An overview of log likelihood ratio cost in forensic science – Where is it used and what values can we expect?

Van Lierop, S.; Ramos, D.; Sjerps, M.; Ypma, Rolf

DOI
10.1016/j.fsisyn.2024.100466

Publication date
2024

Document Version
Final published version

Published in
Forensic Science International: Synergy

Citation for published version (APA):
Van Lierop, S., Ramos, D., Sjerps, M., & Ypma, R. (2024). An overview of log likelihood ratio cost in forensic science – Where is it used and what values can we expect? *Forensic Science International: Synergy*, *8*, Article 100466. https://doi.org/10.1016/j.fsisyn.2024.100466

Forensic Science International: Synergy 8 (2024) 100466

Contents lists available at ScienceDirect

Forensic Science International: Synergy

journal homepage: www.sciencedirect.com/journal/forensic-science-international-synergy

An overview of log likelihood ratio cost in forensic science – Where is it used and what values can we expect?

Stijn van Lierop [a], Daniel Ramos [b], Marjan Sjerps [a,c,*], Rolf Ypma [a]

[a] Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague, 2497GB, the Netherlands
[b] AUDIAS Lab, Universidad Autonoma de Madrid, Escuela Politécnica Superior, Calle Francisco Tomàs y Valiente 11, 28049, Madrid, Spain
[c] University of Amsterdam, KdVI, PO Box 94248, Amsterdam, 1090 GE, the Netherlands

ARTICLE INFO

Keywords:
C_{llr}
Log-likelihood ratio cost
Automated likelihood ratio systems
Benchmarking review
Forensic evaluation
Forensic datasets
Performance of LR system

ABSTRACT

There is increasing support for reporting evidential strength as a likelihood ratio (LR) and increasing interest in (semi-)automated LR systems. The log-likelihood ratio cost (C_{llr}) is a popular metric for such systems, penalizing misleading LRs further from 1 more. $C_{llr} = 0$ indicates perfection while $C_{llr} = 1$ indicates an uninformative system. However, beyond this, what constitutes a "good" C_{llr} is unclear.

Aiming to provide handles on when a C_{llr} is "good", we studied 136 publications on (semi-)automated LR systems. Results show C_{llr} use heavily depends on the field, e.g., being absent in DNA analysis. Despite more publications on automated LR systems over time, the proportion reporting C_{llr} remains stable. Noticeably, C_{llr} values lack clear patterns and depend on the area, analysis and dataset.

As LR systems become more prevalent, comparing them becomes crucial. This is hampered by different studies using different datasets. We advocate using public benchmark datasets to advance the field.

1. Introduction

In forensic science, support for using likelihood ratios (LRs) to assess the strength of forensic evidence has been growing [1–7]. Consequently, the issue of evaluating the performance of LR systems has been addressed in the recent past, focusing on their validation for use in casework [8]. In particular, several performance metrics have been proposed. The scientific community and forensic practitioners are notably utilizing the *Log Likelihood Ratio Cost*, also known as C_{llr}, which was initially introduced in Ref. [9] within the context of likelihood-ratio-based speaker verifiers and subsequently adapted for forensic speaker recognition [10]. However, it is important to note that the use of the C_{llr} extends beyond speech-based systems and can be applied to any method that produces LRs.

The C_{llr} is defined as:

$$C_{llr} = \frac{1}{2}\cdot\left(\left[\frac{1}{N_{H_1}}\sum_{i}^{N_{H_1}} log_2\left(1+\frac{1}{LR_{H_{1_i}}}\right)\right] + \left[\frac{1}{N_{H_2}}\sum_{j}^{N_{H_2}} log_2(1+LR_{H_{2_j}})\right]\right) \tag{1}$$

Here, N_{H_1} is the number of samples for which H_1 is true, N_{H_2} is the number of samples for which H_2 is true, LR_{H1} are the LR values predicted by the system for samples where H_1 is true, and LR_{H2} are the LR values predicted by the system for samples where H_2 is true. Hence, when we have a set of empirical LR values predicted by a certain LR system and the corresponding ground truth labels are available, we can calculate a C_{llr}. Furthermore, the metric can be split into two parts, giving an indication of the error due to imperfect discrimination ('do H_1-true samples get a higher LR than H_2-true samples?') and imperfect calibration ('is the value of the assigned LR correct, not under- or overstating the evidence?'). These two metrics can be calculated by applying the Pool Adjacent Violators (PAV) algorithm [11,12] on the evaluation set, mimicking 'perfect' calibration [9], and re-calculating C_{llr}. The resulting value is taken as an assessment of discrimination (called $C_{llr-min}$), the difference ($C_{llr-cal} = C_{llr} - C_{llr-min}$) as an assessment of calibration [13,14].

Importantly, it is worth highlighting some advantages and drawbacks of the C_{llr} as a performance metric for LR systems. Among its advantages, the C_{llr} is a so-called 'strictly proper scoring rule', possessing favorable mathematical properties such as probabilistic and information-theoretical interpretation. As already mentioned, the metric provides an indication of both calibration and discriminating power of a method, allowing separate estimation of these two aspects of

* Corresponding author. Netherlands Forensic Institute, Laan van Ypenburg 6, The Hague, 2497GB, the Netherlands.
E-mail address: m.j.sjerps@uva.nl (M. Sjerps).

https://doi.org/10.1016/j.fsisyn.2024.100466
Received 23 November 2023; Received in revised form 7 March 2024; Accepted 29 March 2024
Available online 17 April 2024

performance [15]. It thus considers not just whether an evaluation was misleading (i.e. the LR supports the wrong hypothesis), but also its associated evidential strength (e.g. misleading LR = 100 is worse than misleading LR = 2). It furthermore fosters incentives for forensic practitioners to offer accurate and truthful LRs, a critical aspect in a field where inaccurate or biased LRs can have significant implications for the criminal justice system, and imposes strong penalties for highly misleading LRs. In addition, the C_{llr} is a scalar that can be easily thresholded for validation, ensuring comparability between different systems, methods, and setups.

However, the C_{llr} also exhibits limitations. Primarily, as any empirical performance measure, it necessitates an empirical set of LRs, introducing challenges in database selection to generate these LRs. Ideally, these databases should resemble actual casework conditions, but such data is often limited, sometimes requiring a two-stage validation procedure using laboratory-collected data. Additionally, the C_{llr} is affected by small sample size effects, i.e., scarcity in empirically generated LRs, potentially leading to unreliable performance measurements. Data handling is a crucial concern in any validation process [8], as emphasized in many publications in the field [16]. Although certainly impacting C_{llr}, these problems are common to all empirical evaluation metrics.

Another significant limitation of the C_{llr} is that, as a scalar, it provides a highly condensed statistic of model performance. In certain situations, further analysis may be beneficial to detect and correct model issues, leading to the proposal of alternative or specialized measures, as described below. The C_{llr} value can be split as a sum of $C_{llr-cal}$ and $C_{llr-min}$, describing the calibration error and discrimination error separately. While a large $C_{llr-cal}$ value indicates an LR system overstates and/or understates the evidential strength, it does not tell us by how much and for which evidence class this tends to happen [17]. In addition, the C_{llr} value weighs the two different types of misleading evidence (misleadingly supporting H_1 or H_2) symmetrically, using a logarithmic scoring rule. One could debate whether or not this is appropriate in a forensic setting. Finally, the interpretation of the actual numerical value of the C_{llr} is not intuitive. Although lower is better and the value should be below 1, many researchers struggle with whether a value of say 0.3 can be considered 'good'.

Various alternative performance metrics and representations have been proposed in the literature. Firstly, a more comprehensive picture can be obtained by looking at the full distribution of LRs under H_1 and H_2, e.g. using Tippett plots [8]. Similarly, the Empirical Cross-Entropy plot (ECE-plot) is often inspected as it generalizes the C_{llr} to unequal prior odds. Furthermore, many metrics exist that focus solely on either discriminating power or calibration. In the former category fall commonly used scalars such as accuracy or false positive rate, and representations as the Receiver Operating Characteristic curve (ROC) or its normalized version, the Detection Error Tradeoff (DET). ROC curves enable the computation of the Area Under the Curve (AUC) value, summarizing the discriminating power of a given method [15]. More recently, tools for assessing the calibration of LR systems have been proposed in the form of fiducial calibration discrepancies and corresponding plots [17], and the metric devPAV [18].

This paper focuses on the practical use of the C_{llr}. We note an increasing number of scientific studies proposing LRs across various disciplines. Therefore, one may expect a corresponding increase in the use of the C_{llr}. We conduct a systematic review of these studies, aiming to provide the forensic community with some intuition for what numerical values of the metric to expect or aim for. This may provide some guidance to the question as to how good a particular LR system performs beyond the anchors of $C_{llr} = 0$, indicating a perfect system, and $C_{llr} = 1$, indicating an uninformative system equivalent to a system that always returns LR = 1. Additionally, we give an overview of usage of this metric over time and across different forensic disciplines.

2. Materials and methods

We performed a literature search on October 31st, 2022 using a keyword search on the University of Amsterdam library and Scopus databases (Table 1). We took April 2006 as the start of the search range, since one of the first papers proposing the use of the C_{llr} metric to measure LR system performance was published in this month [9]. Hence, only English publications published between April 2006 and October 30th, 2022 were included in this review. In addition, we checked the references and citations of every selected paper to find additional relevant literature. We did this manually, but also performed an automated network search via the online tools Inciteful.xyz and connectedpapers.com. All articles were screened against several inclusion and exclusion criteria (Table 2).

If articles reported multiple C_{llr} values, the most forensically relevant C_{llr} was chosen, meaning the C_{llr} value that was the result of evaluation on data and conditions mostly resembling casework in the view of the authors. In case multiple models were evaluated on the same forensically relevant dataset, the C_{llr} of the best-performing model was selected. In cases where C_{llr} values were reported for multiple, all forensically relevant datasets, these C_{llr} values were included separately as they could be viewed as independent evaluations.

All data processing and visualization was done using Microsoft Excel 2022 [19] and the Pandas [20], NumPy [21], Matplotlib [22], Seaborn [23], SciPy [24] and NFI Likelihood Ratio [25] libraries in Python 3.10 [26]. All code for plotting and analysis as well as the dataset itself is available in the supplementary information.

Table 1
Keywords used for systematic literature search. All combinations of keyword 1, keyword 2 and optionally keyword 3 was used to query the University of Amsterdam and Scopus databases.

Keyword 1	Keyword 2	Keyword 3 (optional)
automated likelihood ratio system	forensic drug	investigation
cllr	forensic biometrics	analysis
cost-log likelihood ratio	blood spatter	evaluation
empirical cross entropy	forensic physics	identification
likelihood ratio	forensic chemistry	comparison
likelihood ratio model	forensic biology	validation
log likelihood ratio cost	digital forensics	individualisation
expert proficiency test	explosives	
automated system	microtraces	
score-based	forensic glass	
feature-based	forensic document	
	forensic text	
	gunshot residue	
	forensic handwriting	
	forensic image	
	forensic fiber	
	forensic hair	
	toolmark	
	footmark	
	forensic odontology	
	forensic firerarm	
	forensic fire	
	taphonomy	
	forensic toxicology	
	forensic soil	
	forensic paint	
	forensic plant	
	wildlife forensics	
	forensic	
	anthropology	
	forensics	
	forensic dna	
	forensic speaker	
	forensic ballistics	
	forensic fingermark	
	forensic fingerprint	
	forensic audio	
	forensic	
	food forensics	

Table 2
Inclusion and exclusion criteria used for systematic literature search.

Inclusion Criteria	Exclusion Criteria
Studies on forensic (semi-)automated LR systems.	Studies on human forensic experts that report LRs.
Studies published between April 2006 and October 30th, 2022.	Studies published before April 2006 or after October 30th, 2022.

3. Results

3.1. Proportion of publications reporting C_{llr}

In total, we found 136 publications on (semi-)automated LR systems [9,10,13,14,27–157], of which 80 (58.8 %) reported a C_{llr} value in some way, either by giving an explicit (list of) value(s) or by plotting an ECE plot from which the C_{llr} value could be read. Out of the 80 publications that reported a C_{llr}, 45 (56.3 %) also reported a $C_{llr-min}$, making it possible to differentiate between discrimination and calibration error. As is shown in Fig. 1, the number of publications on (semi-)automated LR systems as well as the proportion of reported C_{llr} values differs widely per forensic expertise area. The various areas included were chosen based on a manual clustering of the selected literature taking into account common forensic expertise areas in literature and forensic practice. While the exact clustering is necessarily to some degree arbitrary, it is nevertheless helpful to illustrate patterns between various forensic sciences.

From Fig. 1 can be seen that by far the largest number of publications on (semi-)automated LR systems is in the area of speaker recognition. The proportion of publications reporting a C_{llr} in this area is also quite high (93.3 %). This makes sense given that speaker recognition is one of the most developed areas with respect to automated LR systems and is the forensic area in which the application of the C_{llr} was first proposed [9]. A relatively high proportion of publications report C_{llr} values in the area of materials and microtraces (81.0 %), which mainly includes publications on the individualisation of traces such as glass, paint and other substances. The area with the third largest number of publications is forensic DNA analysis. This is mainly due to the large number of probabilistic genotyping models that have been developed for the interpretation of DNA evidence (e.g. STRMix [111], EuroForMix [158], DNAxs [159], etc.). The null proportion of publications reporting a C_{llr} in this area (0 %) is striking. For stylometric analysis (100.0 %) and face comparison (85.7 %) the C_{llr} seems to be frequently used as well, but it should be noted that all publications found in stylometric analysis were written by the same author(s). In fingermark comparison, the C_{llr} seems to be less often used (50.0 %). Publications from the remaining expertise areas where less than 5 publications could be found were pooled together in a single “Other” category. For these categories, a bit less than half (44.0 %) of publications use the C_{llr} metric.

3.2. Proportion of publications reporting C_{llr} over time

Besides looking at the differences between forensic areas, it is also interesting to look at the development of the application of the C_{llr} over time. Fig. 2 shows the number of publications on (semi-)automated LR systems and the proportion reporting a C_{llr} per year. There is a clear upward trend in the number of publications per year on (semi-)automated LR systems. Until 2018, no significant change is visible in the proportion of papers using the C_{llr} metric for evaluation. After 2018 however, the proportion of papers using the metric seems to be much smaller compared to the period before. This is due to some extent to an increased number of publications in the DNA analysis field in the last couple of years. When filtering out all DNA publications (Fig. 3), the difference between the periods becomes smaller, but remains present.

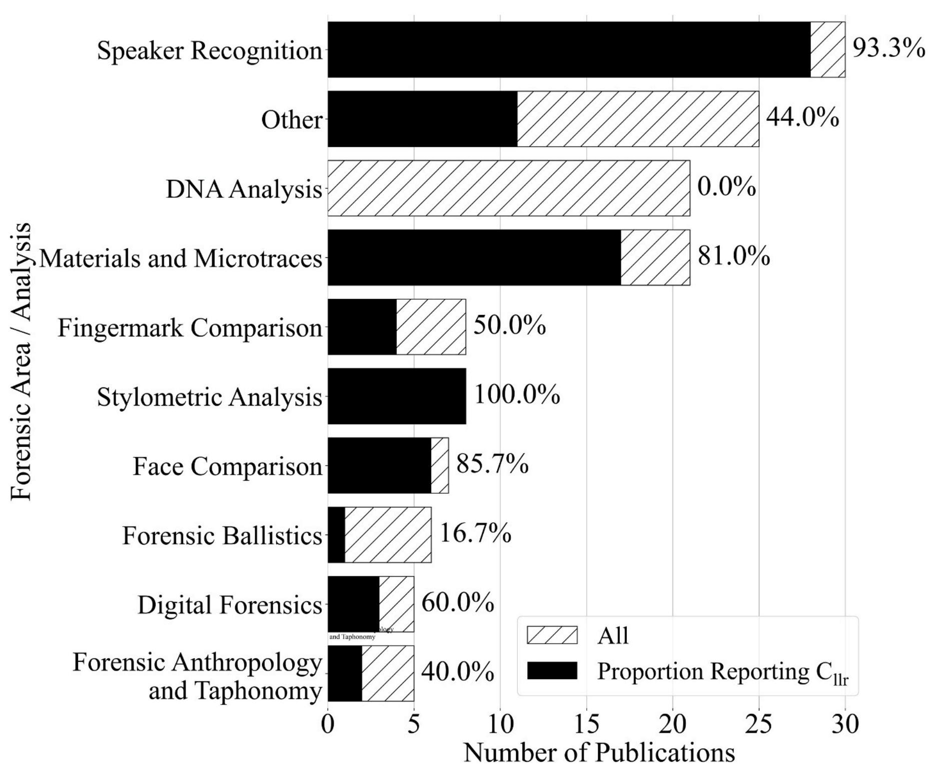

Fig. 1. The number of publications on (semi-)automated likelihood ratio (LR) systems per forensic expertise area. The total number of publications per area is indicated by a dashed bar and the number of publications reporting one or more C_{llr} values per area is indicated by a solid bar. The bars are overlapped, not stacked. The areas are ordered based on absolute number of publications on (semi-)automated LR systems. The percentage on top of each bar indicates the proportion of publications reporting a C_{llr} value in that area. All publications from areas with less than 5 publications were pooled together in a single “Other” category.

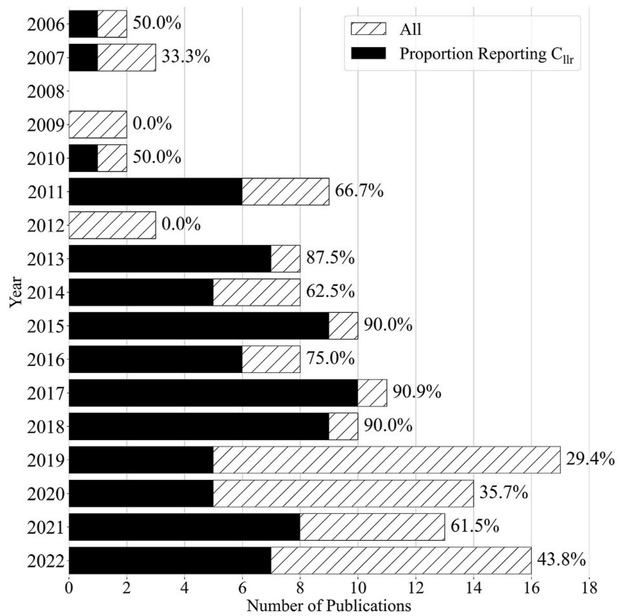

Fig. 2. The number of publications on (semi-)automated likelihood ratio (LR) systems per year. The total number of publications per year is indicated by a dashed bar and the number of publications reporting C_{llr} values per year is indicated by a solid bar. The bars are overlapped, not stacked. The percentage of papers reporting a C_{llr} is printed on top of each bar.

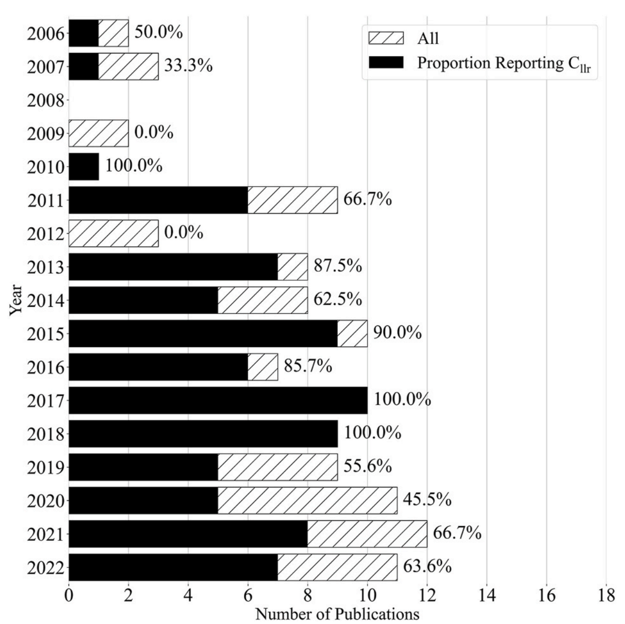

Fig. 3. The number of publications on (semi-)automated likelihood ratio (LR) systems per year, excluding all publications in the area of DNA analysis. The total number of publications per year is indicated by a dashed bar and the number of publications reporting C_{llr} values per year is indicated by a solid bar. The bars are overlapped, not stacked. The percentage of papers reporting a C_{llr} is printed on top of each bar.

3.3. Comparison of C_{llr} values between forensic expertise areas

From the 80 publications on (semi-)automated LR systems that reported one or more C_{llr} values, we selected 95 forensically relevant C_{llr} values. Fig. 4 shows the distribution of forensically relevant C_{llr} values per forensic expertise area, only including the best values per unique dataset. Because different models evaluated on the same dataset can be compared and one would naturally pick the best model for application in practice, only values from the best models per dataset are shown. Note that Fig. 4 shows the best values per distinct dataset in a certain area, whereas Fig. 1 shows the number of unique publications per area. This explains why stylometric comparison, but not signature comparison, is shown separately in Fig. 1, even though the former shows fewer data points in Fig. 4. For the former, we found 8 publications with 3 C_{llr} values on distinct datasets, for the latter this was 3 publications with 8 C_{llr} values on distinct datasets.

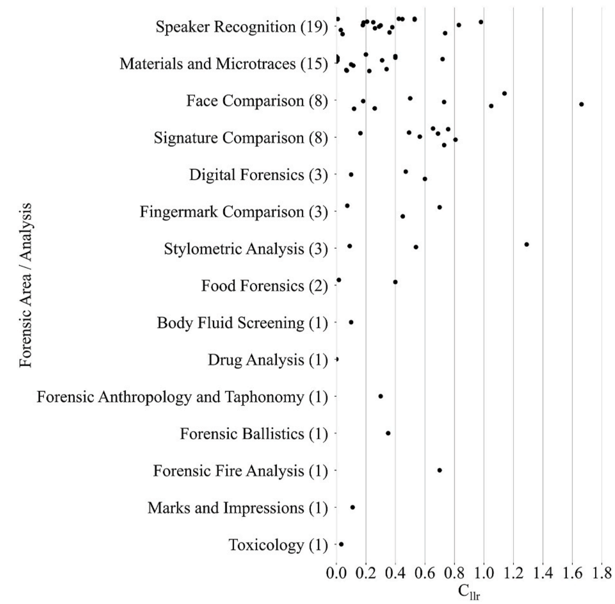

Fig. 4. The best values per dataset (lower C_{llr} values indicate better performance) of selected forensically relevant C_{llr} values per forensic expertise area. The number in between brackets indicates the number of C_{llr} values plotted.

From Fig. 4 it can be seen that there is quite a lot of variance in C_{llr} values within every area. Even in relatively well-defined areas such as speaker recognition and face comparison, C_{llr} values are distributed between [0.0002, 0.98] and [0.12, 1.662]. The values between the areas do follow, to some extent, the expected pattern. For example, we would expect much more discriminating results in materials and microtraces than in speaker recognition, and thus lower C_{llr} values. In contrast, for signature comparisons, we would expect higher C_{llr} values. This is clear from the figure, with C_{llr} values predominantly below 0.4 for the former, ranging across [0, 1] for speaker recognition and mainly around 0.6 for signature comparison. Literature does allow for drawing tentative conclusions on which C_{llr} values one may expect in general for a state-of-the-art system per area. We should be very careful in any conclusions though, as the values reported may fluctuate more because of the dataset queried than the method employed. Unfortunately, for many areas such as drug analysis and forensic ballistics, only one or two C_{llr} values are available, making it hard to draw any conclusions. This also makes it more difficult to observe any clear pattern in C_{llr} values in general.

4. Discussion

There are slight differences visible between C_{llr} values reported in different forensic expertise areas, with fields that generally provide higher discriminating power generally showing lower C_{llr} values. However, the abundance of diverse datasets used for evaluation and the relative scarcity of C_{llr} values reported make it difficult to distinguish any clear patterns. This is a disappointing finding, as the current state of the literature thus not allows for further intuition on what constitutes a 'good' C_{llr}. To address the noise caused by evaluation on these diverse datasets, different systems should be evaluated on standard, forensically relevant, benchmark datasets which would allow for a much more direct comparison between systems. The set of C_{llr} reference values resulting from such evaluation can serve as a valuable resource for assessing the quality of (semi-)automated systems being developed in the future.

There is no objective system for assessing whether a dataset is 'forensically relevant'. For this study, we subjectively assessed whether evaluations were conducted under circumstances resembling casework conditions. In this way, we aimed to get at least a global view of C_{llr} values that can be expected for a system in practice, while mitigating the confounding impact of unrealistic setups. Nevertheless, what constitutes forensic relevance remains a subjective decision that may differ per jurisdiction or organization.

Even though the number of publications on (semi-)automated LR systems has been increasing since 2006, the use of the C_{llr} metric has remained relatively constant. It is challenging to explain this phenomenon, but it is plausible that the stability of the metric usage can be more attributed to an existing user base that has been publishing more over the years, rather than widespread adoption of the metric across the field.

In addition, it is worth noting the seemingly non-existent use of C_{llr} in forensic DNA analysis. Especially, since its use seems particularly well suited for forensic DNA comparison, given the well-established Bayesian reasoning methodology in the field. One possible reason is the fact that there are cases where the C_{llr} may not be so relevant. For instance, in cases without contamination, mixtures, dropout, etc., and with abundant samples, DNA LR values present high discriminating power in general. As the C_{llr} is an empirical measure, generating a sufficiently large number of LR values is necessary to obtain statistically stable C_{llr} values, and in cases with very high discriminating power, all C_{llr} values may be numerically zero or very close to zero. This reduces the usefulness of the metric. Another reason could be that, because the metric has historically been rarely used in the field, subsequent publications may also overlook it.

Although much attention in forensic DNA analysis is devoted to validation, there is little emphasis on specific metrics [160–163]. Usually well-known metrics are employed, e.g. ROC, AUC, false positive rates (given a certain LR threshold), or rates of misleading evidence [164–166]. We have not seen metrics specifically for calibration, even though the concept is certainly recognized in the field [138,167,168]. There thus may be a role to play for strictly proper scoring rules, which can comprehensively indicate the performance of an LR system.

5. Conclusions

The aim of this review was to gain insight into what C_{llr} values can be expected for state-of-the-art systems in forensic sciences, and to investigate the current application of the log-likelihood ratio cost (C_{llr}) in the evaluation of (semi-)automated likelihood ratio (LR) systems. While no claim is made that all existing relevant literature was found, considerable effort was put into getting an as complete picture as possible. We found that the use of the C_{llr} metric heavily depends on the field, with the metric being conspicuously absent in the forensic DNA analysis field. In addition, we found that the number of publications on (semi-)automated LR systems has increased over the years, but the use of the C_{llr} metric has remained more constant over time. The results do not show a clear pattern for what C_{llr} values can be expected and values vary a lot between forensic expertise areas, types of analyses, and datasets. Hence, we cannot establish a clear range of variation in general to assess the goodness of a system computing LR values. We set out to investigate if we could get a feeling of what a 'good' C_{llr} value is. We are not quite there yet. A path forward may be provided by benchmarks, i.e. datasets and evaluation protocols publicly available. Such benchmarks are standard practice in many fields of machine learning, and also used in several forensic fields [169], and allow for a much more direct comparison between systems. Given the trend that the number of LR systems developed and used increases yearly, having the ability to properly evaluate them will be increasingly important. Thus, setting up and curating such benchmark datasets, and establishing suitable evaluation protocols, is an investment in the future of the field.

Data availability

The data that support the findings of this study as well as the code used to analyze the data are available at https://github.com/NetherlandsForensicInstitute/cllr-review.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Stijn van Lierop: Writing – original draft, Investigation, Data curation. **Daniel Ramos:** Writing – review & editing. **Marjan Sjerps:** Writing – review & editing, Supervision. **Rolf Ypma:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgments

We would like to thank Corina Benschop for her help in understanding evaluation metrics used in forensic DNA analysis. Daniel Ramos is supported by grant from the UE and from the Spanish Ministerio de Ciencia e Innovación, Agencia y del Fondo Europeo de Desarrollo Regional, PID2021-125943OB-I00, MCIN/AEI/10.13039/501100011033/FEDER.

Appendix A. Literature Summary

Table A.3
Summary of all literature per forensic expertise area and type of analysis. If a C_{llr} is reported, information about the evaluation dataset is also given.

Ref.	Cllr (Cllr-min)	Dataset
Forensic Biometrics – Speaker Recognition		
[9]	0.379 (0.366)	Detection scores provided by three speaker verification teams.
[10]	0.25 (0.21)	NIST 2006 SRE (608 speakers) and NIST SRE 2005
[34]	0.068	Laboratory-quality audio dataset of 27 Australian male speakers by Yuko Kinoshita.
[36]	0.1845 (0.16)	NIST SRE Data, 2008 data for training and 2010 data for evaluation.
[35]	0.98 (0.43)	Shenyang China Male Speaker set, 64 male speakers of standard Chinese aged between 19 and 23.
[46]	0.042	60 female speakers from Northeastern China

(continued on next page)

Table A.3 (*continued*)

Ref.	Cllr (Cllr-min)	Dataset
[47]	0.3	GFS1 Corpus (German police male telephone tap recordings) containing 39 offender recordings, 21 suspect recordings and 49 reference population recordings.
[44]	0.009	297 male speakers selected to give geographical representative coverage of Japan.
[49]	0.375	NIST SRE Data, 2008 data for training and 2010 data for evaluation
[48]	0.496	Archive of chatlog messages of pedophiles (115 authors)
[14]	0.53 (0.47)	NIST HASR Dataset (150 comparisons)
[53]	0.447 (0.405)	4188 intercepted telephone speech by police in Dutch, Moroccan, Arabic, Berber and Turkish files with speech from 604 different speakers.
[62]	0.1092	background population: 50 speakers with similar background, reference population: 20 speakers with same background as speakers in background population.
[63]	0.29	26 young Australian female speakers, calculated 22 same speaker LRs and 231 different speaker LRs.
[66]	0.53	Telephone recordings of pairs of male speakers from a database of Australian English voice recordings designed for forensic casework.
[70]	0.179 (0.154)	MARP corpus containing 60 speakers (35 male, 25 female).
[71]	0.365	forensic_eval_01 data with 100 speakers in reference population.
[73]	0.423	Australian English Male of 231 Speakers designed for Forensic Research.
[79]	0.03	HKMTR database containing 23 speakers
[80]	0.36	136 Australian Female Speakers
[82]	0.449 (0.148)	forensic_eval_01 data, used 423 recordings from 105 speakers.
[84]	0.5	Corpus of spontaneous Japanese, 118 speakers (40 test, 39 background, 39 development), segments of around 10 min containing audio from presentations and mock conversations.
[86]	0.674 (0.658)	forensic_eval_01 data 105 speakers (train) and 61 speakers (test).
[90]	0.261	Questioned speaker telephone calls, known speaker police interviews, 111 SS and 9720 DS scores, 105 recordings training and 61 testing.
[93]	0.5192	forensic_eval_01 data, 423 recordings from 105 speakers for training, 223 recordings from 61 speakers for testing.
[94]	0.593 (0.198)	forensic_eval_01 data, 100 speakers in reference population.
[95]	0.83	DyVis Corpus, southern British male speakers, 97 SS and 4656 DS comparisons, 60 s mobile telephone calls.
[98]	0.208 (0.098)	forensic_eval_01 data
[100]	0.246 (0.189)	forensic_eval_01 data (223 recordings from 61 speakers).
[99]	0.255 (0.124)	forensic_eval_01 data
[102]	0.737	UCLA Speaker Variability Database, 103 female and 105 male speakers, 3 recordings per speaker.
[110]	Not reported	
[156]	Not reported	
Forensic Biometrics – Fingermark Comparison		
[170]	Not reported	
[28]	Not reported	
[45]	Not reported	
[50]	Not reported	
[52]	0.074 (0.063)	58 fingermarks real case data for evaluation, 25000 simulated fingermarks of 200 individuals for training.
[76]	0.165 (0.11)	Development set for training and validation casework fingerprint dataset for testing.
[77]	0.7	Biometric Collection of People 2008
[81]	0.45 (0.35)	NFI Casework dataset + Simulated dataset + Printcards from Dutch national database (HAVANK2).
Forensic Biometrics – Face Comparison		
[90]	0.104	XM2VTS database with 225 faces (8 images per face).
[97]	0.0935 (0.036)	XM2VTS and Face-3D Datasets
[115]	0.00177	National mugshot database, 51562 records, 9000 facial comparisons.
[117]	0.43	2011 ENFSI Face Proficiency Tests
	0.5	2017 ENFSI Face Proficiency Tests
	0.57	2013 ENFSI Face Proficiency Tests
	1.23	2012 ENFSI Face Proficiency Tests.
[141]	Not reported	
[146]	0.26	FaceScrub database
[144]	0.12	2011 ENFSI Face Proficiency Tests
	1.05	2012 ENFSI Face Proficiency Tests
	0.182	2013 ENFSI Face Proficiency Tests
	0.687	2017 ENFSI Face Proficiency Tests
	0.73	2018 ENFSI Face Proficiency Tests
	1.14	2019 ENFSI Face Proficiency Tests
	1.662	2020 ENFSI Face Proficiency Tests.
Forensic Biometrics – Stylometric Analysis		
[78]	1.29 (0.96)	NUS SMS Corpus (Singapore, India and US messages mostly).
[55]	0.54	Archive of chatlog messages of pedophiles (115 authors).
[83]	0.09 (0.05)	Archive of chatlog messages of pedophiles (115 authors).
[85]	0.683	Archive of chatlog messages of pedophiles (115 authors).
[116]	0.7	Amazon Product Data Authorship Verification Corpus.
[130]	0.66703	Amazon Product Data Authorship Verification Corpus.
[148]	0.5395 (0.51987)	Amazon Product Data Authorship Verification Corpus, 2157 authors, 3 document lengths (700, 1400 and 2100 words).
Forensic Biometrics – Handwriting Comparison		
[41]	Not reported	
[43]	Not reported	
[171]	Not reported	Created handwriting samples similar to casework, 7196 responses from 86 participants on 180 distinct QKsets.
[140]	Not reported	
[143]	Not reported	
Forensic Biometrics – Signature Comparison		
[37]	0.493 (0.238)	Dataset of Dutch online real or simulated (forged) signatures
	0.565 (0.351)	Dataset of Chinese online real or simulated (forged) signatures

(*continued on next page*)

Table A.3 (*continued*)

Ref.	Cllr (Cllr-min)	Dataset
	0.730 (0.573)	Dataset of Dutch offline real or simulated (forged) signatures
	0.758 (0.693)	Dataset of Chinese offline real or simulated (forged) signatures
[65]	0.655 (0.021)	Real life offline signatures from Italian university employees and students over a span of several years + forgeries, 628 total (250 training + 478 testing)
	0.689 (0.052)	240 real Bengali offline signatures from 10 contributors + forgeries from different parts of West-Bengal (India) (120 training and 420 testing)
	0.807 (0.466)	750 German online signatures (300 training + 450 testing) gathered by researchers with digital pen.
[89]	0.1629 (0.1433)	Chinese Signature Database with 1654 signatures.
Forensic Biometrics – Earmark Comparison		
[42]	Not reported	
Forensic Biometrics - Gait Analysis		
[39]	Not reported	
Forensic Biology – DNA Analysis		
[32]	Not reported	
[75]	Not reported	
[87]	Not reported	
[96]	Not reported	
[104]	Not reported	
[106]	Not reported	
[109]	Not reported	
[107]	Not reported	
[172]	Not reported	
[112]	Not reported	
[113]	Not reported	
[114]	Not reported	60 mixed DNA profiles of up to 5 contributors per mixture.
[153]	Not reported	
[125]	Not reported	
[127]	Not reported	
[128]	Not reported	
[138]	Not reported	
[173]	Not reported	
[150]	Not reported	
[151]	Not reported	
[152]	Not reported	
Forensic Biology - Body Fluid Identification		
[131]	0.1	Laboratory mixture data.
Materials and Microtraces – Glass Classification		
[33]	0.07 (0.01)	SEM-EDX Glass Database of 278 objects (79 containers and 199 float glass samples).
Materials and Microtraces – Glass Individualisation		
[31]	0.31 (0.26)	Database of UK and Polish window- and container glass SEM-EDX measurements.
[13]	0.72 (0.42)	Poland SEM-EDX glass dataset containing 165 glass objects (87 car windows and 78 building windows).
[60]	0.043 (0.026)	Raman micro spectroscopy database of solid car paint samples coming from 25 solid blue car paints, each sample measured at least three times.
[61]	0.2 (0.15)	SEM-EDX measurements from Institute of Forensic Technology Krakow.
[88]	0.0081 (0.0045)	Glass database of 979 glass particles from different sources of crime scene glass (known source) from about 10 years of casework at the NFI.
[90]	0.007	LA-ICP-MS glass sample data from over 10 years of casework, 659 sources for training and 320 sources for testing.
[174]	0.067 (0.017)	BKA glass database as background with FIU glass database for testing.
[101]	Not reported	
[139]	0.29 (0.079)	BKA glass database as background with FIU glass database for testing.
[145]	0.12 (0.1)	Swedish SEM-EDX Glass dataset containing 380 glass samples from Swedish casework measured in 5 replicates.
Materials and Microtraces – Ink Individualisation		
[29]	Not reported	
[68]	0.1833 (0.0899)	10 samples each of 40 different ink sources
Materials and Microtraces – Paint Individualisation		
[30]	Not reported	
Materials and Microtraces – Car Fuel Individualisation		
[51]	0.1 (0.01)	Repeated samples from 15 different petrol stations in The Hague.
Materials and Microtraces – Gasoline Individualisation		
[129]	0.028 (0.024)	GC-MS diesel oil dataset, 158 diesel oil samples from Swedish gas stations and oil refineries between 2015 and 2020
Materials and Microtraces – Car Paint Individualisation		
[59]	0.19 (0.1)	Raman micro spectroscopy database of metallic car paint samples from 30 metallic blue car paints, each sample measured at least three times.
[59]	0.07 (0.03)	Raman micro spectroscopy database of solid car paint samples from 25 solid blue car paints, each sample measured at least three times.
[60]	0.227 (0.147)	Raman micro spectroscopy database of metallic car paint samples from 30 metallic blue car paints, each sample measured at least three times.
[69]	0.0002 (0.0002) 0.0005 (0.0005) 0.223 (0.078)	Raman micro spectroscopy database of metallic car paint samples from 30 metallic blue car paints, each sample measured at least three times Raman micro spectroscopy database of solid car paint samples from 25 solid blue car paints, each sample measured at least three times Raman spectroscopy database of polypropelene spectra.
[68]	0.1756 (0.1533)	7 measurements each from 36 different acrylic car paint sources.
Forensic Ballistics – Cartridge Case Comparison		
[56]	Not reported	
Forensic Ballistics – Pattern Matching		
[121]	Not reported	2D measurements from 400 cartridge cases of 200 9 mm Glock pistols 3D measurements from 400 cartridge cases of 200 9 mm Glock pistols.

(*continued on next page*)

Table A.3 (*continued*)

Ref.	Cllr (Cllr-min)	Dataset
[123]	Not reported	
[124]	Not reported	
[122]	Not reported	
[108]	Not reported	
[175]	Not reported	
[176]	Not reported	
[177]	Not reported	
[147]	0.351	3D digital images of the bases of approximately 3000 fired cartidge cases.
Marks and Impressions – Shoemark Comparison		
[38]	Not reported	
Marks and Impressions – Tape Comparison		
[57]	Not reported	
[103]	0.11 (0.08)	LA-ICP-MS data for 90 electrical tapes
Marks and Impressions – Tape Comparison		
[178]	Not reported	
Forensic Drug Analysis – Drug Individualisation		
[40]	Not reported	
[58]	0.00006	GC-MS chemical impurity profiles of MDMA tablets collected by the Netherlands Forensic Institute (NFI) between 2004 and 2011 from 1372 batches, 20 pills each.
Forensic Drug Analysis – Activity Level Drug Analysis		
[142]	Not reported	
Toxicology – Alcohol Abuse Prediction		
[72]	0.033 (0.033)	Database of regional antidoping and toxicology center from 125 subjects (118 males and 7 females).
[120]	0.08	Database of regional antidoping and toxicology center containing 125 subjects (118 males and 7 females).
Food Forensics – Fraud		
[54]	0.4	178 wine samples from 3 brands of Italian wines
[64]	0.017 (0.017)	Olive oil dataset of 572 Italian olive oils described by fatty acid content.
Materials and Microtraces – Car Plastics Individualisation		
[91]	0.4 (0.3)	22 objects from 22 different vehicles/plastic containers 63 SS scores and 1890 DS scores.
Materials and Microtraces – Gunshot Residue Analysis		
[155]	Not reported	
[149]	0.34	1953 stubs from 210 criminal cases, 1557 same-shot pairs and 859.515 different-shot pairs.
Materials and Microtraces – Geolocation		
[92]	0.4 (0.25)	104 wolframite samples (5327 grains) from 45 different mines from 10 countries worldwide.
Explosions and Explosives – Explosive Individualisation		
[27]	Not reported	
[67]	Not reported	
[74]	Not reported	
Forensic Anthropology and Taphonomy – PMI Estimation		
[105]	Not reported	
Forensic Anthropology and Taphonomy – Stature Estimation		
[119]	Not reported	
Forensic Anthropology and Taphonomy – Sex Estimation		
[132]	0.3	Bartholdy et al. Dataset [179]
[133]	0.42 Not reported	FaceScrub database, 100.000+ images of more than 250 different males and 250 different females SCFace Databse.
[118]	Not reported	
Digital Forensics – Camera Identification		
[126]	Not reported	
Digital Forensics – Route Determination		
[157]	0.1 (0.08)	Set of distances measured using iPhone health apps using different iPhones and wearing pace and locations (1015 training set, 677 test set).
Digital Forensics – Mobile Phone Pairing		
[135]	0.47	Call records from 22 phone users (15 had two phones).
Digital Forensics – Camera Identification		
[134]	0.6 (0.57)	Vision Dataset, 34 different devices, around 7000 images, 223 non-stabilized videos and 190 stabilized videos.
[154]	Not reported	
Forensic Fire Analysis – Fire Debris Classification		
[136]	0.7 (0.58)	Fire debris GC-MS dataset, 492 ground truth IL and 299 ground truth SUB samples.
Forensic Odour Evidence – Hand Odour Identification		
[137]	Not reported	

References

[1] G.S. Morrison, Advancing a paradigm shift in evaluation of forensic evidence: the rise of forensic data science. https://doi.org/10.1016/j.fsisyn.2022.100270, 2022.

[2] J.J. Koehler, J.L. Mnookin, M.J. Saks, The scientific reinvention of forensic science. https://doi.org/10.1073/pnas.2301840120, Oct. 2023.

[3] T. Busey, M. Klutzke, Calibrating the perceived strength of evidence of forensic testimony statements. https://doi.org/10.1016/j.scijus.2022.10.003, Jan. 2023.

[4] J.P. Williams, D.M. Ommen, J. Hannig, Generalized fiducial factor: an alternative to the bayes factor for forensic identification of source problems. https://doi.org/10.1214/22-aoas1632, Mar. 2023.

[5] S.P. Lund, H. Iyer, Likelihood ratio as weight of forensic evidence: a closer look. https://doi.org/10.6028/jres.122.027, Oct. 2017.

[6] S. Willis, L. Kenna, S. Dermott, G. Donnell, A. Barrett, B. Rasmusson, A. Nordgaard, C. Berger, M. Sjerps, J. Molina, G. Zadora, C. Aitken, L. Lunt, C. Champod, A. Biedermann, T. Hicks, F. Taroni, Enfsi Guideline for Evaluative Reporting in Forensic Science, 2015.

[7] K. Martire, R. Kemp, M. Sayle, B. Newell, On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68, https://doi.org/10.1016/j.forsciint.2014.04.005.

[8] D. Ramos, D. Meuwly, R. Haraksim, C.E. Berger, Validation of forensic automatic likelihood ratio methods, in: Handbook of Forensic Statistics, CRC Press, 2020.

[9] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Comput. Speech Lang 20 (2–3) (2006) 230–275, https://doi.org/10.1016/j.csl.2005.08.001.
[10] 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, https://doi.org/10.1109/TASL.2007.902747.
[11] T. Fawcett, A. Niculescu-Mizil, Pav and the roc convex hull, Mach. Learn. 68 (1) (2007) 97–106, https://doi.org/10.1007/s10994-007-5011-0.
[12] A. Akmeemana, P. Weis, R. Corzo, D. Ramos, P. Zoon, T. Trejos, T. Ernst, C. Pollock, E. Bakowska, C. Neumann, J. Almirall, Interpretation of chemical data from glass analysis for forensic purposes, J. Chemometr. 35 (1) (Jun. 2020), https://doi.org/10.1002/cem.3267.
[13] D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, C. Aitken, Information-theoretical assessment of the performance of likelihood ratio computation methods, J. Forensic Sci. 58 (6) (2013) 1503–1518, https://doi.org/10.1111/1556-4029.12233.
[14] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, Forensic Sci. Int. 230 (1–3) (2013) 156–169, https://doi.org/10.1016/j.forsciint.2013.04.014.
[15] D. Ramos, J. Franco-Pedroso, A. Lozano-Diez, J. Gonzalez-Rodriguez, Deconstructing cross-entropy for probabilistic binary classifiers, Entropy 20 (3) (2018). https://www.mdpi.com/1099-4300/20/3/208.
[16] P.R. Geoffrey Stewart Morrison, C. Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, Aust. J. Forensic Sci. 44 (2) (2012) 155–167, https://doi.org/10.1080/00450618.2011.630412.
[17] J. Hannig, H. Iyer, Testing for calibration discrepancy of reported likelihood ratios in forensic science, J. Roy. Stat. Soc. Stat. Soc. 185 (1) (2021) 267–301, https://doi.org/10.1111/rssa.12747.
[18] P. Vergeer, Y. van Schaik, M. Sjerps, Measuring calibration of likelihood-ratio systems: a comparison of four metrics, including a new metric devPAV, Forensic Sci. Int. 321 (2021) 110722, https://doi.org/10.1016/j.forsciint.2021.110722.
[19] Microsoft Corporation, Microsoft excel. URL https://office.microsoft.com/excel.
[20] T. pandas development team, pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.3509134, Feb. 2020.
[21] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362, https://doi.org/10.1038/s41586-020-2649-2.
[22] J.D. Hunter, Matplotlib: a 2d graphics environment, Comput. Sci. Eng. 9 (3) (2007) 90–95, https://doi.org/10.1109/MCSE.2007.55.
[23] M.L. Waskom, seaborn: statistical data visualization, J. Open Source Softw. 6 (60) (2021) 3021, https://doi.org/10.21105/joss.03021.
[24] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272, https://doi.org/10.1038/s41592-019-0686-2.
[25] NetherlandsForensicInstitute, Netherlandsforensicinstitute/lir, Lir python likelihood ratio library, URL, https://github.com/NetherlandsForensicInstitute/lir, Apr 2023.
[26] G. Van Rossum, F.L. Drake, Python 3 reference manual, CreateSpace (2009). Scotts Valley, CA.
[27] G. Pierrini, S. Doyle, C. Champod, F. Taroni, D. Wakelin, C. Lock, Evaluation of preliminary isotopic analysis (13c and 15n) of explosives a likelihood ratio approach to assess the links between semtex samples, Forensic Sci. Int. 167 (1) (2007) 43–48, https://doi.org/10.1016/j.forsciint.2006.06.013.
[28] N.M. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—modelling within finger variability, Forensic Sci. Int. 167 (2–3) (2007) 189–195, https://doi.org/10.1016/j.forsciint.2006.06.054.
[29] C. Neumann, P. Margot, New perspectives in the use of ink evidence in forensic science, Forensic Sci. Int. 192 (1–3) (2009) 29–42, https://doi.org/10.1016/j.forsciint.2009.07.013.
[30] N. Farmer, W. Meier-Augenstein, D. Lucy, Stable isotope analysis of white paints and likelihood ratios, Sci. Justice 49 (2) (2009) 114–119, https://doi.org/10.1016/j.scijus.2009.02.003.
[31] G. Zadora, D. Ramos, Evaluation of glass samples for forensic purposes — an application of likelihood ratios and an information–theoretical approach, Chemometr. Intell. Lab. Syst. 102 (2) (2010) 63–83, https://doi.org/10.1016/j.chemolab.2010.03.007.
[32] K. Slooten, Validation of DNA-based identification software by computation of pedigree likelihood ratios, Forensic Sci. Int.: Genetics 5 (4) (2011) 308–315, https://doi.org/10.1016/j.fsigen.2010.06.005.
[33] D. Ramos, G. Zadora, Information-theoretical feature selection using data obtained by scanning electron microscopy coupled with and energy dispersive x-ray spectrometer for the classification of glass traces, Anal. Chim. Acta 705 (1–2) (2011) 207–217, https://doi.org/10.1016/j.aca.2011.05.029.
[34] G.S. Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM), Speech Commun. 53 (2) (2011) 242–256, https://doi.org/10.1016/j.specom.2010.09.005.
[35] G.S. Morrison, C. Zhang, P. Rose, An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system, Forensic Sci. Int. 208 (1–3) (2011) 59–65, https://doi.org/10.1016/j.forsciint.2010.11.001.
[36] D. Van Leeuwen, N. Brummer, A Speaker Line-Up for the Likelihood Ratio, 2011, pp. 501–504, https://doi.org/10.21437/Interspeech.2011-141.
[37] M. Liwicki, M.I. Malik, C.E. van den Heuvel, X. Chen, C. Berger, R. Stoel, M. Blumenstein, B. Found, Signature verification competition for online and offline skilled forgeries (SigComp2011), in: 2011 International Conference on Document Analysis and Recognition, IEEE, 2011, https://doi.org/10.1109/icdar.2011.294.
[38] J. Skerrett, C. Neumann, I. Mateos-Garcia, A bayesian approach for interpreting shoemark evidence in forensic casework: accounting for wear features, Forensic Sci. Int. 210 (1–3) (2011) 26–30, https://doi.org/10.1016/j.forsciint.2011.01.030.
[39] I. Bouchrika, M. Goffredo, J. Carter, M. Nixon, On using gait in forensic biometrics, J. Forensic Sci. 56 (4) (2011) 882–889, https://doi.org/10.1111/j.1556-4029.2011.01793.x.
[40] M. Lopatka, M. Vallat, Surface granularity as a discriminating feature of illicit tablets, Forensic Sci. Int. 210 (1–3) (2011) 188–194, https://doi.org/10.1016/j.forsciint.2011.03.008.
[41] 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, https://doi.org/10.1016/j.forsciint.2011.12.009.
[42] S. Junod, J. Pasquier, C. Champod, The development of an automatic recognition system for earmark and earprint comparisons, Forensic Sci. Int. 222 (1–3) (2012) 170–178, https://doi.org/10.1016/j.forsciint.2012.05.021.
[43] L.J. Davis, C.P. Saunders, A. Hepler, J. Buscaglia, Using subsampling to estimate the strength of handwriting evidence via score-based likelihood ratios, Forensic Sci. Int. 216 (1–3) (2012) 146–157, https://doi.org/10.1016/j.forsciint.2011.09.013.
[44] P. Rose, More is better: likelihood ratio-based forensic voice comparison with vocalic segmental cepstra frontfends, Int. J. Speech Lang. Law 20 (1) (2013) 77–116, https://doi.org/10.1558/ijsll.v20i1.77.
[45] I. Alberink, A. de Jongh, C. Rodriguez, Fingermark evidence evaluation based on automated fingerprint identification system matching scores: the effect of different types of conditioning on likelihood ratios, J. Forensic Sci. 59 (1) (2013) 70–81, https://doi.org/10.1111/1556-4029.12105.
[46] C.C. Huang, J. Epps, E. Enzinger, Separate MAP adaptation of GMM parameters for forensic voice comparison on limited data, in: 2013 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2013, https://doi.org/10.1109/wifs.2013.6707785.
[47] Y. Solewicz, G. Jardine, T. Becker, S. Gfroerer, Estimated Intra-speaker Variability Boundaries in Forensic Speaker Recognition Casework, 2013.
[48] S. Ishihara, A comparative study of likelihood ratio based forensic text comparison in procedures: multivariate kernel density vs. Gaussian mixture model-universal background model, in: Australasian Language Technology Association Workshop, 2013. URL, https://api.semanticscholar.org/CorpusID:5410772.
[49] D.A. van Leeuwen, N. Brümmer, The Distribution of Calibrated Likelihood-Ratios in Speaker Recognition, 2013 arXiv:1304.1199.
[50] N.E. Anthonioz, C. Champod, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems—modeling between finger variability, Forensic Sci. Int. 235 (2014) 86–101, https://doi.org/10.1016/j.forsciint.2013.12.003.
[51] P. Vergeer, A. Bolck, L. Peschier, C. Berger, J. Hendrikse, Likelihood ratio methods for forensic comparison of evaporated gasoline residues, Sci. Justice 54 (6) (2014) 401–411, https://doi.org/10.1016/j.scijus.2014.04.008.
[52] R. Haraksim, Validation of likelihood ratio methods used for forensic evidence evaluation, Ph.D. thesis, https://doi.org/10.3990/1.9789036536486, 2014.
[53] D. van der Vloed, J. Bouten, D. van Leeuwen, NFI-FRITS: a forensic speaker recognition database and some first experiments, in: The Speaker and Language Recognition Workshop (Odyssey 2014), ISCA, 2014, https://doi.org/10.21437/odyssey.2014-2.
[54] A. Martyna, G. Zadora, I. Stanimirova, D. Ramos, Wine authenticity verification as a forensic problem: an application of likelihood ratio test to label verification, Food Chem. 150 (2014) 287–295, https://doi.org/10.1016/j.foodchem.2013.10.111.
[55] S. Ishihara, A comparative study of likelihood ratio based forensic text comparison procedures: multivariate kernel density with lexical features vs. word n-grams vs. character n-grams, in: 2014 Fifth Cybercrime and Trustworthy Computing Conference, IEEE, 2014, https://doi.org/10.1109/ctc.2014.9.
[56] 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, https://doi.org/10.1111/1556-4029.12382.
[57] J.F. Carter, S. Doyle, B.-L. Phasumane, N. NicDaeid, The role of isotope ratio mass spectrometry as a tool for the comparison of physical evidence, Sci. Justice 54 (5) (2014) 327–334, https://doi.org/10.1016/j.scijus.2014.06.010.
[58] A. Bolck, H. Ni, M. Lopatka, Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison, Law, Probability and Risk 14 (3) (2015) 243–266, https://doi.org/10.1093/lpr/mgv009.
[59] A. Michalska, A. Martyna, J. Zikeba-Palus, G. Zadora, Application of a likelihood ratio approach in solving a comparison problem of Raman spectra recorded for

blue automotive paints, J. Raman Spectrosc. 46 (9) (2015) 772–783, https://doi.org/10.1002/jrs.4719.
[60] A. Martyna, A. Michalska, G. Zadora, Interpretation of FTIR spectra of polymers and Raman spectra of car paints by means of likelihood ratio approach supported by wavelet transform for reducing data dimensionality, Anal. Bioanal. Chem. 407 (12) (2015) 3357–3376, https://doi.org/10.1007/s00216-015-8558-9.
[61] G. Napier, A. Nobile, T. Neocleous, An online application for the classification and evidence evaluation of forensic glass fragments, Chemometr. Intell. Lab. Syst. 146 (2015) 418–425, https://doi.org/10.1016/j.chemolab.2015.06.013.
[62] H. Wang, C. Zhang, Forensic automatic speaker recognition based on likelihood ratio using acoustic-phonetic features measured automatically, J. Forensic Sci. Med. 1 (2) (2015) 119, https://doi.org/10.4103/2349-5014.169617.
[63] P. Rose, Forensic voice comparison with monophthongal formant trajectories - a likelihood ratio-based discrimination of vowel acoustics in a close social group of young australian females, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, https://doi.org/10.1109/icassp.2015.7178886.
[64] P. Wlasiuk, A. Martyna, G. Zadora, A likelihood ratio model for the determination of the geographical origin of olive oil, Anal. Chim. Acta 853 (2015) 187–199, https://doi.org/10.1016/j.aca.2014.10.022.
[65] M.I. Malik, S. Ahmed, A. Marcelli, U. Pal, M. Blumenstein, L. Alewijns, M. Liwicki, ICDAR2015 competition on signature verification and writer identification for on- and off-line skilled forgeries (SigWIcomp2015), in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, https://doi.org/10.1109/icdar.2015.7333948.
[66] E. Enzinger, G.S. Morrison, Mismatched distances from speakers to telephone in a forensic-voice-comparison case, Speech Commun. 70 (2015) 28–41, https://doi.org/10.1016/j.specom.2015.03.001.
[67] H. Brust, M. Koeberg, A. van der Heijden, W. Wiarda, I. Mügler, M. Schrader, G. Vivo-Truyols, P. Schoenmakers, A. van Asten, Isotopic and elemental profiling of ammonium nitrate in forensic explosives investigations, Forensic Sci. Int. 248 (2015) 101–112, https://doi.org/10.1016/j.forsciint.2014.11.024.
[68] J. Franco-Pedroso, D. Ramos, J. Gonzalez-Rodriguez, Gaussian mixture models of between-source variation for likelihood ratio computation from multivariate data, PLoS One 11 (2) (2016) e0149958, https://doi.org/10.1371/journal.pone.0149958.
[69] A. Martyna, G. Zadora, T. Neocleous, A. Michalska, N. Dean, Hybrid approach combining chemometrics and likelihood ratio framework for reporting the evidential value of spectra, Anal. Chim. Acta 931 (2016) 34–46, https://doi.org/10.1016/j.aca.2016.05.016.
[70] F. Kelly, J.H.L. Hansen, Score-aging calibration for speaker verification, IEEE/ACM Trans. Audio, Speech Language Proc. 24 (12) (2016) 2414–2424, https://doi.org/10.1109/taslp.2016.2602542.
[71] D. van der Vloed, Evaluation of batvox 4.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 85 (2016) 127–130, https://doi.org/10.1016/j.specom.2016.10.001.
[72] E. Alladio, A. Martyna, A. Salomone, V. Pirro, M. Vincenti, G. Zadora, Evaluation of direct and indirect ethanol biomarkers using a likelihood ratio approach to identify chronic alcohol abusers for forensic purposes, Forensic Sci. Int. 271 (2017) 13–22, https://doi.org/10.1016/j.forsciint.2016.12.019.
[73] E. Enzinger, G.S. Morrison, F. Ochoa, A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case, Sci. Justice 56 (1) (2016) 42–57, https://doi.org/10.1016/j.scijus.2015.06.005.
[74] D.-M.K. Dennis, M.R. Williams, M.E. Sigman, Assessing the evidentiary value of smokeless powder comparisons, Forensic Sci. Int. 259 (2016) 179–187, https://doi.org/10.1016/j.forsciint.2015.12.034.
[75] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int.: Genetics 21 (2016) 35–44, https://doi.org/10.1016/j.fsigen.2015.11.008.
[76] 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://doi.org/10.1016/j.dib.2016.11.008.
[77] S.S. Kamath, AFIS based likelihood ratios for latent fingerprint comparisons, Ph.D. thesis, https://doi.org/10.33915/etd.5927, 2017.
[78] S. Ishihara, A forensic authorship classification in sms messages: a likelihood ratio based approach using n-gram, in: Australasian Language Technology Association Workshop, 2011. URL, https://api.semanticscholar.org/CorpusID:792582.
[79] P. Rose, Likelihood ratio-based forensic voice comparison with higher level features: research and reality, Comput. Speech Lang 45 (2017) 475–502, https://doi.org/10.1016/j.csl.2017.03.003.
[80] E. Enzinger, G.S. Morrison, Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case, Forensic Sci. Int. 277 (2017) 30–40, https://doi.org/10.1016/j.forsciint.2017.05.007.
[81] A.J. Leegwater, D. Meuwly, M. Sjerps, P. Vergeer, I. Alberink, Performance study of a score-based likelihood ratio system for forensic fingermark comparison, J. Forensic Sci. 62 (3) (2017) 626–640, https://doi.org/10.1111/1556-4029.13339.
[82] D.G. da Silva, C.A. Medina, Evaluation of MSR identity toolbox under conditions reflecting those of a real forensic case (forensic_eval_01), Speech Commun. 94 (2017) 42–49, https://doi.org/10.1016/j.specom.2017.09.001.
[83] S. Ishihara, Strength of linguistic text evidence: a fused forensic text comparison system, Forensic Sci. Int. 278 (2017) 184–197, https://doi.org/10.1016/j.forsciint.2017.06.040.
[84] S. Ishihara, Sensitivity of likelihood-ratio based forensic voice comparison under mismatched conditions of within-speaker sample sizes across databases, Aust. J. Forensic Sci. 50 (4) (2017) 307–322, https://doi.org/10.1080/00450618.2016.1259351.
[85] S. Ishihara, Strength of forensic text comparison evidence from stylometric features: a multivariate likelihood ratio-based analysis, Int. J. Speech Lang. Law 24 (1) (2017) 67–98, https://doi.org/10.1558/ijsll.30305.
[86] G.S. Morrison, The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings, Forensic Sci. Int. 283 (2018) e1–e7, https://doi.org/10.1016/j.forsciint.2017.12.024.
[87] S. Manabe, C. Morimoto, Y. Hamano, S. Fujimoto, K. Tamaki, Development and validation of open-source software for DNA mixture interpretation based on a quantitative continuous model, PLoS One 12 (11) (2017) e0188183, https://doi.org/10.1371/journal.pone.0188183.
[88] A. van Es, W. Wiarda, M. Hordijk, I. Alberink, P. Vergeer, Implementation and assessment of a likelihood ratio approach for the evaluation of la-icp-ms evidence in forensic glass analysis, Sci. Justice 57 (3) (2017) 181–192, https://doi.org/10.1016/j.scijus.2017.03.002.
[89] X. hong Chen, C. Champod, X. Yang, S. pei Shi, Y. wen Luo, N. Wang, Y. chen Wang, Q. meng Lu, Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features, Forensic Sci. Int. 282 (2018) 101–110, https://doi.org/10.1016/j.forsciint.2017.11.022.
[90] G.S. Morrison, N. Poh, Avoiding overstating the strength of forensic evidence: shrunk likelihood ratios/bayes factors, Sci. Justice 58 (3) (2018) 200–218, https://doi.org/10.1016/j.scijus.2017.12.005.
[91] A. Martyna, G. Zadora, D. Ramos, Forensic comparison of pyrograms using score-based likelihood ratios, J. Anal. Appl. Pyrol. 133 (2018) 198–215, https://doi.org/10.1016/j.jaap.2018.03.024.
[92] A. Martyna, H.-E. Gäbler, A. Bahr, G. Zadora, Geochemical wolframite fingerprinting – the likelihood ratio approach for laser ablation ICP-MS data, Anal. Bioanal. Chem. 410 (13) (2018) 3073–3091, https://doi.org/10.1007/s00216-018-1007-9.
[93] J. Koschwitz, The Effect of Speaking Style on the Performance of a Forensic Voice Comparison System, Uppsala University, Department of Linguistics and Philology, 2018. Master's thesis.
[94] C. Zhang, C. Tang, Evaluation of batvox 3.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 100 (2018) 13–17, https://doi.org/10.1016/j.specom.2018.04.008.
[95] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, E.S.S. Fernández, The individual and the system: assessing the stability of the output of a semi-automatic forensic voice comparison system, in: Interspeech 2018, ISCA, 2018, https://doi.org/10.21437/interspeech.2018-1649.
[96] H. Swaminathan, M.O. Qureshi, C.M. Grgicak, K. Duffy, D.S. Lun, Four model variants within a continuous forensic DNA mixture interpretation framework: effects on evidential inference and reporting, PLoS One 13 (11) (2018) e0207599, https://doi.org/10.1371/journal.pone.0207599.
[97] N. Susyanto, R. Veldhuis, L. Spreeuwers, C. Klaassen, Semiparametric likelihood-ratio-based biometric score-level fusion via parametric copula, IET Biom. 8 (4) (2019) 277–283, https://doi.org/10.1049/iet-bmt.2018.5106.
[98] M. Jessen, J. Bortlík, P. Schwarz, Y.A. Solewicz, Evaluation of phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 111 (2019) 22–28, https://doi.org/10.1016/j.specom.2019.05.002.
[99] M. Jessen, G. Meir, Y.A. Solewicz, Evaluation of nuance forensics 9.2 and 11.1 under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 110 (2019) 101–107, https://doi.org/10.1016/j.specom.2019.04.006.
[100] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, A. Alexander, Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01), Speech Commun. 112 (2019) 30–36, https://doi.org/10.1016/j.specom.2019.06.005.
[101] S. Park, A. Carriquiry, Learning algorithms to evaluate forensic glass evidence, Ann. Appl. Stat. 13 (2) (Jun. 2019), https://doi.org/10.1214/18-aoas1211.
[102] S.J. Park, A. Afshan, J. Kreiman, G. Yeung, A. Alwan, Target and non-target speaker discrimination by humans and machines, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, https://doi.org/10.1109/icassp.2019.8683362.
[103] A. Gupta, C. Martinez-Lopez, J.M. Curran, J.R. Almirall, Multi-element comparisons of tapes evidence using dimensionality reduction for calculating likelihood ratios, Forensic Sci. Int. 301 (2019) 426–434, https://doi.org/10.1016/j.forsciint.2019.06.002.
[104] M. Boroń, A. B. and, 310/2020 ministry of science and higher education ranking 20 pt. in category of journals listed on the ranking published by ministry of science and higher education. IC journals masters list 2019 94, 04 pt. (index copernicus value) internal validation of LRmix studio biostatistical software, Iss. Forensic Sci. 306 (2019) 71–80, https://doi.org/10.34836/pk.2019.306.4.
[105] M.G. Andersson, A.-S. Ceciliason, H. Sandler, P. Mostad, Application of the bayesian framework for forensic interpretation to casework involving postmortem interval estimates of decomposed human remains, Forensic Sci. Int. 301 (2019) 402–414, https://doi.org/10.1016/j.forsciint.2019.05.050.
[106] C.C. Benschop, J. Hoogenboom, P. Hovers, M. Slagter, D. Kruise, R. Parag, K. Steensma, K. Slooten, J.H. Nagel, P. Dieltjes, V. van Marion, H. van Paassen, J. de Jong, C. Creeten, T. Sijen, A.L. Kneppers, DNAxs/DNAStatistX: development and validation of a software suite for the data management and probabilistic

interpretation of DNA profiles, Forensic Sci. Int.: Genetics 42 (2019) 81–89, https://doi.org/10.1016/j.fsigen.2019.06.015.
[107] C.C. Benschop, A. Nijveld, F.E. Duijs, T. Sijen, An assessment of the performance of the probabilistic genotyping software EuroForMix: trends in likelihood ratios and analysis of type i & II errors, Forensic Sci. Int.: Genetics 42 (2019) 31–38, https://doi.org/10.1016/j.fsigen.2019.06.005.
[108] D. Roberge, A. Beauchamp, S. Lévesque, Objective identification of bullets based on 3d pattern matching and line counting scores, Int. J. Pattern Recogn. Artif. Intell. 33 (11) (2019) 1940021, https://doi.org/10.1142/s0218001419400214.
[109] M. Xu, Q. Du, G. Ma, Z. Chen, Q. Liu, L. Fu, B. Cong, S. Li, Utility of ForenSeq™ DNA signature prep kit in the research of pairwise 2nd-degree kinship identification, Int. J. Leg. Med. 133 (6) (2019) 1641–1650, https://doi.org/10.1007/s00414-019-02003-6.
[110] G.S. Morrison, F. Kelly, A statistical procedure to adjust for time-interval mismatch in forensic voice comparison, Speech Commun. 112 (2019) 15–21, https://doi.org/10.1016/j.specom.2019.07.001.
[111] S. Noël, J. Noël, D. Granger, J.-F. Lefebvre, D. Séguin, STRmix™ put to the test: 300 000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates, Forensic Sci. Int.: Genetics 41 (2019) 24–31, https://doi.org/10.1016/j.fsigen.2019.03.017.
[112] J.S. Buckleton, J.-A. Bright, K. Cheng, H. Kelly, D.A. Taylor, The effect of varying the number of contributors in the prosecution and alternate propositions, Forensic Sci. Int.: Genetics 38 (2019) 225–231, https://doi.org/10.1016/j.fsigen.2018.11.011.
[113] Ø. Bleka, L. Prieto, P. Gill, CaseSolver: an investigative open source expert system based on EuroForMix, Forensic Sci. Int.: Genetics 41 (2019) 83–92, https://doi.org/10.1016/j.fsigen.2019.03.024.
[114] T. Bille, S. Weitz, J.S. Buckleton, J.-A. Bright, Interpreting a major component from a mixed DNA profile with an unknown number of minor contributors, Forensic Sci. Int.: Genetics 40 (2019) 150–159, https://doi.org/10.1016/j.fsigen.2019.02.017.
[115] A.L. Molder, I.E. Astrom, E. Leitet, Development of a score-to-likelihood ratio model for facial recognition using authentic criminalistic data, in: 2020 8th International Workshop on Biometrics and Forensics (IWBF), IEEE, 2020, https://doi.org/10.1109/iwbf49977.2020.9107954.
[116] S. Ishihara, The influence of background data size on the performance of a score-based likelihood ratio system: a case of forensic text comparison, in: Proceedings of the the 18th Annual Workshop of the Australasian Language Technology Association, Australasian Language Technology Association, Virtual Workshop, 2020, pp. 21–31. URL, https://aclanthology.org/2020.alta-1.3.
[117] A.M. Rodriguez, Z. Geradts, M. Worring, Likelihood ratios for deep neural networks in face comparison, J. Forensic Sci. 65 (4) (2020) 1169–1183, https://doi.org/10.1111/1556-4029.14324.
[118] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, R. Shrestha, T. Kanchan, Estimation of sex in forensic examinations using logistic regression and likelihood ratios, in: Forensic Science International: Reports 2, 2020 100118, https://doi.org/10.1016/j.fsir.2020.100118.
[119] R. Verma, K. Krishan, D. Rani, A. Kumar, V. Sharma, Stature estimation in forensic examinations using regression analysis: a likelihood ratio perspective, Forensic Sci. Int.: Report 2 (2020) 100069, https://doi.org/10.1016/j.fsir.2020.100069.
[120] G. Biosa, D. Giurghita, E. Alladio, M. Vincenti, T. Neocleous, Evaluation of forensic data using logistic regression-based classification methods and an r shiny implementation, Front. Chem. 8 (Oct. 2020), https://doi.org/10.3389/fchem.2020.00738.
[121] E.J. Mattijssen, C.L. Witteman, C.E. Berger, N.W. Brand, R.D. Stoel, Validity and reliability of forensic firearm examiners, Forensic Sci. Int. 307 (2020) 110112, https://doi.org/10.1016/j.forsciint.2019.110112.
[122] E.J.A.T. Mattijssen, C.L.M. Witteman, C.E.H. Berger, X.A. Zheng, J.A. Soons, R. D. Stoel, Firearm examination: examiner judgments and computer-based comparisons, J. Forensic Sci. 66 (1) (2020) 96–111, https://doi.org/10.1111/1556-4029.14557.
[123] F. Riva, E.J. Mattijssen, R. Hermsen, P. Pieper, W. Kerkhoff, C. Champod, Comparison and interpretation of impressed marks left by a firearm on cartridge cases – towards an operational implementation of a likelihood ratio based technique, Forensic Sci. Int. 313 (2020) 110363, https://doi.org/10.1016/j.forsciint.2020.110363.
[124] J. Song, Z. Chen, T.V. Vorburger, J.A. Soons, Evaluating likelihood ratio (LR) for firearm evidence identifications in forensic science based on the congruent matching cells (CMC) method, Forensic Sci. Int. 317 (2020) 110502, https://doi.org/10.1016/j.forsciint.2020.110502.
[125] C.C. Benschop, J. Hoogenboom, F. Bargeman, P. Hovers, M. Slagter, J. van der Linden, R. Parag, D. Kruise, K. Drobnic, G. Klucevsek, W. Parson, B. Berger, F. X. Laurent, M. Faivre, A. Ulus, P. Schneider, M. Bogus, A.L. Kneppers, T. Sijen, Multi-laboratory validation of DNAxs including the statistical library DNAStatistX, Forensic Sci. Int.: Genetics 49 (2020) 102390, https://doi.org/10.1016/j.fsigen.2020.102390.
[126] S. Reinders, L. Lin, W. Chen, Y. Guan, J. Newman, Score-based likelihood ratios for camera device identification, Electron. Imag. 32 (4) (2020), https://doi.org/10.2352/issn.2470-1173.2020.4.mwsf-215, 10.2352/issn.2470-1173.2020.4.mwsf-215.
[127] S. Starinsky-Elbaz, T. Ram, L. Voskoboinik, Z. Pasternak, Weight-of-evidence for DNA identification of missing persons and human remains using CODIS, Forensic Science, Med. Pathol. 16 (3) (2020) 389–394, https://doi.org/10.1007/s12024-020-00248-x.
[128] Ø. Bleka, R. Just, J. Le, P. Gill, An examination of STR nomenclatures, filters and models for MPS mixture interpretation, Forensic Sci. Int.: Genetics 48 (2020) 102319, https://doi.org/10.1016/j.fsigen.2020.102319.
[129] J. Malmborg, A. Nordgaard, Validation of a feature-based likelihood ratio method for the SAILR software. part i: gas chromatography–mass spectrometry data for comparison of diesel oil samples, Forensic Chem. 26 (2021) 100375, https://doi.org/10.1016/j.forc.2021.100375.
[130] S. Ishihara, Score-based likelihood ratios for linguistic text evidence with a bag-of-words model, Forensic Sci. Int. 327 (2021) 110980, https://doi.org/10.1016/j.forsciint.2021.110980.
[131] R. Ypma, P.M. van Wijk, R. Gill, M. Sjerps, M. van den Berge, Calculating LRs for presence of body fluids from mRNA assay data in mixtures, Forensic Sci. Int.: Genetics 52 (2021) 102455, https://doi.org/10.1016/j.fsigen.2020.102455.
[132] G.S. Morrison, P. Weber, N. Basu, R. Puch-Solis, P.S. Randolph-Quinney, Calculation of likelihood ratios for inference of biological sex from human skeletal remains, Forensic Sci. Int.: Synergy 3 (2021) 100202, https://doi.org/10.1016/j.fsisyn.2021.100202.
[133] R. Verma, N. Bhardwaj, P.D. Singh, A. Bhavsar, V. Sharma, Estimation of sex through morphometric landmark indices in facial images with strength of evidence in logistic regression analysis, Forensic Sci. Int.: Report 4 (2021) 100226, https://doi.org/10.1016/j.fsir.2021.100226.
[134] P. Ferrara, R. Haraksim, L. Beslay, Performance evaluation of source camera attribution by using likelihood ratio methods, J. Imag. 7 (7) (2021) 116, https://doi.org/10.3390/jimaging7070116.
[135] W. Bosma, S. Dalm, E. van Eijk, R. el Harchaoui, E. Rijgersberg, H.T. Tops, A. Veenstra, R. Ypma, Establishing phone-pair co-usage by comparing mobility patterns, Sci. Justice 60 (2) (2020) 180–190, https://doi.org/10.1016/j.scijus.2019.10.005.
[136] N.A. Thurn, T. Wood, M.R. Williams, M.E. Sigman, Classification of ground-truth fire debris samples using artificial neural networks, Forensic Chem. 23 (2021) 100313, https://doi.org/10.1016/j.forc.2021.100313.
[137] I. Rivals, C. Sautier, G. Cognon, V. Cuzuel, Evaluation of distance-based approaches for forensic comparison: application to hand odor evidence, J. Forensic Sci. 66 (6) (2021) 2208–2217, https://doi.org/10.1111/1556-4029.14818.
[138] S. Riman, H. Iyer, P.M. Vallone, Examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset, PLoS One 16 (9) (2021) e0256714, https://doi.org/10.1371/journal.pone.0256714.
[139] D. Ramos, J. Maroñas, J. Almirall, Improving calibration of forensic glass comparisons by considering uncertainty in feature-based elemental data, Chemometr. Intell. Lab. Syst. 217 (2021) 104399, https://doi.org/10.1016/j.chemolab.2021.104399.
[140] V. Sharma, M. Bains, R. Verma, N. Verma, R. Kumar, Novel use of logistic regression and likelihood ratios for the estimation of gender of the writer from a database of handwriting features, Aust. J. Forensic Sci. 55 (1) (2021) 89–106, https://doi.org/10.1080/00450618.2021.1956587.
[141] W. Xiao, Forensic face recognition based on KDE and evidence theory, MATEC Web of Conf. 336 (2021) 06008, https://doi.org/10.1051/matecconf/202133606008.
[142] M. Russell, G. Wevers, B. Bogun, E. Mayo, M. McKinnel, J. Watson, Establishing likelihood ratios for evaluating opposing propositions concerning the activity causing methamphetamine contamination: smoking or manufacture? Forensic Sci. Int. 326 (2021) 110939 https://doi.org/10.1016/j.forsciint.2021.110939.
[143] M.Q. Johnson, D.M. Ommen, Handwriting identification using random forests and score-based likelihood ratios, Stat. Anal. Data Min.: The ASA Data Sci. J. 15 (3) (2021) 357–375, https://doi.org/10.1002/sam.11566.
[144] A.M. Rodriguez, Z. Geradts, M. Worring, Calibration of score based likelihood ratio estimation in automated forensic facial image comparison, Forensic Sci. Int. 334 (2022) 111239, https://doi.org/10.1016/j.forsciint.2022.111239.
[145] J. Malmborg, A. Nordgaard, Validation of a feature-based likelihood ratio method for the SAILR software. part II: elemental compositional data for comparison of glass samples, Forensic Chem. 27 (2022) 100385, https://doi.org/10.1016/j.forc.2021.100385.
[146] R. Verma, N. Bhardwaj, A. Bhavsar, K. Krishan, Towards facial recognition using likelihood ratio approach to facial landmark indices from images, Forensic Sci. Int.: Report 5 (2022) 100254, https://doi.org/10.1016/j.fsir.2021.100254.
[147] N. Basu, R.S. Bolton-King, G.S. Morrison, Forensic comparison of fired cartridge cases: feature-extraction methods for feature-based calculation of likelihood ratios, Forensic Sci. Int.: Synergy 5 (2022) 100272, https://doi.org/10.1016/j.fsisyn.2022.100272.
[148] S. Ishihara, M. Carne, Likelihood ratio estimation for authorship text evidence: an empirical comparison of score- and feature-based methods, Forensic Sci. Int. 334 (2022) 111268, https://doi.org/10.1016/j.forsciint.2022.111268.
[149] T. Matzen, C. Kukurin, J. van de Wetering, S. Ariëns, W. Bosma, A. Knijnenberg, A. Stamouli, R.J. Ypma, Objectifying evidence evaluation for gunshot residue comparisons using machine learning on criminal case data, Forensic Sci. Int. 335 (2022) 111293, https://doi.org/10.1016/j.forsciint.2022.111293.
[150] C.C. Benschop, M. Slagter, J.H. Nagel, P. Hovers, S. Tuinman, F.E. Duijs, L.J. Grol, M. Jegers, A. Berghout, A.-W. van der Zwan, R.J. Ypma, J. de Jong, A.L. Kneppers, Development and validation of a fast and automated DNA identification line, Forensic Sci. Int.: Genetics 60 (2022) 102738, https://doi.org/10.1016/j.fsigen.2022.102738.
[151] S. Manabe, T. Fukagawa, K. Fujii, N. Mizuno, K. Sekiguchi, A. Akane, K. Tamaki, Development and validation of kongoh ver. 3.0.1: open-source software for dna mixture interpretation in the globalfiler system based on a quantitative

continuous model, Leg. Med. 54 (2022) 101972. https://www.sciencedirect.com/science/article/pii/S134462232100136X.
[152] M.S. Adamowicz, T.N. Rambo, J.L. Clarke, Internal validation of MaSTR™ probabilistic genotyping software for the interpretation of 2–5 person mixed DNA profiles, Genes 13 (8) (2022) 1429, https://doi.org/10.3390/genes13081429.
[153] C. Costa, C. Figueiredo, A. Amorim, S. Costa, P.M. Ferreira, N. Pinto, Quantification of forensic genetic evidence: comparison of results obtained by qualitative and quantitative software for real casework samples, Forensic Sci. Int.: Genetics 59 (2022) 102715, https://doi.org/10.1016/j.fsigen.2022.102715.
[154] S. Reinders, Y. Guan, D. Ommen, J. Newman, Source-anchored, trace-anchored, and general match score-based likelihood ratios for camera device identification, J. Forensic Sci. 67 (3) (2022) 975–988, https://doi.org/10.1111/1556-4029.14991.
[155] K. Menking-Hoggatt, C. Ott, C.V. Pyl, K. Dalzell, J. Curran, L. Arroyo, T. Trejos, Prevalence and probabilistic assessment of organic and inorganic gunshot residue and background profiles using LIBS, electrochemistry, and SEM-EDS, Forensic Chem. 29 (2022) 100429, https://doi.org/10.1016/j.forc.2022.100429.
[156] M. Niu, L. He, Z. Fang, B. Zhao, K. Wang, Pseudo-phoneme label loss for text-independent speaker verification, Appl. Sci. 12 (15) (2022) 7463, https://doi.org/10.3390/app12157463.
[157] M.M. Vink, M.M. Sjerps, A.A. Boztas, J.J.P. van Zandwijk, Likelihood ratio method for the interpretation of iPhone health app data in digital forensics, Forensic Sci. Int.: Digit. Invest. 41 (2022) 301389, https://doi.org/10.1016/j.fsidi.2022.301389.
[158] Ø. Bleka, G. Storvik, P. Gill, EuroForMix: an open source software based on a continuous model to evaluate STR DNA profiles from a mixture of contributors with artefacts, Forensic Sci. Int.: Genetics 21 (2016) 35–44, https://doi.org/10.1016/j.fsigen.2015.11.008.
[159] C.C. Benschop, J. Hoogenboom, P. Hovers, M. Slagter, D. Kruise, R. Parag, K. Steensma, K. Slooten, J.H. Nagel, P. Dieltjes, V. van Marion, H. van Paassen, J. de Jong, C. Creeten, T. Sijen, A.L. Kneppers, DNAxs/DNAStatistX: development and validation of a software suite for the data management and probabilistic interpretation of DNA profiles, Forensic Sci. Int.: Genetics 42 (2019) 81–89, https://doi.org/10.1016/j.fsigen.2019.06.015.
[160] Scientific Working Group on DNA Analysis Methods, Guidelines for the Validation of Probabilistic Genotyping Software, 2015.
[161] ENFSI, Best Practice Manual for the Internal Validation of Probabilistic Software to Undertake DNA Mixture Interpretation, 2017.
[162] J. Buckleton Coble, J.M. Butler, T. Egeland, R. Fimmers, P. Gill, L. Gusmão, B. Guttman, M. Krawczak, N. Morling, W. Parson, N. Pinto, P.M. Schneider, S. T. Sherry, S. Willuweit, M. Prinz, DNA commission of the international society for forensic genetics: recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications, Forensic Sci. Int. Genet. 25 (2016) 191–197.
[163] Forensic Science Regulator, Software Validation for DNA Mixture Interpretation, 2018.
[164] Ø. Bleka, C.C.G. Benschop, G. Storvik, P. Gill, A comparative study of qualitative and quantitative models used to interpret complex STR DNA profiles, Forensic Sci. Int. Genet. 25 (2016) 85–96.
[165] P. Gill, C. Benschop, J. Buckleton, Ø. Bleka, D. Taylor, A review of probabilistic genotyping systems: EuroForMix, DNAStatistX and STRmix™, Genes 12 (10) (Sep. 2021).
[166] M. Susik, I.F. Sbalzarini, Analysis of the Hamiltonian Monte Carlo genotyping algorithm on PROVEDIt mixtures including a novel precision benchmark, Forensic Sci. Int. Genet. 64 (2023) 102840.
[167] J.-A. Bright, M. Jones Dukes, S.N. Pugh, I.W. Evett, J.S. Buckleton, Applying calibration to LRs produced by a DNA interpretation software, Aust. J. Forensic Sci. 53 (2) (2021) 147–153.
[168] J. Buckleton, J.-A. Bright, D. Taylor, R. Wivell, Ø. Bleka, P. Gill, C. Benschop, B. Budowle, M. Coble, Re: Riman et al. examining performance and likelihood ratios for two likelihood ratio systems using the PROVEDIt dataset, Forensic Sci. Int.: Genetics 59 (2022) 102709, https://doi.org/10.1016/j.fsigen.2022.102709.
[169] D.A. van Leeuwen, A.F. Martin, M.A. Przybocki, J.S. Bouten, Nist and NFI-TNO evaluations of automatic speaker recognition, Comput. Speech Lang 20 (2) (2006) 128–158, odyssey 2004: The speaker and Language Recognition Workshop, https://www.sciencedirect.com/science/article/pii/S088523080500032X.
[170] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (1) (2006) 54–64, https://doi.org/10.1111/j.1556-4029.2006.00327.x.
[171] R.A. Hicklin, L. Eisenhart, N. Richetelli, M.D. Miller, P. Belcastro, T.M. Burkes, C. L. Parks, M.A. Smith, J. Buscaglia, E.M. Peters, R.S. Perlman, J.V. Abonamah, B. A. Eckenrode, Accuracy and reliability of forensic handwriting comparisons, Proc. Natl. Acad. Sci. USA 119 (32) (Aug. 2022), https://doi.org/10.1073/pnas.2119944119.
[172] S. Noël, J. Noël, D. Granger, J.-F. Lefebvre, D. Séguin, STRmix™ put to the test: 300 000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates, Forensic Sci. Int.: Genetics 41 (2019) 24–31, https://doi.org/10.1016/j.fsigen.2019.03.017.
[173] J. Alfieri, M.D. Coble, C. Conroy, A. Dahl, D.R. Hares, B.S. Weir, C. Wolock, E. Zhao, H. Kingston, T.W. Zolandz, A new implementation of a semi-continuous method for DNA mixture interpretation, Forensic Sci. Int.: Report 6 (2022) 100281, https://doi.org/10.1016/j.fsir.2022.100281.
[174] R. Corzo, T. Hoffman, P. Weis, J. Franco-Pedroso, D. Ramos, J. Almirall, The use of LA-ICP-MS databases to calculate likelihood ratios for the forensic analysis of glass evidence, Talanta 186 (2018) 655–661, https://doi.org/10.1016/j.talanta.2018.02.027.
[175] K.L. Monson, E.D. Smith, E.M. Peters, Accuracy of comparison decisions by forensic firearms examiners, J. Forensic Sci. 68 (1) (2022) 86–100, https://doi.org/10.1111/1556-4029.15152.
[176] M. Neuman, C. Hundl, A. Grimaldi, D. Eudaley, D. Stein, P. Stout, Blind testing in firearms: preliminary results from a blind quality control program, J. Forensic Sci. 67 (3) (2022) 964–974, https://doi.org/10.1111/1556-4029.15031.
[177] M. Cuellar, C. Gonzalez, I.E. Dror, Human and machine similarity judgments in forensic firearm comparisons, Forensic Sci. Int.: Synergy 5 (2022) 100283, https://doi.org/10.1016/j.fsisyn.2022.100283.
[178] M. Prusinowski, E. Brooks, T. Trejos, Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits, Forensic Sci. Int. 307 (2020) 110103, https://doi.org/10.1016/j.forsciint.2019.110103.
[179] B.P. Bartholdy, E. Sandoval, M.L.P. Hoogland, S.A. Schrader, Getting rid of dichotomous sex estimations: why logistic regression should be preferred over discriminant function analysis, J. Forensic Sci. 65 (5) (2020) 1685–1691, https://doi.org/10.1111/1556-4029.14482.