks from same or from different tools and zones.

		Tool	
		Same	*Different*
Zone	*Same*	W	B
	Different	B	B

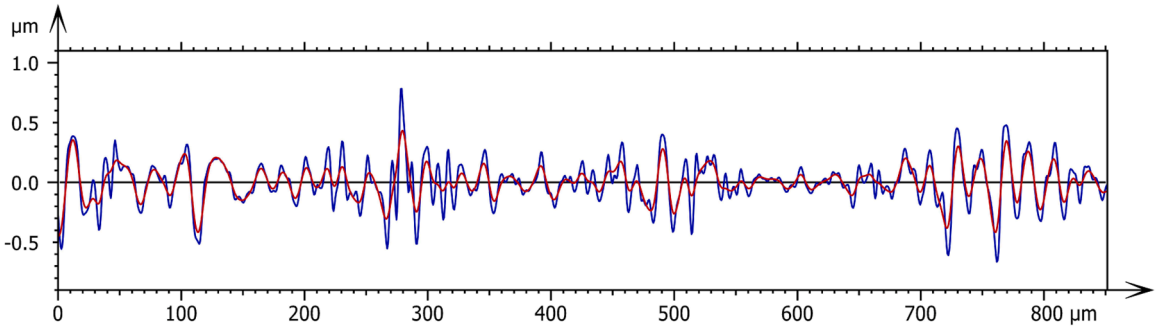


Fig. 8. Example of an original mean profile (in blue) and its filtered profile (in red) for ROI1.

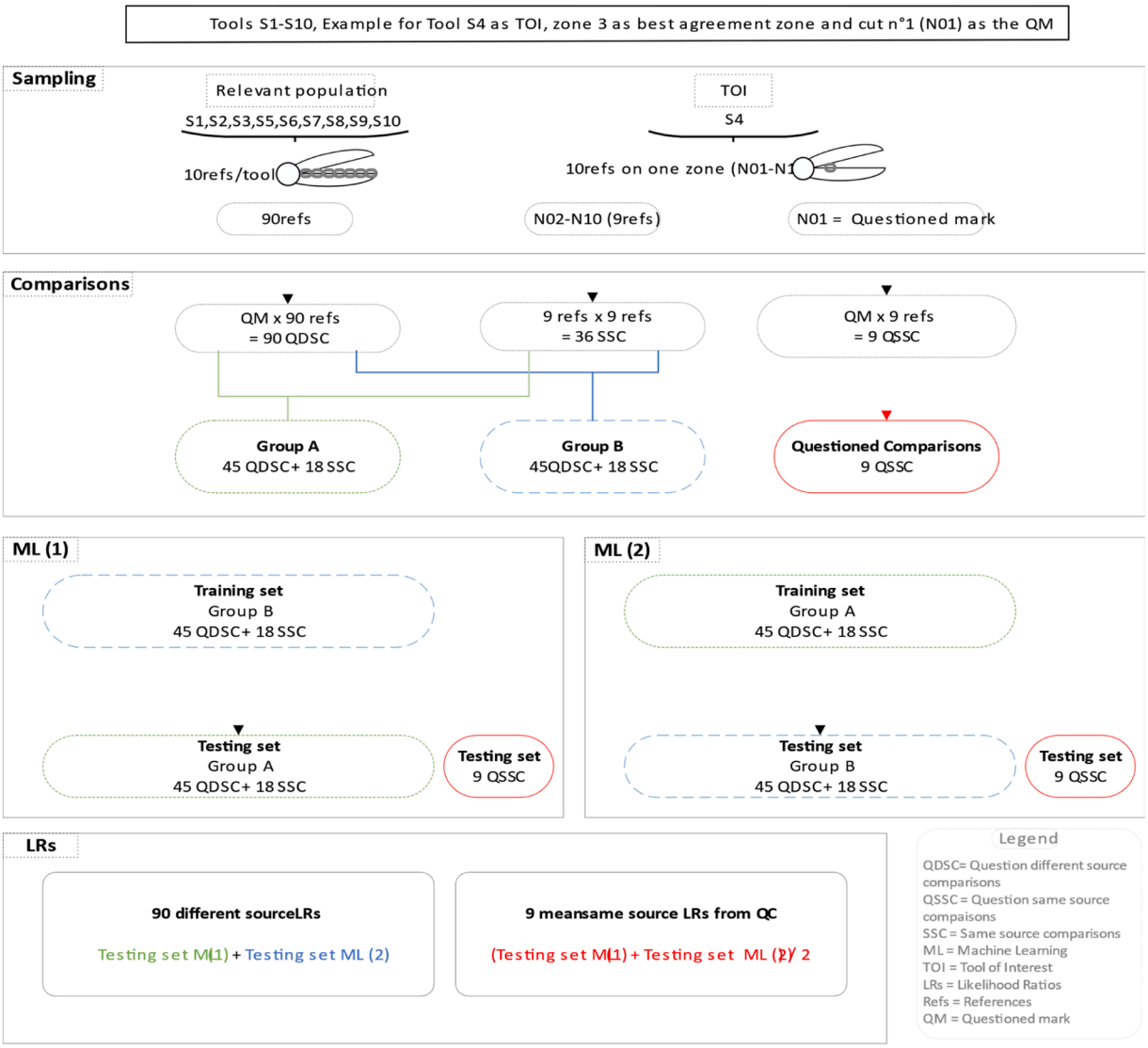


Fig. 9. This diagram shows the case example for tool S4, zone 3 and cut number 1 in the ideal scenario.

comparison *i*, noted P(Wpredict(i)), is used to compute its LR according to the following formula (Eq. 1).

$$LR_i = \frac{P\,(W\,Predict(i))}{1 - P\,(W\,Predict(i))} \div \frac{P\,(Wset)}{1 - P\,(Wset)} \quad (1)$$

The ratio P(Wpredict(i))/1-(P(Wpredict(i))) is the posterior odds in favour of W. The ratio P(Wset)/(1- P(Wset)) is the prior odds in favour of class W considering the training set. P(Wset) is the relative proportion of W comparisons in the training set. For example, if the training set is made of 30 same source comparisons and 30 different source comparisons, P(Wset) will be 0.5.

After the LR computation, a calibration is made using logistic regression model as in Jacquet [24]. Log10 of the LR (LLR) distributions are obtained for all comparisons.

They represent the LLR obtained under both conditions ('W' and 'B' respectively). Values of LLRs over 0 for different source comparisons and under 0 for same source comparisons indicate errors, meaning a LR supporting the hypothesis that the source is the same when it is not or the opposite. In all plots presented below, errors of (LLR) are indicated on the plots. In case a curve shows values over and under 0 but the written error is 0, the error is indeed 0 and the exceeding part of curve is due to the continuity of curves even after the minimum and maximum value they represent. The error is given as the rate of misleading evidence in favor of prosecution / defense (RMEP/RMED) as in Riva et al. [41]. Total error ("Err Tot") percentages (number of errors divided by number of comparisons) are also computed to showcase the overall error. The term "error rates" will be used also for the "rates of misleading evidence" in the remaining of this article.

To ensure calibration, when no error occurred, meaning that all LLR under *Hp* are above zero and all LLR under *Hd* are below zero, we designate the closest LLR to 0 under both propositions as misleading by changing their classes.

As it will be explained in the next part, the plots represent the LR values that can be expected in a specific case with a given TOI and questioned mark. The LR values under the *Hp* distribution displays actual values that can be expected if the TOI is at the origin of the questioned mark. With the same logic, the *Hd* distribution displays values that can be found when comparing the questioned mark to other tools of the relevant population.

In the subsequent section, we will detail how we assembled the datasets, addressing challenges like data volume. Additionally, we will clarify the criteria for comparisons between identical and different sources.

2.5. Sampling and datasets

Datasets are comprised of comparisons from both classes as described earlier. In general, the different source comparisons require the use of different tools coming from the relevant population of other sources [27]. This relevant population represents all the tools that could have been used in the case. Very often in forensic practice, when examiners are facing a questioned mark to be compared against a TOI, they often lack access to an extensive set of other tools from the relevant population.

In the context of pliers, the population of other sources shouldn't be indiscriminately chosen from any type of pliers. Instead, examiners should opt for pliers that yield references with analogous global topographical textures. Such global textures might encompass features like the recurring wavy pattern shown for tool K1 (as seen in Fig. 4d) or other consistent shapes, exemplified by K6 in Fig. 21. These distinctive features can serve as exclusion criteria. In the same way, when a firearm with 4 land engraved areas can be ruled out as the source for a bullet that exhibits 6 land engraved areas. Importantly, this texture is intrinsically tied to the initial filtering phase. Indeed references and toolmarks that do not have a similar texture will not be filtered the same way and should not be part of a population of alternative sources. Therefore only toolmarks processed using identical cut-off values should be pitted against one another for comparison. This reduces the amount of tools that can be considered in a relevant population for that assessment task. The population used to assessed fine striated marks will then be tools of the same brand and model as the TOI. More explanation on the concept of relevant population and the LR calculation are presented in the discussion Section 4.2 'LR calculation taking into account the features due to design' and in Section 4.3 'Revisiting the propositions'.

We aimed to leverage our relatively large sample size to create an *ideal* different source dataset. Indeed if an examiner had unlimited time and resources we would suggest that he/she would select 10 pliers of the same brand and model as TOI (or pliers that cannot be excluded from the topographical texture of their references) to construct that dataset. With these pliers, the examiner will create as many references as he can ensuring none are from the same blade location (zone). This approach defines our *ideal* case. In

our research, we use ten pliers, with a sample cutting diameter of 2 mm and blade lengths of about 20 mm, allowing for 10 references per tool, totaling 100 references for our different source dataset. To understand the variability observed among other sources, a questioned mark will be compared to these 100 references. The distribution of these LRs will represent the extent of values that can be expected when comparing a toolmark to tools other than the TOI.

We do not expect examiners to ever get such an ideal dataset, hence we investigated at alternative ways to obtain a dataset that will approximates the *ideal*. The readily available material to the examiner is the TOI. We focus then on the worst-case scenario, where no tools with a topographical texture similar to the questioned mark are available. We refer to this situation as the *single tool* scenario. Different source transactions using only the TOI will be created with comparisons between all different zones of a same tool. Comparing results obtained in the *single tool* case with the *ideal* case will allow to assess whether or not such an economical mechanism is appropriate. As we will show later using PCA, different zones of a single tool are showing variabilities in striations that are in line with the variabilities observed between tools.

The following datasets were then created for each ground truth state:

The same source transactions will always be constructed in the same way as it represents the variability of the zone of the TOI that showed the best agreement with the questioned mark found during the preliminary steps of examination (first observation under microscope, and research to find what is the zone of best agreement along the blade). 10 cuts is reasonable in terms of work and considered sufficient to represent the variability of a given zone of a tool.

One of these 10 cuts will be considered as a questioned mark and the other 9 as references made in the zone of best agreement. All these marks will be compared to each other resulting in 45 comparisons. The 36 comparisons not comprising the questioned mark will be used to train the algorithm. The 9 comparisons made with the questioned mark will be used as a test set and result in 9 same source LRs that are specific to the case at hand.

The different source transactions will depends on the two scenarios discussed above.

We encourage the reader to read the following text with the Fig. 9 simultaneously to understand the construction of the scenario. For the ideal scenario (Fig. 9), one of the 10 same model tools is selected as the TOI (for examples S4). The other 9 tools (S1, S2, S3, S5, S6, S7, S8, S9 and S10) are used to create different source references ('Sampling' box in Fig. 9). As 10 references are made along the blades of each tool, this gives 90 references. For the TOI, 10 references are made in a single zone. One is considered as the questioned mark (cut n°01 for example) and the other 9 as references that will be used for the same-source comparisons. In the 'Comparisons' box of Fig. 9,the 90 references from other tools are compared to the questioned mark. It will then result in 90 comparisons between the questioned mark and all the references of different sources. The 9 references from the TOI (S4 in this example) are compared to each other to have 36 same source comparisons. Finally, the questioned mark is compared to the 9 references from the TOI resulting in 9 'questioned comparisons'. These 9 comparisons are the ones of interest as they were made between the mark and the references from the TOI on the "best" zone. They will be kept and introduced in the testing set. Two groups A and B are created which randomly takes 50 % of the 90 comparisons 'questioned mark vs different tool references' and 50 % of the 36 same source comparisons. Each group A and B (Fig. 9) comprises 45 + 18 comparisons. From this point, the machine learning (ML(1) and ML(2) boxes in Fig. 9)is conducted and the LRs computed. The colours and dotted lines enable to understand which group is selected for each ML model. In ML(1), group B is used as the training set to build the model and applied to the testing set which contains group A and the 9 'questioned comparisons'. ML(2) switches groups A and B and thus has group A as the training set to build the model that will be used on the testing set which combines group B and the 'questioned comparisons'. LRs from both MLs (ML(1) and ML(2)) are gathered as shown in the 'LRs' box in Fig. 9. The 45 different source comparisons from each ML are gathered resulting in 90 different source LRs. The 9 'questioned comparisons' (in red) are considered as same source comparisons as the questioned mark was taken as one of the references from

the zone chosen in S4. These 9 'questioned comparisons' are the same in ML(1) and ML(2). However, the resulting LR values depend on the training set used. As the training set is different for ML(1) and ML(2), different LRs will be obtained for the same 'questioned comparisons'. These two set of LRs for the same comparisons are averaged to keep 9 LRs. This whole process will be conducted 9 more times by changing the questioned mark, for example taking the reference cut n°02 in S4 as the questioned mark instead of n°01. Note that all the obtained LRs result from a comparisons with the question mark. Comparisons between references of the zone of S4 are used as same-source comparisons to build the models of ML in the training set but are discarded in the testing set. Each iteration is called a case here. The 10 references in a single zone can thus lead to 10 fictive cases. A case will be comprised of 90LRs for different source comparisons and 9LRs for same source comparisons. With our 10 references per zone of a tool, we can construct 10 cases, resulting in a total of 900 different sources LRs and 90 same source LRs for a selected tool and zone. Fig. 9 summarizes the above explanation and illustrate how distributions are built for the ideal dataset.

In the *single tool* scenario, the examiner does not have access to other tools of the same brand and model as the TOI. The transactions from other sources are obtained with the cuts from the TOI only. The working assumption is that the features captured are accidental and have the same variability among zones of a given tool as we can observe between zones from different tools. In other words potential sub-class

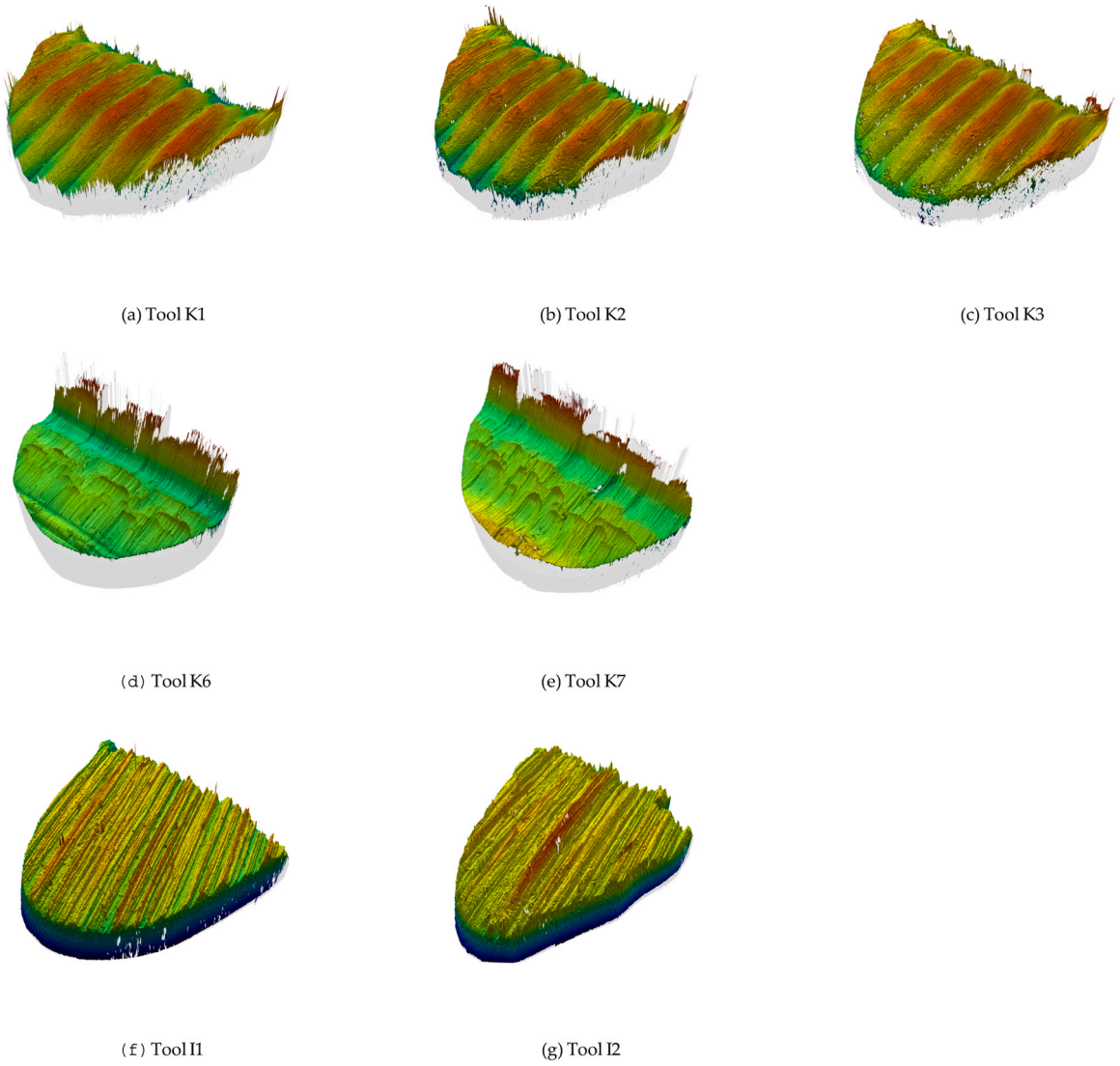


Fig. 10. 3D images for the selected tools. Each row represents a different model of pliers. Topographies were displayed using MountainsMap®, the colors scale from dark blue to red represent the position on the Z axis.

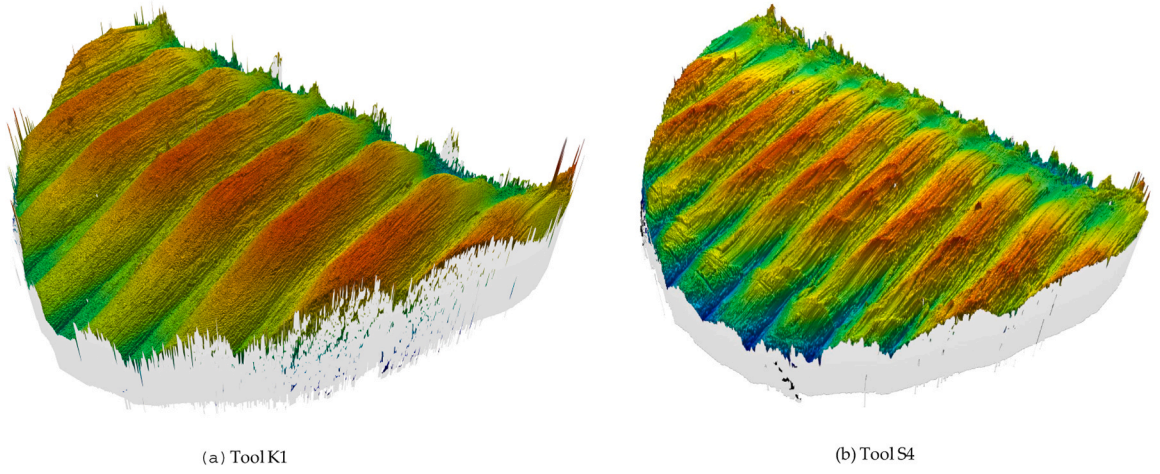


Fig. 11. 3D images the same model of tools but years apart.

characteristics would be filtered out. In our case, a tool can cut in 10 zones without overlap along the blades. One of these 10 zones is selected as the zone to capture the within-source variability. The other 9 zones will be used to capture the different-source dataset (with one reference per zone). These 9 references are compared to each other giving a total of 36 comparisons. Here, a single case will be represented with 36 different sources LRs and 9 same source LRs. The combination of all 10 possible cases result in 360 different sources LRs and 90 same source LRs.

Once datasets are created, we directly obtain the LR values. The combination of 10 cases at the exact same zone of a tool illustrates the variability that can appear in a single zone. Moreover, the errors are reported to indicate the limits of the method.

3. Results

3.1. Variability between different tools

We have observed that the topographical texture is dependent on the tool and the blade. Fig. 10 shows how distinctive the overall topography and pattern can be between tools of different models and how similar they are when sharing the same model.

The first model (K1, K2 and K3) has a repetitive wavy pattern that is approximately 250 μm wide and 15 μm deep. The second model (K6 and K7) is much more flat with less than 5 μm difference between the lowest and the highest point.

Textured pattern can influence the striated marks in various ways. For K6 and K7, it is perpendicular to the striation pattern, while for K1, K2, or K3, it follows the direction of the cut. The last model (I1 and I2) displays continuous striations that are broader and more defined, with depths ranging from 2 to 20 μm and widths between 1 and 3 μm. These observations and measurements show that models can be distinguished without extracting 3D profiles or resorting to correlation functions for comparison.

Consideration of the manufacturing process is crucial, as demonstrated in Fig. 11. The two displayed marks originate from tools of the same model. However, Fig. 11a is from a tool obtained directly from the factory in Germany four years prior to Fig. 11b, which was purchased in a local store. Although both exhibit a wavy-type pattern, their periodicity differ. Specifically, Fig. 11b shows a wave approximately every 200 μm, whereas Fig. 11a displays one every 250 μm. This difference is evident in the images: Fig. 11a has 8 waves compared to the 10 on Fig. 11b. Since the wavy pattern was consistent across the blades of all 15 examined tools, we do not expect variations in periodicity within a single tool's blade. The pattern arises from a continuous milling process of the edges, and deviations manifest when there are changes in the machine's kinematics or alterations in parameters like speed, feed rate, and the width and depth of the milling tool. This was confirmed by a production manager at the Knipex factory in Germany. Therefore, this characteristic can serve as a basis for excluding certain tools, similar to the distinctions shown in Fig. 10.

The digital filtering process is critical in revealing the topographical textures and patterns of toolmarks, as shown in the case of K1. The cut-off value, a key factor in this process, varies not only among different tools but also between blades of the same tool. Such variation can significantly change the pattern's appearance, making it risky to compare profiles filtered with different cut-off values. In our experiments, we observed marked contrasts between the references of various models. The required cut-off value to expose crucial information differed, rendering comparisons of differently filtered profiles counterproductive. Therefore, we decided against these comparisons due to the noticeable differences in the toolmarks' overall topographical texture from our set of tools. In practice, a forensic practitioner would likely dismiss a tool as the source upon observing such disparities during the initial examination stages, barring other factors that might explain these variations. The filtering stage affects what is considered as part of the relevant population, effectively excluding tools with divergent topographical textures. Consequently, constructing a generic distribution encompassing all types of pliers is impractical.

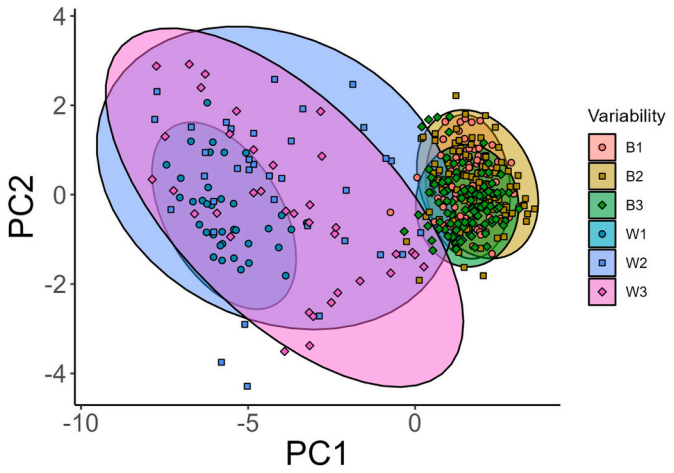


Fig. 12. PCA plot of the tool K6 (same model as tools starting with letter "B"). W1, W2 and W3 are respectively the within-source variabilities in Zone 1, 2 and 3. B1, 2 and 3 are the corresponding between-source variabilities obtained in the *ideal* scenario. Note that the datasets are of a single case and not the combination of all 10 possible cases for each zone. PC1(75 % variance), PC2 (8 % variance). Ellipses contain 95 % of the data of each group.

Does each zone along the blade have the same variability in its features? Does the same zone of two tools have the same features? Answer to these questions will tell us if a generic within-source variability can be used (meaning merging all same source comparisons together) or if the within-source variability requires to be specific to the tool or to the zone itself.

3.2. Zone variability

If the specific zone does not affect within-source variability, it would mean that examiners could utilize generic references from any zone to assess this variability. Conversely, if each zone demonstrates distinct variability from cut to cut, examiners would have to create samples specific to the zone in question. In this context, the results regarding within-source and between-source variability are presented using PCA plots. PCA is selected because it offers a straightforward and rapid method for dimension reduction and for visually representing the differences (or lack thereof) between groups of data. In both PCAs, a total variance over 80 percent were obtained when considering PC1 and PC2 only.

The four scores identified in the earlier described machine learning process are employed as variables to inform the PCA. Before performing PCA, each comparison is assigned a class (B1, B2, B3, W1, W2, and

W3). These classes are represented by different colors and shapes in the plot and different dot shapes were chosen within the B and the W classes to enhance the visualization of potential separation. The two principal components derived from this analysis (PC1 and PC2 respectively) are then used to display the results.

The dispersion of points and the colored areas illustrate the variability within each class. These plots enable to visually assess whether the two classes ("W" and "B") are distinguishable and if the within-source variability applies to each zone. A noticeable separation between classes suggests that each zone exhibits distinct behavior. Consequently, adopting a generic approach to within-source variability may not be applicable in all scenarios.

In Fig. 12, all within-source variabilities are obtained using tool K6. Three different zones are selected, and for each, the between-source variability is calculated using 10 other tools, as per the *ideal* scenario

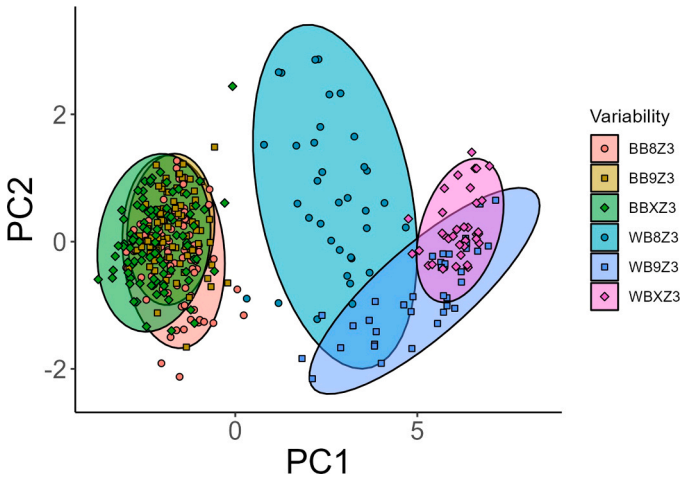


Fig. 13. PCA plot of the variabilities of three different tools in zone 3, WB8Z3, WB9Z3 and WBXZ3 are the within-source variabilities of each tools and BB8Z3, BB9Z3 and BBXZ3 the respective between-source variabilities. The datasets are of a single case and not a combination of all 10 possible cases for each zone. PC1(82 % variance), PC2(5 % variance). Ellipses contain 95 % of the data of each group.

previously described. Notably, the within-source variabilities are distinct from each other. For instance, W2 is less spread compared to W1. This observation shows that the within-source variability ought to be considered at a specific zone and cannot be properly measured outside that zone of interest. W1 is constructed by comparing 10 toolmarks from zone 1, while W2 involves 10 toolmarks from zone 2. This indicates that within-source variability for a single zone cannot be generalized. Zone-specific variability must be established for the zone that most closely aligns with the questioned mark. However, the between-source distributions (B1, B2, and B3) are all located in the same area, suggesting that despite variations in the questioned mark, the dispersion remains similar.

In Fig. 13, the data is obtained from three tools of the same model (B8, B9, and BX). Zone 3 is consistently selected on each tool to establish a within-source variability and the corresponding between-source variability. Our goal is to determine whether the variability of the same zone across tools of an identical model would be similar, suggesting that the manufacturing process introduces repetitive variability. The results indicate that the within-category variabilities (WB8Z3, WB9Z3, WBXZ3) do not fully overlap and exhibit significant differences. This suggests that the within-source variability between tools of the same model differs, even within the same zone, emphasizing the need to construct a specific within-source variability in casework. Therefore, it should not be assumed that two tools of the same model will have similar within-source variabilities. However, consistent with previous findings, the between-source variabilities display a substantial overlap.

From these results, we conclude that the within-source variability has to be built using the TOI at a specific zone. This zone will be selected after preliminary observations and the identification of good agreement characteristics between the questioned mark and references of the TOI at

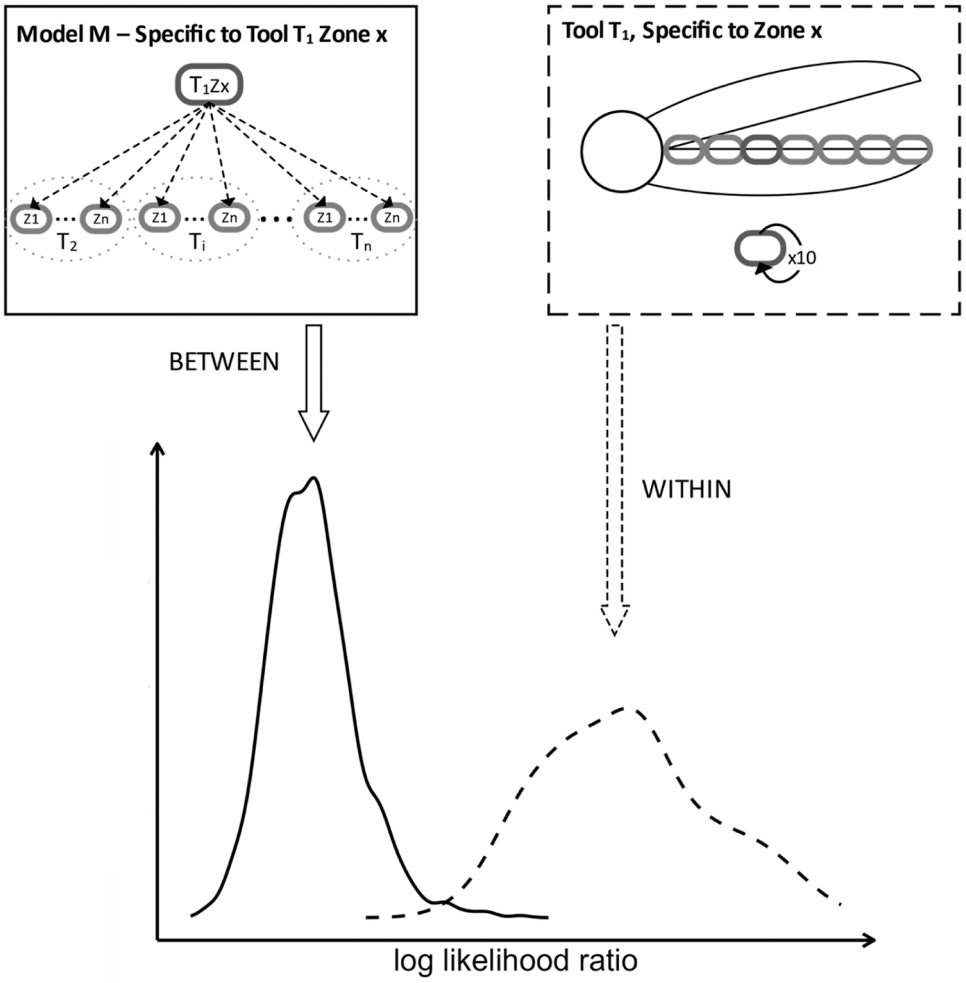


Fig. 14. Representation of the within and between variability distributions of the LRs for the ideal scenario.

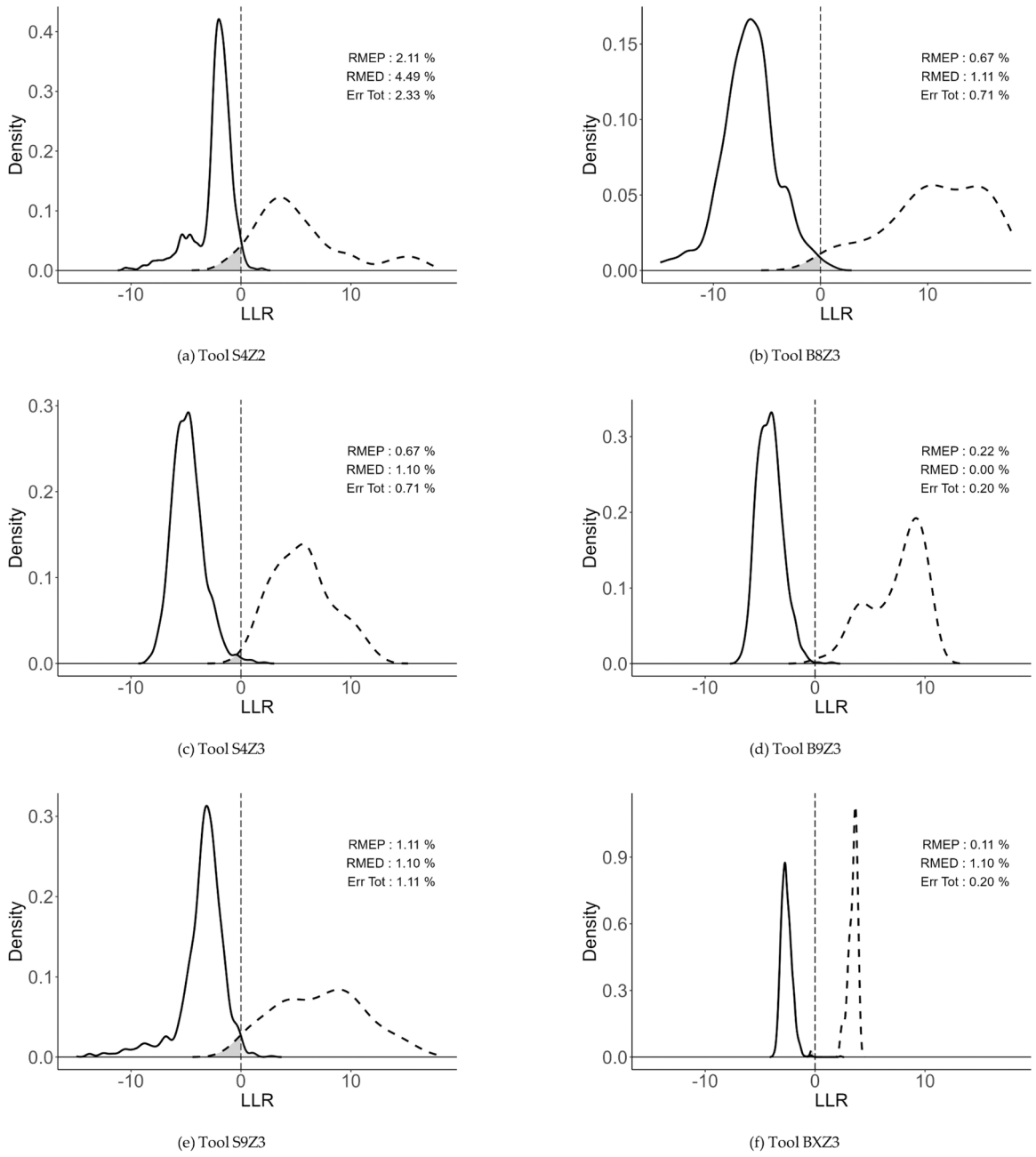


Fig. 15. The calibrated LLR distributions in the *ideal* scenario for 6 tool/zone combinations. Error rates are given by the "RMED" and "RMEP" and "Err Tot" values. The dotted curve represents the 'W' class or *Hp* class ('W' for *within*-source) and the solid curve the 'B' class or *Hd* ('B' for *between*-sources).

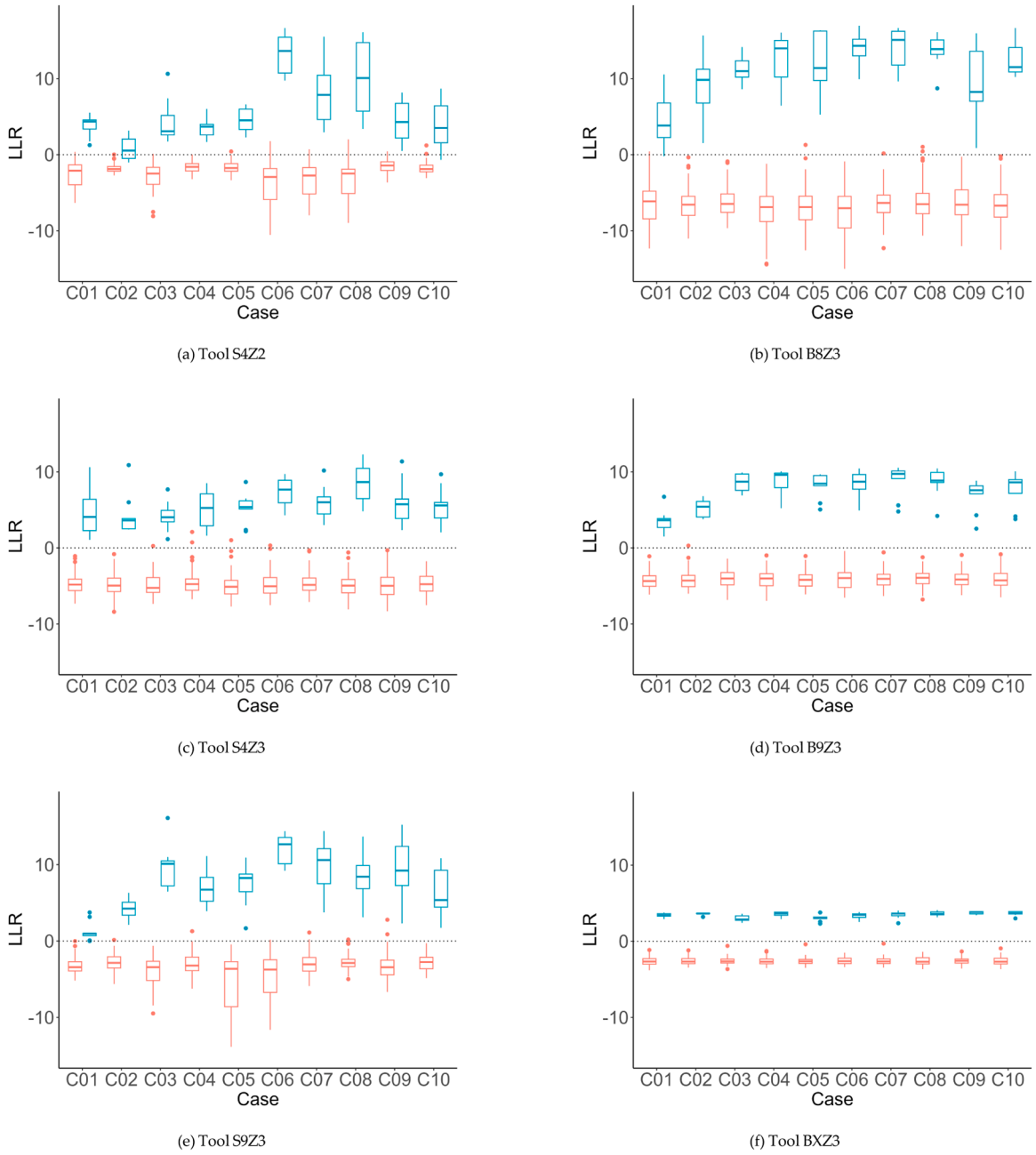
a given location (or zone) on the blade.

3.3. Ideal Scenario

In this *ideal scenario*, the examiner has access to multiple tools of the same brand and model as the TOI. Our study assumes that tools of the same brand, model, and production period share a similar gross topographical texture. The examiner first examines both the questioned mark and reference marks from the TOI, identifying the zone of best agreement, when such a zone exists, and acknowledging the inherent variability in each zone. This selected zone is then utilized to create a same source dataset for the TOI through multiple cuttings within that specific area. The other tools are used to prepare cuts from various zones, with a careful effort to avoid repetition. The different source dataset is derived by comparing one reference from the TOI's best agreement zone (in actual cases, this would be the questioned toolmark) against references from the other tools, across different zones, as illustrated in Fig. 14 and outlined in Fig. 9.

Fig. 15 shows the LLR distribution in the ideal scenario for six different tools/zones. The three cases on the left (letter a, c and e) are made using references from the same side-cutters (S1-S10). For example, Fig. 15a has S4 as the TOI and all the other tools S1-S10 (without S4) as references to build the different sources dataset. The three plots on the right (letter b, d and f) are obtained for the bolt-cutters, tools B1- B10. Each plot is a combination of all 10 cases previously described in the Section 2.5. LRs are presented in log10 (LLRs) on the plots.

The within-source variability depends on the tool and on the zone, whereas the between-source variability is more stable but specific to the (questioned) mark selected. It must be noted that both variabilities are tied together as they are both computed during the same machine



(a) Tool S4Z2

(b) Tool B8Z3

(c) Tool S4Z3

(d) Tool B9Z3

(e) Tool S9Z3

(f) Tool BXZ3

Fig. 16. The calibrated LLR distributions in the *ideal* scenario for 6 tool/zone combinations. Boxplots represent the values of LLR for each of the 10 cases.

learning step.

It shows that same source and different source comparisons can be supported by the computation of a LR derived from machine learning techniques. The overall error rates are all below 3 % (most of the time below 1 %), indicating how accurate the method is.

With glmnet, tools that yield to little to no error are penalized for it and LRs tend to be smaller as the error diminishes (BXZ3). For those cases, the logistic regression calibration tends not to work as it pushes LLR values towards extreme values. To prevent this from happening while maintaining calibration, non-existent error values are introduced in order to keep the benefits of the calibration. One error for each class is added automatically by using the closest values to a LR of 1 in each class and changing the class parameter so that it is counted as an error.

Fig. 16 displays the results for each tool by combining the LLR values from ten cases. In each case, one reference is treated as the questioned mark, revealing significant differences between cases. The within-source variability tends to be less consistent. For some tools, variations emerge when creating references within the same zone, leading to increased variability across cases. This phenomenon is most noticeable in the first two cases, which correspond to the initial cuts made on our tools. This might echo the findings in Djadja's PhD, where the initial shots exhibited distinct characteristics compared to later test fires (Djadja, 2024). For instance, tool BX in zone 3 demonstrates minimal variability, while tool S4 in zone 2 illustrates substantial variability between cases within the same zone.

In some cases, the same source LRs for single questioned mark can vary significantly from values of the order of 10^1 to more than 10^6. The variability introduced when creating references can lead to significant discrepancies, resulting in noticeable differences during comparisons.

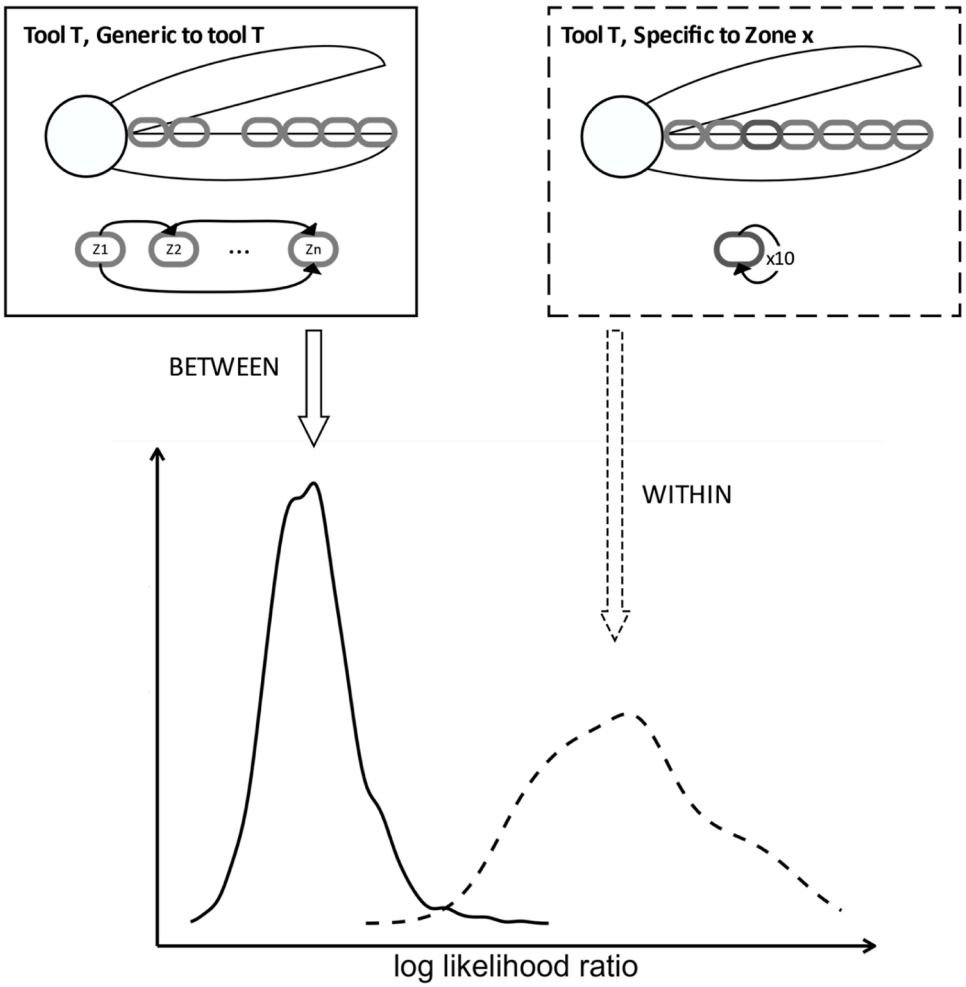


Fig. 17. Representation of the within and between variability distributions of the LRs for the single tool scenario.

3.4. *Single tool scenario*

As observed, the variability within a source is influenced by the specific zone on the blade, indicating that it cannot be generalized across any zone of a tool or to zones of other tools. The *Single Tool Scenario* is designed to investigate whether the between-source variability can be effectively studied by examining the other zones available on the same tool. By comparing these findings with those from the *Ideal Scenario*, we can determine whether this approach of reduced data acquisition offers a viable alternative.

Fig. 17 illustrates the process through which both distributions are generated. The methodology mirrors that of the *ideal scenario*, with the sole distinction lying in how datasets are constructed, as detailed in Section 2.5.

We consider here that marks made on the same zone of a tool are from the same source and marks made in different zones of a tool are from different sources.

Fig. 18 shows the distributions of the LLR in six cases contrasting the ideal scenario and the single tool scenario, lending support to the idea that constructing the between-source variability using the TOI alone is feasible. There's a trend towards reducing the magnitude of the LLR under Hp while increasing them under Hd when moving from the ideal to the single scenario. In a sense the method is more conservative. The maximum total error observed for "single* scenario is slightly above 4 %.

Our observations are not giving definitive arguments to discard this method, quite the contrary. In most cases, the results are either as good as in the *ideal* scenario or more conservative. This *single tool* method offers promising prospect in terms of operational applicability of the model.

3.5. *Combining results from 2 sided mark*

Cutting pliers will leave on a tip of a cut wire a two-sided mark (resulting from each side of the blade) as in Fig. 4a. In this paper we've considered each side as a single mark. In casework, both sides have to be used at the same time and the proposed methodology can be applied simultaneously on both. The question becomes how to combine results and LRs to assess the origin of the whole mark and not only a single side of the mark, keeping in mind that both marks can be dependent.

Even though blades are manufactured separately, they will be positioned in front of each other after the tool is assembled. At a given cutting point, the zone of one blade will always be in front of the same zone of the other blade. With usage of the tool, wear and damages will occur on both blades. Hence, we would expect some dependencies.

Fig. 19 illustrates that dependency. It shows the LLRs for each mark (A1 and A2) separately and for the combined results (A1A2and A1xA2). The LLRs for the combination A1A2 are computed jointly by using the results of both marks comparisons (24 output variables in total) as input of the machine learning process. A1xA2 shows the LRs with the hypothesis that marks are independent. They are obtained by multiplying LRs from A1 and A2. Overall the combination leads to higher LRs on average compared to each mark separately. A1A2 LRs are lower than the LRs that are obtained by a simple multiplication of both marks LRs under

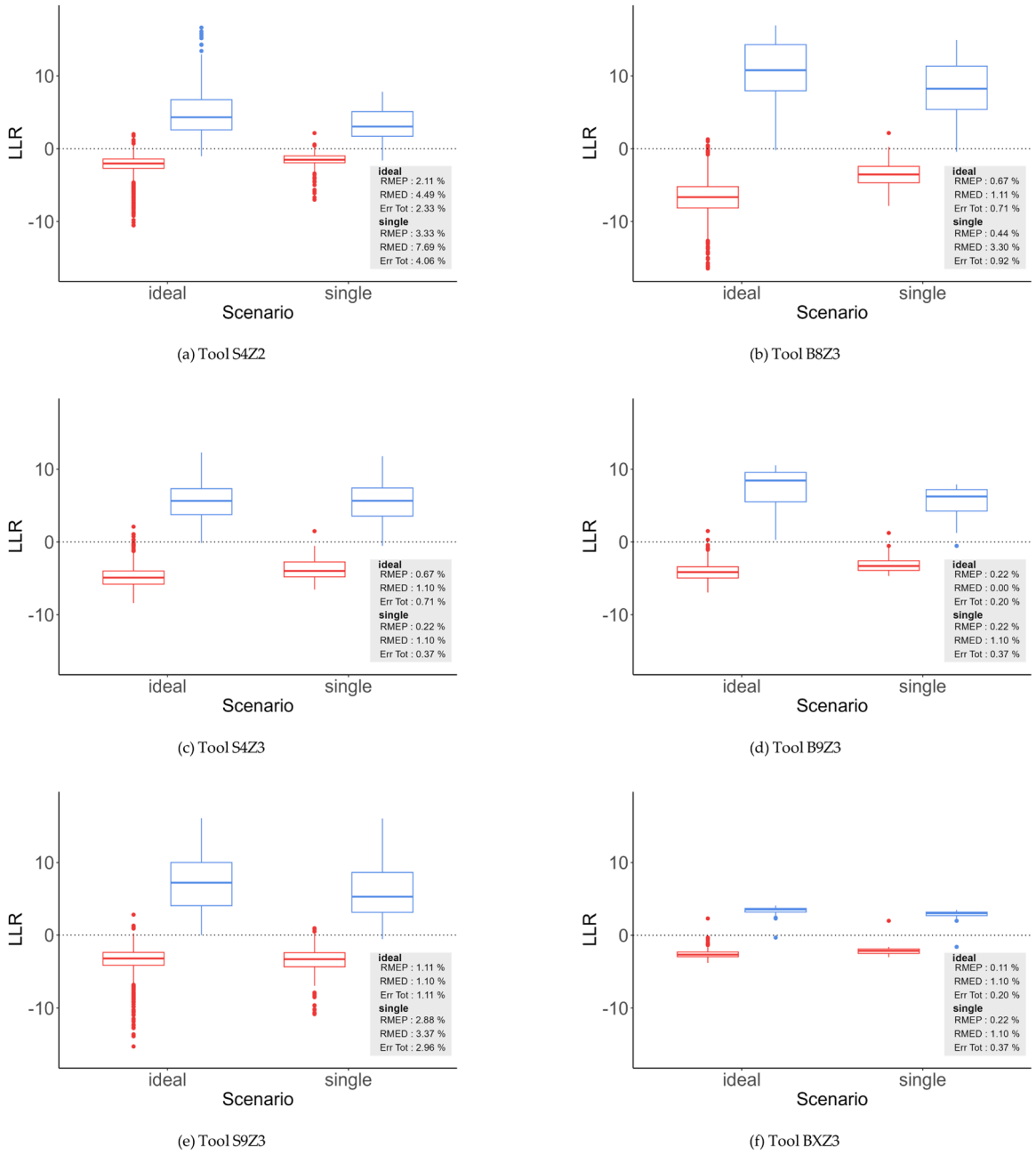


Fig. 18. Comparison of the LLR distribution in 6 different cases using the *single tool* and the *ideal* scenario. The red boxes represent the different source comparisons and the blue the same source comparisons.

the assumption of independence (A1xA2). It is important to stress that this deviation from independence applied only to the tool considered here. Our aim was not the prove the absence of independence, but, using this tool, to show that the easy assumption of independence may not be safe. The observed difference in the median LRs both under Hp and Hd between A1A2 and A1xA2 supports that claim.

With the addition of more variables, the combination (A1A2) gives as expected lower error rates (Table 3) and higher LR values.

4. Discussion

4.1. The developed method

The error rates observed in this study highlight the effectiveness of the developed method. These errors can arise from same source comparisons that yield a limited number of features in agreement. A portion of these errors can be attributed to the operator's failure to cut precisely in the same location, which leads to significant lag values and challenges for the algorithm in accurately aligning corresponding profiles.

Looking ahead, there are several strategies to reduce these error rates. One approach involves refining the algorithms and considering lags beyond those identified by the Cross-Correlation Function (CCF), or excluding references that are not sufficiently aligned to ensure only those with minimal lag are retained. The filtering process is crucial, particularly for features common across multiple tools, such as the waviness observed on S tools. Notably, error rates were higher for S tools compared to B tools, likely due to the filtering process. Striking the right balance in filtering is paramount: excessive filtering may remove small yet critical discriminating features, while insufficient filtering can leave

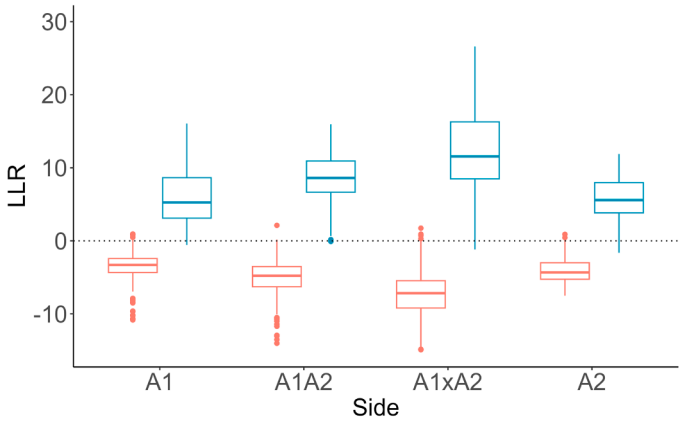


Fig. 19. Boxplots of LLRs for S9Z3 using the same references samples. The legend on the x axis indicates which side(s) was(were) considered: A1 or A2 alone, A1A2 when both marks were considered jointly as an input to the ML and A1xA2 when the joint LR is obtained by the multiplication of the LR associated with A1 and A2 respectively.

Table 3
Table summarizing error rates for the tool S9 at zone 3 in the single tool scenario when considering marks sides separately or combined.

A1	A2	A1A2	A1xA2
RMEP (%) 2.88	2.44	0.44	1.11
RMED (%) 3.37	3.30	1.10	1.12
Err Tot (%) 2.96	2.58	0.55	1.11

waviness-related features intact, potentially increasing correlation during cross-toolmark comparisons.

The misalignment of profiles by the CCF is a notable issue, as depicted in Fig. 20, which demonstrates the appearance of such misalignment. This problem arises in our research primarily for two reasons. First, the computation of CCF correlation considers the length of the profiles, leading to a decreased score when the lag is high due to the limited portion of the profiles being utilized. Typically, the CCF identifies higher values for lower lags, meaning it may converge to a nearer but less accurate alignment over the correct but more distant one. Secondly, the presence of larger features can skew the CCF's alignment focus towards these, minimizing the significance of smaller peaks and valleys and thereby causing misalignment. To mitigate these issues, further algorithmic development could be undertaken, aiming to enhance the precision of alignment despite these challenges.

During our 3D acquisitions, we also realized that even though cutting plier marks are striated patterns, the dynamic of the cut also creates some shapes that are recognizable in 2D or 3D before filtering. These shapes have some degree of repetitivity and could be used for aligning and/or comparing marks. We did not take advantage of these shapes in the comparison step, however they are useful to check if the cuts are correctly aligned by the algorithm (Fig. 21). They were also used to position the ROIs in MountainsMap® when the software did not perform well enough. These shapes could be used as a first step when it comes to the alignment of topographies using a Simplex algorithm for example [40]. It would also prevent misalignment made by the CCF.

The error tables show that global error rates remain low (below 4 % across all scenarios), despite some instances where the RMEP is notably high. A single misclassification from *Hd* to *Hp* can introduce approximately a 1 % error, given that the count of LRs under *Hp* typically stands at 90 in most cases. However, when considering the total of 990 LRs, an individual error contributes only about 0.1 % to the global error rate (Err Tot). While RMEP and RMED serve as accurate indicators of error for each hypothesis, they do not effectively represent the total error magnitude due to their dependency on the volume of data for *Hp* and *Hd*. Therefore, assessing the global error rate is essential, especially in cases of unbalanced data, to accurately gauge the overall methodological error.

This study highlights a unique aspect: we obtained 9 LRs derived from comparing the questioned mark to 9 references, all created from the same zone on a pair of pliers. Traditionally, examiners opt for the comparison that yields the best results for their reports, often choosing the one that they believe best replicates how the tool was used. However, unlike with tools such as screwdrivers [18,6], the manner in which pliers are used has minimal impact on the resulting striations. Instead, the variance between comparisons largely stems from the inherent variability in creating references, which could be influenced by the reference material or the tool itself. Relying solely on the best outcome may lead to an overestimation of the evidentiary weight of that comparison. We argue that averaging the results would provide a more accurate reflection of the evidence's strength.

The validation of our method aligns with the guidelines set forth by Meuwly et al. (2017)[28]. A method attains validation when it exhibits satisfactory performance across various metrics, including accuracy, discriminating power, and calibration. Our methodology demonstrates both high accuracy and strong discriminative ability, as evidenced by the close alignment of observed LRs with the ground truth for both same source and different source comparisons. This is further supported by our distribution plots, which showcase effective classification capabilities alongside low error rates.

ECE (Empirical Crosse-Entropy) plots offer an in-depth view of our method's calibration quality [39], as demonstrated in Fig. 22, which confirms the effective calibration through Cllr(cal) values. These plots specifically highlight one tool from each model that exhibited the highest error rates in both scenarios. Additionally, our study aligns with the secondary characteristics outlined by Meuwly et al. [28], including coherence, robustness, and generalization capabilities. Notably, changes in the dataset did not substantially affect the system's performance, underscoring its reliability across different conditions.

4.2. LR calculation taking into account the features due to design

In this study, our focus is primarily on accidental features (AF) found in striations. However, features determined by design (FDD) also hold potential for comparison. In our context, these typically include the angles formed by the blades on the cut surface or the overall surface roughness, which we algorithmically minimized. When the FDD are observed in agreement, the distinctiveness of these design-based features can significantly enhance the evidentiary value of a comparison. They will then contribute to the overall LR.

For instance, the angle created by both sides of a wire cut is determined by the shape and angle of the tool's blades, as depicted in Fig. 4a. This characteristic is a direct result of the tool's design. However, its manifestation in a toolmark can also be affected by the manner in which the tool is used; for instance, tilting the pliers during use can alter the angle between the two sides. In examinations, such design-based features often serve as initial points of comparison by forensic examiners.

Mattijssen provides guidelines for calculating the LR for these FDD [27]. If the FDD between a mark and the suspected tool are incompatible — for example, if there's an unexplainable discrepancy in the angle — then the numerator of the LR for these FDD is set to 0, rendering the overall LR effectively zero. Conversely, if the FDD are compatible, the numerator is set to 1 or near 1, with the denominator reflecting the rarity of the feature within the relevant population. This rarity is assessed based on specific case factors, such as the time and geographical region of the investigation.

The LR values generated by our system are predicated on the premise that the FDD are consistent be tween the questioned mark and the TOI. If this compatibility were absent, the TOI would be immediately disqualified from further comparison. This precondition means that the computed LR inherently accounts for the presence of compatible FDD, as

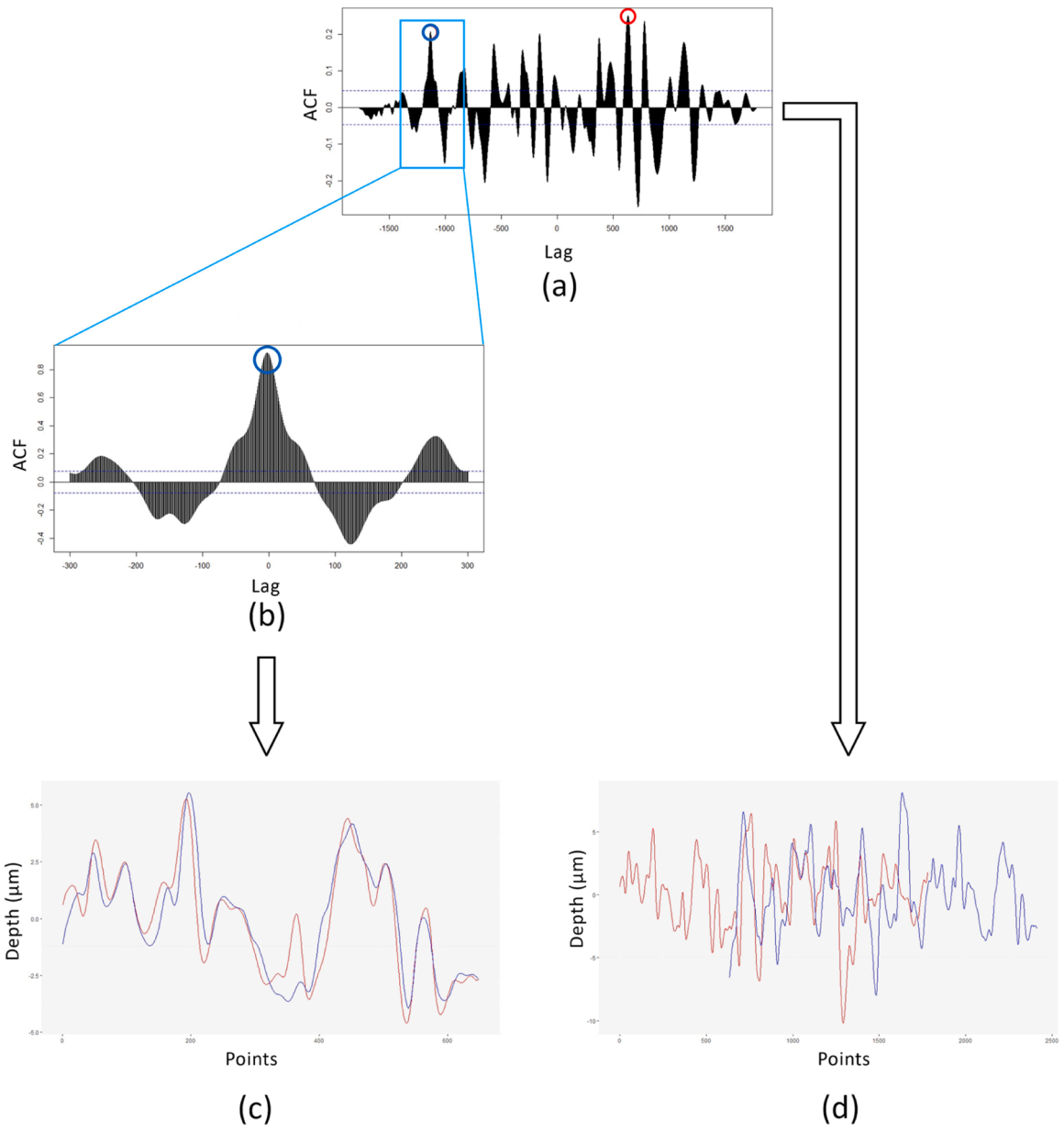


Fig. 20. Explanation of a CCF misalignment. (a) shows the autocorrelation function plot for a comparison between marks made at the same zone of the same tool. The highest value is found at lag=632. However when looking at the profiles with this lag (d), they are misaligned. In (b), we forced the alignment to be at a −1132 lag value which corresponds to the correct alignment of these two profiles. (c) is the visual confirmation that −1132 lag is appropriate.

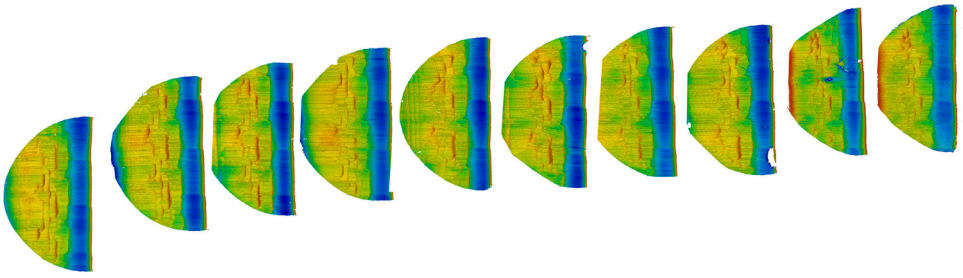
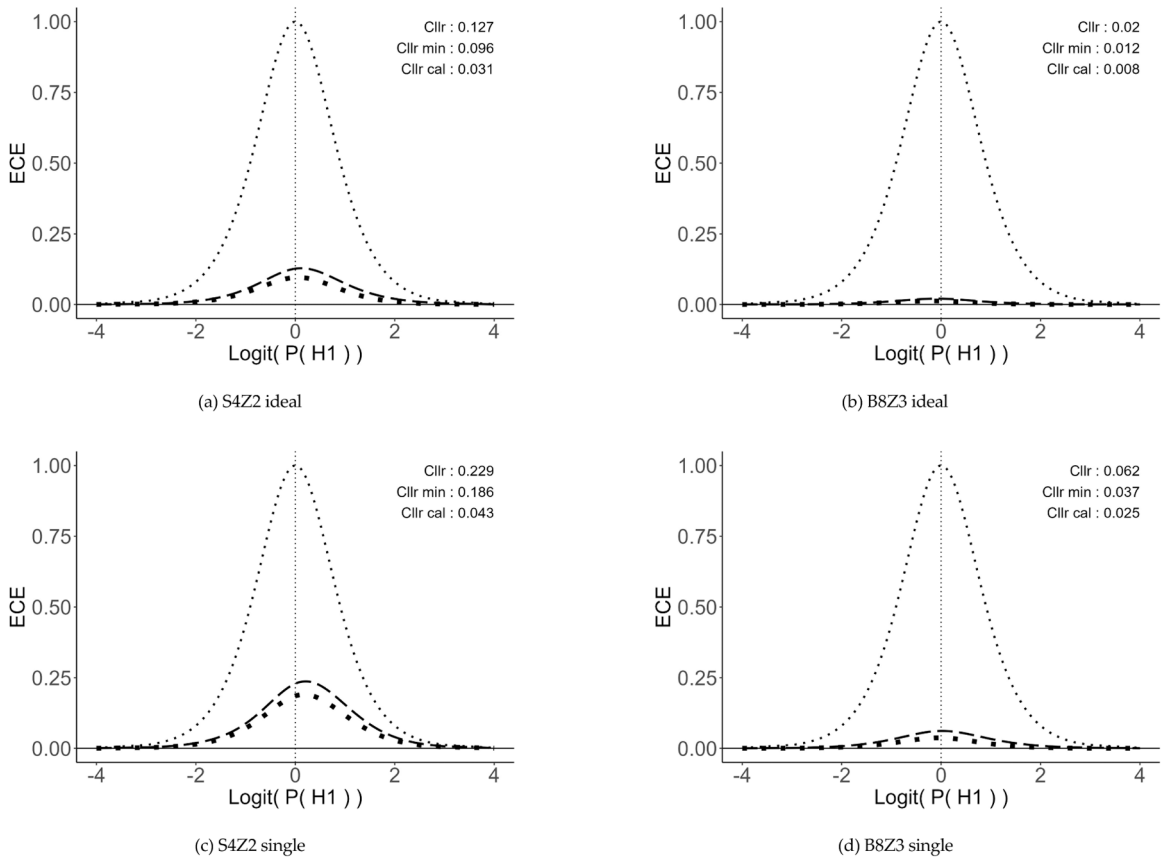


Fig. 21. Aligned marks' topographies from tool K6 zone 2.



(a) S4Z2 ideal

(b) B8Z3 ideal

(c) S4Z2 single

(d) B8Z3 single

Fig. 22. ECE plots in the *ideal* and *single tool* scenarios for tool S4Z2 and B8Z3.

outlined by Mattijssen et al. [27]. To fully integrate the FDD's evidentiary contribution, one could assign a separate LR to these features and then multiply it by the conditional LR derived from the accidental features (AF), as detailed in Eq. 2.

$$LR = LR(FDD) \cdot LR(AF \mid FDD) \quad (2)$$

4.3. Revisiting the propositions

It is worth revisiting the wording of the propositions (or hypotheses) that we have used for *Hp* and *Hd*. In a toolmark case, the propositions will usually be:

Hp: "The tool of interest (TOI) has been used to cut the wire and leave the toolmark"; and

Hd: "Another unknown tool has been used to cut the wire and leave the toolmark".

In this study, we have calculated within-source distributions for a single tool at a specific zone. It means that the likelihood ratios computed are lending support for that particular zone or another zone only. Due to the unique examination requirements for pliers, emphasis on a specific zone may be necessary, especially if an exhaustive consideration of all other zones from a tool could not be made. Without that precision, within-source variability would require comparisons between cuts from different zones of the same tool, leading to expected discrepancies and undermining the analysis. It is essential to recognize that our focus is on a single zone, even though we treat the tool as the overall potential cause of the mark.

This approach is analogous to how examiners focus on a zone of best agreement when analyzing toolmarks, effectively dismissing other zones.

Under these conditions, we could argue that *Hp* should read as "The tool of interest (TOI) has been used to cut the wire *at that zone* and leave the toolmark".

When focusing on the alternative proposition (*Hd*), the "another unknown tool" referred to is not the TOI. Hence the *ideal* scenario described above should prevail. However access to a collection of tools is difficult to implement in casework and time consuming. We've showed however that using the other zones of the TOI itself is a viable option to approximate the *ideal* situation.

4.4. Future perspectives

More data could be processed by the algorithm with different models of pliers and different types of wires. The amount of cuts for the within-source variability could also be increased in order to have more balanced data and investigate if the within-source variability increases or decreases with the number of cuts. The methodology can be applied to any type of toolmark, but it will require some adaptions for example for the definition of the regions of interest or the limits for the alignment of the CCF.

At present, the result of a comparison between a mark and reference is a set of 4 scores. Additional distance metrics could be added and implemented without difficulty. The method is flexible in the sense that new outputs from matching algorithms could be added to the set of variables used as input for the machine learning.

5. Conclusion

This paper outlines a comprehensive methodology designed to enhance the objectivity of toolmark examinations. By integrating 3D acquisition techniques with a mix of comparison metrics and machine learning, Log-Likelihood Ratios (L-LRs) are computed directly for each comparison, allowing the system's performance to be assessed under

various conditions. Our findings suggest that datasets can be constructed in multiple ways without significantly affecting the LR values, making this methodology versatile and applicable in actual casework.

Furthermore, the development of this methodology reveals the importance of specifying within- and between-source variabilities, particularly when applying data filters that require a predetermined cut-off value. Although our results are based on a limited number of pliers, with a focus on cases that exhibited the most errors, we anticipate similar outcomes with other plier models. However, the error rates and values presented should be considered specific to the tools studied. We recommend that examiners create their

own datasets using this procedure in their casework. This approach not only provides a case-specific error rate but also validates the methodology based on the accuracy demonstrated in this research.

CRediT authorship contribution statement

Christophe Champod: Writing – review & editing, Validation, Supervision, Methodology. **Jean-Alexandre Patteet:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

None

References

[1] AFTE, Theory of identification, range of striae comparison reports, and modified glossary definitions-an afte criteria for identification committee report, AFTE J. 24 (1992) 336–340.
[2] AFTE, Theory of identification as it relates to toolmarks, AFTE J. 30 (1998) 86–88.
[3] P. Ahvenainen, I. Kassamakov, K. Hanhijärvi, J. Aaltonen, S. Lehto, T. Reinikainen, E. Hæggström, Csi helsinki: Swli in forensic science: Comparing toolmarks of diagonal cutting pliers, : Am. Institue Phys., AIP Conf. Proc. (2010) 2084–2091.
[4] B. Bachrach, A. Jain, S. Jung, R.D. Koons, A statistical validation of the individuality and repeatability of striated tool marks: screwdrivers and tongue and groove pliers, J. Forensic Sci. 55 (2010) 348–357.
[5] Bachrach, B., 2002. Development of a 3D-based automated firearms evidence comparison system. Journal of Forensic Science 47, 1253–1264. .
[6] M. Baiker, I. Keereweer, R. Pieterman, E. Vermeij, J. van der Weerd, P. Zoon, Quantitative comparison of striated toolmarks. (doi:), Forensic Sci. Int. 242 (2014) 186–199, https://doi.org/10.1016/j.forsciint.2014.06.038.
[7] Baldwin, D., Morris, M., Bajic, S., Zhou, Z., Kreiser, J., 2004. Statistical tools for forensic analysis of toolmarks. Report. Ames Lab., IA (US). URL: https://www.osti.gov/servlets/purl/825030.
[8] A.A. Biasotti, Rifling methods - a review and assessment of the individual characteristics produced, AFTE J. 13 (1981) 34–61.
[9] R. Bolton-King, J.P.O. Evans, C.L. Smith, J.D. Painter, D.F. Allsop, W.M. Cranton, What are the prospects of 3D profiling systems applied to firearms and toolmark identification? AFTE J. 42 (2010) 23–33.
[10] Cadevall, 2018. New bullet and cartridge case 3D measurements tools. 25th Annual Meeting of ENFSI Firearms/GSR Working Group.
[11] F. Cassidy, Examination of toolmarks from sequentially manufactured tongue-and-groove pliers, J. Forensic Sci. 25 (1980) 796–809.
[12] C. Champod, C.J. Lennard, P.A. Margot, M. Stoilovic, Boca Raton. Fingerprints and other Ridge Skin Impressions, 2nd ed.,, CRC Press, 2016.
[13] Chumbley, L.S., Morris, M., 2013. Significance of Association in Tool Mark Characterization. Report 243319. National Institute of Justice. URL: ⟨https://www.ncjrs.gov/pdffiles1/nij/grants/243319.pdf⟩.
[14] L.S. Chumbley, M.D. Morris, M.J. Kreiser, C. Fisher, J. Craft, L.J. Genalo, S. Davis, D. Faden, J. Kidd, Validation of tool mark comparisons obtained using a quantitative, comparative, statistical algorithm, J. Forensic Sci. 55 (2010) 953–961.
[15] Digital Surf, 2024. Mountainsmap® for profilometry. URL: ⟨https://www.digitalsurf.com/fr/produits-solutions/profilometrie/⟩.
[16] L. Ekstrand, S. Zhang, T. Grieve, L.S. Chumbley, M.J. Kreiser, Virtual tool mark generation for efficient striation analysis, J. Forensic Sci. 59 (2014) 950–959.
[17] R.A. Freeman, Consecutively rifled polygon barrels. AFTE J. 10 (1978) 40–42.
[18] D.L. Garcia, R. Pieterman, M. Baiker, Influence of the axial rotation angle on tool mark striations., URL: http://www.sciencedirect.com/science/article/pii/S0379073817303201https://www. sciencedirect.com/science/article/pii/S0379073817303201?via%3Dihub, doi: Forensic Sci. Int. 279 (2017) 203–218, https://doi.org/10.1016/j.forsciint.2017.08.021.
[19] T. Grieve, L.S. Chumbley, J. Kreiser, M. Morris, L. Ekstrand, S. Zhang, Objective comparison of toolmarks from the cutting surfaces of slip-joint pliers, AFTE J. 46 (2014) 176.
[20] Hastie, T., Qian, J., Tay, K., 2023. An introduction to glmnet. URL: https://glmnet.stanford.edu/articles/glmnet.html.
[21] V.V. Heikkinen, I. Kassamakov, C. Barbeau, S. Lehto, T. Reinikainen, E. Hæggström, Identifying diagonal cutter marks on thin wires using 3D imaging, J. Forensic Sci. 59 (2014) 112–116, https://doi.org/10.1111/1556-4029.12291.
[22] M. Heizmann, Strategies for the automated recognition of marks in forensic science, in: Z.J. Geradts, L.I. Rudin (Eds.), SPIE International Symposium on Law Enforcement Technologies – Investigative Image Processing II, SPIE, 2002, pp. 68–69.
[23] Heizmann, M., 2002a. Automated comparison of striation marks with the system ge/2, in: Geradts, Z.J., Rudin, L.I. (Eds.), SPIE International Symposium on Law Enforcement Technologies – Investigative Image Processing II, SPIE. pp. 80–91.
[24] Jacquet, M., 2021. Interprétation des scores de reconnaissance faciale automatique pour l'investigation et le tribunal. Phd thesis. University of Lausanne, School of Criminal Justice.
[25] Kuhn, M., 2022. caret: Classification and Regression Training. URL: ⟨https://CRAN.R-project.org/package=caret⟩. r package version 6.0-93.
[26] C. Macziewski, R. Spotts, S. Chumbley, Validation of toolmark comparisons made at different vertical and horizontal angles, J. Forensic Sci. 62 (2017) 612–618. URL: https://doi.org/10.1111/1556-4029.13342http://onlinelibrary.wiley. com/store/10.1111/1556-4029.13342/asset/jfo13342.pdf?v=1&t=j3sylgmf&s=1b8f8572e2540f2c087ec7e9685291e71c0a0392, doi: 10.1111/1556-4029.13342.
[27] Mattijssen, E.J., Berger, C.E., Vergeer, P., Kerkhoff, W., Stoel, R.D., 2021. Firearm evaluation at source level: How to define the relevant population and how to apply an unrestrictive alternative proposition. PDF hosted at the Radboud Repository of the Radboud University Nijmegen URL: ⟨https://repository.ubn.ru.nl/bitstream/handle/2066/233894/233894.pdf⟩.
[28] D. Meuwly, D. Ramos, R. Haraksim, A Guideline for the Validationof Likelihood Ratio Methods Used for Forensic Evidence Evaluation, Forensic Sci. Int 276 (2017) 142–153.
[29] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, 2009 (URL), ⟨https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf⟩.
[30] R.G. Nichols, Firearm and toolmark identification criteria: a review of the litterature, J. Forensic Sci. 42 (1997) 466–474.
[31] R.G. Nichols, Firearm and tool mark identification: The scientific reliability and validity of the afte theory of identification discussed within the framework of a study of ten consecutively manufactured extractors, AFTE J. 36 (2004) 67–88.
[32] Nichols, R.G., 2013. Tools. Academic Press, Waltham. book section Tools. pp. 60–68. URL: http://www.sciencedirect.com/ science/article/pii/B9780123821652002828https://ac.els-cdn.com/B9780123821652002828/3-s2.0-B9780123821652002828- main.pdf?_tid=27728b94-c75a-4b27-9973-cb6eff078cbc&acdnat=1539615570_7e0b956d94039438674e34b56e2c65ba, doi: https://doi.org/10.1016/B978-0-12-382165-2.00282-8.
[33] NIJ, 2015. FINAL REPORT: Forensic Optical Topography Working Group. Report. National Institute of Justice. URL: ⟨https://forensiccoe.org/private/5d937bdece0c6⟩.
[34] J.A. Patteet, C. Champod, Striated toolmarks comparison and reporting methods: review and perspectives. Forensic Sci. Int. (2024) https://doi.org/10.1016/j.forsciint.2024.111997.
[35] PCAST, 2016. Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods. Report. Executive Office of the President President's Council of Advisors on Science and Technology. URL: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.
[36] People v. Genrich, Case No 1019COA132. 2019. Court Case.
[37] Petraco, N.D.K., Gambino, C., Kammerman, F.L., 2011. Application of Machine Learning to Toolmarks: Statistically Based Methods for Impression Pattern Comparisons. Report. U.S Department of Justice. URL: https://www.ncjrs.gov/pdffiles1/ nij/grants/239048.pdf.
[38] N.D.K. Petraco, L. Kuo, H. Chan, E. Phelps, C. Gambino, P. McLaughlin, F. Kammerman, P. Diaczuk, S. Peter, N. Petraco, J.E. Hamby, Estimates of striation pattern identification error rates by algorithmic methods, AFTE J. 45 (2013) 235–244.
[39] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (2018) 208.
[40] Riva, F., 2011. Etude sur la valeur indicielle des traces présentes sur les douilles. Phd thesis. University of Lausanne, School of Criminal Justice.
[41] F. Riva, E.J.A.T. Mattijssen, R. Hermsen, P. Pieper, W. Kerkhoff, C. Champod, Comparison and interpreta- tion of impressed marks left by a firearm on cartridge cases – towards an operational implementation of a likelihood ra- tio based technique, URL: http://www.sciencedirect.com/science/article/pii/S0379073820302255, doi: Forensic Sci. Int. 313 (2020) 110363 https://doi.org/10.1016/j.forsciint.2020.110363.
[42] RStudio Team, 2020. RStudio: Integrated Development Environment for R. RStudio, PBC. Boston, MA. URL: http://www.rstudio.com/.
[43] J. Song, T. Vorburger, T. Renegar, H. Rhee, A. Zheng, L. Ma, J. Libert, S. Ballou, B. Bachrach, K. Bogart, Correlation of topography measurements of nist srm 2460 standard bullets by four techniques, Meas. Sci. Technol. 17 (2006) 500–503.
[44] R. Spotts, L.S. Chumbley, L. Ekstrand, S. Zhang, J. Kreiser, Optimization of a statistical algorithm for objective comparison of toolmarks. J. Forensic Sci. 60 (2015) 303–314, https://doi.org/10.1111/1556-4029.12642.

[45] N. Volkov, N. Finkelstein, Y. Novoselsky, T. Tsach, Bolt cutter blade's imprint in toolmarks examination, J. Forensic Sci. 60 (2015) 1589–1593, https://doi.org/10.1111/1556-4029.12867.
[46] T.V. Vorburger, J.F. Song, W. Chu, L. Ma, S.H. Bui, A. Zheng, T.B. Renegar, Applications of cross-correlation functions, URL: http://www.sciencedirect.com/science/article/pii/S0043164810001407, doi:https://, Wear 271 (2011) 529–533, https://doi.org/10.1016/j.wear.2010.03.030.
[47] Wangqian Ju, H.H., 2021. cmpsR: R Implementation of Congruent Matching Profile Segments Method. URL: https://CRAN.R- project.org/package=cmpsR. r package version 0.1.2.
[48] D.J. Watson, The identification of toolmarks produced from consecutively manufactured knife blades in soft plastics, AFTE J. 10 (1978) 43–45.
[49] D.J. Watson, The identification of consecutively manufactured crimping dies. AFTE J. 10 (1978) 19–21.
[50] H. Wickham, M. Averick, J. Bryan, W. Chang, L.D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T.L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D.P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse, J. Open Source Softw. 4 (2019) 1686, https://doi.org/10.21105/joss.01686.
[51] F. Xie, S. Xiao, L. Blunt, W. Zeng, X. Jiang, Automated bullet-identification system based on surface topography techniques, Wear 266 (2009) 518–522. ⟨https://eprints.hud.ac.uk/id/eprint/31746/1/wear09.pdf⟩ (URL).