



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2023

FORENSIC OFFLINE SIGNATURE HANDWRITING EXAMINATION BASED ON THREE-DIMENSIONAL AND PSEUDO-DYNAMIC FEATURES

Chen Xiaohong

Chen Xiaohong, 2023, FORENSIC OFFLINE SIGNATURE HANDWRITING EXAMINATION
BASED ON THREE-DIMENSIONAL AND PSEUDO-DYNAMIC FEATURES

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_77FA00CC86135

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.

FACULTE DE DROIT, DES SCIENCES CRIMINELLES ET
D'ADMINISTRATION PUBLIQUE
ECOLE DES SCIENCES CRIMINELLES

FORENSIC OFFLINE SIGNATURE HANDWRITING
EXAMINATION BASED ON THREE-DIMENSIONAL
AND PSEUDO-DYNAMIC FEATURES

THESE DE DOCTORAT

présentée à la Faculté de Droit, des Sciences Criminelles et
d'Administration Publique de l'Université de Lausanne pour
l'obtention du grade de Docteur en science forensique par

CHEN XIAOHONG

Directeur de thèse : Prof. Christophe Champod

Lausanne

2023



UNIL | Université de Lausanne

FACULTE DE DROIT, DES SCIENCES CRIMINELLES ET
D'ADMINISTRATION PUBLIQUE
ECOLE DES SCIENCES CRIMINELLES

FORENSIC OFFLINE SIGNATURE
HANDWRITING EXAMINATION BASED ON
THREE-DIMENSIONAL AND PSEUDO-
DYNAMIC FEATURES

THÈSE DE DOCTORAT

présentée à la

Faculté de droit, des sciences criminelles et d'administration publique
de l'Université de Lausanne

pour l'obtention du grade de

Docteur en sciences en science forensique

par

CHEN Xiaohong

Directeur de thèse

Prof. Christophe Champod

Jury

Prof. Thomas Souvignet, Président du jury

Prof. Charles Berger, expert externe (Leiden University & Netherlands Forensic Institute)

Dr. Linton A. Mohammed, expert externe (Forensic Science Consultants, Poway, USA)

Prof. Silvia Bozza, expert externe (Università Ca' Foscari Venezia, Italy)

LAUSANNE

2023

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Madame Chen Xiaohong, candidate au doctorat en science forensique, intitulée

Forensic Offline Signature Handwriting Examination Based on Three-dimensional and Pseudo-dynamic Features



Professeur Thomas Souvignet
Président du Jury

Lausanne, le 28 octobre 2022

Abstract

The scientific foundation of forensic handwriting examination has been a subject of debate and controversy for many years. This is mainly attributable to the subjective approach adopted by experts who do not take advantage of modern computer-based techniques to support their decision-making. One reason for their reluctance is that computer-based systems focused on two-dimensional and static features remain vulnerable to skilled forgeries. This research aimed to overcome this limitation by enriching the features used by an offline recognition system. In addition to two-dimensional features, pseudo-dynamic three-dimensional information was included, measured using three-dimensional microscopy synchronized with white-light imaging. Three-dimensional and pseudo-dynamic features were acquired from a dataset comprising 23,624 signatures. The sample comprised genuine signatures and a range of forgeries (freehand, random, or traced). For forensic purposes, score-based likelihood ratios were adopted to estimate the strength of handwriting evidence. The rates of misleading evidence were low enough (RMEP = 0, RMED = 0.0002) to warrant implementation in forensic laboratories. In addition, the testing of machine learning techniques demonstrated the feasibility and robustness of the system for commercial application and consideration of feature selection. For the most accurate classification algorithm, a precision of 99.91% is reported for an equivalent recall rate.

Résumé

Les fondements scientifiques de l'examen forensique de l'écriture manuscrite font l'objet de débats et de controverses depuis de nombreuses années. Cela est principalement dû à l'approche subjective adoptée par les experts qui ne tirent pas parti des techniques informatiques modernes pour assister leur prise de décision. L'une des raisons de leur réticence est que les systèmes informatiques axés sur des caractéristiques bidimensionnelles et statiques restent vulnérables aux contrefaçons. Cette recherche vise à surmonter cette limitation en enrichissant les caractéristiques utilisées par un système de reconnaissance. Outre les caractéristiques bidimensionnelles, des informations tridimensionnelles pseudo-dynamiques ont été incluses, mesurées à l'aide d'une microscopie tridimensionnelle synchronisée avec l'imagerie obtenue en lumière blanche. Les caractéristiques tridimensionnelles et pseudo-dynamiques ont été acquises à partir d'un ensemble de données comprenant 23'624 signatures. L'échantillon comprenait des signatures authentiques et une série de contrefaçons (à main levée, sans modèle ou tracées). La mesure de rapports de vraisemblance basés sur des scores a été adopté pour estimer la force associées aux comparaisons. Les taux d'orientation fallacieuses sont suffisamment faibles (RMEP = 0, RMED = 0,0002) pour justifier une mise en œuvre dans les laboratoires de police scientifique. En outre, un test des techniques d'apprentissage automatique a démontré la faisabilité et la robustesse du système pour une application commerciale et la prise en compte de la sélection des caractéristiques. Pour l'algorithme de classification le plus précis, une précision de 99,91 % est rapportée pour un taux de rappel équivalent.

Acknowledgements

One day, 19 years ago, my master's adviser, Professor Yu-wen Jia¹, called me to his office and asked me to conduct research on handwriting quantification because he believed handwriting quantification was the only way to break the bottleneck in the development of handwriting examination. "Newborn calves are not afraid of tigers". With no background in computer science and statistics, I taught myself computer programming and statistical analysis, and used online signatures as the research object to undertake research completely differently from traditional handwriting identification. Three years later, my graduation thesis, entitled "A Preliminary Study of the Quantification of the Dynamic Features of Signature Handwriting", was awarded for an "Excellent Master's Degree Thesis" by the Liaoning Province Government of China. Later, I was honoured by the "Young Scientist Award" from the International Association of Forensic Science (IAFS). However, after having made progress, I encountered a development bottleneck in my research path. To improve my abilities, I considered studying for a PhD. I thank Professor Shen Min, who was the leader of my institute at the time, who told me, 'What you need is an improvement of ability, not just a doctorate degree'. Therefore, I did not go to study for a doctorate right away at that time.

Five years ago, I met my PhD adviser, Professor Christophe Champod. During my doctoral research stage, he taught me research skills and spared no effort in helping me gain more knowledge, making my research more convenient, and encouraging me to improve my abilities in every possible aspect. When I was eager for success, he would tell me to slow down and carefully study whether the details were correct at each step. When I encountered difficulties, he worked with me to find a solution. He is both a teacher and a friend. I thank him for discovering my potential and stimulating my creativity so I could successfully complete this research project. Professor Christophe Champod is knowledgeable, rigorous, sincere, and frank. Under his guidance, I felt an infinite desire for knowledge. With his help, I experienced another stage of growth. Professor Yu-wen Jia and Professor Champod are two great forensic scientists and educators who take the development of forensic science as their mission and strive to contribute to it. I am lucky to have had them as advisers; indeed, they have become my role models. I would like also to thank my thesis committee: Professors Silvia Bozza, Charles Berger and Mohammed Linton for their valuable input and feedback on the manuscript.

Nineteen years ago, I made a difficult choice. Since then, after years of perseverance, the door to scientific exploration has been opened for me. What I want now is to persist in doing this difficult but correct thing.

Finally, I also want to thank Kevin for his support in allowing me to pursue my dreams without distraction. He is also willing to listen to my ideas in scientific research and has given me a lot of inspiration and help from his unique perspective. I also thank my two lovely sons, Jason and Ethan, who allow me to enjoy the happiness of family outside of work.

¹ Yu-wen Jia, 1936–2016, chief professor of China Criminal Police University, founder of questioned document examination of China.

To Prof. Yu-wen Jia in heaven

To my family

Table of Contents

Table of Contents.....	1
Glossary.....	3
Chapter 1 Introduction.....	1
Chapter 2 State of the Art.....	5
2.1 Forensic handwriting examination.....	5
2.2 Offline handwriting verification.....	4
2.3 Signature databases.....	7
2.4 Quantitative measurement of handwriting features.....	15
2.5 Assessment of handwriting verification systems.....	17
2.6 Limitations of current research.....	18
2.6.1 Signature databases are not comprehensive enough.....	19
2.6.2 Three-dimensional and pseudo-dynamic attributes of handwriting are ignored.....	19
2.6.3 Flaws in the method using 2D static handwriting features.....	19
2.7. Three-dimensional research in forensic science.....	20
Chapter 3 Methods and Materials.....	23
3.1 Chinese signatures datasets.....	23
3.1.1 Signatures from volunteers.....	23
3.1.2 Signature dataset from proficiency tests.....	26
3.1.3 Real forensic cases.....	28
3.2 Acquisition and reconstruction of 3D images of signatures.....	28
3.2.1 Reflection holographic microscopes: Lyncee Tec R2200.....	29
3.2.2 Wide-area 3D measurement system: Keyence VR 3200.....	31
3.2.3 Two-dimensional white-light images and 3D image acquisition.....	31
3.3 Extraction and post-processing of three-dimensional and pseudo-dynamic features.....	32
3.3.1 Writing sequence tracing.....	32
3.3.2 Features extraction in white-light and 3D signature images.....	35
3.3.3 Pseudo-dynamic features visualization.....	37
3.3.4 Post-processing of three-dimensional and pseudo-dynamic features.....	43
3.4 Statistical description of features.....	44
3.4.1 Multivariate analysis of variables and discriminant analysis.....	45
3.4.2 Descriptive and Comparative measurement.....	46
3.5 ML method for signature verification.....	48
3.6 Estimation of the strength of the signature evidence.....	50
3.6.1 Probability density distribution.....	50
3.6.2 Score-based LR calculation.....	57
3.6.3 Performance evaluation.....	65
3.7 Validation tests.....	68

3.7.1 Competition test.....	68
3.7.2 CNAS proficiency test (PT).....	68
3.7.3 Real forensic cases test.....	69
Chapter 4 Results.....	70
4.1 Between- and within-writer variations.....	70
4.1.1 Statistical description and descriptive analysis.....	70
4.1.2 Probability density distribution.....	73
4.2 Machine Learning (ML).....	77
4.2.1 <i>K</i> -nearest neighbour.....	77
4.2.2 Discriminant analysis.....	78
4.2.3 Naive Bayes (<i>NB</i>).....	80
4.2.4 Tree-based models.....	80
4.2.5 Random Forest (<i>R-Forest</i>).....	85
4.2.6 Support vector machines (SVM) with radial basis function kernel.....	86
4.2.7 Neural networks.....	86
4.2.8 Comparing models.....	91
4.2.9 Application to test data.....	94
4.2.10 Variable importance.....	98
4.3 Performance evaluation of score-based LR system.....	100
4.3.1 PAVA calibration of score-based LR using MKDE based on dataset_3.....	101
4.3.2 PAVA calibration of score-based LR using DST based on dataset_3.....	108
4.3.3 Comparison of performance between MKDE and DST.....	118
4.4 Validation tests.....	123
4.4.1 Proficiency tests (PT).....	123
4.4.2 Test of real forensic cases.....	124
4.4.3 Impact of different writing conditions on signatures.....	126
Chapter 5 Discussion and Future Perspectives.....	130
5.1 Scientific basis for handwriting comparison and assessment in this research.....	130
5.2 An effort to change the operative model of handwriting examination.....	132
5.3 Adaptation of the claim of uniqueness of handwriting.....	133
5.4 Technical contributions of this research.....	134
5.5 Open-ended questions for an optimal DTW algorithm.....	136
5.5 Future perspectives.....	146
Chapter 6: Conclusion.....	149
Achievements.....	152
Publications.....	152
Journals.....	152
Book.....	152
Patents.....	152
Support from foundations and research institutions.....	153
References.....	154
Appendix.....	173

Glossary

C_{llr}	Calibrated log likelihood ratio
C_{llr}^{min}	C_{llr}^{min} represents the minimum possible value of C_{llr} that can be achieved by the optimally calibrated system. The smaller the C_{llr}^{min} , the better the system.
2D	Two-dimensional
3D	Three-dimensional
AER	Average error rate
APE	Applied probability of error
AvNN	Averaged neural network
CNAS	China National Accreditation Service for Conformity Assessment
DET	Detection error trade-off
DHM	Digital holographic microscope
DST	Dempster–Shafer theory
DTW	Dynamic time warping
ECE	Empirical cross-entropy
EER	Equal error rate
FAR	False accept rate
FF	Freehand forgery (for a signature)
FHE	Forensic handwriting examiner
FRR	False reject rate
GBM	Gradient boosting machine
GE	Genuine (for a signature)
KDE	Kernel density estimation
KNN	K-nearest neighbour
KST	Kolmogorov–Smirnov Test
LDA	Linear discriminant analysis
LR (LRs)	Likelihood ratio (Likelihood ratios)
LLR (LLRs)	Log10 likelihood ratio (Log10 likelihood ratios)
MANOVA	Multivariate analysis of variance
MDA	Mixture discriminant analysis
MKDE	Multivariate kernel density estimation
ML	Machine learning
MVKD	Multivariate kernel density procedure
MVN	Multivariate normal
NB	Naïve Bayes
NN	Neural networks
nnet	Neural net
PAVA	Pool-adjacent violators algorithm
PCA	Principal component analysis
pcaNNet	Neural networks with PCA feature extraction
PDE	Probability density estimation
PR	Pattern recognition
PT	Proficiency test
QDA	Quadratic discriminant analysis
RDA	Regularized discriminant analysis
RF	Random forgery
R-Forest	Random forest
RMED	Rate of misleading evidence in favour of H_d
RMEP	Rate of misleading evidence in favour of H_p
SigCom2011	Signature Verification Competition for Online and Offline Skilled Forgeries
SVM	Support vector machine
TF	Traced forgery
WD	Writer dependent
WI	Writer independent
XGB_linear	eXtreme gradient boosting based on linear model

Chapter 1 Introduction

Handwriting identification is an important task in both forensic science and computer science. Today, forensic expertise still lies with forensic handwriting examiners (FHEs) who apply their observational skills and make decisions about authorship based on their training and experience. Meanwhile, computer-based handwriting verification systems have yet to successfully make their way into forensic processes and are only used for civil biometric applications. Currently, computer-based recognition techniques and FHEs operate in silos, as if their recognition objectives were distinct. First, the two methods for measuring features are different. The computer community uses image-space domain or frequency domain measurements; FHEs visually and holistically identify features through observation and experience. Second, the features used by both communities are not the same, and their tools and methods are different. The computer community relies on systematic measures and statistical methods, while FHEs rely more on training and empirical judgment. In addition, the presentation of the output of their tasks is different; the computer science community generally gives a binary outcome (recognized or not), while FHEs need to convey their forensic results on a scale of evidential strength. Finally, the two groups have different value orientations regarding their tasks. The computer community pays more attention to the balance of system performance, cost, and efficiency, while FHEs will favour system performance.

Meanwhile, debates over the validity and reliability of forensic handwriting examination and other forensic techniques have been ongoing for many years (e.g., Saks & Koehler, 2005; Saks, 2010; Saks, 1998). The basic tenets traditionally used to justify the variability and individuality of handwriting have been challenged and criticized, and there is limited supporting research. In a 2009 report, the US National Research Council proposed that the scientific basis for handwriting comparison and assessment in forensic handwriting examination should be strengthened (NRC, 2009). While guidelines aiming to standardize FHE processes do exist (ASTM, 2007; De Baere et al., 2016; SAMR & SAC, 2018), the methods of forensic handwriting analysis still stand on weak foundations when it comes to relying on objective methods independent of the FHEs. According to this research, FHEs would benefit greatly from embracing computer-based methods. That said, current computer-based handwritten signature verification systems, based on measurements of static (offline) two-dimensional (2D) images, are themselves vulnerable to skilled forgery (Soleimani, 2016). In addition, research in this area tends to focus on datasets that lack in size and diversity and does not cover all forensic

scenarios, especially when there is an allegation of forgery. Indeed, forgeries can be produced using different methods (e.g., tracing the forged entity through a translucent overlay placed on the genuine or a freehand simulation based on a study of an available model).

These shortcomings have hindered the application of computer-based techniques to forensic handwriting examination. Studies have shown that offline signature verification² is much more difficult than online signature verification. Indeed, EER ranges from 2% to 5% for online signature verification³ and from 10% to 30% for offline verification systems (Mohammed et al., 2015).

The computer science community treats handwriting as a 2D static image for offline verification. The three-dimensional (3D), dynamic nature of handwriting is often overlooked. However, handwriting is the product of a behavioural process; its morphology mainly depends on dynamic handwriting actions. Thus, handwriting is the trace of a dynamic writing action, rather than a static mark, that is left on paper by a writing instrument. At first glance, the depth of the indented trace does not seem distinctive in a 2D image, but, in reality, the depth allows for discrimination, as this research will show. Handwriting embeds dynamic and discriminant information in this third dimension. This research goes beyond the 2D, static measurements usually made of handwriting and embracing the new capabilities offered by 3D acquisition systems. In recent years, significant progress has been made in the application of 3D measurements to questioned document examination. Some researchers have reconstructed the indentation of handwriting on paper by means of laser holographic microscope to help establish stroke order (Spagnolo, 2006) or to help FHEs visually analyse handwriting strokes (Spagnolo et al., 2013). Based on such features, 3D measurement techniques should allow FHEs to better distinguish the handwriting of different writers. This research will show that 3D features offer increased discriminating power compared to 2D features, allowing FHEs to better distinguish genuine from forged entries.

This research aims also to enrich the features used by an offline verification system in its applications to forensic science. This project takes advantage of dynamic time warping techniques to capture features while maintaining the writing sequence. We have qualified these additional features as “pseudo-dynamic” because they are extracted while considering writing sequence. They are not extracted at the time of capture but acquired after the writing act from

² Offline signature verification is a process of verifying signatures using static images.

³ Online signature verification systems is a process of verifying signatures using a digitizer to extract information, such as x, y coordinates, time, and pressure.

the image. These features are different from well-known dynamic features extracted from online handwriting but still reflect the writing sequence; these have been then called “pseudo-dynamic features”.

In addition to 2D features, this work will add 3D information of a pseudo-dynamic nature measured using 3D microscope synchronized with white-light imaging. Three-dimensional and pseudo-dynamic features were acquired from a dataset comprising 23,624 signatures. The sample includes genuine signatures and a range of forgeries (freehand, random, or traced). For forensic purposes, score-based likelihood ratios (LRs) were adopted to assess the strength of the signature evidence. This work aims to not only show the increased discriminating power obtained using these features but also position the discipline within a proper evaluative framework.

The remainder of this thesis is organized as follows:

In chapter 2: State of the Art, the status of research in both forensic handwriting examination and handwriting verification is summarized, focusing on signature databases, the quantitative measurement of handwriting features, and the assessment of handwriting verification systems.

In chapter 3: Methods and Materials, Chinese signature databases, proficiency tests, and real forensic cases used in this study are introduced. Signature databases are used to train and validate the system, and proficiency tests and real forensic cases are used to test the system. The acquisition of data and reconstruction of 3D profiles of signatures is based on the use of digital holographic microscope and wide-area 3D measurement systems. Writing sequence tracing allows FHEs to extract 3D and pseudo-dynamic features from white-light and 3D images. Then, the LR approach is described. The application of machine learning (ML) techniques is presented because they will be used to assess the feasibility and robustness of these features for commercial applications that are deployed in traditional biometric verification systems. Finally, we describe the validation tests used to assess the performance and to identify the limits of the system.

In chapter 4: Statistical Analysis, we will present statistical descriptions of the features (e.g., MVN, discriminant analysis, and probability density estimation) that are used to explain and illustrate between- and within-writer variations. The results of the application of two methods of likelihood ratio (LR) calculation, multivariate kernel density estimation (MKDE) and Dempster–Shafer theory (DST), are given. The results of 13 ML methods are presented. The comparison of these ML methods shows that random forest (*R-Forest*) is the most accurate

system; variable importance is also provided according to the *R-Forest* method. Performance evaluation is conducted on the MKDE and DST LR systems, and two calibrations — Pool-adjacent violators algorithm (PAVA) and logistic methods—are used to evaluate the performance of the systems. Finally, validation test results will show the performance of the system and will highlight the need to explore within-writer variations under different writing conditions.

In chapter 5: Discussion and Perspective, the system of quantitative measurement, statistical analysis, and LR evaluation is discussed as a paradigm for both forensic handwriting examination and computer-based handwriting verification. In addition, the technical contributions of this research are summarised. The optimal algorithms of Dynamic time warping (DTW) are presented to demonstrate that optimisation is an open-ended question. Perspective is given for future research efforts.

In chapter 6: Conclusion, a route is suggested for bringing forensic handwriting examination in line with a rigorous documented methodology, based on data and not only on the personal appraisal by FHEs.

Four Chinese signature datasets from volunteers are used in this research: dataset_1, dataset_2, dataset_3, and dataset_4. Two papers (Chen X. 2015 and Chen X. H. et al. 2018) included in the appendix are research outputs based on dataset_1 and dataset_2. Dataset_2 in Chen X. et al. (2018) that included 20 volunteers was an expanded version based the dataset_1 on Chen (2015). The core dataset (dataset_3) of this thesis is a new dataset including 100 volunteers. In addition, dataset_4 including another 20 individuals was collected to identify the influence of signatures under different writing conditions to explore the limitation of this research.

Chapter 2 State of the Art

Writer verification helps determine whether two handwriting samples were written by the same or by different writers. This is an important task in forensic handwriting examination.

Disputed handwritten signatures often appear in civil or criminal cases and are submitted to forensic handwriting experts. This is also an active area in biometric research. Offline computer-based handwriting verification systems do have similar objectives. Handwritings, especially handwritten signatures, are used as behavioural biometric characteristics for security or authorization purposes (Bhattacharyya et al., 2009). This is mainly because signatures have been established as the most widespread means of personal authentication. Signatures are generally recognized as a legal means of authentication by administrative and financial institutions. Additionally, signature verification does not require any invasive measurements, and people are familiar with the use of signatures in their daily life (Impedovo & Pirlo, 2008). However, this verification task is still conducted manually by trained forensic handwriting examination. The state of practice and the increasing demand for objective and reproducible methods in HSV and forensic handwriting examination are reviewed below.

2.1 Forensic handwriting examination

Since the birth of writing, crimes involving handwriting have been commonplace, and handwriting identification (or examination) has also emerged as a field of expertise. In France, for example, Demelle's treatise dates back to 1604 (Demelle, 1604). In China, the earliest handwriting case dates back to 119 (Chen, 280). Today, forensic handwriting examination remains an active speciality of forensic science. Forensic handwriting examiners usually provide evidence relating to the identity of the author of a disputed handwriting and signature. A protocol named ACE was introduced by Huber (1959; 1972) to describe the process underlying handwriting identification: first, analysis (A) of reproducible and discriminating elements both on questioned and reference samples; then, the comparison (C) of the known discriminating elements with the unknown; finally, the evaluation (E) of the similarities or differences in discriminating elements (Harralson & Miller, 2017). The comparative examination of handwriting emphasizes the need to consider all of the characteristics of the handwriting in question and to use logic and sound reasoning when drawing conclusions. Simply adding the similarities and differences does not necessarily lead to a suitable result. On the contrary, it is important to consider holistically variables, such as the writer's age, health,

signature style, and other external factors (Bisesi, 2006). Allen (2018, p. 55) described the scientific method of the discipline as follows:

To conclude that two writings were made by one person, it would be necessary to show that no other explanation is possible. The hypothesis that two writings are by one person must be tested by observation of the writings and by reference to the resemblances and variations found within and between those of members of the relevant population. It is not sufficient to note that the writings are similar, assume that everyone writes differently, and therefore conclude that they were written by one person. To do this is to ignore the possibilities of coincidence and of simulation. Only when the findings have been assessed against all the possible alternative hypotheses and these have been ruled out as practically impossible would the conclusion be justified. This is the fundamental principle for the reaching of conclusions for questioned handwriting; the same principle applies throughout forensic science.

The conventional forensic handwriting practice has established a basic theoretical system for handwriting comparison and has played an important role in judicial disputes. A process map of forensic handwriting is shown in Figure 1.⁴

⁴ In the original text, the resolution of the image is so low that the details are not legible. Thanks to Linton A. Mohammed for the original graphic.

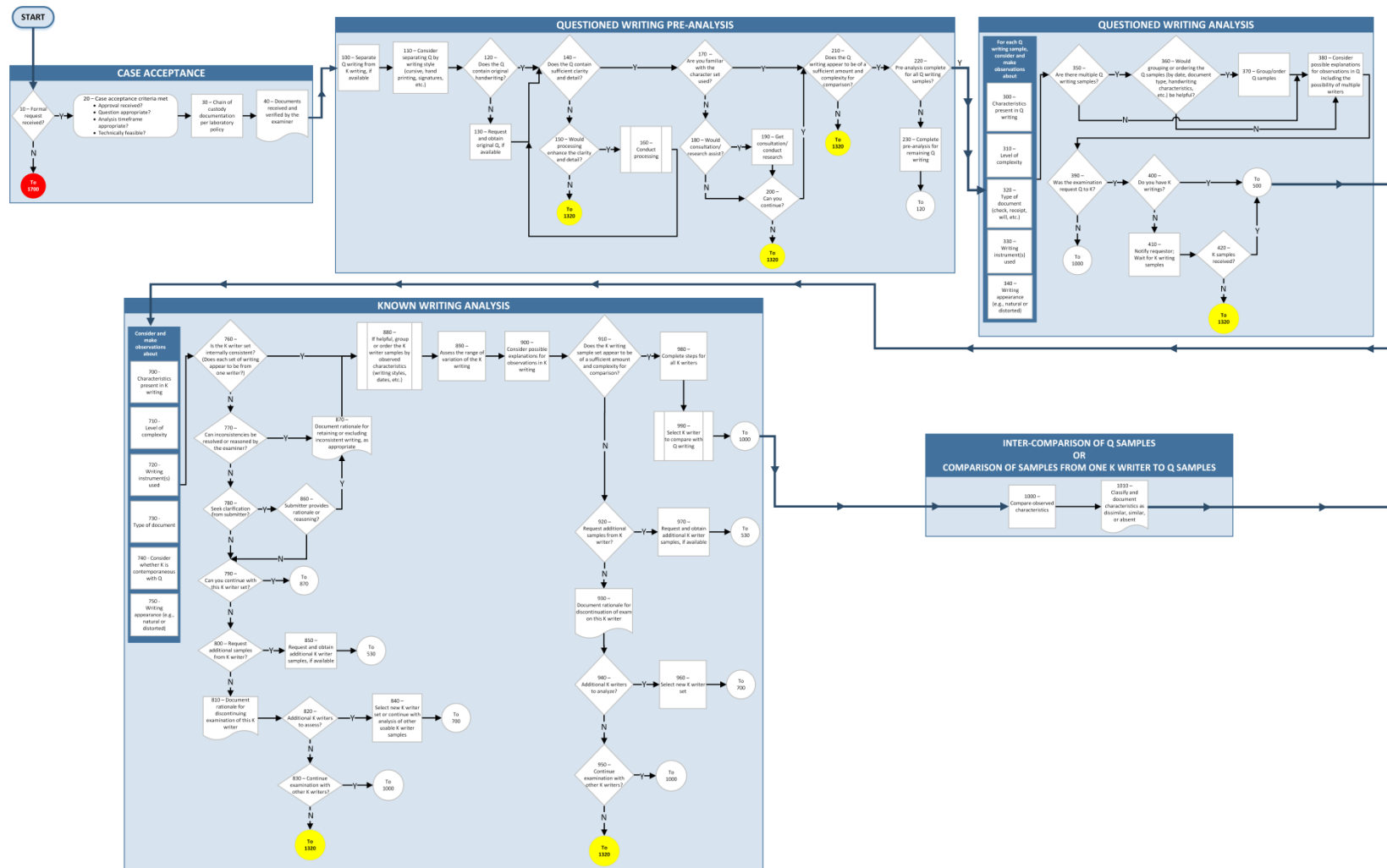


Figure 1: Handwriting examination process map (Figure 1.1 in Taylor et al., 2020)

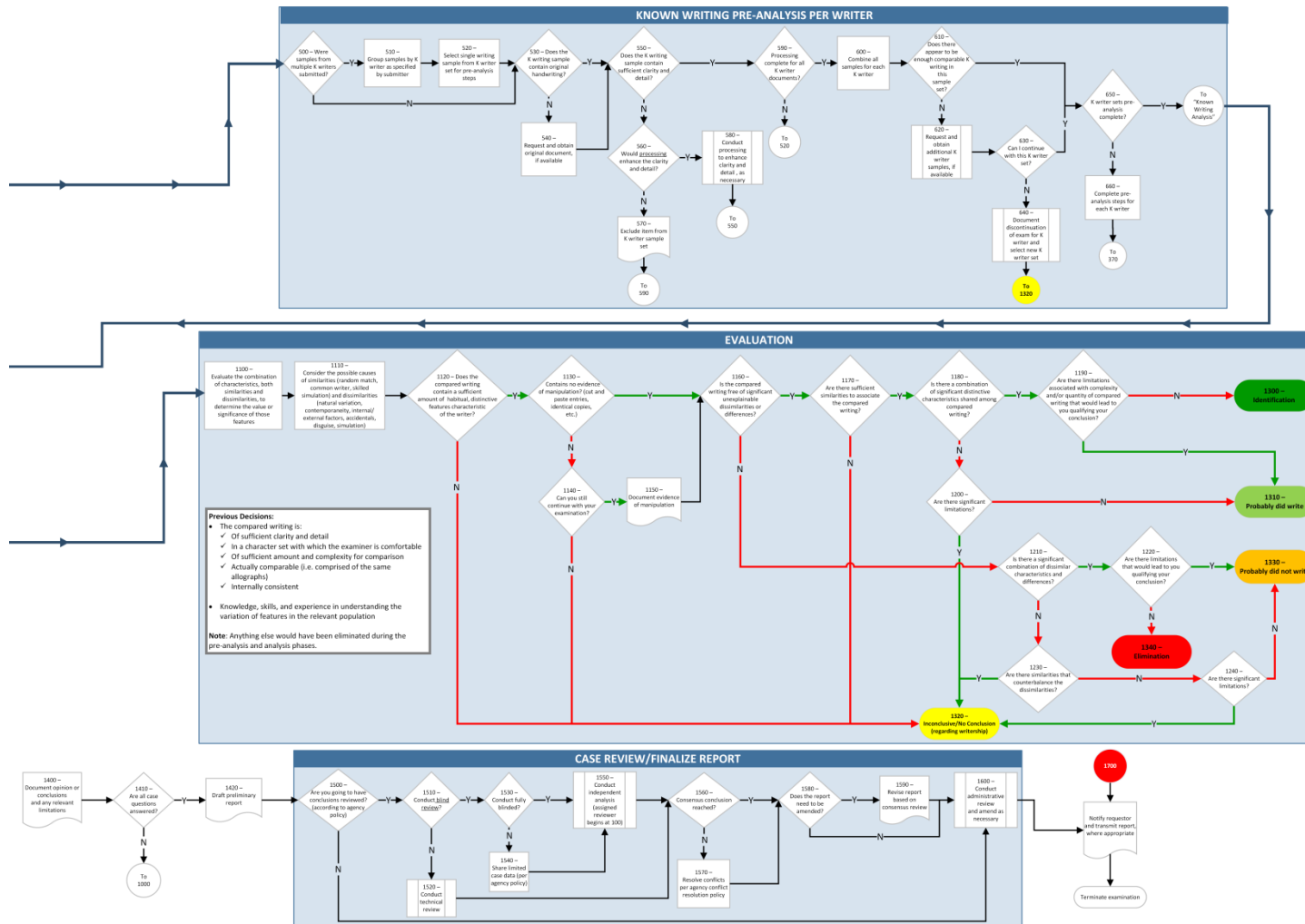


Figure 1 (continued): Handwriting examination process map (Figure 1.1 in Taylor et al., 2020)

The process map is essentially descriptive but captures the various steps of the examination distinguishing Analysis, Comparison, Evaluation, and Verification (ACE-V). To establish and maintain working practices in the field of forensic handwriting examination that will deliver reliable results, maximize the quality of the information, and produce robust evidence, technical standards have been proposed. For example, the Scientific Working Group for Forensic Document Examination (SWGDOC) provides a standard of forensic handwriting examination procedures that should be used by forensic document examiners for examinations and comparisons (SWGDOC, 2013). The European Network of Forensic Science Institutes (ENFSI) provides a framework of procedures, quality principles, training processes, and approaches to the forensic examination of handwriting (ENFSI, 2018). A specification for forensic identification of handwriting was released by the State Administration for Market Regulation (SAMR) and Standardization Administration of the People's Republic of China (SAC) in 2018 (SAMR & SAC, 2018). This standard specifies the terms and definitions of handwriting identification, the classification of handwriting features, the steps and methods of handwriting examination, the production of handwriting feature comparison tables, the technical points of abnormal handwriting examination, and signature handwriting examination. In addition, the document suggests the types, basis, and expressions of expert opinions. Note that these guidelines and standards are descriptive in nature and set a common language and procedure to conduct handwriting examination. When it comes to how experts come to a specified conclusion, the documents leave that task entirely to the training and experience of the FHEs.

Studies have shown that trained FHEs perform significantly better than lay people in the analysis, comparison, and evaluation of handwriting (e.g., Bird et al., 2010b). Conscious that leaving the assessment to FHEs is hampering the transparency of the examination process, research to demonstrate that FHEs outperform novices (or laypersons) was key in helping the field to gain admissibility in court. A recent paper deals with experts in South Korea (Kang et al., 2022) and shows that the holistic approach brings added value to the decision maker, above what could be expected from a layperson. But the detailed inferential process remains obscure and does not rely on any systematic, quantitative measures, nor statistical analysis. Some authors also have highlighted that the difference between experts and novices is modest (Martire et al. 2018). In addition, it is well-documented that cognitive bias can negatively affect the decision-making (for a review, see Li & Ma, 2018). As highlighted by Sulner (2018), the *“problems associated with the reliability of handwriting identification opinion evidence [...] still prevail.”*

There is a growing body of research in forensic science on the application of quantitative extraction methods to handwriting features (Marquis et al., 2005; Marquis et al., 2006; Marquis et al., 2011a; Marquis et al., 2011b; Ling, 2002; Found et al., 1994) and on statistical methods for forensic handwriting examination assessment (Taroni et al., 2014; Taroni et al., 2012; Bozza et al., 2008; Hepler et al., 2012; Johnson & Ommen, 2021; Crawford et al. 2021). These efforts aim at reducing the subjective aspect of forensic handwriting examination or at least providing an objective mechanism to support them. Almost all research focuses on the measurement of 2D static images on paper. However, 2D systems are vulnerable to skilled forgeries. Because of the influence of subjective factors (e.g., intention to disguise or simulate), writing conditions (e.g., writing instrument, carrier, cushion, posture), physiological factors (e.g., alcohol, drugs, pathology, age), writing time, and other factors, signature handwriting can reflect complex between- and within-individual variations. The actual written utterance may be the result of a combination of multiple factors, and the degree of complexity should not be underestimated. Given the lack of sufficient diversity in databases used on systematic research, although there are many methods for solving the problem of individual identification in certain situations, there is still no method that can systematically solve all problems of signature handwriting verification. Handwriting verification systems fail to reach commercial application, unlike facial verification or fingerprints, reflecting the significant gap between practical application and theoretical research.

The reliability of forensic handwriting examination has been debated for decades. Courts assess expertise by looking for indices of validity. We will focus for illustrative purposes on the US judicial practice. In *Frye v. United States* [293 F. 1013 (D.C. Cir. 1923)], the federal appellate court noted that:

Somewhere in this twilight zone the evidential force of the principle must be recognized, and while courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.

For sixty years, the *Frye* test had become the dominant expert evidence filter used in American courts. In 1993, in *Daubert v. Merrell Dow Pharmaceuticals* (509 U.S. 579 (1993)), the US Supreme Court cited *Frye v. United States*, 54 App. D.C. 46,47,293 F. 1013, 1014, affirming that expert opinion based on a scientific technique is inadmissible unless the technique is “generally accepted” as reliable in the relevant scientific community (as in *Frye*). In addition, however, the US Supreme Court introduced additional factors to help judges decide on

admissibility (Table 1: Factors from the *Daubert* decision (TABLE 1 in Grivas et al., 2008)). Under *Daubert*, proffered scientific testimony must be shown to stand on a defensible foundation. Judges had an additional duty to act as gatekeepers in deciding the admissibility of scientific expert witness testimony (Grivas et al., 2008).

Table 1: Factors from the *Daubert* decision (TABLE 1 in Grivas et al., 2008)

The content of testimony must

1. Be testable and have been tested through the scientific method;
 2. Have been subject to peer review;
 3. Have established standards;
 4. Have a known or potential error rate;
 5. Have widespread acceptance by the relevant scientific community.
-

Conversely, a second decision in relation to handwriting evidence, *United States v. Starzecpyzel* (880 F. Supp. 1027 (S.D.N.Y. 1995)), offered an early indication of how *Daubert* could change judicial views, complicating the debate over the admissibility of expert handwriting testimony. In this case, FHEs' expertise was described as 'not properly characterized as scientific, but as practical in character'. The court, therefore, placed forensic document expertise under the category of "technical, or other specialized knowledge", which was apparently not covered by *Daubert*. This case represented an acknowledgement by the court that science is too complex to evaluate using a single set of standards (*Kumho Tire Company, Ltd. v. Carmichael*, 526 US 137., 1999, Haack S., 2005) (Table 2: Factors from the *Kumho* decision (TABLE 2 in Grivas et al., 2008)).

Table 2: Factors from the *Kumho* decision (TABLE 2 in Grivas et al., 2008)

-
1. Expert witnesses can develop theories based on their observation and experience and then apply those theories to the case before the court.
 2. All forms of expert witness testimony should be evaluated with the same level of rigor.
 3. The *Daubert* standards are flexible guidelines that may not be applicable in every instance of expert witness testimony.
-

Saks et al. (2005) commented as follows:

Ironically, then, fields that initially gained entry to the courts by declaring themselves to be 'sciences' now sought to remain in court by denying any connection with scientific methods, data, or principles' (Saks et al., 2005, p. 684).

The *Kumho* decision led to the consideration of handwriting examination as a technical skill but one that ought to be assessed in light of the *Daubert* factors. It is fair to say that the essential methods of forensic handwriting stand on a weak foundation regarding objective methods. Forensic handwriting examination mainly depends on the expert's experience. Due to a lack of

systematic measures and quantitative methods, it is difficult to ensure that each trainee will gain sufficient knowledge of the corresponding skills during the training process. As a result, it is not uncommon for different experts to have conflicting opinions about the same handwriting comparison (Bird et al., 2010a).

2.2 Offline handwriting verification

Generally, biometric systems (Bhattacharyya et al., 2009; Yager et al., 2010) can be divided into two types. This first rests on behavioural biometrics, in which users perform certain actions for data acquisition (e.g., speech, signature verification, and keystroke dynamics). The second rests on physiological biometrics, in which users do not need to perform any actions since the system derives data from the direct measurement of parts of the body, such as fingerprints, palm prints, or irises. Handwritten signature verification (HSV) is a behavioural biometric technique. HSV systems can be divided into two main types: *online* signature verification and *offline* signature verification. Offline handwritten signature systems deal with static images, whereas online handwritten signature systems deal with data obtained from acquisition hardware. In general, HSV systems comprise six main stages to complete their task: data acquisition, pre-processing, feature extraction and selection, comparison, verification, and performance evaluation. For related studies using the three most common datasets, we refer to Tables 3 to 5. According to whether or not the system depends on the writer, HSV can be divided into Writer-dependent (WD) and Writer-independent (WI) systems. In other words, WI systems do not need the signature dataset of the writer who was suspected of writing the signature in question. Conversely, WD systems do require the signature dataset from the writer in question. WD and WI are quite similar to “inner-individual” and “inter-individual,” respectively, terms that we will use in this research (see Section 3.6.3).

Table 3: Related studies using MYCT signature dataset

Feature	Classifier	Type*	Performance	Related research
Operates a family of six groups of grids lattices (GoGs)	SVM	WD	EER=4.01	Zois et al. (2016)
Textual features	SVM	WD	EER=9.12	Diaz et al. (2016a)
Texture features	SVM	WD	EER=6.10	Bhunja et al. (2019)
Sparse representation techniques	SVM	WD	EER=1.37	Zois et al. (2019)
Structural and directional features	RNN	WD	EER=0.01	Ghosh (2021)

* WD means writer dependent; WI means writer independent.

Table 4: Related studies using GPDS960 signature dataset

Feature	Classifier	Type*	Performance	Related research
Curvelet transform features	OC-SVM	WI	EER=15.07	Guerbai et al. (2015)
Deep CNN features	SVM	WD	EER=10.70	Hafemann et al. (2016)
Operates a family of six groups of grids lattices (GoGs)	SVM	WD	EER=3.24	Zois et al. (2016)
Textual features	SVM	WD	EER=14.58	Diaz et al. (2016a)
Deep CNN features	SVM	WD	EER=1.72	Hafemann et al. (2017)
Deep CNN features	SVM	WD	EER=0.41	Hafemann et al. (2018)
Sparse representation techniques	SVM	WD	EER=0.70	Zois et al. (2019)
Structural and directional features	RNN	WD	EER=1.46	Ghosh (2021)

* WD means writer dependent; WI means writer independent.

Table 5: Related studies using CEDAR signature dataset

Feature	Classifier	Type*	Performance	Related research
Global, statistic	Naive Bayes	WD	AER=23.5	
Distance statistics	Geometrical and Topologic	WD	AER=21.7	Srihari et al. (2004)
Surroundedness	Neural network	WI	AER=8.33	Kumar et al. (2012)
Curvelet transform	OC-SVM	WI	AER=5.6	Guerbai et al. (2015)
Directional Code Co-occurrence matrix (DCCM) feature generation method	Feature Dissimilarity Measures (FDM)	WI	AER=2.63	Hamadene & Chibani (2016)
Operates a family of six groups of grids lattices (GoGs)	SVM	WD	EER=3.02 AER=2.74	Zois et al. (2016)
KAZE features	SVM	WI	EER=1.6	Okawa (2016)
Geometrical features + Genetic Algorithm	SVM	WD	AER=4.67	Sharif et al. (2018)
Texture features	SVM	WD	EER=1.64	Bhunja et al. (2019)
Texture features	Capsule Network	WD	Accuracy=98.8	Gumusbas & Yildirim (2019)
Sparse representation techniques	SVM	WD	EER=0.79	Zois et al. (2019)
Structural and directional features	RNN	WD	EER=0.01	Ghosh (2021)

* WD means writer dependent; WI means writer independent.

Recently, there has been remarkable progress in quantitative measurement and pattern recognition. Could such advances provide a solution for strengthening the basis of handwriting comparison and assessment in forensic handwriting examination? HSV systems have a higher error rate than other biometric systems (Tahezadeh et al., 2011). Furthermore, HSV systems based on the measurement of 2D images are vulnerable to forgery (Fierrez et al., 2008). Such limitations have hindered the application of HSV. However, HSV presents a promising area for quantitative feature measurement and statistical assessment in forensic handwriting examination. Nevertheless, a gap still exists between signature verification methods and the requirements for their application in forensic science. There is indeed a need for training data and pattern recognition (PR) systems that are compatible with the forensic requirements. Current PR output reporting schemes do not fit the needs of forensic science (Malik, 2015). The underlying issue is that most state-of-the-art handwriting/signature analysis systems cannot be directly applied to forensic cases. The gap between FHEs and the PR community is summarized in Table 6.

Table 6: Gap between FHEs and the PR community

Lack of common terminology	In the PR community, different names are often used for the same forgery type, and sometimes the same name is used to refer to different types of forgeries (Malik, 2015).
Non-representative and non-diverse databases	Systems trained on databases collected in controlled environments are not well suited to forensic applications (Malik, 2015).
Result interpretation	The binary output provided by a PR system is not acceptable as a presentation method in courts. In general, the PR community has not adopted LR. (Found & Rogers, 2003; Malik, 2015).
Performance evaluation	Log-likelihood-ratio cost (Cllr) is used to measure the validity and reliability of forensic likelihood-ratio systems, which has gained little attention in the PR community (Morrison, 2011).

In the case of Chinese signature verification, the process might be much more difficult than is the case for Western-language signatures. For instance, two offline signature datasets were collected in the context of the Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011; Liwiki et al., 2011): a Chinese dataset and a Dutch dataset. The performance results (accuracy) of offline signature verification systems for Dutch signatures ranged from 72.02% to 97.67%. For Chinese signatures, the accuracy was lower, ranging from 51.95% to 80.04%. These results for Chinese signatures

were not nearly as good as results presented in the literature reporting accuracies ranging from 91% to 98.5% (Pal et al., 2011).

In forensic practice, several PR systems have been developed, aiming to integrate FHE knowledge into interactive computerized solutions, where a subset of features employed by experts are assessed using algorithms. Such projects include Trigraph (Niels et al., 2005), Wanda (Franke et al., 2004), CEDAR-FOX (Srihari et al., 2007), and the well-known Forensic Information System for Handwriting (FISH) (Hecker, 1993). Meanwhile, the use of these PR systems is very limited in FHEs' daily work. For instance, CEDAR-FOX has proven successful in narrowing down lists of candidates in a database comparison. However, these are only as good as their databases allow and might not consider handwriting simulations and forgeries (Ellen et al., 2018). A recent research study examined the relationship between two systems: an automated handwriting/black box system (FLASH ID; Miller et al. 2017) and a system (MovAlyzeR⁵). The former uses measurements extracted from a static image of handwriting, the later captures kinematic features from pen strokes. The comparison between these two systems validated biometric matching algorithms in FLASH ID (Fuglsby et al., 2021). FLASH ID (handwriting biometric)⁶ is similar to CEDAR-FOX and Wanda, and it strives to promote the application of computer graphics image processing and ML technology in the field of forensic handwriting identification. Such a system may be capable of solving the verification of one or some single types of handwriting, but it lacks sufficient understanding of the complex situations faced by forensic handwriting identification (e.g., various ways of simulated handwriting, disguised handwriting, and handwriting under different writing conditions). By perfecting the system and making it truly suitable for the practice of handwriting examination, FLASH ID may be like Cedar-Fox and Wanda, the certification and application in forensic handwriting examination is limited.

2.3 Signature databases

To train classifiers and test the performance of verification systems, various handwriting datasets are available for handwriting verification research. Only a few, however, provide forensic relevant signatures for training and testing. Three databases (GPDS, MCYT, and CEDAR) are often used in offline signature verification research⁷ (Table 7). The GPDS synthetic signature database is the largest; it is not made up of real signatures but of synthetic ones. Since the scientific soundness of the synthetic generation algorithm is not clear, synthetic signature databases should be distinguished from real ones.

⁵ www.neuroscript.net

⁶ <https://www.sciometrics.com/flashid.html>

⁷ <https://gpds.ulpgc.es/> and <http://atvs.ii.uam.es/atvs/databases.jsp>

Table 7: GPDS, MCYT, and CEDAR signature databases

DATABASE	Number of writers	Number of signatures per writer	Total mount
GPDS Synthetic Duplicator Engine for Static Signatures (Diaz et al., 2016a; Diaz et al., 2016b)	75	(13 or 10 genuine+15 forged)*20	>37,500
	100	(22 or 19 genuine+10 forged)*20	>58,000
	100	(22 or 19 genuine+10 forged)*20	>58,000
GPDS Bengali and Devanagari Synthetic Signature Databases (Diaz et al., 2016c; Ferrer et al., 2017)	100	(12 genuine+10 forged)*3 forged (different model pens)	5,400
GPDS Synthetic OnLine and OffLine Signature Database (Ferrer et al., 2016)	10,000	24 genuine+30 forged (different model pens)	540,000
GPDS Synthetic Signature Database (Ferrer et al., 2014)	4000	24 genuine+30 forged (10 forged)	216,000
GPDS960 signature database (Blumenstein et al., 2010)	960	24 genuine+30 skilled forgeries (10 forged)	51,840
MCYT BiosecurID-SONOF DB (Galbally et al., 2015; Fierrez et al., 2010)	132	16 genuine+12 skilled forgeries	1,584
MCYT Bimodal Biometric Database (MCYT-SignatureOff-75) (Fierrez-Aguilar et al., 2004)	75	15 genuine+15 forged (3 forged)	2,250

The GPDS960 signature database (Ferrer, 2012) is no longer available because of the General Data Protection Regulation (EU) 2016/679 ('GDPR'). To comply with GDPR and increase data collection, GPDS provides synthetic signatures in recent signature datasets; that is, a synthesis algorithm generates new samples from those of an existing user. Two synthetic signature datasets—the Dual Offline and Online Databases of Bengali and Devanagari Signatures—contain data from 100 synthetic individuals: 24 genuine signatures for each individual; all of the static signatures are generated with different pens. Synthetic users in the GPDS synthetic signature database are generated following the procedure described in Ferrer et al. (2015). The GPDS Synthetic OnLine and OffLine signature database contains data from 10,000 synthetic individuals: 24 genuine signatures for each individual, plus 30 forgeries of

his/her signature. All of the static signatures are generated with different pens. The synthetic users in this database were generated following procedures described in Ferrer et al. (2014), Diaz et al. (2016c), Galbally et al. (2012a), and Galbally et al. (2012b). Offline synthetic signatures are shown in Figure 2.

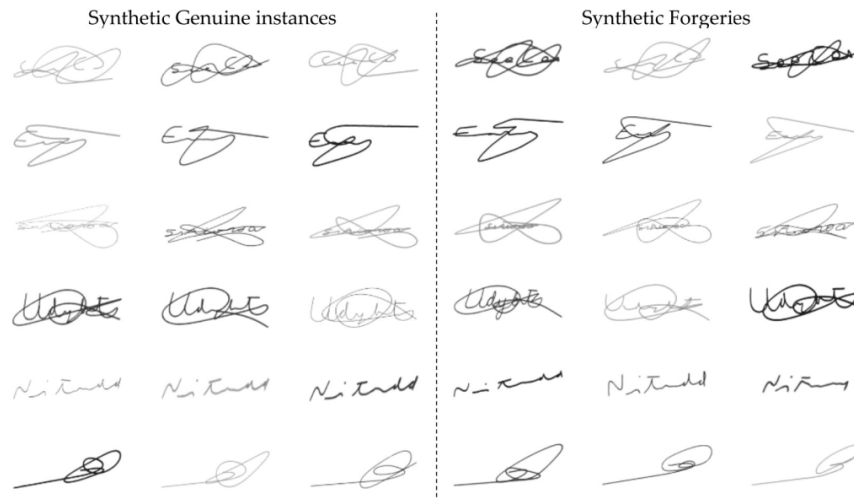


Figure 2: Six possible synthetic identities with three genuine specimens (first three columns) and three possible forged signatures (last three columns). The first four signatures are composed of test plus flourish, the fifth example has only text, and the sixth is a simple flourish (Figure 11 in Ferrer et al., 2014).

Though the authors of this study claim that these synthetic databases show a remarkably high degree of similarity with real databases, several researchers note the following limitations: “a common pitfall is to underestimate the data scientist’s influence during the generation process on the resulting intrinsic properties of the generated synthetic data”; this is because ‘datasets need to be transformed by numerous pre-processing and configuration procedures to make them accessible to generative models. During these preparatory steps, our assumptions about the data play a fundamental role’ (Robin, 2019). Another comment was as follows:

Synthetic data-generation methods score very high on cost-effectiveness, privacy, enhanced security and data augmentation, to name a few measures. However, they come with their own limitations, too. While synthetic data can mimic many properties of real data, by their very design the synthetic data-generation models we have discussed do not recreate the original data exactly. So, any analysis on synthetic data needs to be verified on the real data set. Synthetic data-generative models look for common trends in the real data when creating synthetic data but may not capture any anomalies present in the real data. In some instances, this may not be a critical issue.

However, in other scenarios, it will severely limit the capabilities of the model and negatively impact the output accuracy (Joshi, 2019).

When it comes to synthesizing biometric samples, the premise is that the model of this biometric trait, such as a DNA profile, which has a clear biological model, or a human face sample, which has a defined geometry model, ought to be known. However, the model for signature handwriting features is not yet fully defined. Therefore, synthetic signature handwriting databases lack adequate theoretical support and practical verification. Hence, in forensic research, such synthetic databases need to be treated with caution. The signature synthesis method is essentially a means of forging, and the synthetic signature is nothing other than a forged signature. Thus, there is a paradox if the intention is to use a forged signature as a true signature to train a model that can recognize a forged signature.

There have been signature competitions intended to reduce the gap between forensic handwriting examination by FHEs and PR by computer scientists. Such efforts reveal aspects of practical cases of forensic handwriting examination. For computer scientists, the signature in forensic practice is much more complicated than any signature datasets used in PR. The signatures datasets used in research were indeed never designed to cover the types of signatures encountered in forensic casework. Genuine signatures, disguised forgeries, random forgeries, simulated forgeries, and skilled forgeries are absent. In addition, likelihood ratio based probabilistic evaluation is becoming more widely used in forensic science (Morrison & Enzinger, 2018; Jacquet & Champod, 2020) including areas such as handwriting (Hepler et al., 2012), signatures (Chen et al., 2018), forensic MDMA comparison (Bolck et al., 2015), fingerprints (Egli et al., 2007; Gonzalez-Rodriguez et al., 2005; Leegwater et al., 2017), speech recognition (Gonzalez-Rodriguez et al., 2006; Brümmer & Du Preez, 2006; Morrison, 2011), and marks left on gun cartridges (Riva, 2011; Riva & Champod, 2014; Riva et al., 2017). Yet, in the field of computer science, attention is rarely paid to the probabilistic interpretation and the assessment of the weight of forensic findings. Meanwhile, for FHEs, computer-based quantitative measurement and analysis could be a possible avenue for making forensic handwriting examination more robust and transparent, considering that current methods are still in need of improvement.

There have been forensic signature competitions using collected signature datasets (Blankers et al., 2009; Liwiki et al., 2010; Liwiki et al., 2011; Liwiki et al., 2012; Malik et al., 2013; Malik et al., 2015) (Table 8). Typically, there are no more than 100 writers and 64 forgers. The largest dataset had 5108 signatures.

Table 8: Forensic signature competitions

Dataset	Training dataset						Testing dataset				
	Language	Writers	Genuine signatures	Disguised signatures	Forgers	Forgeries	Writers	Genuine signatures	Disguised signatures	Forgers	Forgeries
SigComp2009 (Blankers et al.,	Dutch	12	60	0	31	1860	100	1200	0	33	792
4NSigComp2010 (Liwiki et al., 2010)	English	1	85	20	27	104	1	28	7	34	90
SigComp2011 (Liwiki et al., 2011)	Chinese	10	235	0	/	340	10	116	0	/	367
	Dutch	10	240	0	/	123	54	648	0	/	638
4NSigcomp2012 (Liwiki et al., 2012)	English	2	113	27	61	194	3	30–63	45–90	2–31	70–775
	Japanese	11	462	0	4	396	20	840	0	4	720
SigWiComp2013 (Malik et al., 2013)	Dutch	66	1356	0	>64	2508	27	270	0	9	974
	Italian	50	250	0	0	0	50	229	0	/	249
SigWiComp2015 (Malik et al., 2015)	Bengali	10	120	0	0	0	10	120	0	/	300

There are various metrics that can help measure the performance of biometric (and forensic) systems (Bhattacharyya et al., 2009; Meuwly & Haraksim, 2017; Naika, 2018) (Table 9). We will refer to these metrics in the next description of system performance in Section 3.6.3.

Table 9: Metrics used to measure the performance of biometric authentication systems

Factors of evaluation	Description
False accept rate (FAR)	Probability that the system incorrectly declares a successful match between the input pattern and a non-matching pattern in the database
False reject rate (FRR)	Probability that the system incorrectly declares a failure of match between the input pattern and the matching template in the database
Equal error rate (EER)	Rates at which both accept and reject errors are equal. This corresponds to the threshold value where the false acceptance rate and false rejection are same.
Accuracy	Percentage of correct decisions with respect to all questioned patterns.
C_{lir}	Measure of log-likelihood ratio, which can properly evaluate the discrimination of all log-likelihood ratio cores. This also evaluates the quality of the calibration. Log-likelihood-ratio cost (C _{lir}) is proposed as a metric of accuracy related to the average cost of the LR method used.
C_{lir}^{min}	C_{lir}^{min} represents the minimum possible value of C_{lir} that can be achieved by an optimally calibrated system. The smaller the C_{lir}^{min} , the better the indication of the system.

The results of these offline forensic signature competitions show that the performance of verification systems is much lower than what has been reported in the literature (Table 10). This indicates that research on forensic signature verification is not at the stage of meeting practical forensic requirements. Such requirements include the stable and accurate verification of signatures under actual forensic conditions, including different writing conditions or different subjective intentions (disguise, simulation). The stability and repeatability of the methods still need to be verified. Researchers need to make more effort toward finding methods that meet these requirements for forensic handwriting examination (see Table 11).

Table 10: Results of offline forensic signature competitions

Dataset	Language	EER (%)	Accuracy (%)	FRR	FAR	C_{lfr}	C_{lfr}^{min}
SigComp2009 (Blankers et al., 2009)	Dutch	9.15–43.02	/	/	/	/	/
4NSigComp2010 (Liwiki et al., 2010)	English	55–80	20–92	0–87.0	10–90	/	/
SigComp2011 (Liwiki et al., 2011)	Chinese	/	51.95–80.04	21.01–50.00	19.62–47.41	0.76–6.23	0.69–0.95
	Dutch	/	71.02–97.67	2.47–29.17	2.19–28.79	0.42–4.13	0.08–0.79
4NSigcomp2012 (Liwiki et al., 2012)	English	/	27.98–86.79	13.27–68.14	13.19–73.63	0.50–6.49	0.36–0.77
SigWiComp2013 (Malik et al., 2013)	Japanese	/	66.67–90.72	9.74–33.33	9.72–33.33	0.79–4.69	0.40–0.78
	Dutch	/	67.90–76.83	23.70–31.11	23.10–32.14	0.88–3.94	0.64–0.86
SigWiComp2015 (Malik et al., 2015)	Italian	/	/	/	/	0.66–13.11	0.02–0.99
	Bengali	/	/	/	/	0.69–2.84	0.04–0.98

Table 11: Requirements for forensic handwriting examination

Item	Requirement
Signature	Actual types (genuine, disguise and forgery), under different writing conditions (writing instruments, surface, position, and more)
Method	Quantitative measurement, features and distribution visualization that is compatible with or supplies supplement for FHEs' examination.
Result format	LR
Performance	Stable and accurate

The concept of “forgery” as referred to in the forensic literature requires clarification. As Harralson and Miller described, “forgery” is used to represent “simulation” in handwriting (Harralson & Miller, 2017). For others, the word “forgery” implies intent to deceive and is best avoided when describing simulated writings (Ellen et al., 2018). Ellen et al. also distinguished some types of simulations, such as freehand simulation, slowly made simulation, poorly made simulation, rapidly made simulation, and traced simulation. In addition, more general forgeries may refer to nongenuine handwriting, regardless of whether or not imitation is intentional (Nguyen et al., 2007; see Figure 3). In this research, according to forensic practice, we selected all the types of forgeries detailed except the case in which the forged signature is produced without knowledge of the genuine writer’s name (random forgery with name unknown). The signature dataset in this research is fully presented in Section 3.1.

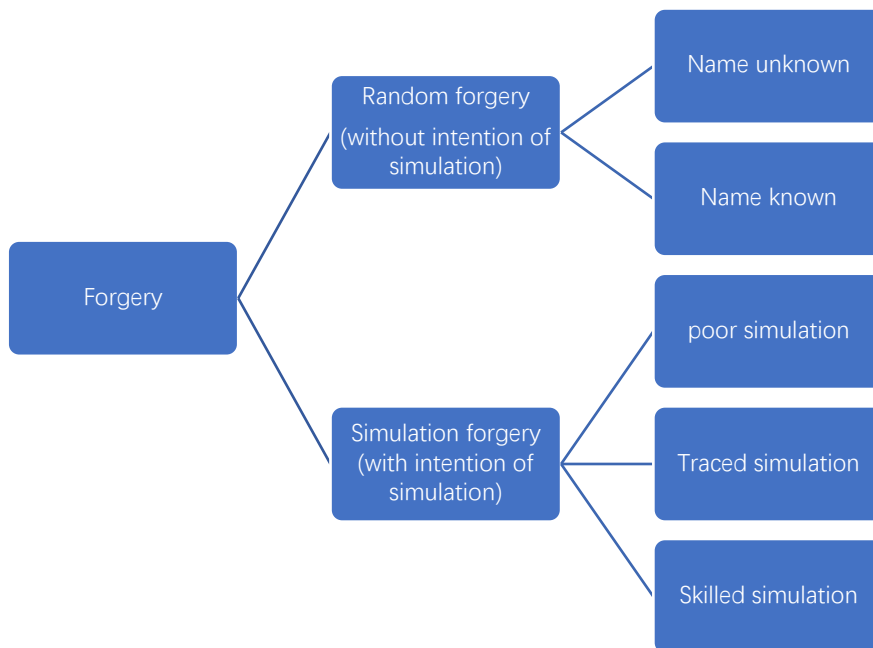


Figure 3: Types of forgery

People’s ability to simulate (or forge) a signature varies greatly. Dewhurst et al. (2008) reported the number of misleading authorship opinions for the calligraphers’ forgeries were approximately four times that of the laypersons’ forgeries. Houmani et al. (2012) presented algorithms to measure forgeries of

different quality. The results using the three classifiers (i.e., DTW, HMM, and GMM) showed an equal error rate (EER) of good quality forgeries was at least 1.5 times that of bad quality forgeries. In this research, to explore simulation forgeries of different quality, we used a deep learning technique to select skilled forgeries (see Section 3.6.1.3).

2.4 Quantitative measurement of handwriting features

In current research, handwriting features are obtained from 2D static images. Handwritten signatures are scanned and imported into a computer in digital form; thus, 2D static images are used as the raw data for feature extraction and analysis. Discrimination can take advantage of two kinds of features: global and local features, measured from the digital image. Global features describe the signature image as a whole (e.g., the ratio of signature height to width, slant features, stroke width distribution, centre of gravity, direction, and the trajectory of the signature). They can be divided into two groups: statistical and geometrical features. Statistical features are taken from the pixel distribution of the signature image. Geometrical features describe the geometrical characteristics of the signature image (Mohammed et al., 2015).

However, the above-mentioned features for offline signature verification are 2D static features reflecting the overall aspects of signatures. Research to improve signature verification systems remains based on 2D static features, using gradient local binary patterns, longest run features (Serdouk et al., 2016), texture features (Hannad et al., 2016), and geometric structure features measured by grid templates (Zois et al., 2016). Some studies have presented pseudo-dynamic features for offline signature verification (Vargas et al., 2011). In reality however, their pseudo-dynamic features comprise the grey-level distribution in the 2D static signature image. This is different from real dynamic features combined with writing sequences.

It leads to the following question: Is it reasonable to treat handwriting as a 2D image? Most research has focused on static features, such as contour, gradient, and direction, while neglecting the potential of dynamic and 3D features. However, handwriting is the product of behavioural processes. The contours of handwriting mainly depend on the writing action, which is different from a stamp, whose contours mainly depend on the printing surface. Handwriting presented to FHEs is the trace left following a writing action rather than a static mark or imprint on the writing surface. Therefore, FHEs try to reconstruct the pen tip movement sequence, called the writing sequence, during analysis. In addition, in 2D images, the depth of the handwriting does not seem to be as distinctive as the contour, but it does exist. In this sense, handwriting embeds dynamic information in the third dimension. In Figure 4, for instance, the purple stroke

(with labelled 1084) is highlighted in 3D profile, grayscale profile, and radian plots. The present research thus goes beyond 2D measurement.

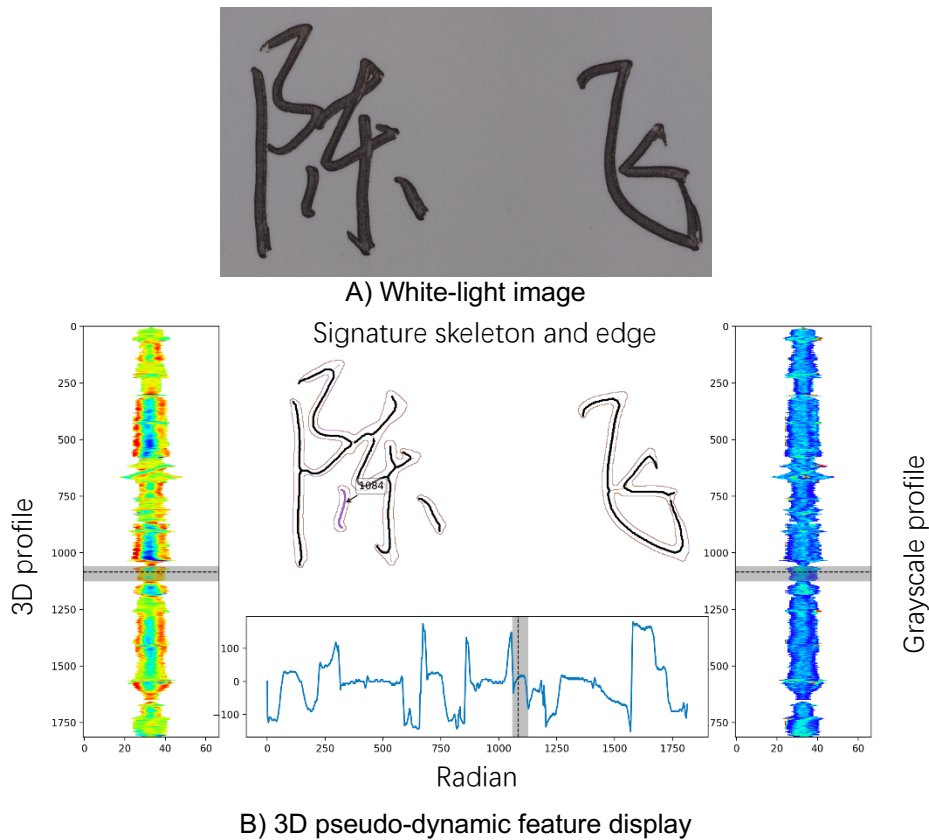


Figure 4: 3D pseudo-dynamic feature for purple strokes labelled 1084

Previous research by the author (Chen, 2015) laid out the process of feature detection and analysis in forensic handwriting examination. A threshold was applied in image binarization after the signatures were digitalized. The skeleton and signature edges were extracted after image processing. Then, a program for writing-sequence recovery processed the skeleton of the signatures. The width, grayscale, and radian were automatically extracted from the writing sequence. Thus, the features of width, grayscale, and radian combined with writing sequence were automatically extracted. Next, a DTW method was applied to cope with differences in writing speeds. The pairwise comparative measurement of the correlation coefficient, distance in DTW and Kolmogorov–Smirnov test (KST) was used to characterize and express the similarities between signatures (see Section 4.3.2). Finally, multivariate analysis of variation confirmed significant differences between genuine and forged signatures, and discriminant analysis showed a high score in the cross-validation rate, with a mean value of 95.8%.

The established quantitative feature extraction and statistical analysis methods give less freedom in feature selection to FHEs and have helped to

develop a transparent assessment framework for forensic handwriting examination. The work presented in this thesis expands the signature database and features and upgrades the analysis methods.

Research has shown that the shape of loops in certain letters can be used to distinguish writers (Marquis et al., 2011a; Marquis et al., 2011b) and to draw inferences about the gender identity and handedness of a writer (Taroni et al., 2014). Progress has been made in the search for formalized handwriting examination. Montani (2015) and Gaborini et al. (2017) measured the distances (x and y axis) and angles of line strokes in signature images to construct an evaluation framework to support expert examinations. Bennour et al. (2019) evaluated the effectiveness of an implicit shape codebook technique to recognize writers from digitized handwriting images. They applied the Harris key-point detector (Harris & Stephens, 1998) to extract junctions and corners to obtain a set of informative regions for each handwriting image. Meanwhile, Agius et al. (2018) extracted a set of writer characteristics, spatial characteristics, and construction characteristics to predict a writer's country of origin. All of the features obtained from handwriting images were measurements of spatial distance or spatial distance ratios. Most of the handwriting features used in the above-mentioned work are similar to those used in research on handwriting verification.

2.5 Assessment of handwriting verification systems

Today, the so-called Bayesian approach provides a unified and logical framework for analysing evidence and presenting results to courts (and other decision makers) in the form of LR's. Forensic scientists should not usurp the role of judge and jury but present their findings in the form of focused LR's; this approach offers a valid theoretical framework for any forensic discipline. Meanwhile, handwriting verification systems (in biometric applications) are used to deliver acceptance or rejection decisions. There is indeed a contrast between a continuous approach assigning weights of evidence and the decision framework used in biometric systems. Having a clear underpinning logic is a decisive mechanism that creates greater transparency in the way scientific findings are presented in court (Evetts et al., 1998; Curran et al., 1998; Champod & Meuwly, 2000; Champod et al., 2004; Champod, & Evetts, 2009). It is suggested that biometric scientists or laboratories should adapt their conventional biometric systems or technologies to work according to the LR approach (Gonzalez-Rodriguez et al., 2005; Srihari et al., 2005; Gonzalez-Rodriguez et al., 2006; Gonzalez-Rodriguez et al., 2007). Bayesian inference provides an appropriate interpretive framework (Taroni et al., 2002). The formal reasoning method based on (Bayesian) decision theory is not only conceptual, but rather it gives the basis of actual decision-making procedures and is a

normative method (Biedermann et al., 2018). As far as signature or handwriting is concerned, some applied examples show the progress made in this area. For example, to quantify the probative value of well-defined measurements performed on questioned signatures, Gaborini et al. (2017) presented a way that has been formalised and is part of a coherent approach to evaluation. Chen et al. (2018) showed an applied example of signature using a score-based LR to assess signature evidence. The limitations of existing probabilistic solutions for dynamic signature evidence have been noted and explained in detail in the literature. In particular, the need to eliminate the assumption of independence between the questioned material and the reference material has been emphasized (Linden et al., 2021). As described by Robertson et al. (2016, p. 90),

We now have a system, either automated or human, that compares trace and reference specimens, reporting a LR to give the evidential value for the same-source and different-source hypotheses. We can study how much information the system is able to extract from the trace and reference material, and whether the value of that information is properly represented by the reported LR. This can tell us whether we can expect to benefit from the system at all, or it can help us to choose between different systems. For forensic scientists, it can also help measure improvement as they develop a system. Up to this point we have discussed how to assess and handle LRs and we have assumed that the LRs we are dealing with have the values that give the most rational update of the prior odds. We are now discussing reported LRs. We cannot necessarily assume that reported LRs have the properties we would expect LRs to have. A reported LR is not only a statement about the evidence but also implies a claim about how well the comparison system performs.

In forensic science, tiny pieces of evidence tend to be hidden in a mostly chaotic environment. Consequently, reasoning and deduction must be performed based on partial knowledge, approximations, uncertainties, and conjectures (Franke & Srihari, 2008). In the literature and in the above-mentioned competitions, the signature verification community has been encouraged to enable their systems to compute LRs instead of computing the usual statistics used in biometrics research (e.g., EER, FAR, and FFR).

2.6 Limitations of current research

For a long time, writer identification has mainly depended on FHEs' training and experience. The application of handwriting verification systems to casework has never taken off. The limitations of existing offline handwriting

verification approaches are summarized below.

2.6.1 Signature databases are not comprehensive enough

Current databases are limited in size or content. A database should be comprehensive and strongly representative of the forensic scenarios. In recent years, following the proposal of FHEs, several competitions collected skilled forgeries and disguised signatures. Nevertheless, existing Chinese signature databases are generally limited to 10–50 writers with 100–4800 signatures. Natural variations in genuine signatures are rarely taken into consideration or fully represented. No two signatures are exactly alike because the act of writing cannot be precisely repeated. For patterns such as face images, fingerprints, or stamps, inner variations stem from geometric changes. Handwriting is quite different, however, in the sense that within-writer variations among signatures cannot be entirely explained by geometric mapping. On the contrary, the causes of dynamic writing processes are numerous and difficult to control.

Several types of forgeries may be encountered in casework, such as skilled freehand simulation forgeries, traced simulation forgeries, and random forgeries. Again, current research lacks breadth when it comes to the nature of forgeries used to challenge the systems. In addition, factors such as the writing skill and educational background of the forgers might affect the performance of the verification system.

2.6.2 Three-dimensional and pseudo-dynamic attributes of handwriting are ignored

Previous handwriting verification research has characterized handwriting as a 2D static image. However, handwriting is the product of a behavioural process, and the contours of handwriting mainly depend on the handwriting action. Moreover, in addition to the obvious handwriting morphology, indentation (due to pressure) in the paper also exists. Thus, handwriting is essentially three-dimensional with dynamic attributes—a dimension that is absent from measurements in current research efforts.

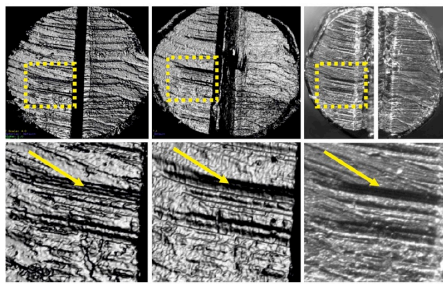
2.6.3 Flaws in the method using 2D static handwriting features

The morphology of 2D static features can be easily simulated, especially in traced simulation forgeries. As a result, high consistency is observed between genuine signatures and forgeries. The performance of handwriting verification systems based on 2D static features is severely reduced when dealing with simulation forgeries (Das et al., 2016).

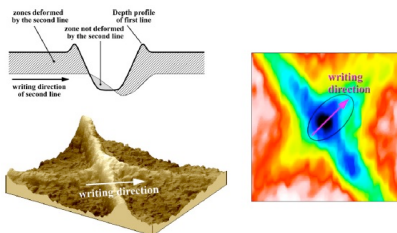
In summary, handwriting verification research is still young, and there is a wide gap between theoretical research and practical forensic applications. It is then necessary to expand offline handwriting verification research in breadth and depth, focusing on the practical needs of forensic handwriting verification. The present research will present a new method for forensic signature verification based on 3D and pseudo-dynamic features and apply it to a larger than normal corpus of signatures. We allow for the consideration of the within- and between-writer variability and develop an interpretation framework allowing the assignment of LR's.

2.7. Three-dimensional research in forensic science

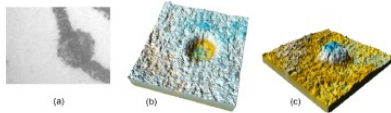
Optical imaging technology is an important tool in forensic science. It has unique advantages for the investigation, extraction, and analysis of moulded marks and striated impressions. Digital holographic technology is a new type of imaging technique, with fast, non-destructive, high-precision imaging capabilities. It has gained increasing attention in the field of forensic science. As shown in Figure 5, 3D detection and imaging has been applied to tool marks (Heikkinen et al., 2014), bullet marks (Sakaya et al., 2008; Chu et al., 2013), handwriting and paper (Spagnolo, 2006; Spagnolo et al., 2013), and other documents. Optical imaging technology can therefore provide forensic experts with an effective and enhanced tool for detection and analysis.



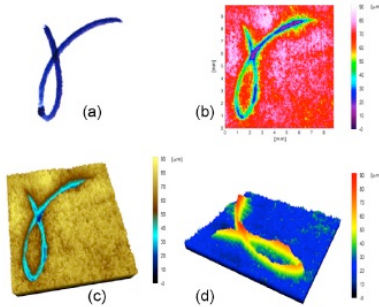
Cutter mark
(Heikkinen et al., 2014)



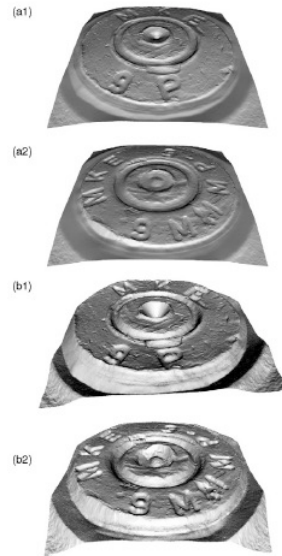
Crossing line
(Spagnolo, 2006)



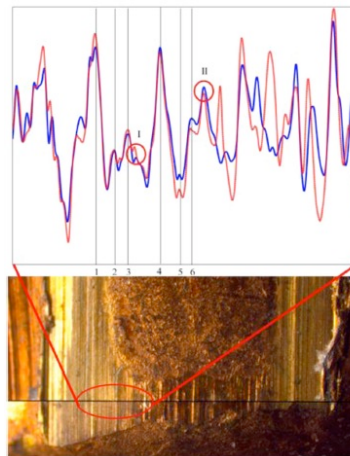
Impressions in paper
(Spagnolo et al., 2013)



Handwriting strokes
(Spagnolo et al., 2013)



Cartridge case
(Sakaya et al., 2008)

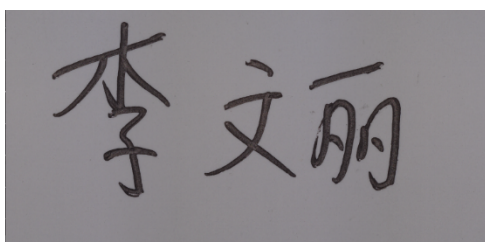


Bullet marks
(Chu et al., 2013)

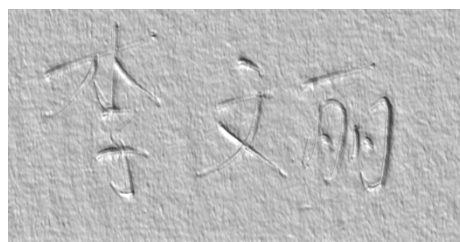
Figure 5: Applications of three-dimensional techniques in forensic science

For document examination, in particular, progress has been made in the application of 3D measurement. Spagnolo et al. (2006; 2013) reconstructed handwriting indentations in paper using laser holographic microscopy to help experts investigate handwriting strokes, identify the stroke order in crossing, and reveal the relationships between indentation depth and time. They further studied the 3D detection of handwriting, discussing variations in different writing tools and among different writers. Their results indicated that 3D features are attributable not only to different writing tools but also to differences among writers. Thus, they proposed the idea of distinguishing writers based on 3D features.

When writing instruments are used to write on paper, their tips act on the surface and reshape it microscopically, forming deep or shallow, wide or narrow groove marks on the paper. This constitutes the 3D appearance of the handwriting. By means of 3D optical measurement, the 3D profile of handwriting can be detected and acquired (Figure 6). The 3D profile measurements provide a 3D image as well as a white-light image. It is possible to obtain the stroke width, grayscale, and radian from the white-light image and depth feature in 3D images in the order of the writing sequence.



A) White-light signature image



B) Three-dimensional signature image

Figure 6: White-light and 3D signature images

Chapter 3 Methods and Materials

3.1 Chinese signatures datasets

A database characterized by representativeness, high quality, and considerable quantity was an important prerequisite of this research. The dataset included more than 140 Chinese volunteers (individuals) who agreed to participate in the study. Almost all had forensic knowledge, although not all were trained in handwriting examination. Nine proficiency tests of forensic signature handwriting examination and materials associated with 55 questioned signatures from real forensic cases were also used.

3.1.1 Signatures from volunteers

For a description of dataset_1 and dataset_2 refer to Chen (2015) and Chen et al. (2018) in the appendix). Dataset_3 is a collection of 302,000 signatures obtained from 100 volunteers. Dataset_4 was collected to explore the impact of different writing conditions on the developed system. A sample associated with a given individual consisted of a set of both genuine and forged signatures. Signatures from volunteers, both genuine signatures and forgeries, were from known sources. One hundred individual signatures were collected, including GE (5000), RF (99,000), FF (99,000), and TF (99,000) (total: 302,000). To increase the diversity of the dataset, the educational levels of the volunteers comprised three levels: primary school, high school, and university. The writing skill of the volunteers was separated into three levels: high, medium, and low.

For a given volunteer, 3,020 signatures consisting of 50 GE, 990 RF, 990 FF, and 990 TF were acquired as follows:

1. Each writer was asked to produce a total of 50 GE over the span of five writing sessions (10 signatures each). Each session was spaced by one or two days.
2. Each writer was instructed to produce forgeries for all other 99 individuals involved in the study. For each targeted individual, the forger prepared 10 RF, 10 FF, and 10 TF. In Chinese, signatures are constructed in a way similar to the writing act. Hence, knowledge of the name of the person is sufficient to produce a freehand forgery that will be legible and potentially

recognized as genuine. To prepare TF, the writer took advantage of a model of the GE and traced the forgeries on a lighted table through a transparent sheet. FFs were produced without tracing assistance after some training by the forger referring to a model GE. The amount of practice time before the production of the forgeries was not controlled.

It must be stressed at the outset that the large number of signatures per participant (3020) means that they were mass-produced to some degree. The risk of such a production is that these signatures may not have the same within-writer variation as signatures produced in forensic case circumstances. In reality writers will use different pens, different paper, different support for that paper (soft/hard). We sign on lines or in boxes, in different positions, and with varying amounts of time in between signatures (sometimes years). This limitation in the within-writer variation may lead to easier discrimination between genuine and simulated signatures within the database, and thus to performance measures that may be unrealistically high.

A specific dataset (dataset_4, see Section 4.4.3) will later investigate the use different pens and different paper to better assess that sample limitation.

The dataset taken from dataset_3 is a selection of 23,624 signatures taken from the above population according to the following process:

1. For RF signatures, since most forgeries produced by a given forger were very similar, only one forgery per forger was randomly selected.
2. For FF, RF, and TF signatures, only forgeries judged by an FHE to be close to the target signature were selected.

A sample associated with a given individual consisted of a set of both genuine and forged signatures. Table 12 gives the totals for each individual by type of signature.

Table 12: Chinese signature dataset_3

Individual ID	Genuine (GE)	Traced simulation			Total
		Freehand simulation forgery (FF)	simulation forgery (TF)	Random forgery (RF)	
1	50	99	98	98	345
2	50	30	99	33	212
3	50	31	98	52	231
4	50	35	99	30	214
5	49	30	99	26	204
6	49	47	86	92	274
7	50	34	99	54	237
8	51	30	99	30	210
9	50	39	99	56	244
10	57	41	99	42	239

11	54	35	99	36	224
12	50	99	99	99	347
13	50	31	99	33	213
14	50	30	99	32	211
15	50	34	97	48	229
16	50	43	30	71	194
17	50	30	99	71	250
18	54	99	97	99	349
19	54	31	98	35	218
20	50	50	99	40	239
21	50	30	99	39	218
22	50	32	95	43	220
23	50	37	99	43	229
24	50	31	99	42	222
25	50	98	99	98	345
26	52	35	99	38	224
27	50	32	99	32	213
28	50	24	97	66	237
29	50	65	99	56	270
30	51	99	98	106	354
31	50	40	99	40	229
32	50	32	98	47	227
33	50	22	98	68	238
34	50	40	99	41	230
35	50	99	98	99	346
36	49	41	98	41	229
37	/	/	/	/	/
38	50	99	99	97	345
39	/	/	/	/	/
40	50	33	97	41	221
41	49	31	98	36	214
42	52	47	98	81	278
43	60	56	99	74	289
44	50	30	99	30	209
45	55	38	98	47	238
46	50	45	99	53	247
47	50	51	97	66	264
48	50	27	97	48	222
49	56	32	98	36	222
50	39	25	99	29	192
51	48	33	99	33	213
52	50	41	98	38	227
53	50	25	99	37	211
54	50	33	99	30	212
55	50	97	99	56	302
56	/	/	/	/	/
57	50	35	97	42	224
58	60	32	98	41	231
59	60	32	99	29	220
60	53	47	97	43	240
61	50	38	99	67	254
62	50	33	99	38	220
63	51	30	99	31	211
64	50	30	99	31	210
65	50	32	99	33	214
66	50	61	76	64	251

67	60	99	99	98	356
68	50	30	99	31	210
69	50	99	99	98	346
70	49	99	98	107	353
71	50	30	99	33	212
72	49	54	97	29	229
73	49	37	99	30	215
74	52	30	99	30	211
75	49	33	99	46	227
76	53	40	99	41	233
77	51	43	99	46	239
78	52	29	98	45	224
79	59	26	99	33	217
80	50	33	99	33	215
81	50	33	99	32	214
82	50	31	99	34	214
83	50	32	99	32	213
84	59	27	99	30	215
85	50	99	97	95	341
86	56	31	99	32	218
87	50	98	99	98	345
88	50	28	92	52	222
89	49	37	98	44	228
90	50	41	99	42	232
91	60	30	98	39	227
92	53	29	99	34	215
93	52	33	99	40	224
94	50	29	98	37	214
95	50	59	99	67	275
96	49	38	98	40	225
97	50	49	99	87	285
98	58	33	99	32	222
99	50	65	98	60	273
100	60	38	96	45	239

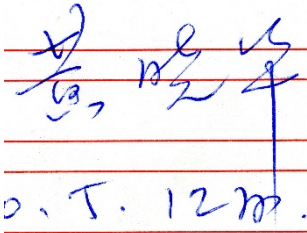
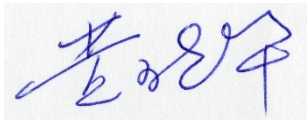
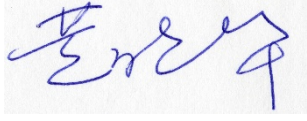
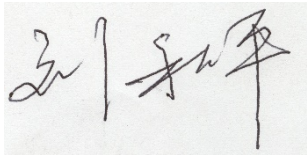
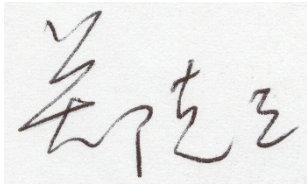
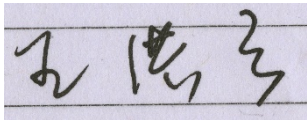
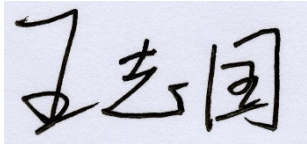
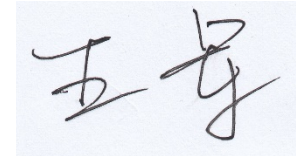
Note: / means damaged or illegible signature.

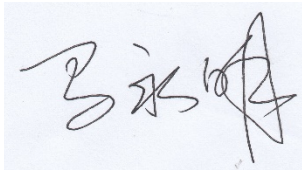
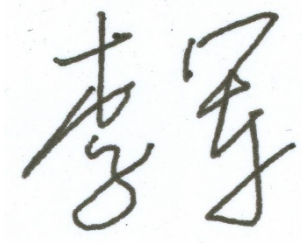
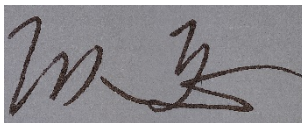
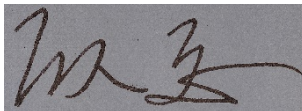
3.1.2 Signature dataset from proficiency tests

A proficiency test is an important instrument used by accreditation bodies to assess the technical competence of laboratories. Proficiency test providers prepared simulated cases for the participants. All the information about their production is known. The Forensic Science Academy is the proficiency test provider for the China National Accreditation Service for Conformity Assessment (CNAS). CNAS signature proficiency tests from 2011 to 2020 were used to assess the performance of the developed offline system under realistic but known forensic conditions (Table 13).

Table 13: Proficiency test of Chinese signatures

PT provider	Year	Questioned signature	Signature image	Expected result
-------------	------	----------------------	-----------------	-----------------

CNAS	2011	1		Different sources
CNAS	2013	1 st of 2		Different sources
		2 nd of 2		Different sources
CNAS	2014	1 st of 2		Same source
		2 nd of 2		Same source
CNAS	2015	1		Same source
<u>PT provider</u>	<u>Year</u>	<u>Questioned signature</u>	<u>Signature image</u>	<u>Expected result</u>
CNAS	2016	1		Different sources
CNAS	2017	1 st of 2		Different sources

		2 nd of 2		Same source
CNAS	2019	1		Different sources
CNAS	2020	1 st of 2		Different sources
		2 nd of 2		Different sources

3.1.3 Real forensic cases

Fifty-five questioned signatures from real forensic cases were selected randomly from the Academy of Forensic Science, China. The developed system will be used to provide results for these questioned signatures. A comparison of the results of the system with those of the FHEs will allow for an assessment of how FHEs can be of assistance in their daily casework.

3.2 Acquisition and reconstruction of 3D images of signatures

Two types of 3D measurement systems were tested to meet the requirement for 3D acquisition and reconstruction (Figure 7):

- A Lyncee Tec DHM reflection-configured digital holographic microscope R series and;
- A Keyence wide-area 3D measurement system VR3000 series.

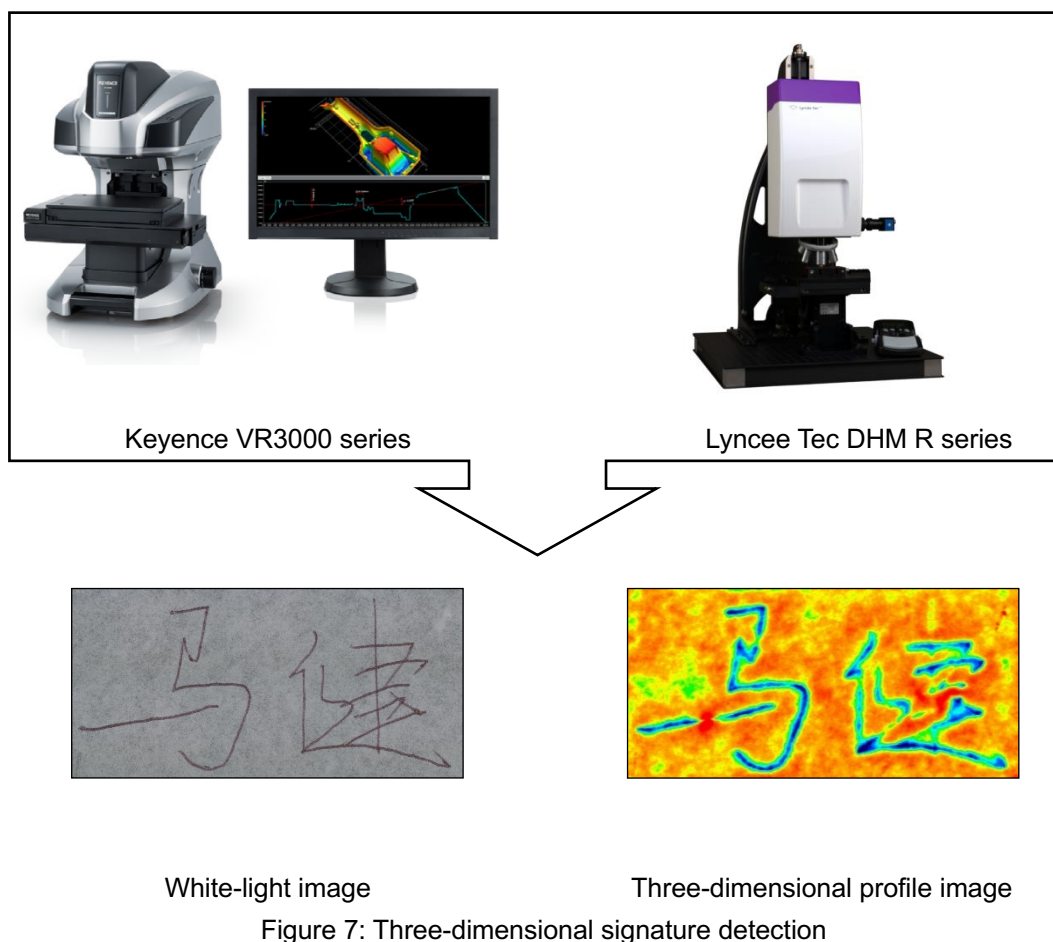


Figure 7: Three-dimensional signature detection

3.2.1 Reflection holographic microscopes: Lyncee Tec R2200⁸

The Lyncee Tec DHM reflection-configured digital holographic microscope R series is manufactured by Lyncee Tec. This customized system offers two simultaneous measurement modes: DHM and colour-intensity images (Figure 8).

⁸ <https://www.lynceetec.com/reflection-dhm/#tab-1>

Lyncée tec 4D Profilometry Biological Imaging Products Applications Technology Company Downloads

FORENSIC DOCUMENT EXAMINATION

Signature 3D mapping for forensic document examination

Signature 3D topography is becoming a new trend in forensic research. It is a powerful method for forensic document examination. The surface to measure for characterizing an entire specimen is often very large. Considering the structure and the signature imprint depth on paper, a relatively good lateral resolution is needed. Therefore, stitching over many images to retrieve the full field of view is necessary.

In this application, a dedicated system and user interface have been developed for the Institute of Forensic Sciences in Shanghai, China. This customized system offers two simultaneous measurement mode: DHM and color intensity images. The benefit of it is to acquire real color information and height information at each pixel, providing a new perspective and dream tool for forensic document examination.

Description:

- > **Courtesy of:** Institute of Forensic Sciences in Shanghai, China
- > **Material:** Normal paper
- > **Typical signature imprint depth:** 1-20 microns
- > **Automated Process**
 - > Loading and pre-positioning of the paper at a right corner of the signature area
 - > Automated scan of the signature (typ. 50x200 individual measurements)
 - > Automated stitching of both 3D topography and color image
- > **Instrument:**
 - > DHM® R-2200 with Motorized stage 300x300x38 mm
 - > Dual camera (Monochrome and white light)
- > **Software**
 - > Customized user interface dedicated to the application.
- > **Time scale** for full signature scanning (20 mm x 30 mm): <60 s
- > **Magnification:** 10x

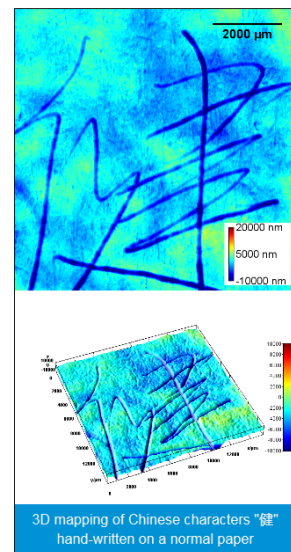
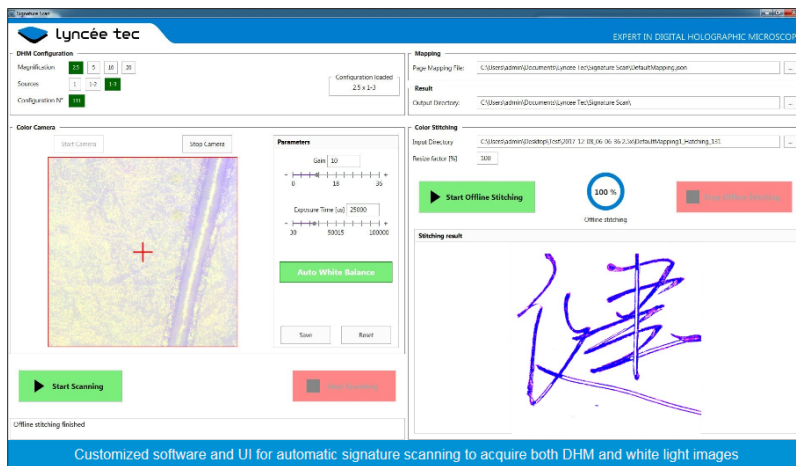
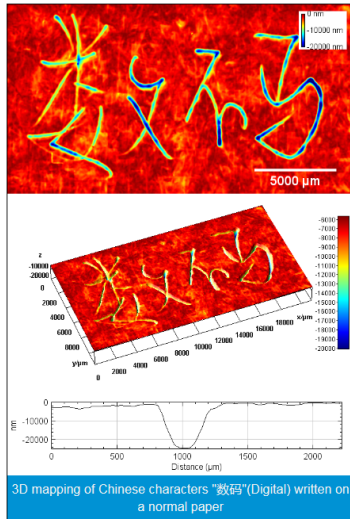


Figure 8: Digital hologram microscopy for forensic document examination: 3D signature mapping

The benefit of DHM is that it can acquire real colour information and height information at each pixel, providing a new perspective and tool for quantitative measurement. Colour images provide 2D images of signatures. The skeletons and edges of the signatures can be obtained by image processing, and writing sequence tracing can be performed on the signature colour image. Then, width, grayscale, and radian can be obtained on the writing sequence order from the colour image. Finally, thanks to the white-light and the synchronous imaging of the 3D profile, data of the 3D profile can also be measured on the writing

sequence order in 3D profile images.

3.2.2 Wide-area 3D measurement system: Keyence VR 3200⁹

In the Keyence wide-area 3D measurement system, structured light is emitted from the transmitter lens and projected onto the surface of the object. The reflected light is then detected by the receiver lens and appears banded and bent based on changes in the topography of the surface. Triangulation is then used to calculate and measure the height of the surface. To enable high-accuracy measurements throughout the field of view, the VR Series uses a telecentric lens with low lens aberration. Objects can be captured as they appear and at their actual size, ensuring high measurement accuracy anywhere on the screen. Based on the light-section method of measurement, the VR Series calculates data down to 1 pixel or less using proprietary light-projection patterns. This results in highly accurate, ultra-precise measurement. The VR Series' ability to accurately measure height differences of only 1 μm has been confirmed through the measurement of a calibrated height difference gauge.

3.2.3 Two-dimensional white-light images and 3D image acquisition

The results obtained on different types of paper surfaces allow us to evaluate the complementarity between the two instruments. Three-dimensional detection results showed that the Lyncee Tec DHM is good at measuring handwriting on specular reflection surfaces, and the Keyence wide-area 3D measurement system is better at measuring handwriting on diffuse reflective surfaces.

Two-dimensional white-light images and 3D images were acquired for all signatures using the Lyncee Tec DHM reflection digital holographic microscope R series and the Keyence wide-area 3D measurement system VR3000 series. The Keyence device offers a lower resolution (0.5 μm) than the Lyncee Tec DHM (0.3 μm) but operates with a higher acquisition speed (10 seconds versus five minutes). The first batches of signatures (6040 signatures from two individuals) were acquired using the Lyncee Tec. Then, it was switched to the Keyence instrument to reduce acquisition time without losing much in terms of accuracy. The higher resolution acquisitions from Lyncee Tec were reduced to the resolution of the Keyence device. In that configuration, the data were fully comparable.

⁹ <https://www.keyence.com/>

3.3 Extraction and post-processing of three-dimensional and pseudo-dynamic features

As explained before, handwriting, including the production of signatures, is the result of a dynamic behaviour. This behaviour materializes on paper in the form of static traces that are submitted to FHEs. FHEs then reconstruct the dynamic writing sequence based on an optical analysis of the traced images. This means that the operator-independent features used to characterize the handwriting should capture the dynamic nature of the behaviour and not rest only on static measurements, such as the relative proportions, sizes, and shapes of letters. Prior research has focused mostly on static features such as contour, gradient, and slope direction but has neglected the writing sequence.

Writing sequence is a new feature that will be measured in this study. This study also takes advantage of dynamic time warping techniques to capture features while maintaining the writing sequence. As already mentioned these are qualified as “pseudo-dynamic features” because they are extracted while considering the writing order. They are not extracted at the time of capture but acquired after the writing act from the images themselves. These features are different from dynamic features extracted from traditional online handwriting systems, such as writing tablets, but they still reflect the writing sequence; hence the name ‘pseudo-dynamic features.’

Writing sequence recovery is performed on the 2D image, and the skeleton and edge of the strokes in signatures are obtained. Given a starting point, a trace of the writing sequence of the strokes will be obtained automatically. The width, grayscale, and radian characterize the order of the writing sequence. After aligning the 3D image and the 2D image, the 3D feature can also be associated with the order of the writing sequence.

3.3.1 Writing sequence tracing

In previous research (Chen X., 2015), signatures were digitalized into a computer by means of an Epson Perfection V700 Photo Scanner with a resolution of 400 dpi. MATLAB 7.0 software was used to extract the features. The same process has been used in this research. First, a threshold is set for the image binarization. Then, the skeletons and edges of the signatures were extracted by separately skeletonizing and edging the images (Figure 9). The stroke order tracing was automatically used in the signature skeletons after a beginning point was manually provided (Figure 10). If any error occurred, it was corrected manually. The stroke order or the sequence of signatures (S) was assigned x coordinates (X) and y coordinates (Y).

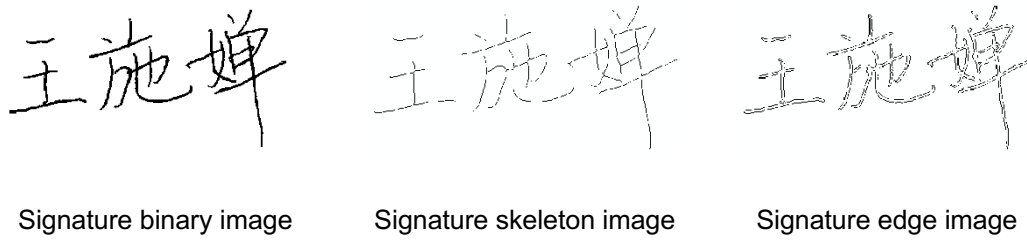


Figure 9: Image pre-processing

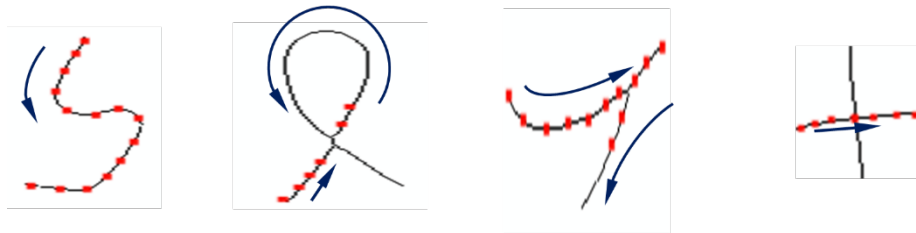
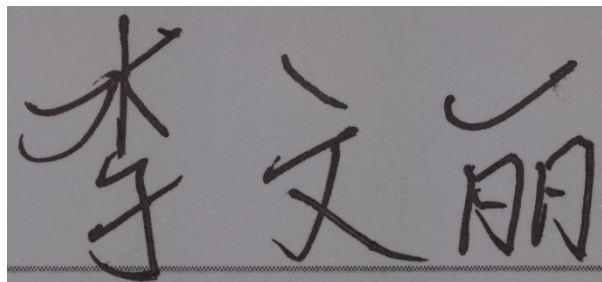


Figure 10: Basic stroke order recovery processing from an image

Writing instruments have various types of tips and can produce different strokes, mainly in terms of line width and depth. A previously used toolkit (Chen, 2015) could not determine the skeletons of handwriting done using big tips and wide lines. It did not work until the image was resized to be smaller. Meanwhile, ink does not always distribute evenly, which also affects the quality of the extracted skeleton. This research improved upon previously developed algorithms to meet these new requirements, a pre-processing step was added before skeletonization, such as image erosion using disk-shape structuring element, 2D image filtering to filter noise, and image dilation with structuring element. To obtain a suitable image for skeletonization, a series of parameters in pre-processing can be tried and set manually according to different aspects of the signature (Figure 11).



A) White-light image

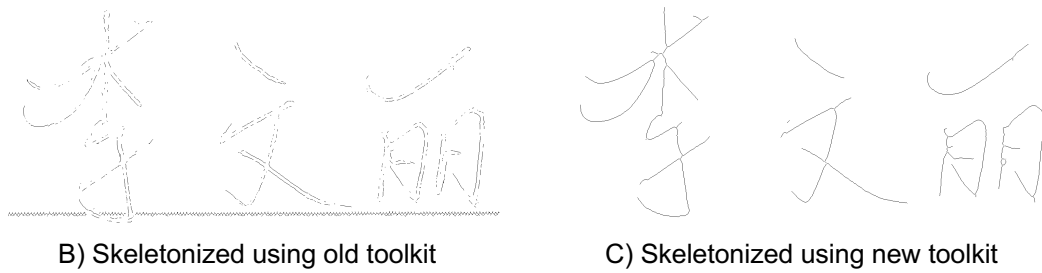


Figure 11: Improvement in the skeletonization process

Different writing surfaces lead to different image backgrounds in actual situations. This research used additional image-processing techniques to filter the background.

With the previous toolkit (Chen X., 2015), the tracing of the writing sequence was done pixel by pixel. When the resolution is high, and number of images increases to millions, the tracing procedure takes quite a lot of time. This project split the strokes and aligned the pixels of each stroke in advance (Figure 12). When a starting pixel is given, it traces the stroke one by one, which makes tracing faster (see Table 14: The comparison between the old and new toolkit.).

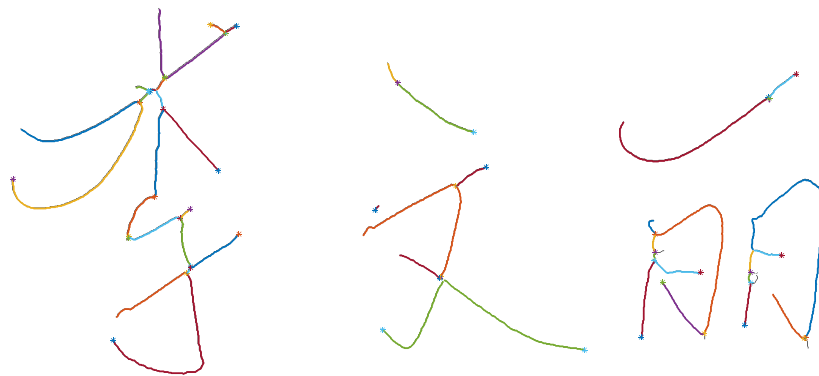


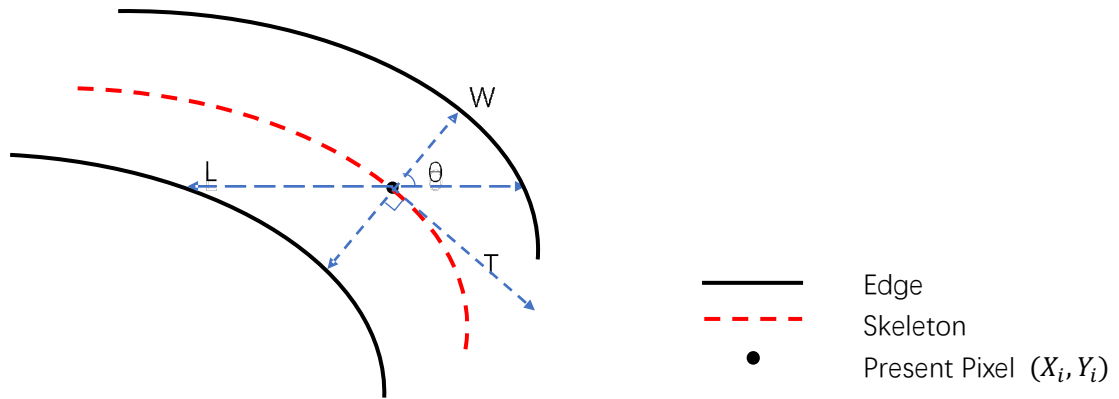
Figure 12: Split strokes in a skeletonized image

Table 14: The comparison between the old and new toolkit.

Toolkit	Pre-processing	Parameters	Stroke Tracing
Previous research (Chen X., 2015)	No	Fix	Pixel by pixel
Current research	Yes	manually	Stroke by stroke

3.3.2 Features extraction in white-light and 3D signature images

After writing sequence acquisition, the width (W), grayscale (G), and radian (R) values were automatically extracted in the stroke order. The grayscale values of the points in the skeleton were used as the grayscale data. The width and radian data were calculated per the following functions, as shown in Figure 13.



$$S = \sum_{i=1}^n (X_i, Y_i); \quad \text{Equation 1}$$

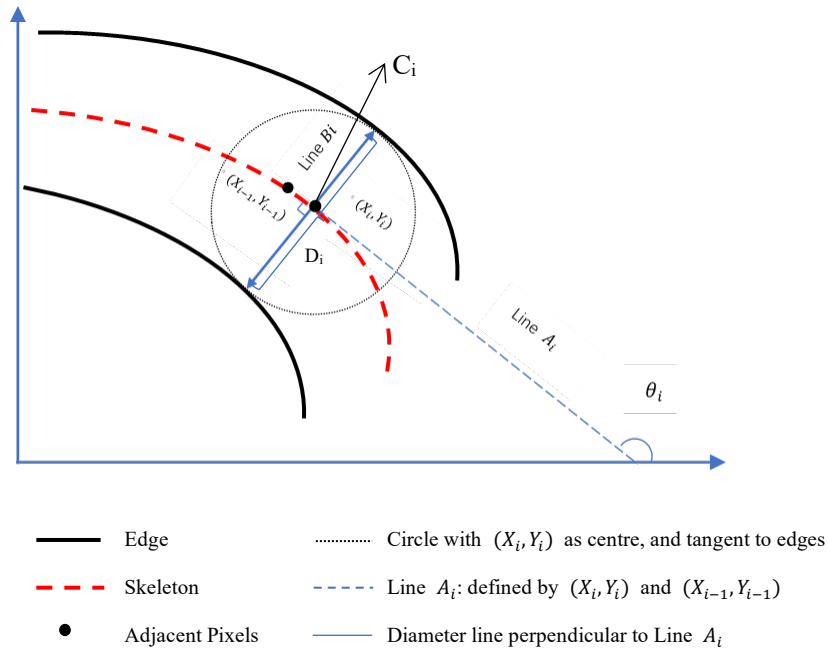
$$W = \sum_{i=1}^n (w_i); \quad \text{Equation 2}$$

$$R = \sum_{i=1}^n \left(\theta_i \cdot \frac{\pi}{180} \right); \quad \text{Equation 3}$$

$$F = \sum_{i=1}^n (W_i, G_i, R_i). \quad \text{Equation 4}$$

Figure 13: Previous features detection. (Stroke order (S); x coordinate (X); y coordinate (Y); width (W), grayscale (G), radian (R); feature (F); tangent line (T). n =length (X) or length (Y); $i=1,2,3\dots n$)

In the previous features detection, grayscale was calculated on the skeleton pixels of the signatures. In this research, grayscale and depth were measured for all pixels of the signatures (Figure 14). This provides more specific features analysis.



$$W = \sum_{i=1}^n D_i. \quad \text{Equation 5}$$

$$C_i = \sum_{i=1}^n (x_i, y_i) \quad \text{Equation 6}$$

$$G = \sum_{i=1}^n G_{B_i}; G_{skel} = \sum_{i=1}^n G_{C_i} \quad \text{Equation 7}$$

$$H = \sum_{i=1}^n H_{B_i}; H_{skel} = \sum_{i=1}^n H_{C_i} \quad \text{Equation 8}$$

$$R = \sum_{i=1}^n \left(\theta_i \cdot \frac{\pi}{180} \right) \quad \text{Equation 9}$$

$$F = \sum_{i=1}^n (W_i, G_i, G_{skel_i}, R_i, H_i, H_{skel_i}) \quad \text{Equation 10}$$

Figure 14: New features detection. (Stroke order (S); x coordinate (X); y coordinate (Y); width (W), grayscale (G), radian (R); tangent line (T); height (H); n=length (X) or length (Y); i=1, 2, 3...n)

3.3.3 Pseudo-dynamic features visualization

Data associated with the signatures include a combination of 2D white-light and 3D information, and were obtained using a 3D imaging system coupled

with a 2D white-light imaging system operating in a synchronous manner. Hence, at each point of the acquired images, information regarding grey levels and 3D can be extracted. Data acquisition was followed by feature extraction, which entailed the following image-processing steps:

1. A signature is composed of multiple strokes corresponding to movements used by the signatory. These strokes and their sequences were identified by operators (using a dedicated interface). Each signature stroke was defined by a skeleton obtained from the grayscale acquisition. Strokes could then be stitched together by their order on a line defined by the skeletons. This approach ensures that the data fully maintain the writing sequence (see Supplemental Video).
2. From each stroke, pseudo-dynamic features were obtained from white-light images in the form of the width, radian, and grayscale distribution measured at each point of the stroke. These features are called pseudo-dynamic because they indirectly reflect the dynamics of each stroke but are visualized in 2D. These features have been detailed in previous work (Chen X., 2015; Chen X. et al., 2018)
3. Taking advantage of the synchronous acquisition of the 3D image, for each point on the stroke, the height information was measured in addition to the pseudo-dynamic features (Chen X. et al., 2018).

Figure 15 shows an example of the acquisition, showing two genuine signatures (GE-1 and GE-2 from the same writer) alongside a corresponding freehand forgery (FF) and a traced forgery (TF), obtained in white-light (left) and with 3D measurements (right).

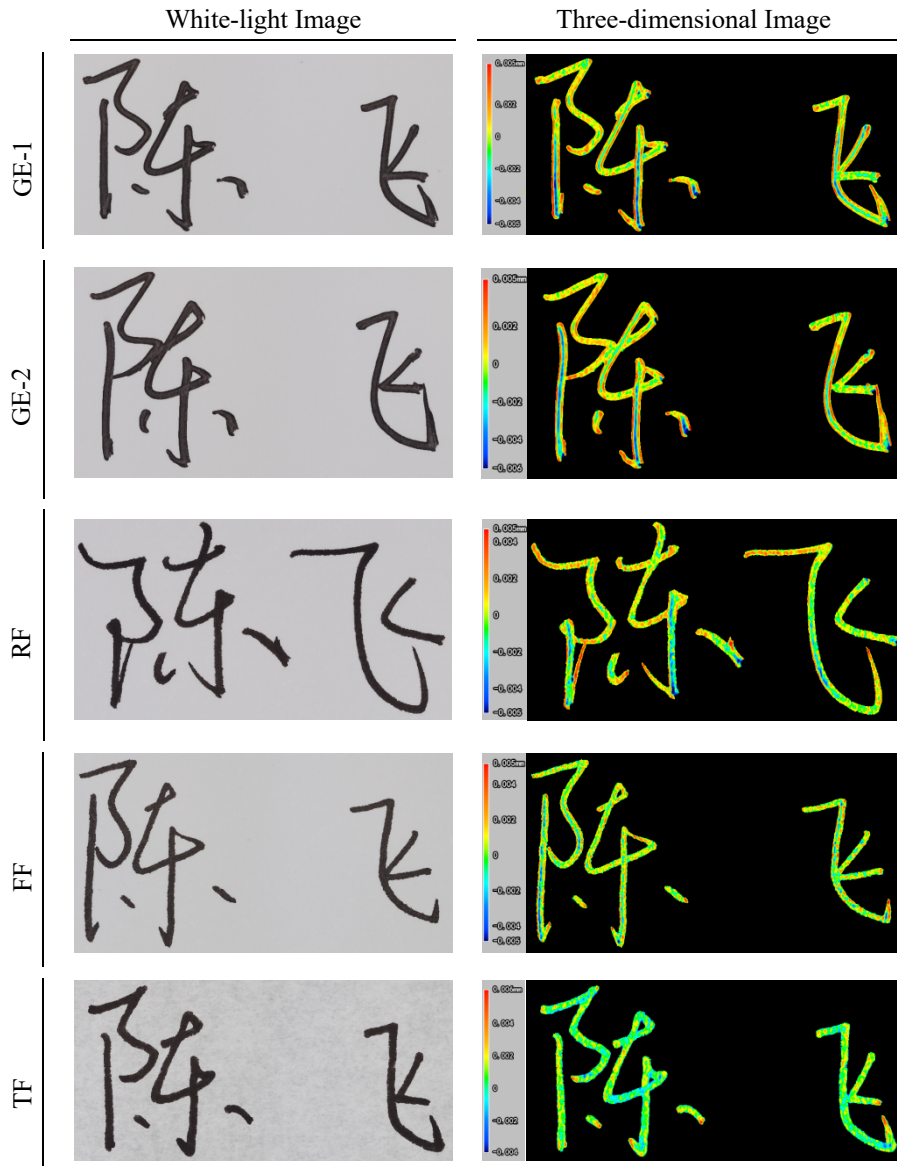


Figure 15: White-light and three-dimensional profile of a signature
(Background of the paper was omitted in the three-dimensional profile image)

Some features are obtained from the white-light images, others from the 3D measurements. More specifically,

1. Figure 16 shows the radian measurements obtained from the white-light images for genuine, random forgery, freehand simulation forgery, and traced simulation forgery.

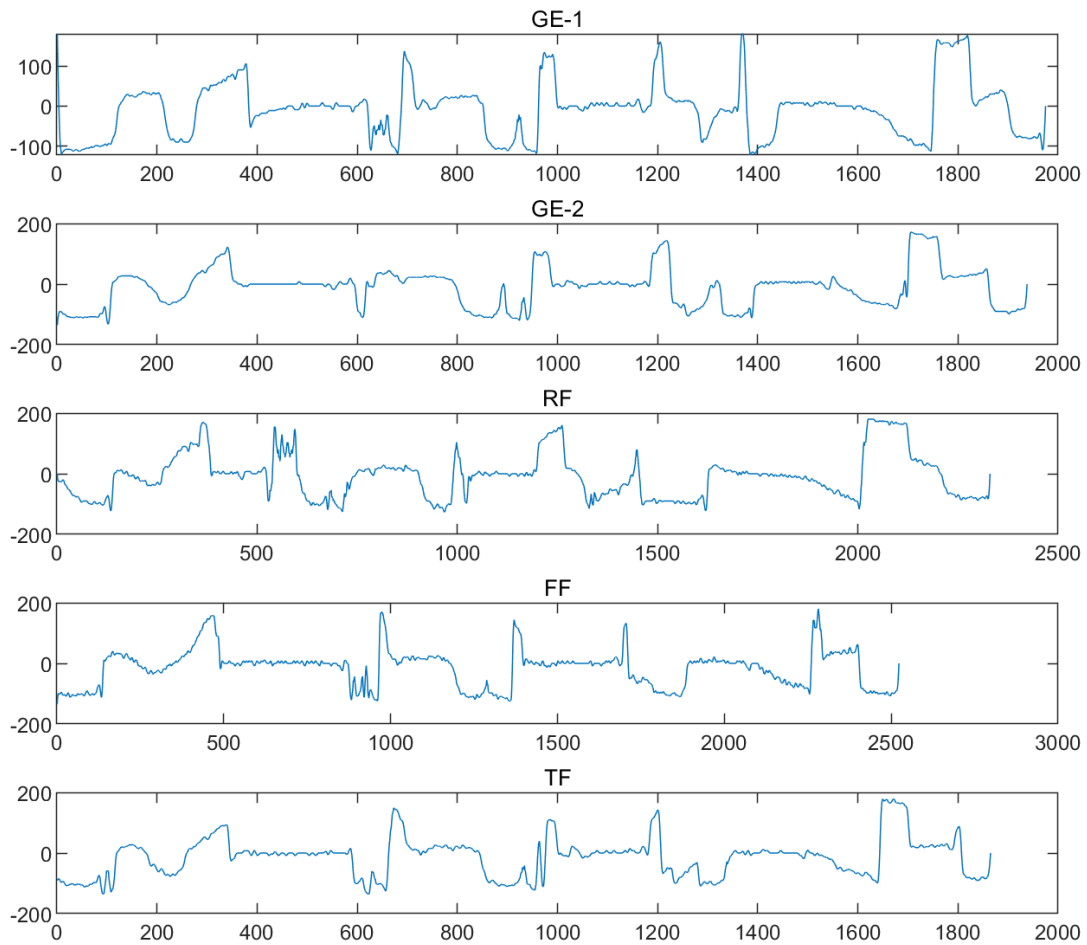


Figure 16: Radian features from the white-light images

2.

Figure 17(a) shows the grayscale distribution on the stroke measured from the white-light images. The y-axis gives the width of the stroke.

Figure 17(b) shows a normalized representation.

3.

Figure 18(a) shows the 3D measurement of height made along the strokes, with the indication of stroke width on the y-axis.

Figure 18(b) shows the normalized measures.

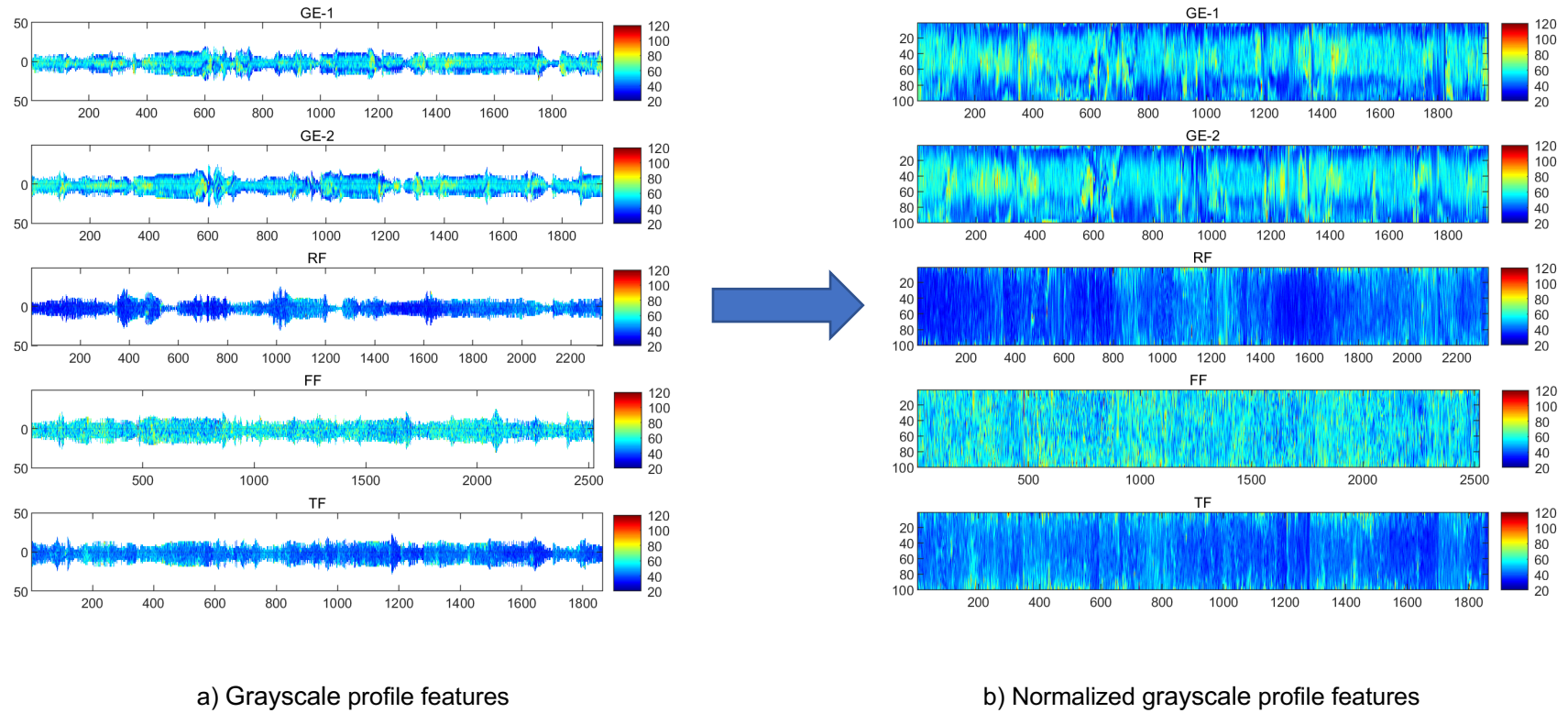
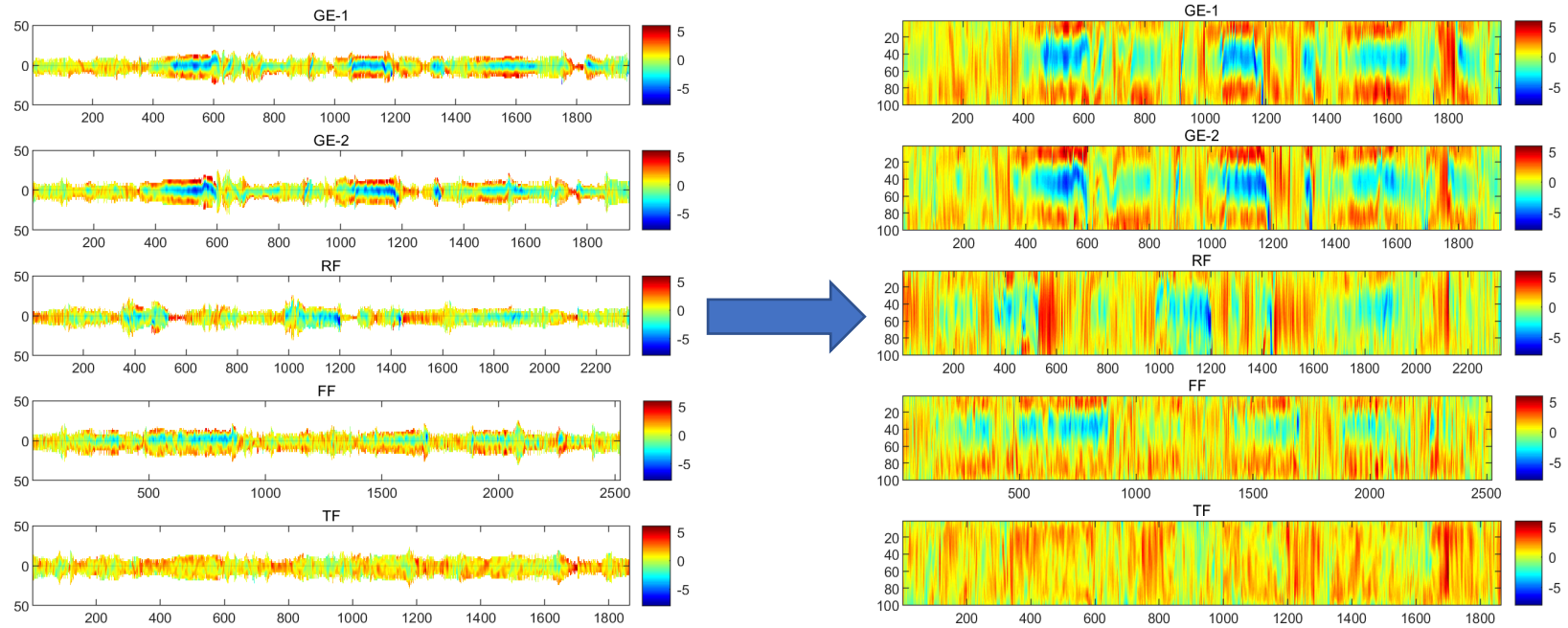


Figure 17: Grayscale profile features of strokes measured from white-light images



a) Three-dimensional profile features

b) Normalized three-dimensional profile features

Figure 18: Three-dimensional profile features of strokes measured from the three-dimensional profile images.

Of particular interest is the difference between the 3D profile images (Figure 15, right) of the two genuine signatures and the forgeries. It is more difficult to observe these differences based on the white-light images alone (Figure 15, left). For the grayscale and height distribution, the genuine signatures show more similarities compared to the forgeries. The similarities between the genuine signatures are maintained, while differences with the forgeries are enhanced. As seen on GE-1 and GE-2 in Figure 15 to

Figure 18, in the white-light image, the morphology of the signature for a given writer is variable in shape and stroke length and in the specific positions between strokes. In the pseudo-dynamic images, they show more within-writer similarities. For the between-writer variations, a random forgery (without the model, see RF in Figure 15 to

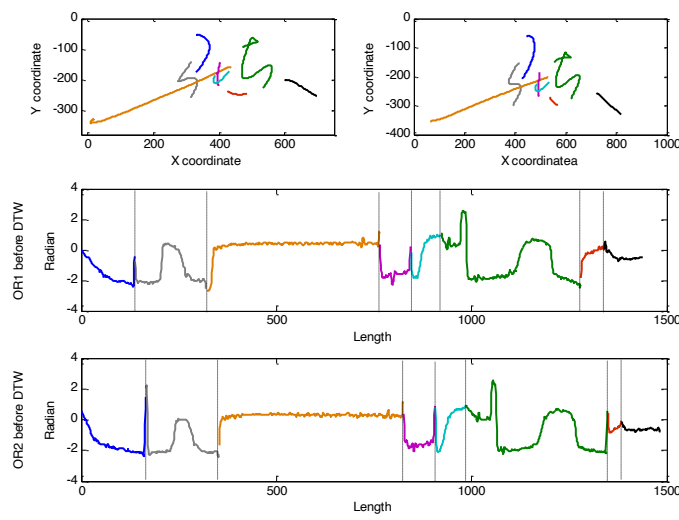
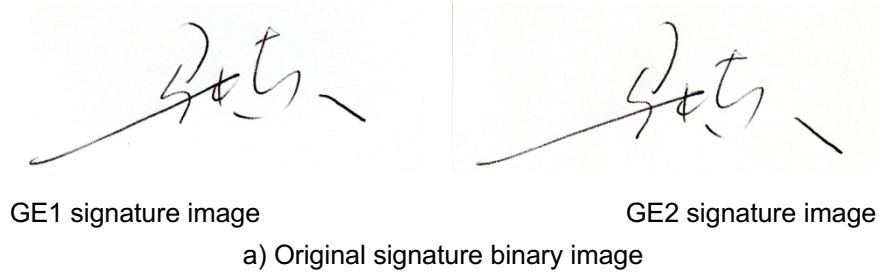
Figure 18) shows a different morphology in the white-light image, whereas a freehand simulation forgery and traced forgery (with a model, see FF and TF in Figure 15 to

Figure 18) could present high similarities with the genuine signature. This is even more true with a traced forgery, which can present almost the same morphology as the genuine signature. The similarities in contour between a traced forgery and a genuine signature can be higher than between genuine signatures. In pseudo-dynamic images, however, between-writer differences are enhanced.

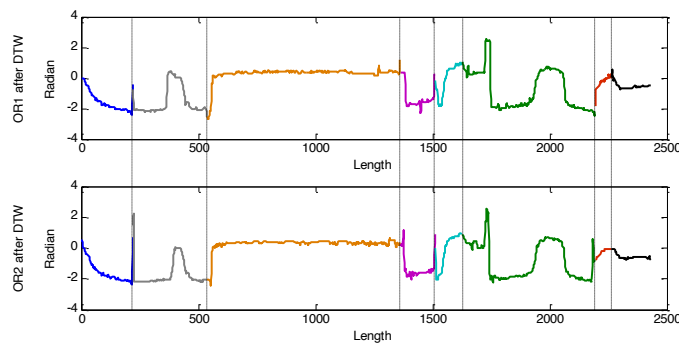
3.3.4 Post-processing of three-dimensional and pseudo-dynamic features

Writing speed can be different, even when the signatures come from the same writer (Figure 19(a) and 19(b)). Dynamic time warping (DTW) provides a flexible and non-linear telescoping to get the minimum distance between two time series with different lengths.

DTW is used to cope with different writing speeds between two signatures. As a result of DTW, the metric obtained gives the maximum similarity. Corresponding strokes should be aligned for comparative analysis. The shape of the strokes could be reflected by the radian feature. DTW is used to make the corresponding strokes aligned based on the radian feature (Figure 19(c)).



b) Stroke alignments before DTW



c) Stroke alignment after DTW

Figure 19: DTW processing

3.4 Statistical description of features

This research has been ongoing for a few years and has led to an expansion

of the size and diversity of the database and the addition of new features to characterize the characteristics of handwriting. Table 15 presents the databases and features used in different stages of the research.

Table 15: Variables and datasets used at the different stages of the research.

Variables	Denoted	Dataset 1 (Chen X., 2015)	Dataset 2 (Chen X. et al., 2018)	Dataset 3
Width	W	√	√	√
Radian	R	√	√	√
Grayscale	G	×	×	√
Height	H	×	×	√
DTW	DTW	×	×	√
Grayscale_skeleton	G_skel	√	√	√
Height_skeleton	H_skel	√	√	√
KST_Width*	KST_W	×	×	√
KST_Radian*	KST_R	×	×	√
KST_Grayscale*	KST_G	×	×	√
KST_Height*	KST_H	×	×	√
KST_Grayscale_skeleton*	KST_G_skel	×	×	√
KST_Height_skeleton*	KST_H_skel	×	×	√

*: KST means statistics in Kolmogorov–Smirnov test

In statistics, a correlation coefficient measures the strength and direction of a linear relationship between two variables. In this research, the correlation coefficient is used to characterize similarities between signatures. For GE, for example, the correlation coefficients of one genuine signature were calculated between each possible pair of this genuine signature and other genuine signatures. The maximum value is used as the final correlation coefficient of the genuine signature. For forgeries (FF, RF, and TF signatures), for example, the correlation coefficients of one forgery signature were calculated between each possible pair of forgeries and all genuine signatures. The maximum value is then used as the final correlation coefficient of this forgery signature. All statistical analyses described below were performed on the correlation coefficients of the width, grayscale, and radian data obtained. MATLAB 2020a was used for this correlation analysis.

3.4.1 Multivariate analysis of variables and discriminant analysis

For the multivariate analysis of variance (MANOVA), we invite the reader to refer to Chen (2015) in the appendix. A graphical representation per writer of their within-genuine variability and their genuine-against-forgery variability using

2D KDE after PCA is presented in Chen et al. (2018).

To measure the ability of these features to carry out forensic tasks, cases of known ground truth from available materials were produced, as described in Chen (2015) and Chen et al. (2018). This involved the random selection of a questioned signature with a genuine signature or any forgeries compared with five genuine references from that individual. In Chen et al. (2018), care was taken to construct within-writer distribution using only signatures from the individual under consideration. Besides, between-writer distribution was obtained by comparing all genuine signatures against the forgeries of that individual. In this study, this approach is referred to as ‘inner-individual’. Given that in forensic practice, it can be difficult to obtain sufficient samples from a given individual to carry out this first method, a second method was tested, referred to as “inter-individual”, where within-writer distribution is made up of all the pairwise comparisons between genuine signatures of all the individuals in the dataset. The between-writer distribution is obtained from the comparisons between all genuine and forged material from all individuals in the dataset. Thus, for the second method, there is a generic within-writer distribution and a generic between-writer distribution that do not depend on the individuals used for the simulated cases. By comparing the two methods, it can be assessed if, in cases in which limited references are available, generic underpinning distributions could still be used.

3.4.2 Descriptive and Comparative measurement

Using data directly measured from subjects for analysis is common in research in many disciplines. In our case, variables such as width, grayscale, radian, and height can be described as descriptive measurement.

As described in Chen (2015), DTW is used to make the corresponding strokes aligned and matched for correlation analysis. The correlation coefficient between signatures is then calculated. For instance, a number m of genuine signatures was used as the reference signatures; a number n of unknown signatures was used as the questioned signatures. For each genuine signature, there were $m-1$ measurements of width, grayscale, radian, and height between reference signatures. In other words, the proximity between two signatures is obtained by the computation of the correlation on the paired variables. Additionally, the distance between two signatures during DTW processing is used as one feature.

The two-sample Kolmogorov–Smirnov test (Massey, 1951; Miller, 1956; Marsaglia et al., 2003) is used to further describe differences between the distributions of features in genuine signatures vs forgeries. More specifically, the

KSTest function in MATLAB 2020a allows to evaluate the difference between the cumulative distribution functions of the distributions of the two sample data vectors. The test statistic *KS2stat* is used to describe the similarity between the two vectors. For each genuine signature, there were $m-1$ measurements of the *KSTest* in width, grayscale, radian, and height between reference signatures.

To summarize the metrics used, for each questioned signature, there are m measurements of correlation with regards to width, grayscale, radian, and height. Additionally, there were m *KSTest* measurements with regards to width, grayscale, radian, height, and DTW between the questioned signatures and m reference signatures.

Hence, for a comparison between one questioned signature (Q) and three specimens (S1–S3), the following matrix is obtained.

	Width*	grayscale*	radian*	height*	DTW**
Q – S1	CW _{Q-S1}	Cg _{Q-S1}	Cr _{Q-S1}	Ch _{Q-S1}	D _{Q-S1}
Q – S2	CW _{Q-S2}	Cg _{Q-S2}	Cr _{Q-S2}	Ch _{Q-S2}	D _{Q-S2}
Q – S3	CW _{Q-S3}	Cg _{Q-S3}	Cr _{Q-S3}	Ch _{Q-S3}	D _{Q-S3}

* Features calculated by correlation coefficient.

** Euclid distance feature calculated by DTW.

KSTest	width	grayscale	radian	height
Q – S1	KSTw _{Q-S1}	KSTg _{Q-S1}	KSTr _{Q-S1}	KSTh _{Q-S1}
Q – S2	KSTw _{Q-S2}	KSTg _{Q-S2}	KSTr _{Q-S2}	KSTh _{Q-S2}
Q – S3	KSTw _{Q-S3}	KSTg _{Q-S3}	KSTr _{Q-S3}	KSTh _{Q-S3}

The feature vector is based on the above measurements but obtained in a comparative way after the comparison between two signatures. When two signatures are compared, the following nine variables are computed: four correlations respectively for the radian, width, grayscale, and height, and a normalized DTW using the radian measures from both signatures as described in Chen (2005) and Chen et al. (2018), and four Kolmogorov–Smirnov distance measures (KST) for the comparison of respectively width, grayscale, radian, and height.

Statistical analysis methods are used to measure within- and between-writer variation. Analysis of variance is used to validate significant between-writer differences in pseudo-dynamic and 3D features. This was the basic initial step towards determining the specificity of 3D and pseudo-dynamic features. Multivariate distributions, kernel-based distributions, and other distributions are used to estimate the probability density of 3D and pseudo-dynamic features to investigate within- and between-writer differences. This is the key component of

this research and is used in concert with the evaluation, provided it was considered reasonable to take 3D and pseudo-dynamic features as the basis for individual discrimination.

The comparative measurements allowed for measuring variability in handwriting, both within genuine sources and compared to forgeries.

3.5 ML method for signature verification

In addition to the next investigation of score-based LRs using the above variables, the nine selected features informed a Machine Learning (ML) strategy in which the ability to distinguish comparisons arising from a common source (a given individual) from comparisons involving forgeries (which also are associated with each individual) has been investigated. That use of ML corresponds to the use of the system as a biometric system for automatic identification. A range of ML classifiers were tested, from low complexity (high level of explainability) models to high complexity (less explainability) models. ML and subsequent statistical analysis were carried out in *R* version 4.0.2 (R core team, 2022) coupled with *RStudio* version 1.3.959 (RStudio Team, 2020) using the following packages: *tidyverse* (Wickham, 2019) for data wrangling, *caret* for ML, computing confusion matrices, and associated error statistics (Kuhn, 2021). The following models were tested: K-nearest neighbour (*knn*), four options of discriminant analysis (*lda*, *rda*, *mda*, *qda*), a naive Bayes classifier (*nb*), tree-based models (*rpart*, *gbm*, *C5.0*, *xgbLinear*, *R-Forest*), support vector machine (*svm Radial*), and neural networks (*nnet*, *avNNet*, *pcaNNet*). The abbreviations in parentheses correspond to the methods used in *caret*. The associated packages were loaded as required. Nine variables were used (setting aside the *_skeleton* variables). The predictor variables were scaled and centred as part of the pre-processing. Parameter tuning was achieved on a training set composed of half of the individuals using leave-one-out cross-validation. The leave-one-out strategy was done by individual; hence, for 48 individuals in the training set, parameter tuning was carried out on 47 folds. For computing efficiency, and to have an equal representation of each state of the target class (common source versus forgery), the number of cases in the training set was limited to 50,000 by random sampling without replacement. For each model, variable importance was investigated using the *vip* package (Greenwell and Boehmke, 2020). The final measure of accuracy in the retained models was obtained from a test set comprising the remaining 49 individuals.

There were two test sets (Figure 20):

A) Named test1: one fully independent, as their IDs were not taken in the

training set and
B) Named test2: one with the remaining cases whose IDs were in the training set.

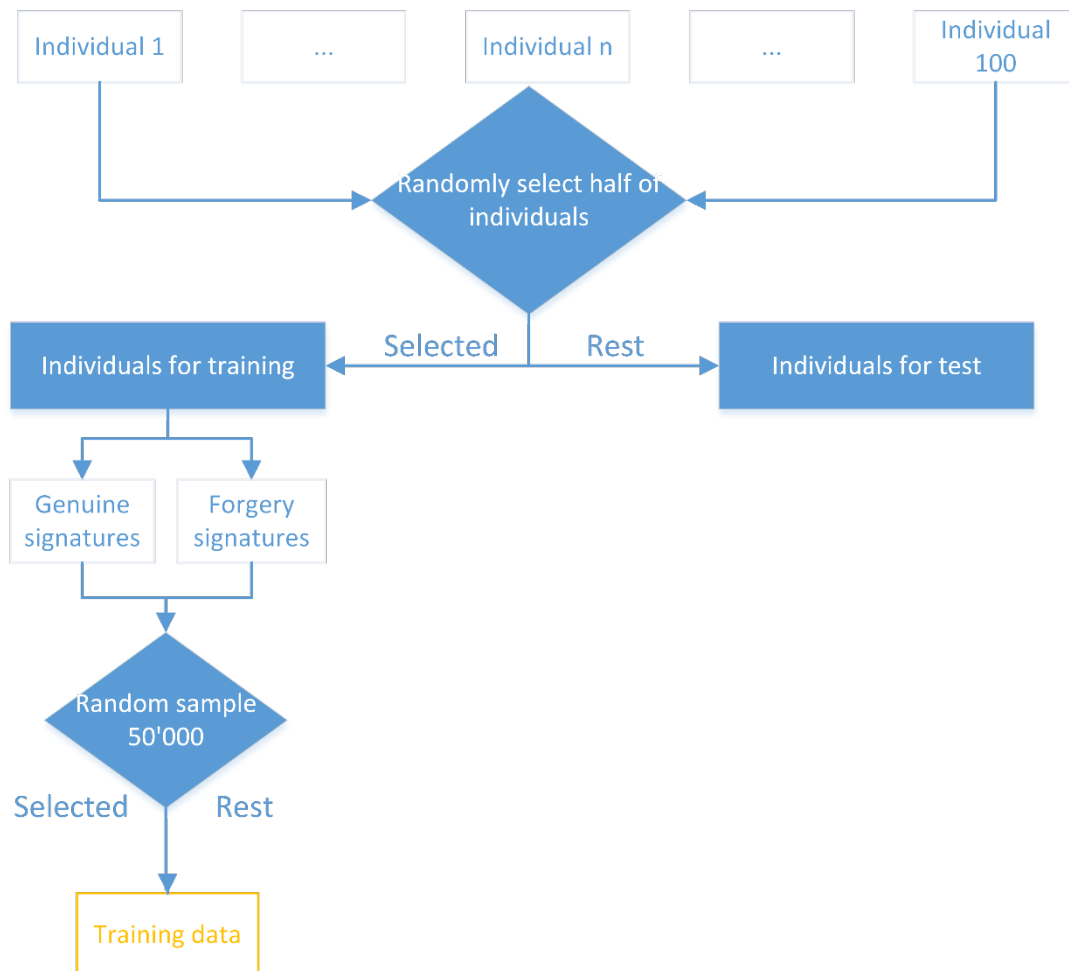


Figure 20: Training and test data for ML

Model-dependent variable importance was obtained directly from some of the models, such as neural net, random forest, and gradient boosting machine. For other models, model-independent variable importance was computed using the *vip* package (Greenwell and Boehmke, 2020).

3.6 Estimation of the strength of the signature evidence

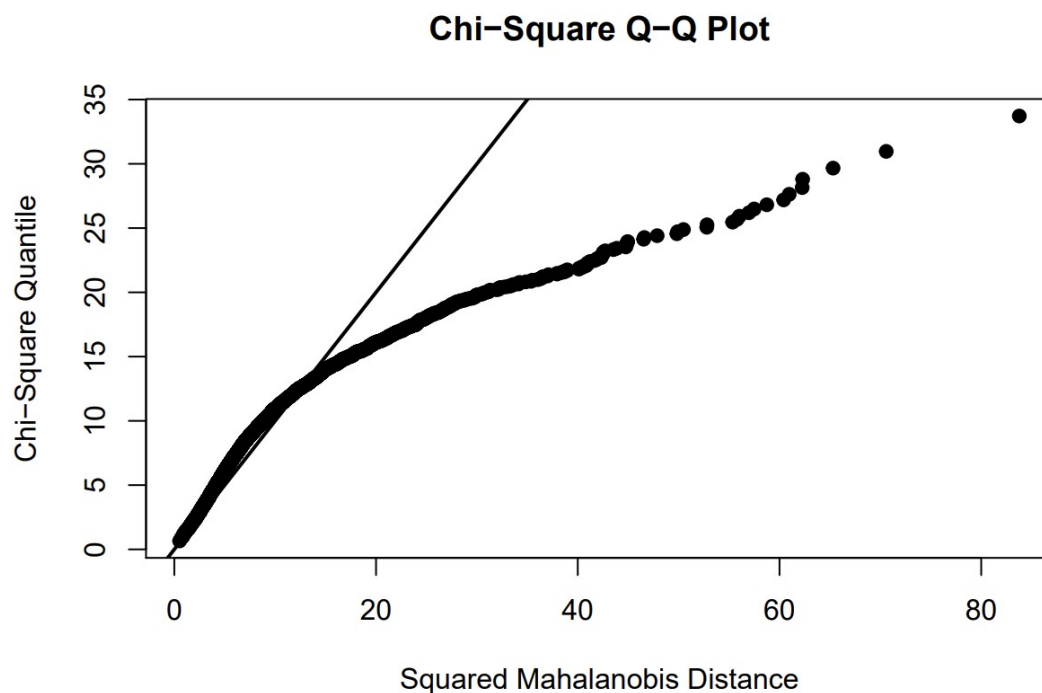
3.6.1 Probability density distribution

Before estimating the probability density distribution, we investigated the type of appropriate distribution.

3.6.1.1 Test for multivariate normal distribution (MVN)

The purpose of testing for MVN was to explore whether 3D pseudo-dynamic features data can be considered as normal multivariate distributions.

The *MVN* package¹⁰ was used without considering the IDs, but a segmentation of the data based on type was done between genuine–genuine comparisons (H_p) and genuine–forgery comparisons (H_d). Tests were conducted with the plotting options available in the *MVN* package.

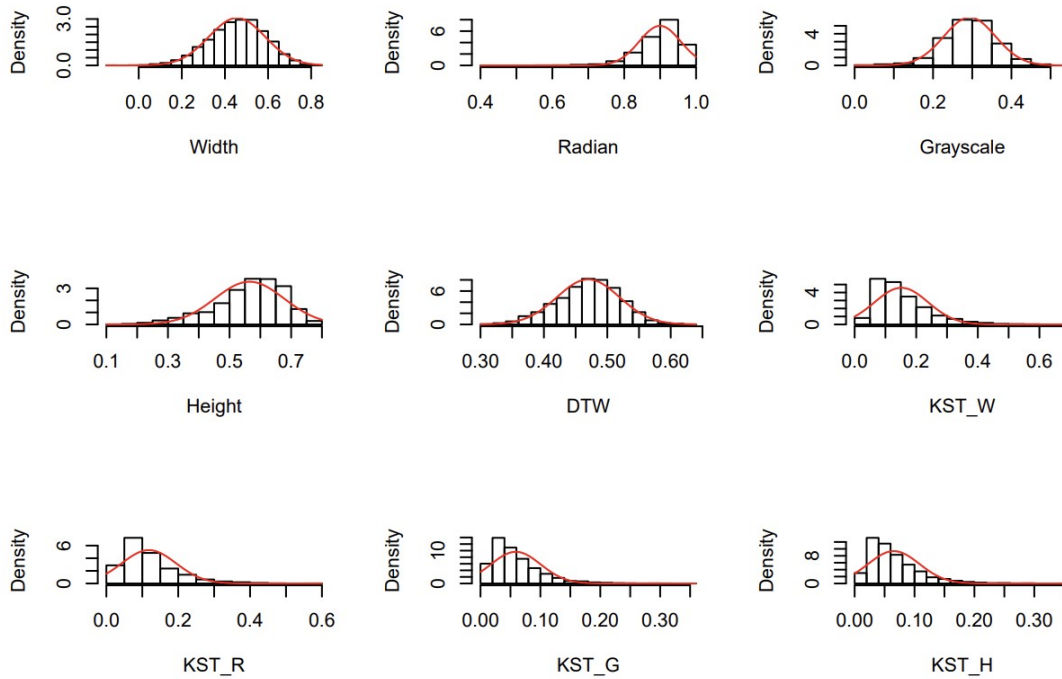


```
Mardia_Test_Hp$multivariateNormality
```

##	Test	Statistic	p value	Result
## 1	Mardia Skewness	12494.659249338	0	NO
## 2	Mardia Kurtosis	78.0269327475341	0	NO
## 3	MVN	<NA>	<NA>	NO

```
HZ_Test_Hp <-mvn(data= SampleResFEA_Hp, mvnTest="hz", univariatePlot="histogram")
```

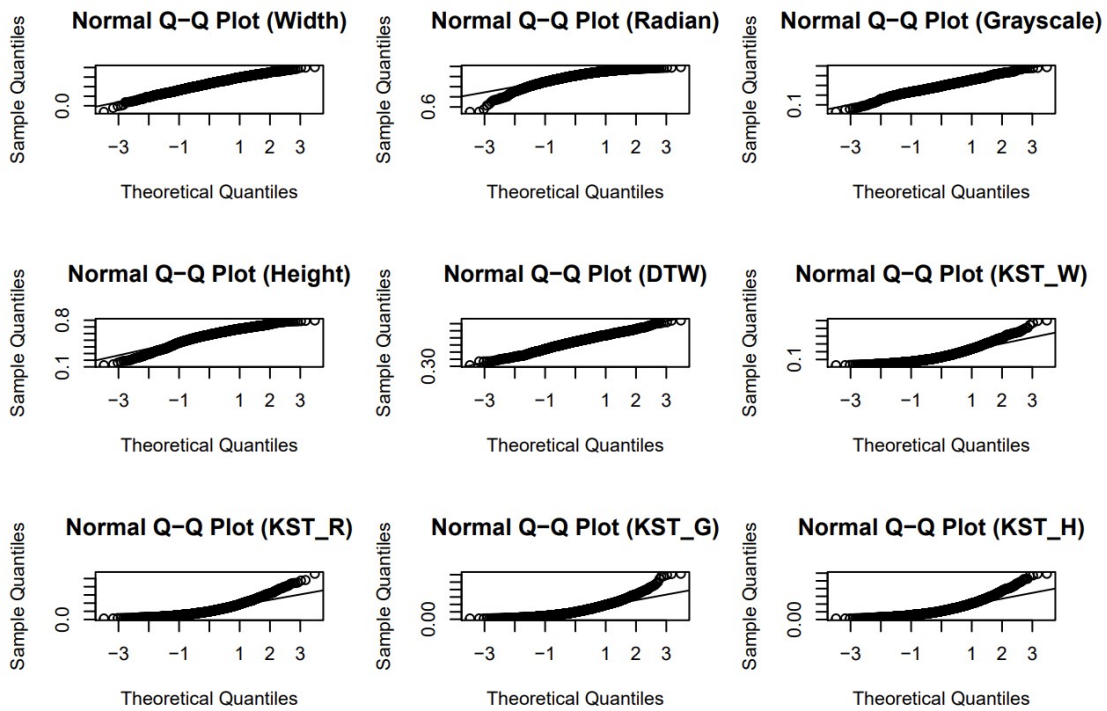
¹⁰ [CRAN - Package MVN \(r-project.org\)](https://cran.r-project.org/web/packages/MVN/index.html)



```
HZ_Test_Hp$multivariateNormality
```

```
##           Test      HZ p value MVN
## 1 Henze-Zirkler 3.508335      0 NO
```

```
Royston_Test_Hd <- mvn (data= SampleResFEA_Hd %>% slice_sample (n = 1999, replace
= FALSE), mvnTest="royston", univariatePlot="qqplot")
```



```
Royston_Test_Hp$multivariateNormality
```

```
##      Test      H      p value MVN
## 1 Royston 935.6094 1.082658e-195 NO
```

```
DH_Test_Hp <- mvn(data= SampleResFEA_Hp , mvnTest="dh")
DH_Test_Hp$multivariateNormality
```

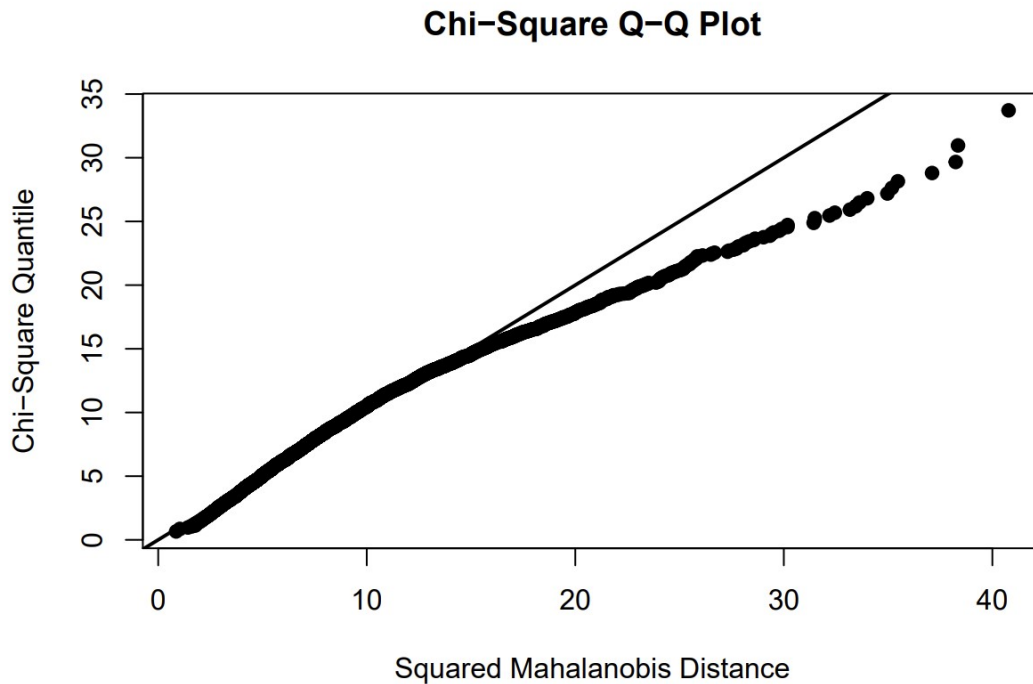
```
##      Test      E df p value MVN
## 1 Doornik-Hansen 9046.261 18      0 NO
```

```
# MVN tests and plots for comparisons under Hd
```

```
Mardia_Test_Hd <- mvn(data= SampleResFEA_Hd, mvnTest="mardia", multivariatePlot = "qq")
```

```
# MVN tests and plots for comparisons under Hd
```

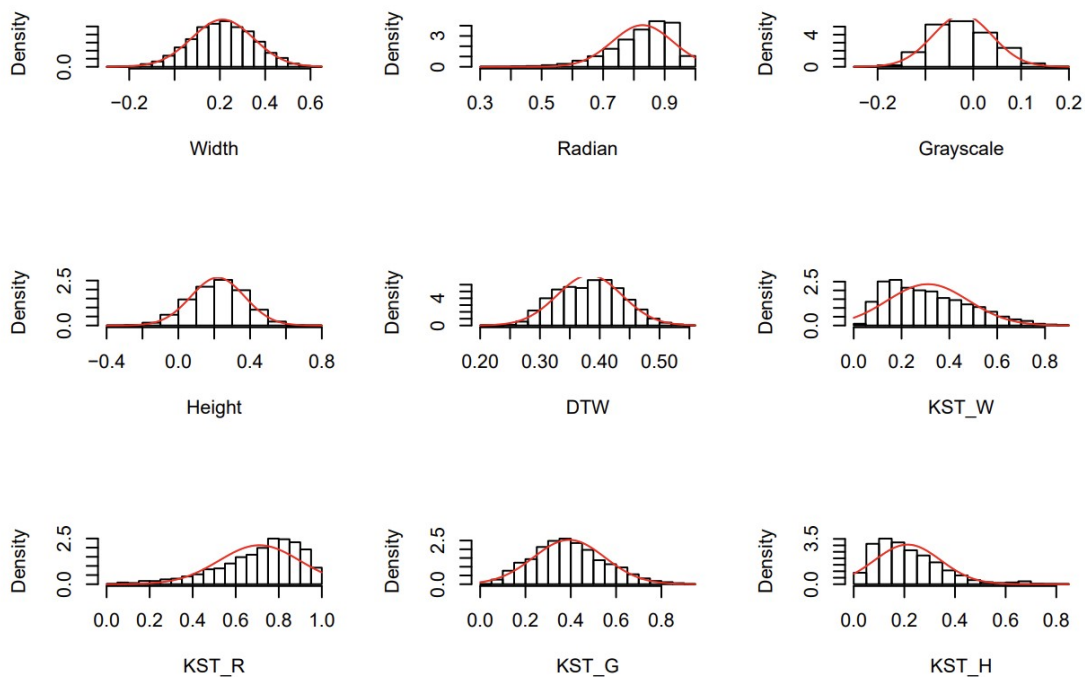
```
Mardia_Test_Hd <- mvn (data= SampleResFEA_Hd, mvnTest="mardia",
multivariatePlot = "qq")
```



```
Mardia_Test_Hd$multivariateNormality
```

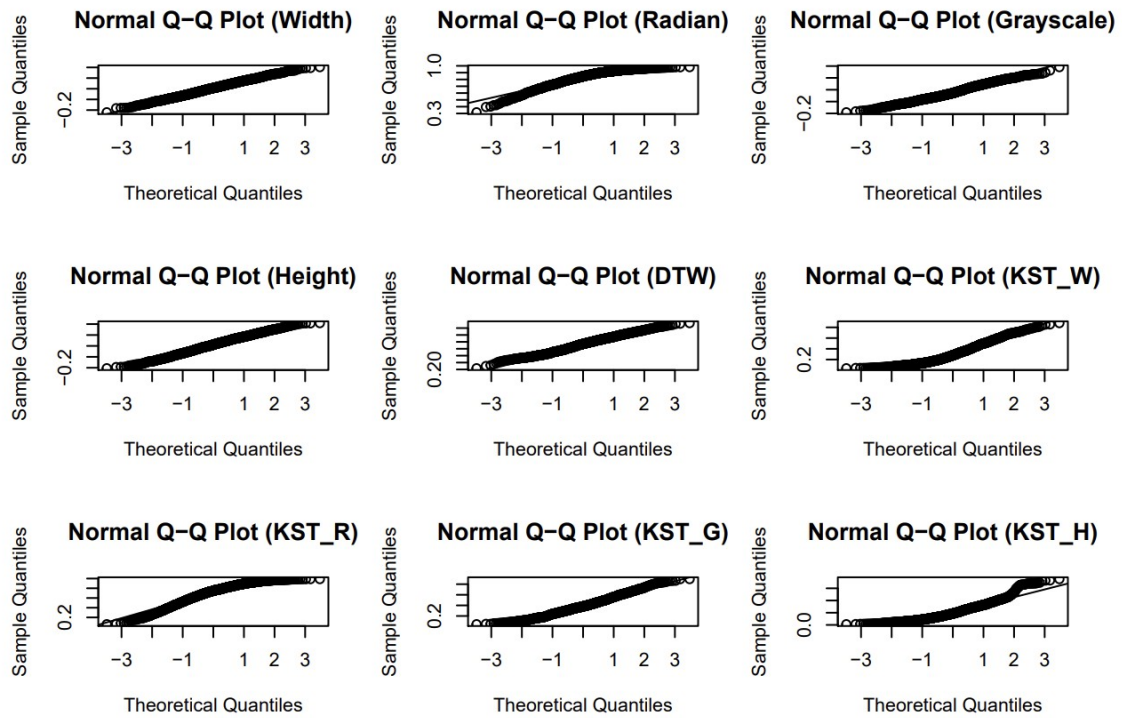
```
##           Test           Statistic p value Result
## 1 Mardia Skewness 6805.05553866982      0      NO
## 2 Mardia Kurtosis 12.2626962716355      0      NO
## 3           MVN           <NA>      <NA>      NO
```

```
HZ_Test_Hd <-mvn(data= SampleResFEA_Hd, mvnTest="hz", univariatePlot="histogram")
```



```
HZ_Test_Hd$multivariateNormality
```

```
##          Test          HZ p value MVN
## 1 Henze-Zirkler 2.348255          0 NO
```

```
Royston_Test_Hd$multivariateNormality
```

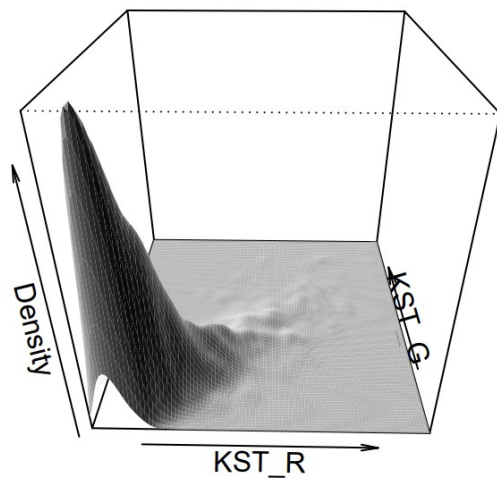
```
##      Test      H      p value MVN
## 1 Royston 606.0543 1.115691e-124 NO
```

```
DH_Test_Hd <-mvn(data= SampleResFEA_Hd, mvnTest="dh")
DH_Test_Hd$multivariateNormality
```

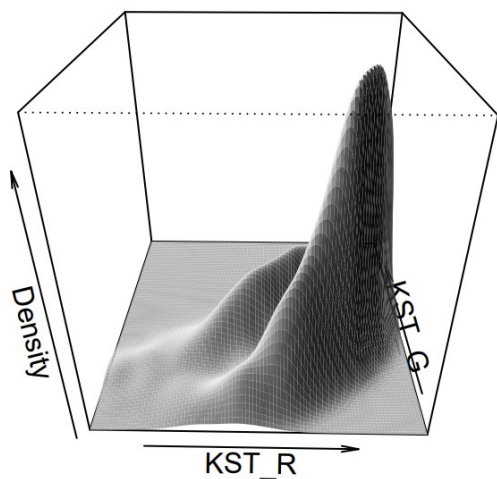
```
##          Test      E df p value MVN
## 1 Doornik-Hansen 3021.745 18      0 NO
```

We can plot an example of density plot, both under H_p and H_d

```
#Density plot (2 variables only)
par(mfrow = c(1,1))
a <- mvn(data= as.matrix(SampleResFEA_Hp[,7:8]), mvnTest="hz", multivariatePlot="persp")
```



```
b <- mvn(data= as.matrix(SampleResFEA_Hd[,7:8]), mvnTest="hz", multivariatePlot="persp")
```



The MVN test results clearly rejected the hypothesis of considering 3D pseudo-dynamic feature data as normal multivariate distribution. Therefore, a multivariate kernel density estimation will be used to subsequently estimate 3D pseudo-dynamic features data.

3.6.1.2 Kernel density estimation

The kernel smoothing function estimate for multivariate data in MATLAB 2020 was used to *compute* a probability density estimate of the sample data in a matrix x , evaluated at the points in pts using the required name-value bw for bandwidth value. The estimation was based on a product Gaussian kernel function.

Furthermore, in Dempster–Shafer theory (DST) calculation, 1-D kernel density estimation was utilized to predict a probability for each feature; *kde* in *ks* package¹¹ was used in estimation. For the DST calculation, see Section 3.6.2.2.

3.6.1.3 Skilled forgery selection

As explained in Section 2.3, people’s ability to simulate a signature varies, and some skilled forgeries are difficult to distinguish from genuine signatures. To explore the performance of the system on skilled forgeries, deep learning was used to select the skilled forgeries from the FF and TF samples. Transfer Learning Using AlexNet¹² (Krizhevsky et al., 2012; BVLC AlexNet Model) in MATLAB 2020 was used to automatically select the skilled forgeries.

For transfer learning, a pretrained network can be taken and used as a starting point to learn a new task. Fine-tuning a network with transfer learning is usually much faster than training a network with randomly initialized weights from scratch. It can transfer learned features to a new task using a smaller number of training images.

Using GE (as target) and RF (as non-target) as training data, freehand simulations and traced simulations were used as test data. Transfer learning for each individual was performed 30 times, and the forgeries that were always misclassified as GE were selected as skilled forgeries.

3.6.2 Score-based LR calculation

3.6.2.1 Score-based LR calculation using MKDE

In previous published research by the author, score-based LR was used to measure the strength of signature comparison findings using subsampling based on dataset_2 (Chen et al., 2018). The forensic findings consist of three parts:

E_U = questioned signatures

E_S = signatures known to have been written by the person of interest (POI)
(and hence genuine), leading to a template for the POI

E_A = signatures obtained from writers other than the POI (and hence forgeries)

¹¹ [CRAN - Package ks \(r-project.org\)](https://cran.r-project.org/package=ks)

¹² [Transfer Learning Using AlexNet - MATLAB & Simulink \(mathworks.com\)](https://www.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html)

The variables expanded from three to nine and were extracted from dataset_3.

In this thesis, we present only the results associated with dataset_3. For previous work, the reader can refer to the publications listed in the appendix.

For a given questioned signature, compared to a set of known specimens (E_S), the findings are represented by a set of scores ($C_w, C_g, C_r, C_h, DTW, KST_w, KST_g, KST_r, KST_h$), denoted as $s(E_U, E_S)$.

The estimated score-based LR \widehat{SLR} (equation 11) is obtained by the ratio between the probability density observed for $s(E_U, E_S)$ given two alternative propositions H_p and H_d . H_p stands for the proposition that the questioned signature shares common authorship with the signatures from the POI. The signature is then genuine. H_d stands for the proposition that the questioned signature is not from the POI's hand but is a forgery. The distributions of the scores under both propositions are needed to obtain the respective densities. This was done by conducting comparisons between signatures from E_S (under H_p) and from E_A against E_S (under H_d).

$$\widehat{SLR} = \frac{\hat{g}[s(E_U, E_S)|H_p]}{\hat{g}[s(E_U, E_S)|H_d]} \quad \text{Equation 11}$$

To simulate operational conditions, simulated cases were generated based on the signature dataset. For instance, for a given individual (writer) composed of m genuine signatures and n forgery signatures, forensic cases were generated by taking one questioned signature (E_S) and five reference signatures (E_U) from the genuine signatures; a set of $\{s(E_U, E_S)\}$ is then obtained representing one forensic case. The choice of five genuine references is motivated by the fact that although the FHE will ask for as many known genuine signatures as possible for comparison purposes, the amount of genuine material is often limited. Based on our practice, the value of five is considered to be a reasonable number of known references.

As introduced in Section 3.4.3.2, two conditions are possible depending on the proposition considered (H_p or H_d) and how the variability will be computed:

Inner-individual mode: Inner-individual mode involves the random selection of a questioned signature that can be a genuine or any one of the forgeries that is compared with the five randomly selected genuine references from a given individual. For each transaction, a score-based LR is calculated and the final LR

associated with the case itself will be the mean of the five score-based LRs. For a given simulated case, the within-writer distribution is obtained using the pairwise comparisons between all of the remaining genuine signatures (leading to $44 \times 43/2$ comparisons) from the individual under consideration. The between-writer distribution is obtained by comparing the remaining genuine signatures against the remaining forgeries of that individual.

For the process of score-based LR calculation, see Figure 21.

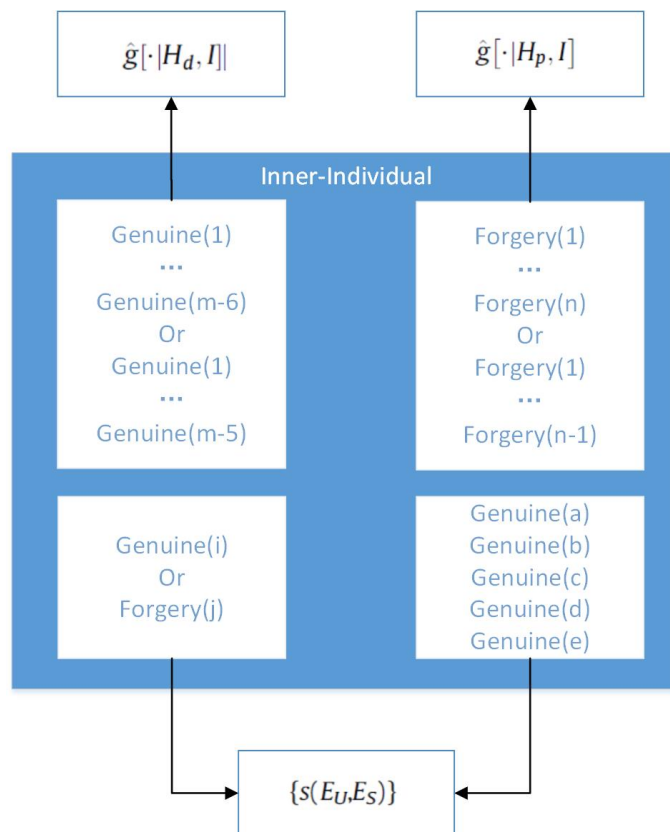


Figure 21

Figure 21: Simulation of cases for inner-individual

Inter-individual mode: In practice, it can be difficult to obtain as many reference samples as signature dataset from each individual as in this study. Hence, we then tested a second method in which the within-writer distribution is based on all of the pairwise comparisons among all of the genuine signatures for the individuals in the dataset. Likewise, the between-writer distribution is based on all the comparisons between genuine and forged entries of all the individuals in the dataset. In the latter method, we have then a generic within-writer distribution and a generic between-writer distribution that do not depend

on the specific individual used to produce the simulated case (see Figure 22).

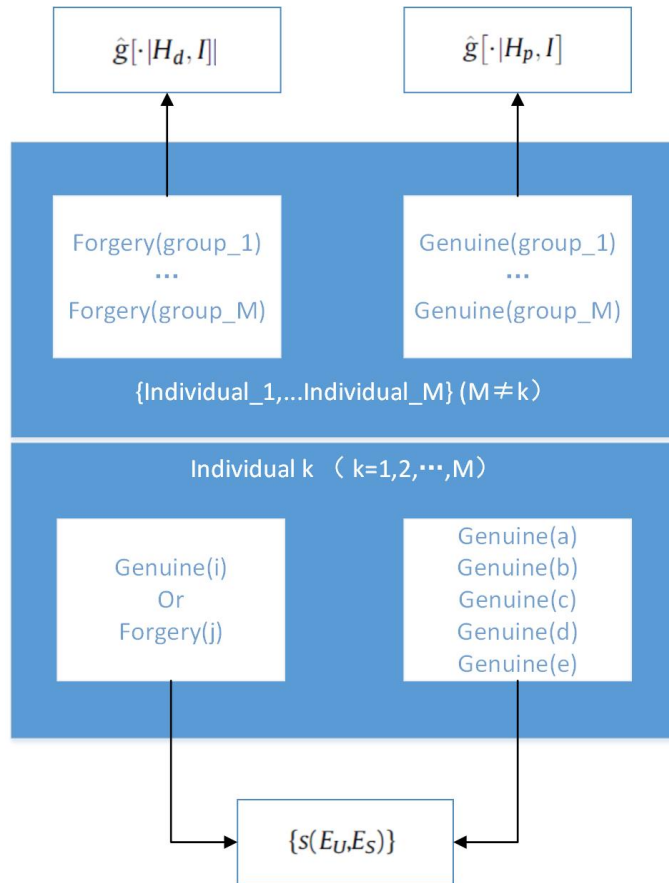


Figure 22: Simulation of cases for inter-individuals

For each individual, simulation cases were generated as shown in Figure 23. Random selection of five references was performed 10 times for each individual, and then a score-base LR was calculated on 231,390 simulation cases in both modes. All results were gathered for inner- and inter-individual modes, respectively. Comparing both methods allowed us to assess whether or not, in cases with limited known references, generic underlying distributions could still be used.

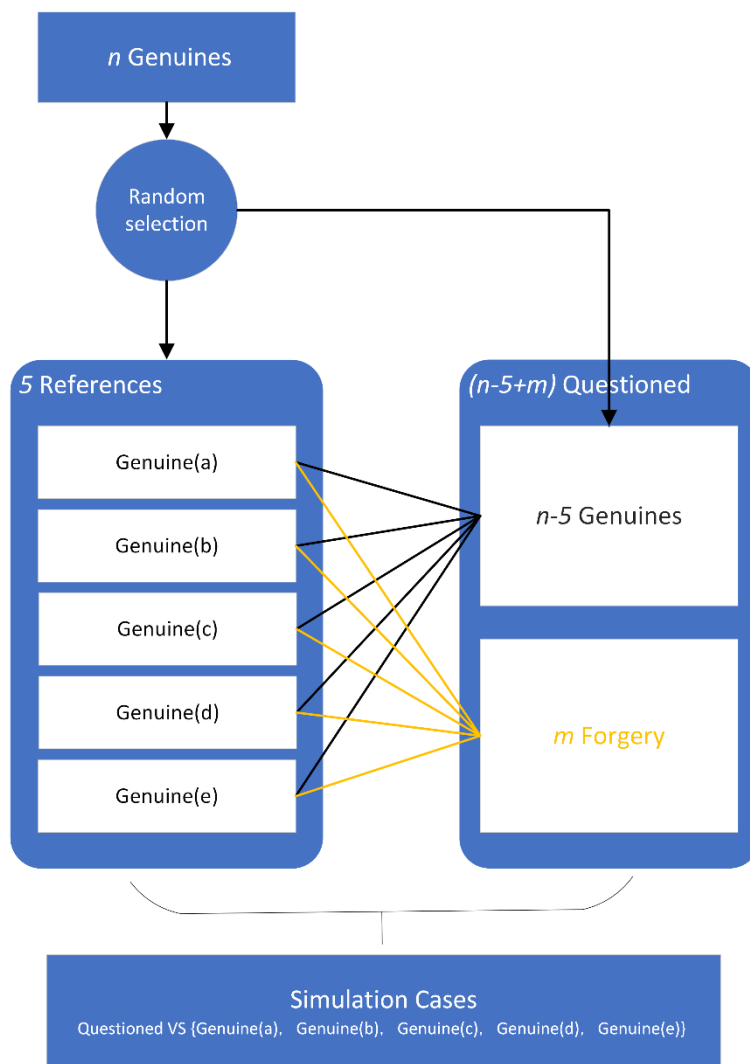


Figure 23: Generation processing for the simulation of cases.

To choose a proper bandwidth for MKED, the parameter varying between 0.1 and 0.2 was tested as follows: 20 individuals were randomly selected in the dataset, as questioned signatures, and the other 22 individuals in the dataset were kept as background information. Ultimately, 0.2 was selected as the optimized bandwidth, for estimating the reasonable LR distribution without too many infinite values (see Figure 24). Note that the choice of the bandwidth impact on the magnitude of the LR obtained both under H_p and H_d .

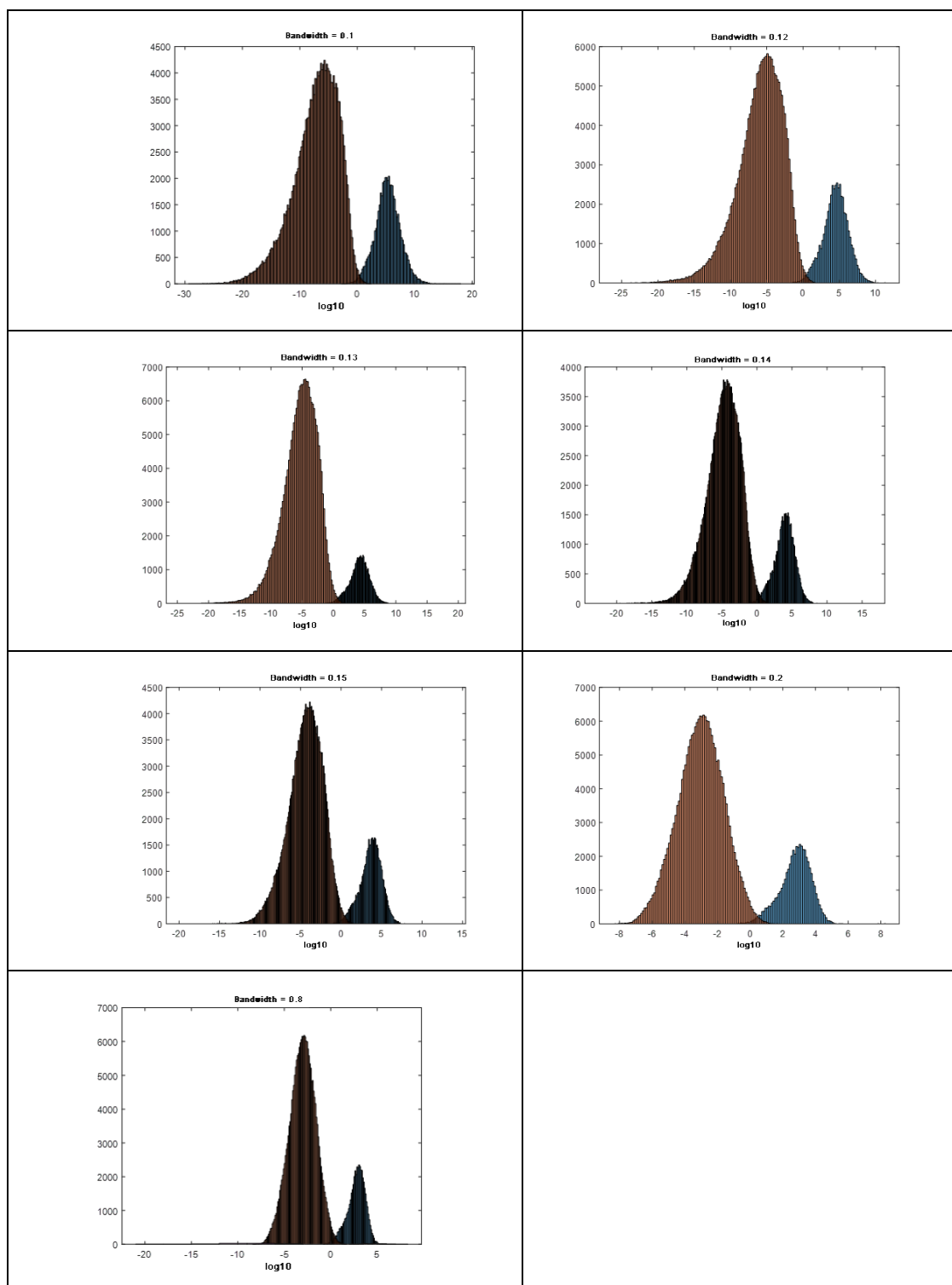


Figure 24: Bandwidth comparison and its selection for MKED. In red the distribution of $\log_{10}(\text{LR})$ under H_p . In blue the distribution of $\log_{10}(\text{LR})$ under H_d .

3.6.2.2 Score-based LR calculation using Dempster–Shafer theory

Dempster–Shafer theory (DST) is a generalization of the Bayesian theory of subjective probability, also referred to as the theory of belief functions. Belief functions base degrees of belief (or confidence or trust) for one question on the probabilities for a related question. They are often used as a method of sensor fusion, which is based on two ideas: obtaining degrees of one question from subjective probabilities for a related question, and Dempster’s rule (Dempster, 1968) for combining such degrees of belief when they are based on independent items of evidence. In essence, the degree of belief in a proposition mainly depends on the number of answers (to the related questions) containing the proposition and the subjective probability of each answer. The combination rules reflect the general assumptions about the data that contributed as well.

Here, a degree of belief (also called a mass) is represented as a belief function rather than a probability distribution. Probability values are assigned to sets of possibilities rather than single events: their appeal rests on the fact they naturally encode evidence in favour of propositions.

DST has the following advantages: The required prior data are more intuitive and easier to obtain than in standard probabilistic reasoning theory. DST satisfies a weaker condition than standard probability—that is, “it is not necessary to meet the probability additivity”. Various data and knowledge can be integrated. DST has the ability to directly express “uncertain” and “don’t know”. This information is expressed in the mass function and is retained during the evidence synthesis process.

In the first step, subjective probabilities (masses) are assigned to all subsets of the frame. In this research, the hypothesis was set to four masses: null (neither genuine nor forgery), genuine, forgery, and either (genuine or forgery). Setting the mass of the null hypothesis as 0, the remaining mass (the gap between the supporting evidence on H_p and the contrary evidence on H_d) is “indeterminate”, which means that it could be either genuine or forgery (Table 16).

Table 16: Hypothesis and mass of Dempster–Shafer theory

Hypothesis	MASS of 3D and pseudo-dynamic features
Null (neither genuine nor forgery)	0
Genuine	supporting evidence on H_p ($SEHP_{\{F\}}^*$)
Forgery	supporting evidence on H_d ($SEHD_{\{F\}}^*$)
Either (genuine or forgery)	$1 - (SEHD_{\{F\}} + SEHP_{\{F\}})^*$

* F means features used in calculation.

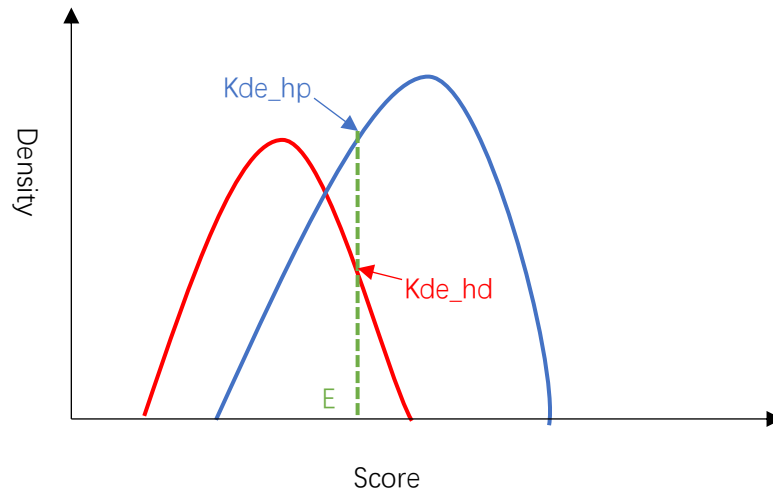


Figure 25: Estimation of density under H_p and H_d denoted as kde_hp , and kde_hd respectively. Score of evidence denoted as E .

For each feature, first, 1-D kde in the ks package is used to estimate the density under H_p and H_d (kde_hp and kde_hd , respectively) (

Figure 25). Second, $kde_hp / (kde_hp + kde_hd)$ is used as probability of supporting evidence on H_p ($SEHP$); $kde_hd / (kde_hp + kde_hd)$ was used as probability of supporting evidence on H_d ($SEHD$). Then, the package *ibelief* was invoked (using the Dempster–Shafer criterion) to combine masses of supporting evidence on H_p and supporting evidence on H_d ($DST(SEHP_{\{F\}})$ and $DST(SEHD_{\{F\}})$, respectively). Finally, LRs were calculated by the ratio of $DST(SEHP_{\{F\}})$ vs $DST(SEHD_{\{F\}})$; see Equations 12–15.

$$SEHP_{\{F\}} = \frac{kde_hp_{\{F\}}}{(kde_hp_{\{F\}} + kde_hd_{\{F\}})}; \quad \text{Equation 12}$$

$$SEHD_{\{F\}} = \frac{kde_hd_{\{F\}}}{(kde_hp_{\{F\}} + kde_hd_{\{F\}})}; \quad \text{Equation 13}$$

$$\{F\} = \{W, R, G, G_{skel}, H, H_{skel}, DTW, W_{kst}, R_{kst}, G_{kst}, G_{skel_{kst}}, H_{kst}, H_{skel_{kst}}\}; \quad \text{Equation 14}$$

$$LLR = \log_{10} \left(\frac{DST(SEHP_{\{F\}})}{DST(SEHD_{\{F\}})} \right). \quad \text{Equation 15}$$

Parallel computing in the packages *doSNOW* (Microsoft Corporation and Stephen Weston, 2020) and *foreach* (Microsoft Corporation and Steve Weston, 2020) is used to save on computation time.

The methods of weighting and no weighting for variables were used in DST calculation. Weighing parameters (Table 17) were assigned based on the relative importance of each variable obtained from the *R-Forest* model (see Section 3.5).

Table 17: Variables and weighting parameters

Variables	Description	Weighting parameter
Width	Width of stroke	0.325
Radian	Radian of stroke	0.001*
Grayscale	Grayscale of stroke	1
Height	H	0.41
DTW	DTW	0.21
Grayscale_skeleton	G_ske	0.001*
Height_skeleton	H_ske	0.001*
KST_Width	Kst_w	0.0436
KST_Radian		0.809
KST_Grayscale		0.762
KST_Height		0.243
KST_Grayscale_skeleton		0.001*
KST_Height_skeleton		0.001*

Note: * indicates variables that were *not* calculated or got a 0 value in ML.

3.6.3 Performance evaluation

Inter- and inter-individual modes

Four distinct properties are mentioned regularly in the context of assessing scientific evidence: reliability (Royall, 2000; Taylor, 2014), validity (Ramos-Castro & Gonzalez-Rodriguez, 2013), accuracy, and precision (Biedermann et al., 2016). Performance measures are obtained through the study of LR

distributions, which are used to assess the method for the evaluation of findings. If the histograms of two probability distributions under the respective hypotheses (H_p and H_d) show an overlap of the distributions; this reflects the discriminating power of a method at a particular value of $\log(\text{LR})$. Tippett plots generalize the rates of misleading evidence in comparison. A detection error trade-off (DET) plot presents false positives versus false negatives in the function of a decision threshold (Aitken et al. 2021).

Performance metrics are represented graphically through Tippett plots, detection error trade-off (DET) plots, empirical cross-entropy (ECE) plots, and applied probability of error (APE) plots. Classification accuracy is measured based on Cllr and ECE. Log-likelihood ratio cost (C_{llr}) is another measure of performance. Discriminative power is measured using C_{llr}^{min} and EER; calibration is assessed using C_{llr}^{cal} and the difference between ECE and ECE-after-PAV (Haraksim et al., 2015).

2.6.3.1 Calibration

Robertson et al. (2016) described “calibration” as follows:

For comparison systems, there is no ascertainable true value for the LR; only in experiments we can know for sure that a hypothesis is true. We therefore define the property ‘calibration’ in a different way. Suppose we are given a comparison system that reports LRs. Rather than simply accepting them at face value, we might instead evaluate the LRs reported for each hypothesis (Robertson et al., 2016, p. 91).

There are many examples of calibration in forensic science, forensic speaker recognition (Morrison, 2018), glass evidence (Corzo et al., 2018), mRNA profiling (Ypma et al., 2021), DNA profiling (Gonzalez-Rodriguez et al., 2007), inkjet classification (Chen et al., 2021), and signatures (Chen et al., 2018). Factors such as a bad choice of databases or of statistical models, or a limited quantity or quality of signatures, can lead to misleading LRs (meaning they may provide support for the wrong proposition). Calibration is a way to mitigate this problem and ensure that the LRs can be probabilistically interpreted as representing the evidential value of the comparison in a Bayesian evaluation framework (Lucena-Molina, 2015; Haraksim, 2015). Several measures for the performance of LR-based system and calibration methods have been proposed in the literature (Brümmer, 2006; Lucena-Molina, 2015; Ramos, 2013a; Ramos, 2013b; Martin, 1997).

3.6.3.1 Logistic and PAVA calibration

The solution proposed in Brümmer (2006) and Brümmer (2010) by means of the Pool Adjacent Violators Algorithm (PAVA) has been used to measure the calibration of LR_s (Ramos, 2013a; Ramos, 2013b). Morrison (2021) suggested that PAVA-based metrics of degree of calibration do not actually measure the degree of calibration; rather, they measure sampling variability between the calibration data and the validation data, as well as overfitting on the validation data.

Morrison (2013) provided a tutorial on logistic regression calibration and fusion at a practical conceptual level with minimal mathematical complexity. A traditional-style phonetic-acoustic forensic-speaker-recognition analysis was conducted on Australian English /o/ recordings calibrated using logistic regression. Different parametric curves were fitted to the formant trajectories of the vowel tokens, and cross-validated LR_s were calculated using a single-stage generative MVKD formula (Morrison & Kinoshita, 2008). Aitken (2004) investigated the calibration of scores generated by the MVKD formula.

3.6.3.4 Performance evaluation in datasets

In this research, a validation toolkit¹³ (Beta v1.06) was initially used to measure the performance and PAVA calibration based on dataset_2. The purpose of this toolkit was to enable end users to effortlessly measure the performance of log-likelihood-ratio values coming from their experiments. Performance representations include Tippet plots, limit Tippet plots, DET plots, ECE plots, and APE plots. Given the dependence on the writer, analysis was carried out per writer (an individual in data collection includes genuine and forged signatures of a given individual). The toolbox calculates C_{llr} , C_{llr}^{cal} , C_{llr}^{min} , EER, rate of misleading evidence in favour of H_d (RMED), and rate of misleading evidence in favour of H_p (RMEP). The above computation involves an exceptionally large number of cases, and we explored whether performance metrics could be estimated with a lower number of cases. To reduce computing time, three percentages (50%, 30%, and 10%) of the data were randomly selected, and the metrics were bootstrapped (1000 bootstrap samples). These results are presented in Chen et al. (2018), which is provided in the appendix.

In this study and for dataset_3, the *comparison* package (Lucy et al., 2020) is used to evaluate the performance of the LR system. This package includes two-level functions to calculate LR assuming multivariate normality; another drops this assumption and uses a multivariate kernel density estimate. This package

¹³ <https://sites.google.com/site/validationtoolbox/home>

also contains code for performing the ECE calibration of LRs. In this research, it is used to perform multivariate LR calculation and evaluation based on dataset_3. It calculates the calibrated set of LRs with logistic regression. Additionally, to show performance based on two methods of regression, PAVA regression is also used to calculate the calibrated set of LRs.

To answer the question as to whether or not the skilled forgeries selected through CNN deep learning necessarily tend to mislead, density distribution will be used to observe the differences between skilled forgeries (as selected by CNN) and other forgeries based on the calibrated log LR.

3.7 Validation tests

3.7.1 Competition test

In the past, many competitions have been organized to measure the detection rate of several classifiers. As we noted, however, most of the current research in the field of signature verification does not take the real needs of FHEs into account. In their case work, they often work with signatures produced in various real-world environments. The Signature Verification Competition for Online and Offline Skilled Forgeries (SigCom2011) was the first verification competition focused on Chinese signatures. This study will use the SigCom2011 dataset as the blind test dataset.

3.7.2 CNAS proficiency test (PT)

Since 2014, the China National Accreditation Service for Conformity Assessment (CNAS) has released PT projects every year. There are 10 PT projects: raw documents could be obtained from 9 out of 10 projects, which provide 3D and pseudo-dynamic features. One out of 10 projects scanned documents in 9 projects, which provided only 2D and pseudo-dynamic features.

In each PT case, using dataset_3 as the background information, the nine-variable option was chosen for the PT case in which original documents were available. The seven-variable option was chosen for the PT case that only provided scanned images (without access to the 3D variables). The retained options for these variables are given in Section 4.1.4.2.

A comparison is carried out between reference signatures and questioned signatures and the LRs are calculated. The comparison between reference signatures also provides a baseline to see if the background information was adequate or suitable for the specific PT case.

3.7.3 Real forensic cases test

Finally, 50 real cases were collected randomly from the Academy of Forensic Science, covering 2019 to 2020. These provided original documents that allowed us to obtain 3D and pseudo-dynamic features. The writing conditions of the real cases—including writing instruments, surfaces, and positions—varied to a large degree between the cases. Performance will be evaluated by comparison between the obtained LRs with the reported expert opinions in the case.

Chapter 4 Results

4.1 Between- and within-writer variations

4.1.1 Statistical description and descriptive analysis

4.1.1.1 Descriptive measurement

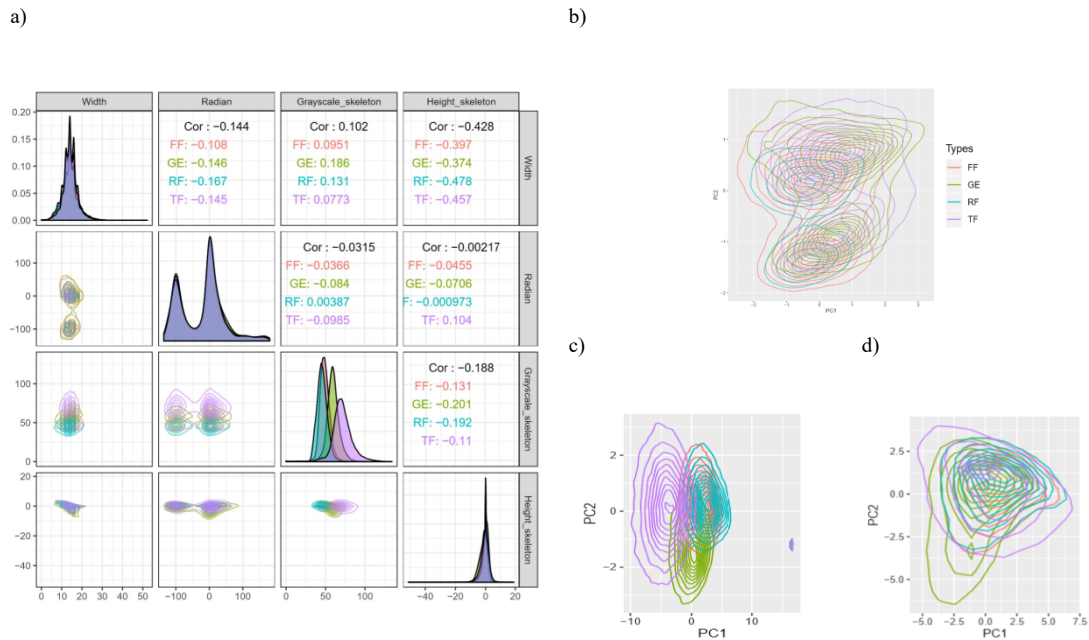


Figure 26: Pair plots based on four 1D features (width, radian, skeleton grayscale, and skeleton height) and kernel 2D distribution after PCA based on distribution features from all individuals in the signature dataset. a) For each plot density in the diagonal, correlation in upper, density_2d in lower. b) Two-dimensional density distribution plot after PCA based on four 1D features (width, radian, skeleton grayscale, and skeleton height) and height (skeleton and distribution). c) Grayscale distribution and d) height distribution. Red lines and areas represent freehand forgery (FF), green lines and areas represent genuine signature (GE), blue lines and areas represent random forgery (RF), and purple lines and areas represent tracing forgery (TF).

As shown in Figure 26 above, when signature data of different prediction modes (typically genuine signatures vs forgeries) were put together, the directly measured features had different degrees of similarity, and no difference was found in the width, height, and radian features. This phenomenon revealed two

things: first, the direct measurement features were greatly affected by the signature text; and, second, the direct measurement features from different source were not discriminable. In other words, description measurement is sensitive to the character morphology of signature.

4.1.1.2 Comparative measurement

The statistical description of features, MANOVA, and discriminant analysis using dataset_1 is documented in Chen (2015) in the appendix. The MANOVA result confirmed significant differences between GE signatures (genuine signature was denoted as original signature (OR) in the published paper) and forgeries (forgeries was denoted as non-original signature (NON-OR) in the published paper) with respect to width and grayscale. Moreover, significant differences between GE and FF and between GE and TF were shown with respect to radian data. The mean distances between the observations showed that the imitation signatures, such as the FF and TF signatures, were close to the GE signatures in width, grayscale, and radian values. The differences between the GE and TF signatures were not significant with regards to radian data. The imitation signatures were more similar to GE signatures than to RF signatures considering their width, grayscale, and radian data. The MANOVA result on dataset_1 indicated that the width, grayscale, and radian information are effective features for discriminant analysis. GE, FF, RF, and TF signatures were classified in the discriminant analysis. The discriminant analysis between GE signatures and forgeries showed high scores in the cross-validation rate with a mean of 95.8% (see Table 18).

Table 18: Discriminant analysis based on width, grayscale, and radian: cross-validation between OR and non-OR in each group

Group	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
C.V ^a	97.6%	97.6%	95.1%	96.4%	94%	97.6%	95.2%	95%	97.4%	93.9%	90.5%	98.8%

Note: ^a, C.V., cross-validation.

The variables are presented in Figure 27, in which each variable is presented by its histograms, pairwise plots, and correlation measures with other variables. For each plot histogram in the diagonal part, correlation in the upper part, and density_2d in the lower part was based on width, radian, skeleton grayscale, skeleton height, grayscale distribution, and height distribution. The distance between signatures during DTW processing is another variable.

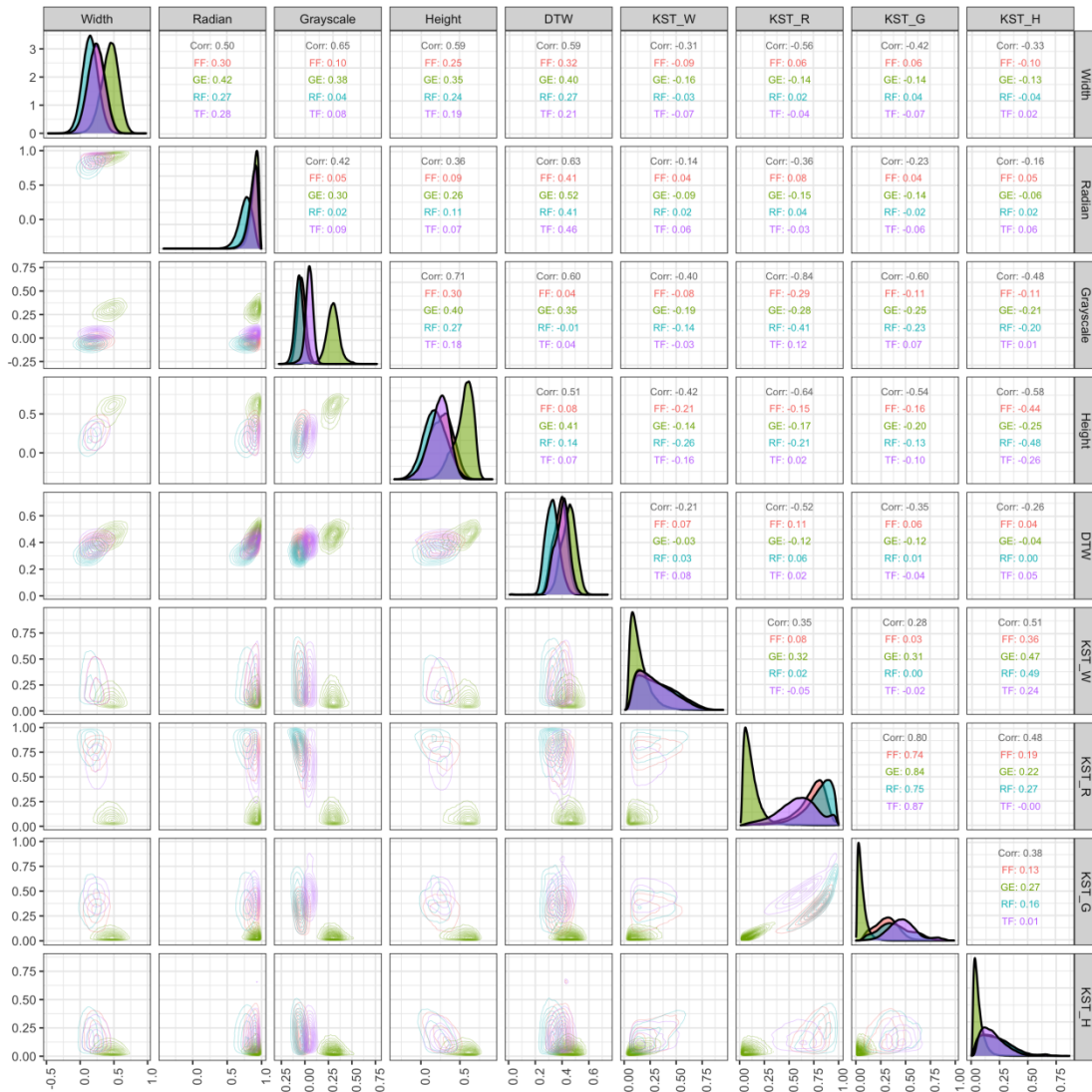


Figure 27: Pair plots based on comparative measurement from all individuals in Dataset_3: Red lines and areas represent freehand forgery (FF), green lines and areas represent genuine signature (GE), blue lines and areas represent random forgery (RF), and purple lines and areas represent tracing forgery (TF).

Comparative measurement provides a novel method to investigate handwriting in a systematic way. For features of width, skeleton height, and height distribution, the similarities between genuine signatures were higher than those between genuine signatures and forgeries. For skeleton grayscale and grayscale distribution features, the similarities between genuine signatures were much higher than those between genuine signatures and forgeries. For radian features, random forgery showed lower similarity with genuine signatures in handwriting morphology. In the distances between signatures following DTW, however, the distance between random forgery and genuine signature was much shorter than that between other types of forgeries. Based on the

interpretation of correlation coefficients (Lippmann, 1987), there were high positive relationships (correlation coefficients: 0.7–1.0) between the correlation coefficient of skeleton height and height distribution and weak positive linear relationships (correlation coefficients: 0.0–0.3) between the correlation coefficients of radian and skeleton grayscale, skeleton height, and height distribution. The remaining pairwise features showed moderate (correlation coefficient: 0.3–0.5) or strong (correlation coefficient: 0.5–0.7) positive linear relationships. In addition, there was a strong positive linear relationship between the correlation coefficient of radian and distance of DTW—that is, a higher similarity in signature morphology corresponded to further DTW distance.

Compared with descriptive measurement features, when the signature data of different character morphology were combined, the comparative measurement features did not reveal the same phenomenon because of the different texts but they showed stable differences between GE and forgeries (GE, RF, FF, and TF). This result tells us that, even in the absence of background data corresponding to the questioned signatures, the signature data for different types of languages might provide reliable background data for real cases. This behaviour is especially important in actual cases, and it can be used to propose a text-independent background data for the evaluation of the findings.

Width, skeleton height, and height distribution showed that similarities between genuine signatures are higher than those between genuine signatures and forgeries. For grayscale, the similarities between genuine signatures are higher than between genuine signatures and forgeries. For radian features, random forgeries show lower similarity with genuine signatures in handwriting morphology. Unexpectedly, the distance between random forgery and genuine signature is lower than for other types of forgeries. The correlation coefficients were high for some combinations of variables. This testifies to the correlated nature of the variables.

4.1.2 Probability density distribution

In the second published paper by Chen et al. (2018), also in its appendix, we have shown that relevant dynamic features (width, grayscale, and radian features measured as a function of the writing sequence derived from static images) allow to discriminate between genuine and forged signatures from Dataset_2 (an expanded dataset based on Dataset_1). These features are important additions to traditional features measured statistically on the images. Our data confirmed that some signatures are easier to forge than others, but reasonable discrimination can be achieved.

The two published papers (Chen, 2015; Chen et al., 2018) initially proposed and verified the signature pseudo-dynamic features and its specificity. The features were obtained from scanned 2D images in these two papers. In this research, 3D information was added to the features. Dataset_3 is a brand-new and larger database that provides signatures for a larger population of people. Compared with Dataset_2, the number of individuals has increased from 20 to 100, and the number of forgers of various forgeries has increased from 3 to 99, which currently could be the largest Chinese signature database. We hope to use Dataset_3 to re-verify the previous 2D pseudo-dynamic feature research and to provide sufficient signature data for the 3D pseudo-dynamic feature research. Therefore, four options for variable combinations were selected to observe the performance of different options in terms of number of variables considered (Table 19). Option 1 represents the first three important variables, according to the variable importance parameters in *R-Forest* (see later Section 4.2.5). Option 2 represents the first six important variables, still according to the variable importance parameters in *R-Forest* importance. Option 3 represents seven variables, excluding height information. Option 4 represents all variables available. Option 5 represents 13 variables, by adding 2D pseudo-dynamic features associated to published papers. To find the best variable combinations, four options were used for comparison.

Table 19: Five options for variable combinations

Option	LR	NM*	Width	Radian	Grayscale	Height	DTW	KST_W	KST_R	KST_G	KST_H	Additional**
1	LR1	3			√				√	√		
2	LR2	6		√	√	√	√		√	√		
3	LR3	7	√	√	√		√	√	√	√		
4	LR4	9	√	√	√	√	√	√	√	√	√	
5	LR5	13	√	√	√	√	√	√	√	√	√	√

* NM: number of variables.

** Additional grayscale, height in skeleton and KST, respectively.

In dataset_3, nine features were measured: width, grayscale, radian, height, DTW, KST_width (KST_W), KST_grayscale (KST_G), KST_radian (KST_R), and KST_height (KST_H). The KDE plot based on the six, seven, and nine features after PCA showed significantly different distributions between GE and forgeries (RF, FF, and TF) than that of the three features (Figure 28).

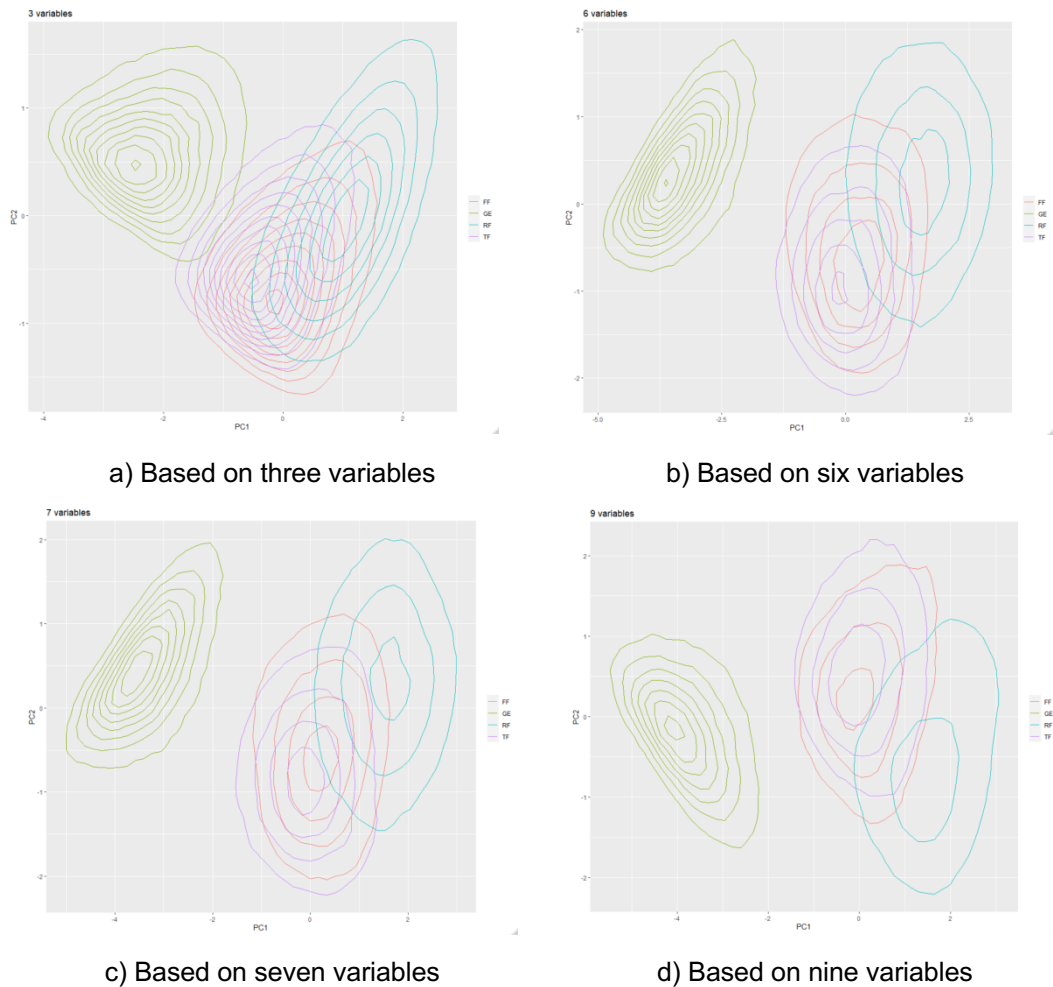


Figure 28: Two-dimensional density distribution plot for all individuals after principal component analysis (PCA) based on different variables options: three variables, six variables, seven variables and nine variables.

A Shiny web application was created to show the KDE plot after PCA for each individual and 100 individuals together (see <https://cchampod.shinyapps.io/ChineseSignatures/>) (Figure 29). Nine variables were used in this application.

PCA plots for any user can be consulted in the dedicated Shiny app. The Shiny app Score-based LR estimates the strength of the signature evidence.

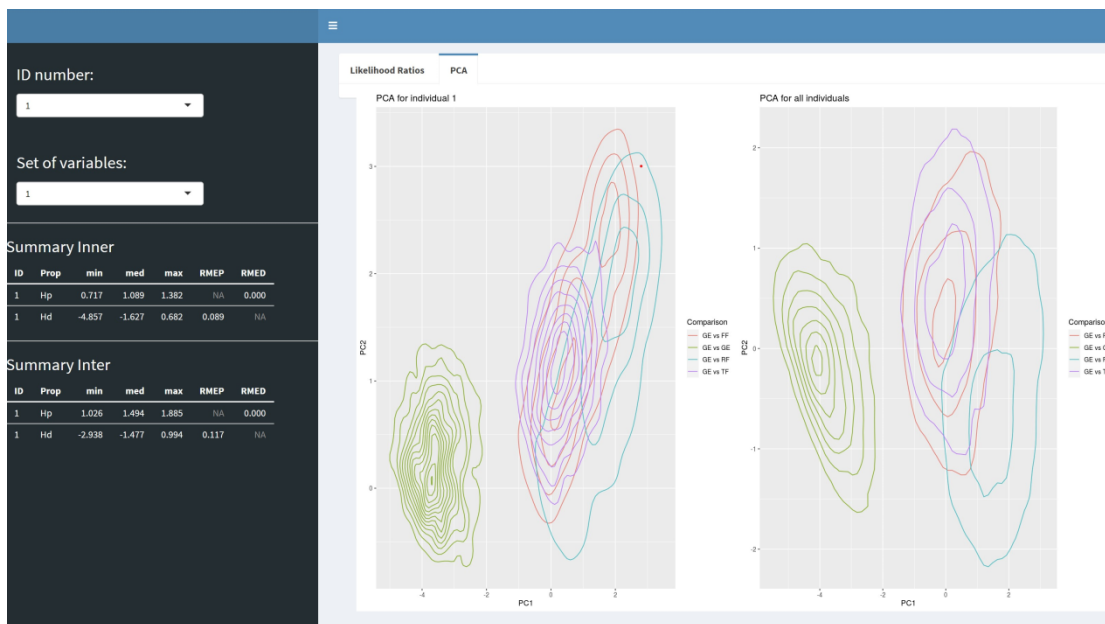


Figure 29: Shiny app for the Chinese signature project

To further illustrate how comparisons between genuine signatures can be distinguished from comparisons between genuine and forgeries, Figure 30 presents a 2D kernel density distribution plot of the first two principal components obtained after PCA. The top left plot shows the joint results for all 97 individuals. This can be contrasted with the results for three individuals: 22, 67, and 89. Globally, it is observed that genuine comparisons are separated from comparisons with forgeries. The distance with genuine comparisons is higher in comparisons involving RF and lower for cases involving other types of forgeries. The individual examples show that the magnitude of these trends depends on the donor. Some individuals (e.g., 89) allow for a distinction between types of forgeries; others (e.g., 67) have all their comparisons involving forgeries in the same cluster. The level of discrimination between genuine signatures and forgeries also depends on the donors (e.g., 22 compared to 89).

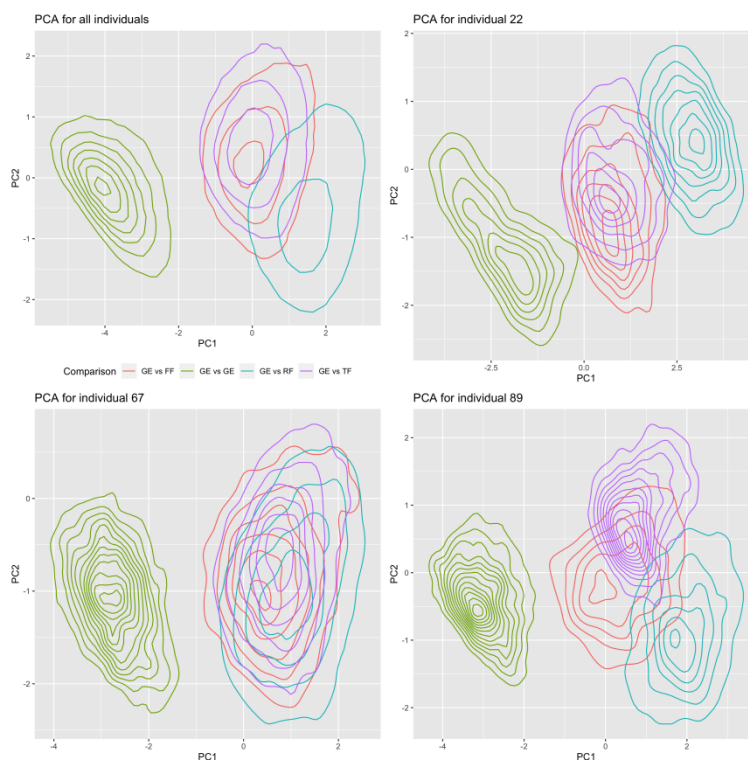


Figure 30: 2D kernel density distribution plot after PCA: PCA for all individuals, for individual 22, for individual 67, and for individual 89, respectively.

4.2 Machine Learning (ML)

In addition to the investigation of the score-based LRs, these nine features have informed a ML strategy in which we investigated the capability to distinguish comparisons arising from a common source (a given individual) from comparisons involving forgeries (associated with each individual). The use case for ML should correspond to the use of the biometric system for automatic identification. A range of ML classifiers have been tested from low complexity (higher explainability) models to high complexity (lower explainability) models. Refer to chapter 3 for methodological details.

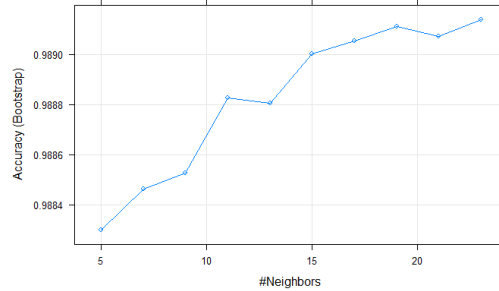
4.2.1 *K*-nearest neighbour

Nearest neighbour is based on the labels of the *K*-nearest patterns in the data space. *KNN* classifier is used to classify unknown observations by assigning them to the class of the most similar labelled examples. Nearest neighbour techniques are strong in case of large datasets and low dimensions as a local method. Variants for multilabel classification, regression, and semi-supervised learning settings can be applied to a broad scope of ML (Kramer, 2013; Zhang,

2016). Table 20 shows the resampling results across tuning parameters for *KNN*.

Table 20: Resampling results across tuning parameters for *KNN*

<i>k</i>	Accuracy
5	0.9883
7	0.9885
9	0.9885
11	0.9888
13	0.9888
15	0.9890
17	0.9891
19	0.9891
21	0.9891
23	0.9891



The final value used for the model was $k = 23$.

4.2.2 Discriminant analysis

Discriminant analysis is utilized to predict the probability of belonging to a given class based on one or multiple predictor variables. It is suitable for continuous and/or categorical predictor variables. Compared with logistic regression, discriminant analysis is more suitable for predicting the observed category when the outcome variable contains more than two classes. (Kassambara, 2018).

4.2.2.1 Linear discriminant analysis (LDA)

The *LDA* algorithm first finds directions that maximize the separation between classes and then uses these directions to predict the individual's class. These directions are called linear discriminants, which are linear combinations of predictor variables (Kassambara, 2018). This model obtained the following results: 0.990 in accuracy and 0.980 in Kappa.

4.2.2.2 Mixture discriminant analysis (MDA)

The *LDA* classifier assumes that each class is from a single normal distribution. This restricts other distributions too much. For *MDA*, there are classes, and each class is assumed to be a Gaussian mixture of subclasses, where each data point has a probability of belonging to each class. We assumed that the covariance matrix between classes is equal (Kassambara, 2018). Table 21 shows the *MDA* resampling results across tuning parameters. The final value used for the model was subclasses = 2.

Table 21: *MDA* resampling results

Subclasses	Accuracy
2	0.9904
3	0.9890
4	0.9899

4.2.2.3 Quadratic discriminant analysis (*QDA*)

QDA is a bit more flexible than *LDA* and *QDA* does not assume the equality of variance or covariance. *LDA* tends to be better than *QDA* when you have a small training set (Kassambara, 2018). By contrast, if the training set is exceptionally large, *QDA* is recommended so that the variance of the classifier is not a major issue, or if the assumption of a common covariance matrix for the K classes is clearly untenable (James et al., 2014). This model obtained the following results: 0.988 in accuracy and 0.976 in Kappa.

4.2.2.4 Regularized discriminant analysis (*RDA*)

RDA builds a classification rule to regularize the group covariance matrices (Friedman, 1989) allowing for a more robust model against multicollinearity in the data. This might be particularly useful for large multivariate datasets that contain highly correlated predictors. Recall that, *LDA* assumed an equality of covariance matrix for all classes, whereas *QDA* assumes different covariance matrices. Regularized discriminant analysis is a middle level between *LDA* and *QDA*. *RDA* reduces the individual covariances of *QDA* to the common covariance, just like *LDA*. In the case in which the number of predictors is greater than that of the number of samples in the training data, this improves the estimate of the covariance matrices, which may increase the accuracy of the model (Kassambara, 2018).

Table 22 shows the resampling results across tuning parameters. The final values used for the model were $\gamma = 0$ and $\lambda = 1$. Accuracy was used to select the optimal model by its largest value; the *MDA* model was slightly better than the others.

Table 22: *RDA* resampling results

Gamma	Lambda	Accuracy
0.0	0.0	0.9879
0.0	0.5	0.9901
0.0	1.0	0.9902
0.5	0.0	0.9864
0.5	0.5	0.9891
0.5	1.0	0.9897
1.0	0.0	0.9784
1.0	0.5	0.9817
1.0	1.0	0.9821

4.2.3 Naive Bayes (*NB*)

The naive Bayes (*NB*) classifier greatly simplifies learning by assuming that the features are independent of a given class; it is based on the Bayes theorem and assumes that the features are independent of a given class. It is a resource-efficient algorithm with fast speed and good scalability (Rish, 2001). Table 23 shows the *NB* results. The final values used for the model were $fL = 0$, $usekernel = \text{TRUE}$, and $adjust = 1$.

Table 23: *NB* resampling results

Usekernel	Accuracy
FALSE	0.9882
TRUE	0.9894

Distribution Type	Accuracy (Bootstrap)
Gaussian	0.9882
Nonparametric	0.9894

4.2.4 Tree-based models

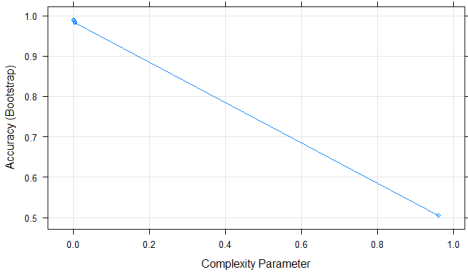
4.2.4.1 Classification and regression tree (*CART*)

The most basic type of tree-structure model is a decision tree or classification and regression tree (*CART*). *CART* analysis is different from the

traditional establishment technology data analysis method. It is suitable for generating clinical decision rules. In addition, *CART* often reveals complex interactions between predictor variables (Lewis, 2000). Table 24 shows the *CART* resampling results across the tuning parameter. The final value used for the model was $cp = 4e-04$.

Table 24: *CART* resampling results

cp	Accuracy
0.0002	0.9871
0.0003	0.9873
0.0003	0.9877
0.0004	0.9878
0.0004	0.9878
0.0012	0.9867
0.0050	0.9832
0.0057	0.9823
0.9619	0.5054



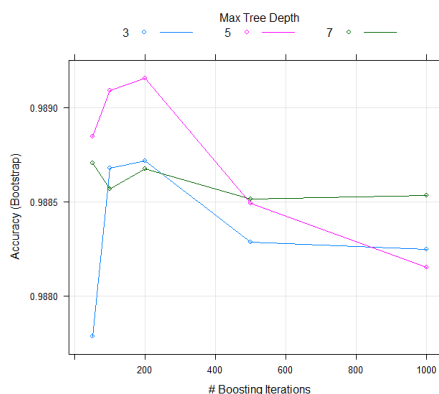
4.2.4.2 Stochastic gradient-boosting machine (*GBM*)

Gradient boosting constructs an additive regression model by sequentially fitting a simple parameterized function (basic learner) to the current pseudo-residuals through the least squares method in each iteration. The pseudo-residual is the gradient of the minimized loss function, relative to the model value of each training data point evaluated in the current step. The results show that both the approximation accuracy and execution speed of gradient boosting can significantly improve stochastic gradient boosting. The stochastic gradient boosting method also increases the robustness to excess capacity of the basic learner (Friedman, 2002). Table 25 shows the *GBM* resampling results across the tuning parameters. The tuning parameter “shrinkage” held constant at a value of 0.1. The tuning parameter “*n.minobsinnode*” held constant at a value of 20. The final values used for the model were $n.trees = 200$, $interaction.depth = 5$, $shrinkage = 0.1$, and $n.minobsinnode = 20$.

Table 25: *GBM* resampling results

Interaction depth	n.trees	Accuracy
3	50	0.9878
3	100	0.9887
3	200	0.9887
3	500	0.9883

3	1000	0.9882
5	50	0.9888
5	100	0.9891
5	200	0.9892
5	500	0.9885
5	1000	0.9882
7	50	0.9887
7	100	0.9886
7	200	0.9887
7	500	0.9885
7	1000	0.9885



4.2.4.2 C5.0 decision trees (C50)

Applying the lifting procedure to the decision tree algorithm can produce an accurate classifier. These classifiers take the form of majority voting on many decision trees. Unfortunately, these classifiers are usually large, complex, and difficult to interpret. *C50* is proven to be competitive with enhanced decision tree algorithms, and the generated rules are usually smaller in size and therefore easier to interpret (Pandya & Pandya, 2015). Table 26 shows the *C50* decision tree resampling results across tuning parameters. The final values used for the model were *trials* = 60, *model* = rules, and *winnow* = TRUE.

Table 26: *C50* resampling results

Model	Winnow	Trials	Accuracy	AccuracySD
tree	FALSE	10	0.9883	0.0370
tree	FALSE	20	0.9882	0.0375
tree	FALSE	30	0.9881	0.0378
tree	FALSE	40	0.9880	0.0383
tree	FALSE	50	0.9881	0.03837
tree	FALSE	60	0.9880	0.0387
tree	FALSE	70	0.9879	0.0388
tree	FALSE	80	0.9878	0.0392
tree	FALSE	90	0.9879	0.0389
tree	FALSE	100	0.9879	0.0389
tree	TRUE	10	0.9883	0.0370
tree	TRUE	20	0.9882	0.0375
tree	TRUE	30	0.9881	0.0378
tree	TRUE	40	0.9880	0.0383
tree	TRUE	50	0.9881	0.0384
tree	TRUE	60	0.9880	0.0387

tree	TRUE	70	0.9879	0.0388
tree	TRUE	80	0.9878	0.0392
tree	TRUE	90	0.9879	0.0389
tree	TRUE	100	0.9879	0.0389
rules	FALSE	10	0.9883	0.0369
rules	FALSE	20	0.9883	0.0378
rules	FALSE	30	0.9882	0.0379
rules	FALSE	40	0.9883	0.0381
rules	FALSE	50	0.9883	0.0380
rules	FALSE	60	0.9884	0.0379
rules	FALSE	70	0.9883	0.0381
rules	FALSE	80	0.9882	0.0383
rules	FALSE	90	0.9883	0.0380
rules	FALSE	100	0.9882	0.0382
rules	TRUE	10	0.9884	0.0369
rules	TRUE	20	0.9883	0.0378
rules	TRUE	30	0.9882	0.0379
rules	TRUE	40	0.9882	0.0381
rules	TRUE	50	0.9883	0.0380
rules	TRUE	60	0.9884	0.0379
rules	TRUE	70	0.9883	0.0381
rules	TRUE	80	0.9882	0.0383
rules	TRUE	90	0.9883	0.0380
rules	TRUE	100	0.9882	0.0382

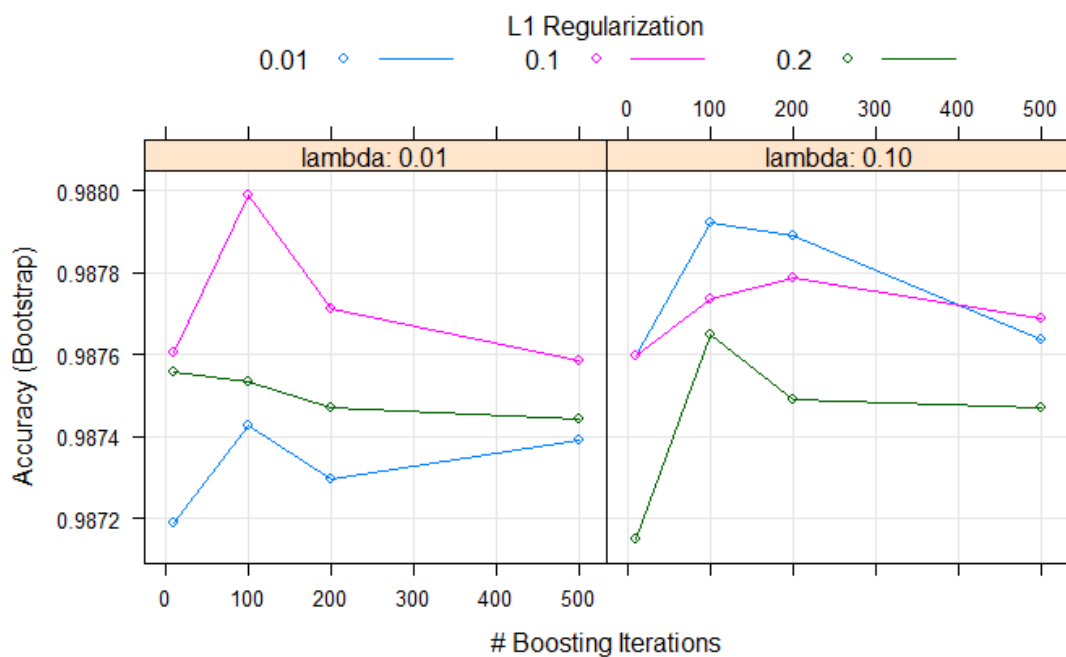
4.2.4.3 eXtreme gradient boosting based on linear model (*XGB_linear*)

eXtreme Gradient Boosting is an efficient and scalable implementation of the gradient boosting framework (Friedman, 2001; Friedman et al., 2000). It includes an efficient linear model solver and tree learning algorithm (Chen, 2015). Table 27 shows the *XGB_linear* resampling results across the tuning parameters. The tuning parameter “eta” held constant at a value of 0.3. The final values used for the model were $nrounds = 100$, $lambda = 0.01$, $alpha = 0.1$, and $eta = 0.3$.

Table 27: *XGB_linear* resampling results

<u>Nrounds</u>	<u>Lambda</u>	<u>Alpha</u>	<u>Eta</u>	<u>Accuracy</u>	<u>AccuracySD</u>	<u>KappaSD</u>
10	0.01	0.01	0.3	0.9872	0.0400	0.0802
10	0.01	0.1	0.3	0.9876	0.0390	0.0781
10	0.01	0.2	0.3	0.9876	0.0383	0.0766
10	0.1	0.01	0.3	0.9876	0.0389	0.0779
10	0.1	0.1	0.3	0.9876	0.0386	0.0772

10	0.1	0.2	0.3	0.9872	0.0401	0.0803
100	0.01	0.01	0.3	0.9874	0.0396	0.0792
100	0.01	0.1	0.3	0.9880	0.0379	0.0757
100	0.01	0.2	0.3	0.9875	0.0390	0.0780
100	0.1	0.01	0.3	0.9879	0.0383	0.0766
100	0.1	0.1	0.3	0.9877	0.0385	0.0770
100	0.1	0.2	0.3	0.9876	0.0390	0.0780
200	0.01	0.01	0.3	0.9873	0.0399	0.0798
200	0.01	0.1	0.3	0.9877	0.0386	0.0771
200	0.01	0.2	0.3	0.9875	0.0392	0.0783
200	0.1	0.01	0.3	0.9879	0.0381	0.0762
200	0.1	0.1	0.3	0.9878	0.0386	0.0771
200	0.1	0.2	0.3	0.9875	0.0395	0.0790
500	0.01	0.01	0.3	0.9874	0.0397	0.0795
500	0.01	0.1	0.3	0.9876	0.0388	0.0775
500	0.01	0.2	0.3	0.9874	0.0391	0.0782
500	0.1	0.01	0.3	0.9876	0.0389	0.0778
500	0.1	0.1	0.3	0.9877	0.0388	0.0776
500	0.1	0.2	0.3	0.9875	0.0395	0.0790



4.2.5 Random Forest (*R-Forest*)

The random forest algorithm is successful as a general classification and regression method. This method combines several random decision trees and aggregates their predictions by averaging. It shows excellent performance in settings in which the number of variables is much larger than the number of observations (Biau & Scornet, 2016). Table 28 shows the R-Forest resampling results across the tuning parameters. The final value used for the model was $mtry = 4$.

Table 29 shows the confusion matrix and statistics for performance and variable importance.

Table 28: *R-Forest* resampling results across tuning parameters

$mtry$	Accuracy
2	0.9894
4	0.9896
6	0.9892
9	0.9886

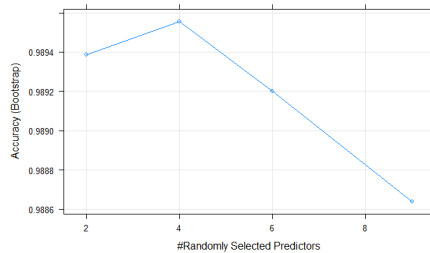
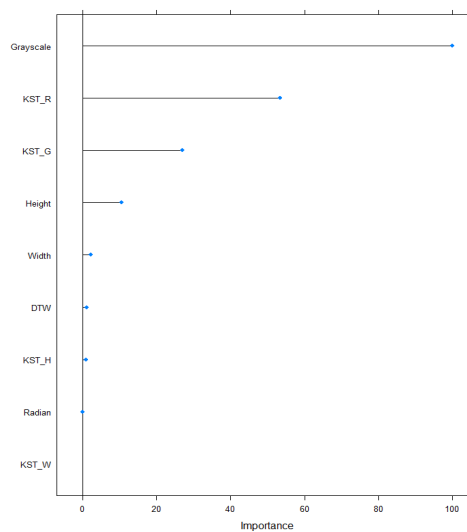


Table 29: Confusion matrix and statistics

Accuracy	0.9968
95% CI	(0.9967, 0.9969)
No information rate	0.8184
p -value [Acc > NIR]	< 2.2e-16
Kappa	0.9894
Mcnemar's test p -value	< 2.2e-16
Sensitivity	0.9967
Specificity	0.9972
Pos pred value	0.9994
Neg pred value	0.9855
Prevalence	0.8184
Detection rate	0.8158
Detection prevalence	0.8163
Balanced accuracy	0.9970

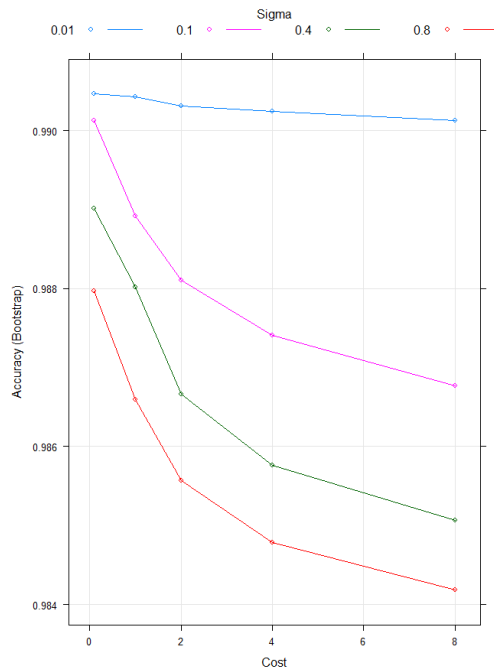


4.2.6 Support vector machines (SVM) with radial basis function kernel

Support vector machine (SVM) is a new type of learning machine based on statistical learning theory. It includes polynomial classifiers, neural networks, and radial basis function (RBF) networks as special cases (Scholkopf et al., 1997). Table 30 shows the SVM resampling results across the tuning parameters. The final values used for the model were $\sigma = 0.01$ and $C = 0.1$.

Table 30: SVM resampling results

Sigma	C	Accuracy	AccuracySD
0.01	0.1	0.9905	0.0312
0.01	1	0.9904	0.0312
0.01	2	0.9903	0.0310
0.01	4	0.9903	0.0309
0.01	8	0.9901	0.0313
0.1	0.1	0.9901	0.0321
0.1	1	0.9889	0.0335
0.1	2	0.9881	0.0351
0.1	4	0.9874	0.0366
0.1	8	0.9868	0.0375
0.4	0.1	0.9890	0.0339
0.4	1	0.9880	0.0355
0.4	2	0.9867	0.0378
0.4	4	0.9858	0.0393
0.4	8	0.9851	0.0403
0.8	0.1	0.9880	0.0354
0.8	1	0.9866	0.0369
0.8	2	0.9856	0.0379
0.8	4	0.9848	0.0381
0.8	8	0.9842	0.0379



4.2.7 Neural networks

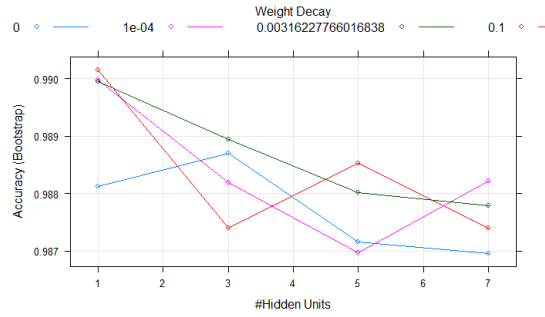
Neural networks (NN) are a large class of flexible nonlinear regression and discriminant models, data reduction models, and nonlinear dynamic systems. They usually are composed of a large number of neurons, that is, simple linear or nonlinear computational elements, which are usually connected to each other in complex ways and often organized into layers (Sarle, 1994). Three NN models were used: neural net (NN), averaged neural Network (AvNN), and neural network with feature extraction (pcaNNet). Table 31–Table 35 and Figure 31: Difference in accuracy between models. show the resampling results across the

tuning parameters.

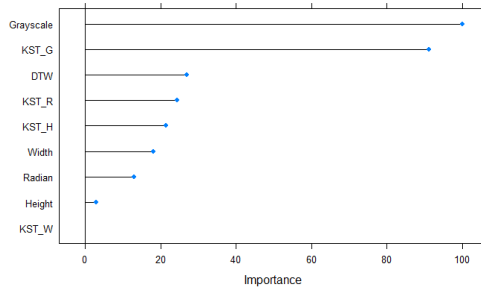
4.2.7.1 Neural net (*nnet*)

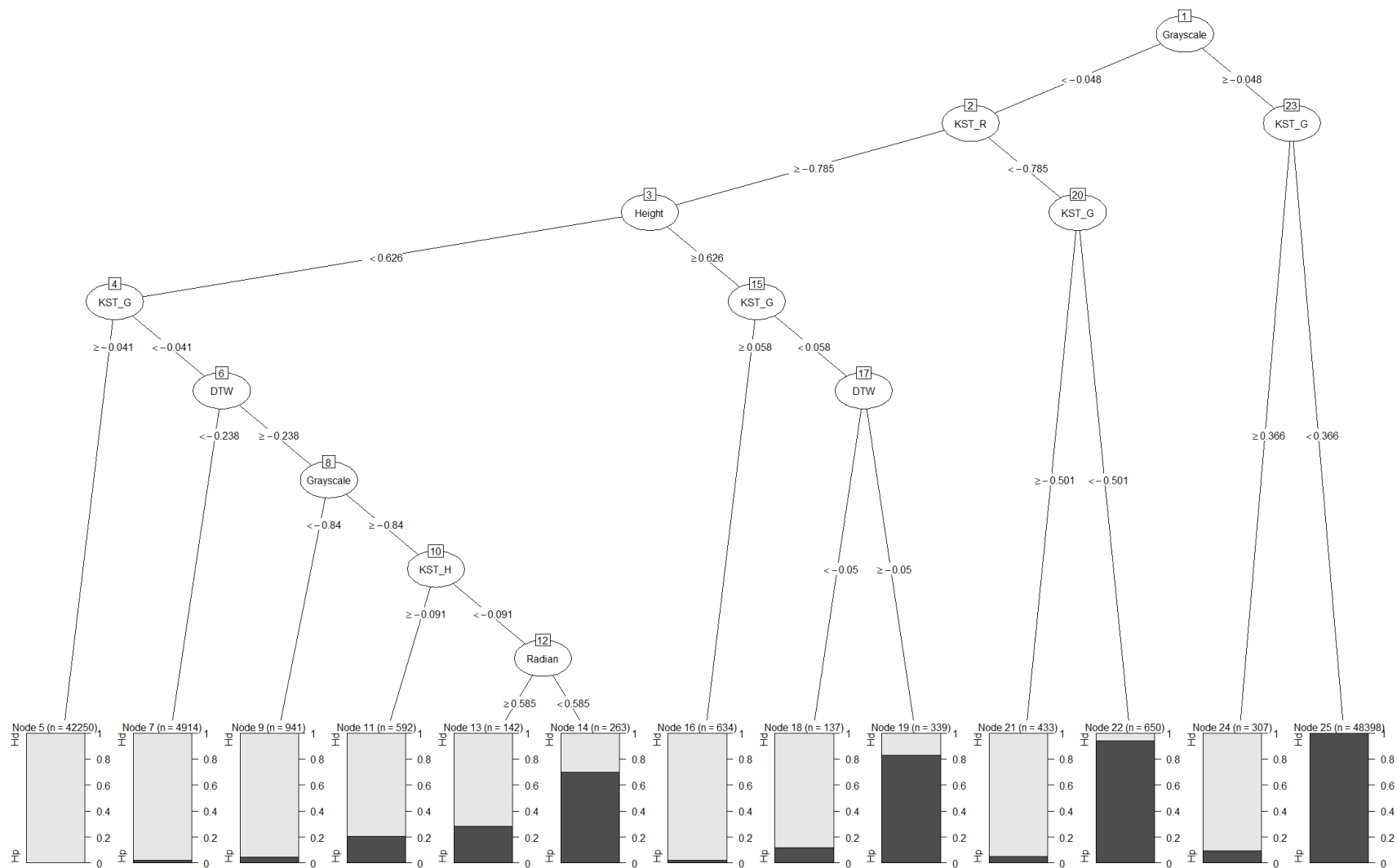
Table 31: *NN* resampling results

Size	Decay	Accuracy
1	0.0000	0.9881
1	0.0001	0.9900
1	0.0032	0.9899
1	0.1000	0.9902
3	0.0000	0.9887
3	0.0001	0.9882
3	0.0032	0.9889
3	0.1000	0.9874
5	0.0000	0.9872
5	0.0001	0.9870
5	0.0032	0.9880
5	0.1000	0.9885
7	0.0000	0.9870
7	0.0001	0.9882
7	0.0032	0.9878
7	0.1000	0.9874



The final values used for the model were size = 1 and decay = 0.1.



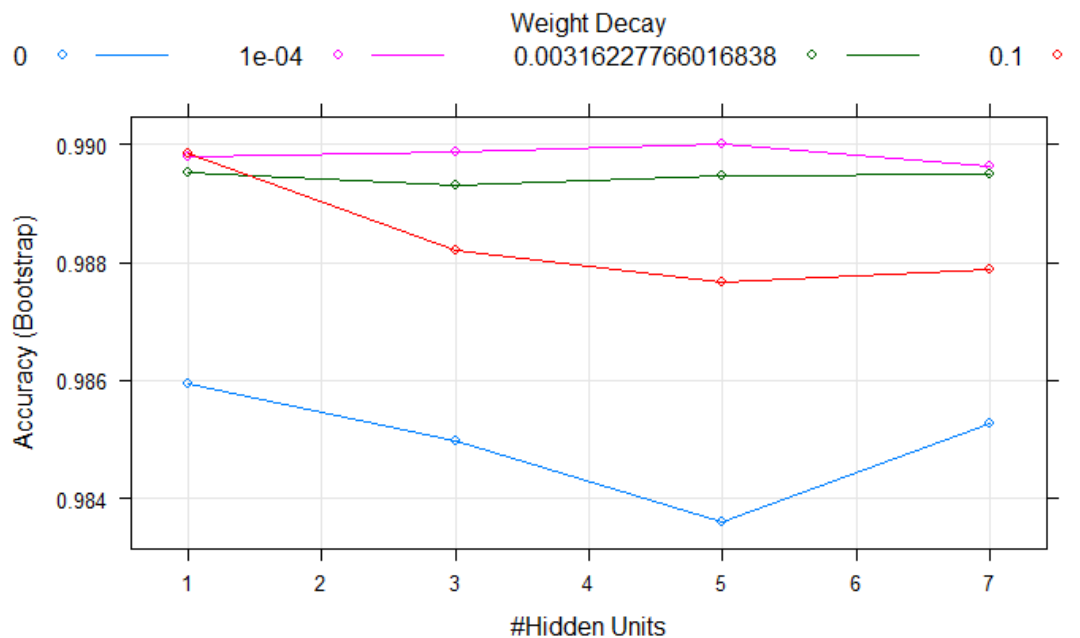


4.3.7.2 Averaged neural network (AvNN)

The tuning parameter “bag” held constant at a value of FALSE. The final values used for the model were $size = 5$, $decay = 1e-04$, and $bag = FALSE$.

Table 32: AvNN resampling results

Size	Decay	Bag	Accuracy	AccuracySD
1	0	FALSE	0.9859	0.0330
1	0.0001	FALSE	0.9898	0.0320
1	0.003162278	FALSE	0.9895	0.0334
1	0.1	FALSE	0.9898	0.0309
3	0	FALSE	0.9850	0.0282
3	0.0001	FALSE	0.9899	0.0330
3	0.003162278	FALSE	0.9893	0.0346
3	0.1	FALSE	0.9882	0.0370
5	0	FALSE	0.9836	0.0310
5	0.0001	FALSE	0.9900	0.0320
5	0.003162278	FALSE	0.9895	0.0339
5	0.1	FALSE	0.9877	0.0380
7	0	FALSE	0.9853	0.0313
7	0.0001	FALSE	0.9896	0.0328
7	0.003162278	FALSE	0.9895	0.0344
7	0.1	FALSE	0.9879	0.0372

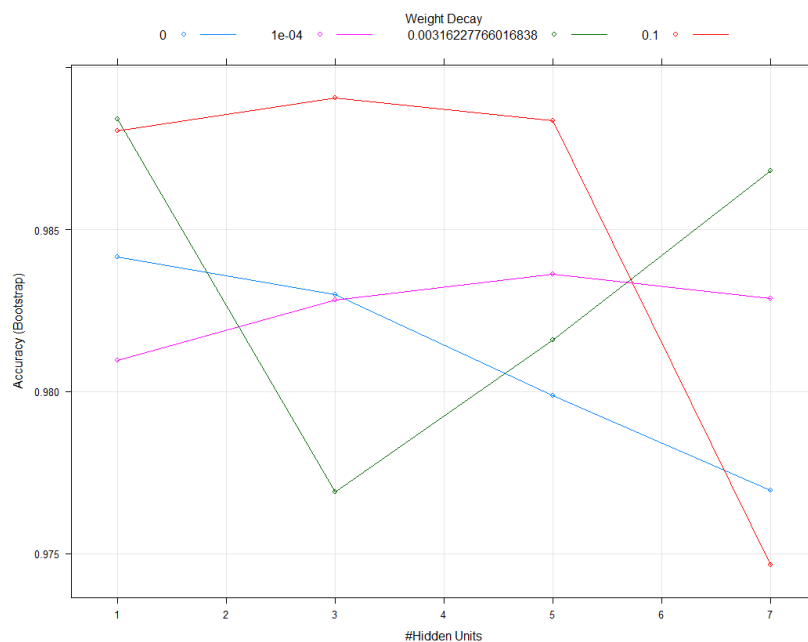


4.3.7.3 Neural networks with PCA feature extraction (*pcaNNet*)

The final values used for the model were *size* = 3 and *decay* = 0.1.

Table 33: *pcaNNet* resampling results

Size	Decay	Accuracy	AccuracySD
1	0.0000	0.9842	0.0370
1	0.0001	0.9810	0.0740
1	0.0032	0.9884	0.0348
1	0.1000	0.9880	0.0354
3	0.0000	0.9830	0.0302
3	0.0001	0.9828	0.0539
3	0.0032	0.9769	0.0778
3	0.1000	0.9891	0.0339
5	0.0000	0.9799	0.0400
5	0.0001	0.9836	0.0471
5	0.0032	0.9816	0.0511
5	0.1000	0.9884	0.0358
7	0.0000	0.9769	0.0641
7	0.0001	0.9829	0.0496
7	0.0032	0.9868	0.0374
7	0.1000	0.9746	0.0866



4.2.8 Comparing models

Thirteen models were selected for the previous analysis: *LDA*, *MDA*, *QDA*, *RDA*, *CART*, *KNN*, *R-Forest*, *NN*, *AvNN*, *C50*, *GBM*, *XGB_linear*, and *NB*. As described at the beginning of Section 4.2, the number of individuals for resamples was 48. See Table 34: Model accuracy comparison for the accuracy results, and Table 35: Kappa of model comparison for the Kappa results. There is no significant difference between models; however, the accuracy of *R-Forest* showed smaller variation in accuracy than other models. Therefore, *R-Forest* was selected as the best model.

Table 34: Model accuracy comparison

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
<i>LDA</i>	0.8210	0.9971	0.9995	0.9902	1.0000	1	0
<i>MDA</i>	0.8228	0.9961	0.9995	0.9904	1.0000	1	0
<i>QDA</i>	0.8230	0.9973	0.9998	0.9879	1.0000	1	0
<i>RDA</i>	0.8210	0.9971	0.9995	0.9902	1.0000	1	0
<i>CART</i>	0.8224	0.9949	0.9984	0.9878	0.9991	1	0
<i>KNN</i>	0.8343	0.9941	0.9982	0.9891	0.9995	1	0
<i>R-Forest</i>	0.8228	0.9983	0.9992	0.9896	1.0000	1	0
<i>NN</i>	0.8425	0.9957	0.9984	0.9902	0.9995	1	0
<i>AvNN</i>	0.8348	0.9959	0.9992	0.9900	1.0000	1	0
<i>C50</i>	0.8239	0.9978	0.9992	0.9884	1.0000	1	0
<i>GBM</i>	0.8302	0.9959	0.9986	0.9892	0.9993	1	0
<i>XGB_linear</i>	0.8260	0.9968	0.9989	0.9880	0.9996	1	0
<i>NB</i>	0.8348	0.9959	0.9984	0.9894	1.0000	1	0

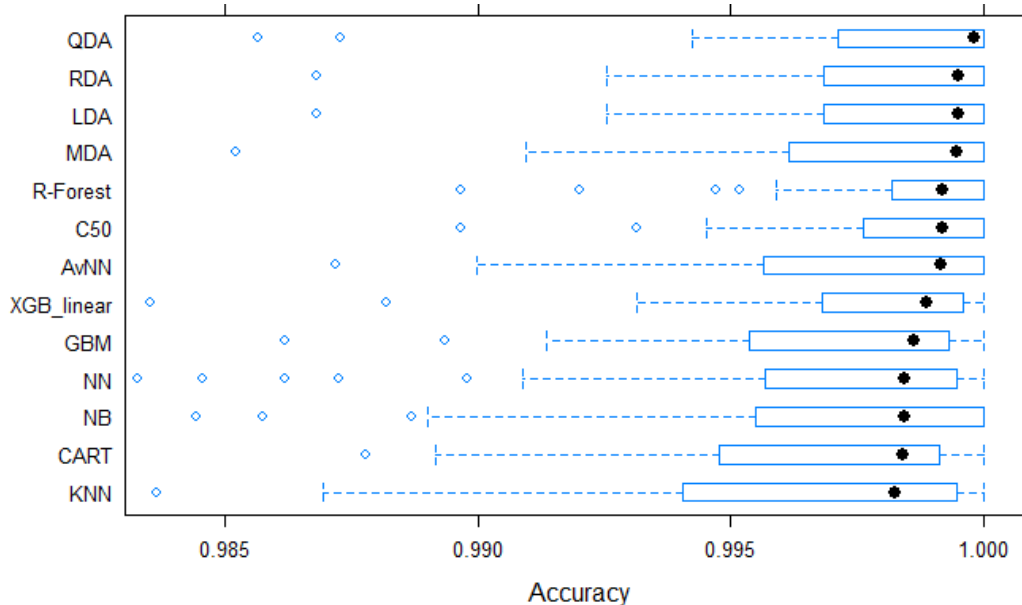


Table 35: Kappa of model comparison

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
<i>LDA</i>	0.6442	0.9938	0.9990	0.9805	1.0000	1	0
<i>MDA</i>	0.6477	0.9922	0.9990	0.9809	1.0000	1	0
<i>QDA</i>	0.6489	0.9945	0.9996	0.9759	1.0000	1	0
<i>RDA</i>	0.6442	0.9938	0.9990	0.9805	1.0000	1	0
<i>CART</i>	0.6472	0.9898	0.9968	0.9756	0.9982	1	0
<i>KNN</i>	0.6712	0.9882	0.9964	0.9783	0.9989	1	0
<i>R-Forest</i>	0.6478	0.9966	0.9984	0.9791	1.0000	1	0
<i>NN</i>	0.6874	0.9914	0.9968	0.9803	0.9990	1	0
<i>AvNN</i>	0.6723	0.9914	0.9983	0.9801	1.0000	1	0
<i>C50</i>	0.6500	0.9955	0.9984	0.9768	1.0000	1	0
<i>GBM</i>	0.6632	0.9917	0.9973	0.9783	0.9985	1	0
<i>XGB_linear</i>	0.6542	0.9937	0.9977	0.9759	0.9992	1	0
<i>NB</i>	0.6723	0.9918	0.9969	0.9788	1.0000	1	0

Call:

```
summary.diff.resamples(object = rocDiffs)
```

p-value adjustment: bonferroni

Upper diagonal: estimates of difference

Lower diagonal: *p*-value for H0: difference = 0

Kappa

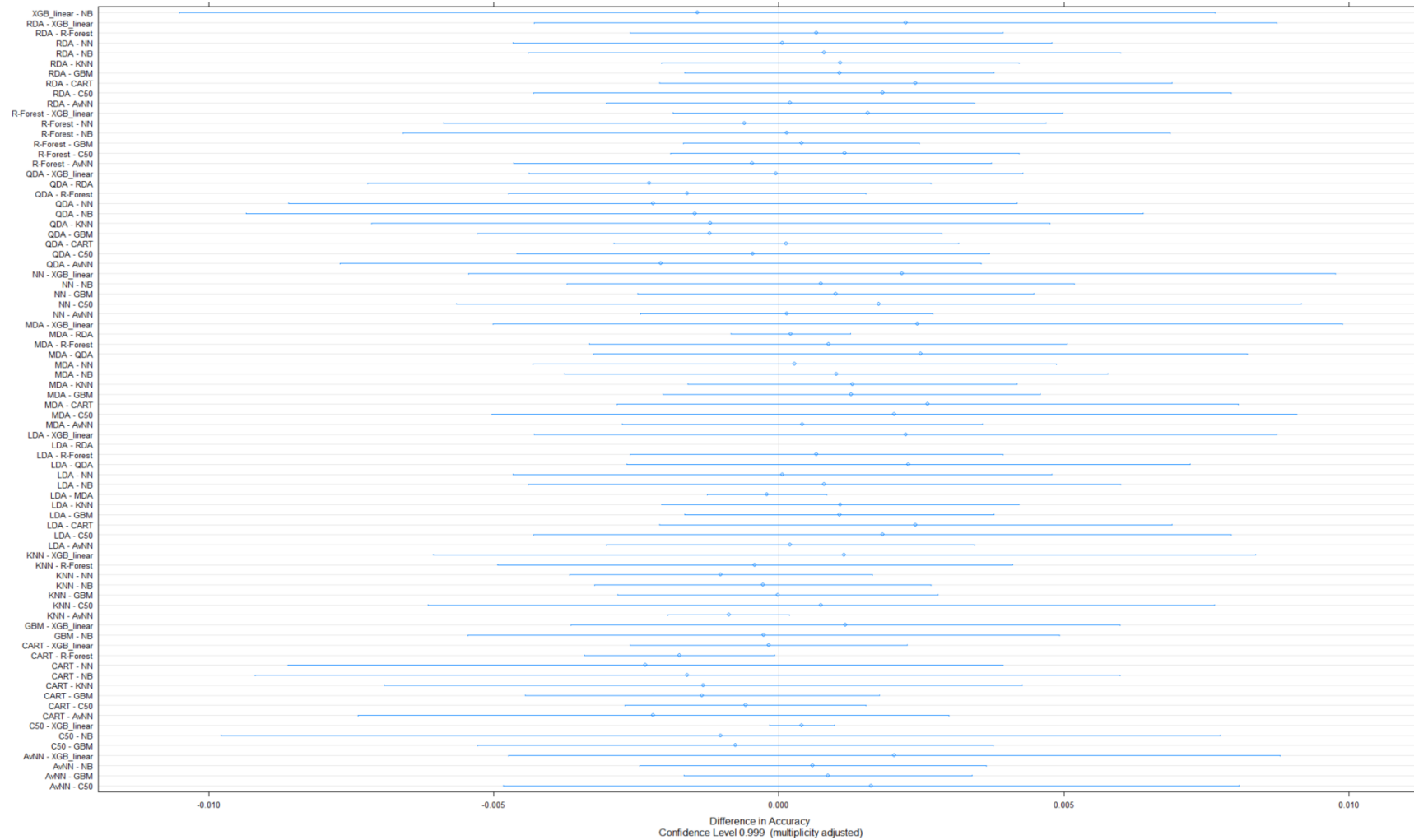


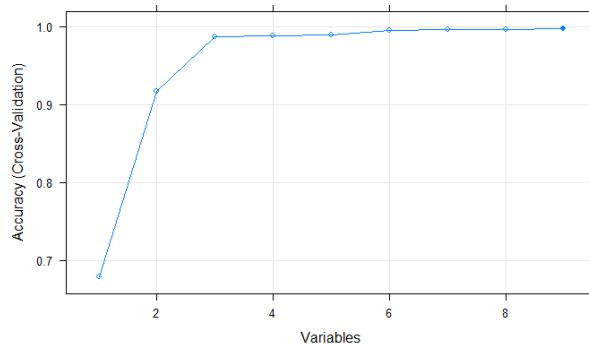
Figure 31: Difference in accuracy between models.

4.2.9 Application to test data

As described in comparison of models, the *R-Forest* model is very efficient. Thus, the *R-Forest* model was used to validate the performance on test data. For recursive feature selection based on training data, Recursive feature elimination (RFE) provides a simple backward selection of predictors based on predictor importance ranking, function of *ref*¹⁴ in *caret* (Kuhn, 2021). The cross-validation (tenfold) was used as the outer resampling method to select the best predictors (Table 36).

Table 36: *R-Forest* resampling result used to select the best predictors

Variables	Accuracy	Kappa	AccuracySD	KappaSD	Selected
1	0.6794	0.3589	0.0039429	0.007886	
2	0.9169	0.8337	0.0885523	0.177105	
3	0.9873	0.9746	0.0008587	0.001717	
4	0.9883	0.9767	0.0008140	0.001628	
5	0.9893	0.9787	0.0006980	0.001396	
6	0.9956	0.9912	0.0006596	0.001319	
7	0.9962	0.9923	0.0007367	0.001473	
8	0.9965	0.9930	0.0008386	0.001677	
9	0.9970	0.9939	0.0007166	0.001433	*



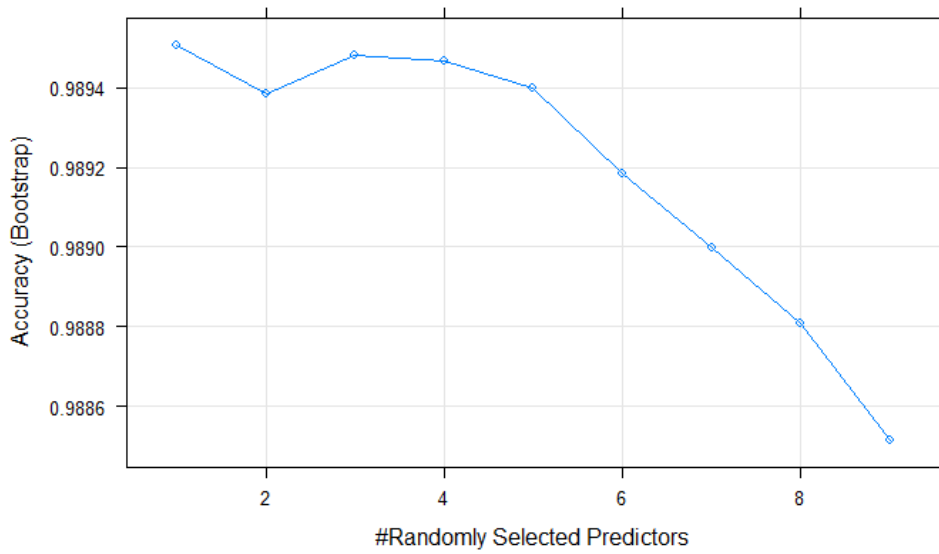
The top five variables (out of nine) were radian, grayscale, DTW, KST_H, and width.

Next, the *R-Forest* model was retrained based on the selected set of predictors (Table 37).

¹⁴ Model details about these functions, including examples, are available at <http://topepo.github.io/caret/recursive-feature-elimination.html>

Table 37: *R*-Forest resampling result (retrained)

Mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.9895	0.9790	0.0361	0.0717
2	0.9894	0.9788	0.0365	0.0726
3	0.9895	0.9790	0.0364	0.0725
4	0.9895	0.9790	0.0362	0.0720
5	0.9894	0.9788	0.0363	0.0723
6	0.9892	0.9784	0.0364	0.0723
7	0.9890	0.9780	0.0363	0.0722
8	0.9888	0.9776	0.0364	0.0724
9	0.9885	0.9770	0.0366	0.0728



The final value used for the model was mtry = 1.

The dataML_test1 and dataML_test2 with all models were predicted. All results have been compiled (Table 38 and Table 39).

Table 38: Performance of MLs_test1 (sample size = 100,000)

MLmodel	Accuracy	Sensitivity	Specificity	Pos pred	Neg pred	Precision	Recall	F1	Prevalence	Detection	Detection	Balanced	Pred_Hd_	Pred_Hp		
				value	value					rate	prevalence		accuracy	when_Hp	Sum_Hp	_when_Hd
AvNN	0.9980	0.9998	0.9699	0.9980	0.9973	0.9980	0.9998	0.9989	0.9381	0.9379	0.9398	0.9849	711	23652	61	358453
R-Forest	0.9979	0.9994	0.9755	0.9984	0.9901	0.9984	0.9994	0.9989	0.9381	0.9375	0.9390	0.9874	579	23652	231	358453
MDA	0.9976	0.9989	0.9791	0.9986	0.9828	0.9986	0.9989	0.9987	0.9381	0.9370	0.9383	0.9890	495	23652	405	358453
QDA	0.9974	0.9993	0.9696	0.9980	0.9884	0.9980	0.9993	0.9986	0.9381	0.9374	0.9393	0.9844	720	23652	269	358453
RDA	0.9970	0.9983	0.9778	0.9985	0.9737	0.9985	0.9983	0.9984	0.9381	0.9365	0.9378	0.9881	524	23652	625	358453
LDA	0.9970	1.0000	0.9517	0.9968	0.9995	0.9968	1.0000	0.9984	0.9381	0.9381	0.9411	0.9758	1142	23652	11	358453
NB	0.9969	1.0000	0.9507	0.9968	0.9999	0.9968	1.0000	0.9984	0.9381	0.9381	0.9411	0.9754	1165	23652	3	358453
C50	0.9907	0.9912	0.9844	0.9990	0.8801	0.9990	0.9912	0.9950	0.9381	0.9298	0.9308	0.9878	370	23652	3171	358453
KNN	0.9895	1.0000	0.8300	0.9889	0.9999	0.9889	1.0000	0.9944	0.9381	0.9381	0.9486	0.9150	4020	23652	1	358453
XGB_linear	0.9327	0.9288	0.9909	0.9994	0.4788	0.9994	0.9288	0.9628	0.9381	0.8713	0.8719	0.9599	215	23652	25509	358453
GBM	0.8549	0.8454	0.9979	0.9998	0.2987	0.9998	0.8454	0.9162	0.9381	0.7931	0.7932	0.9217	49	23652	55409	358453
NN	0.8181	0.8062	0.9992	0.9999	0.2538	0.9999	0.8062	0.8927	0.9381	0.7563	0.7563	0.9027	19	23652	69483	358453
CART	0.7573	0.7414	0.9985	0.9999	0.2030	0.9999	0.7414	0.8514	0.9381	0.6955	0.6956	0.8699	36	23652	92705	358453

Table 39: Performance of MLs_test2 (sample size = 100,000)

Rank	MLmodel	Accuracy	Sensitivity	Specificity	Pos		Precision	Recall	F1	Prevalence	Detection	Detection	Balanced	Pred_Hd_	Pred_Hp		
					pred	Neg pred					rate	prevalence	accuracy	when_Hp	Sum_Hp	_when_Hd	Sum_Hd
1	R-Forest	0.9988	0.9995	0.9960	0.9990	0.9979	0.9990	0.9995	0.9992	0.7962	0.7958	0.7966	0.9977	507	127282	267	497256
2	MDA	0.9987	0.9990	0.9972	0.9993	0.9963	0.9993	0.9990	0.9992	0.7962	0.7954	0.7960	0.9981	360	127282	474	497256
3	AvNN	0.9986	0.9998	0.9942	0.9985	0.9990	0.9985	0.9998	0.9991	0.7962	0.7960	0.7972	0.9970	734	127282	122	497256
4	QDA	0.9984	0.9995	0.9942	0.9985	0.9981	0.9985	0.9995	0.9990	0.7962	0.7958	0.7970	0.9969	738	127282	245	497256
5	RDA	0.9982	0.9987	0.9966	0.9991	0.9948	0.9991	0.9987	0.9989	0.7962	0.7951	0.7958	0.9976	432	127282	664	497256
6	LDA	0.9977	1.0000	0.9890	0.9972	0.9999	0.9972	1.0000	0.9986	0.7962	0.7962	0.7984	0.9945	1396	127282	14	497256
7	NB	0.9976	1.0000	0.9885	0.9971	0.9999	0.9971	1.0000	0.9985	0.7962	0.7962	0.7985	0.9943	1460	127282	13	497256
8	C50	0.9930	0.9917	0.9980	0.9995	0.9686	0.9995	0.9917	0.9956	0.7962	0.7896	0.7900	0.9948	260	127282	4112	497256
9	KNN	0.9812	1.0000	0.9076	0.9769	1.0000	0.9769	1.0000	0.9883	0.7962	0.7962	0.8150	0.9538	11755	127282	1	497256
10	XGB_linear	0.9379	0.9223	0.9988	0.9997	0.7670	0.9997	0.9223	0.9595	0.7962	0.7344	0.7346	0.9606	148	127282	38616	497256
11	GBM	0.8640	0.8294	0.9994	0.9998	0.5999	0.9998	0.8294	0.9067	0.7962	0.6604	0.6605	0.9144	76	127282	84834	497256
12	NN	0.8079	0.7589	0.9993	0.9998	0.5147	0.9998	0.7589	0.8628	0.7962	0.6042	0.6043	0.8791	88	127282	119905	497256
13	CART	0.7814	0.7256	0.9993	0.9998	0.4825	0.9998	0.7256	0.8409	0.7962	0.5777	0.5779	0.8625	84	127282	136441	497256

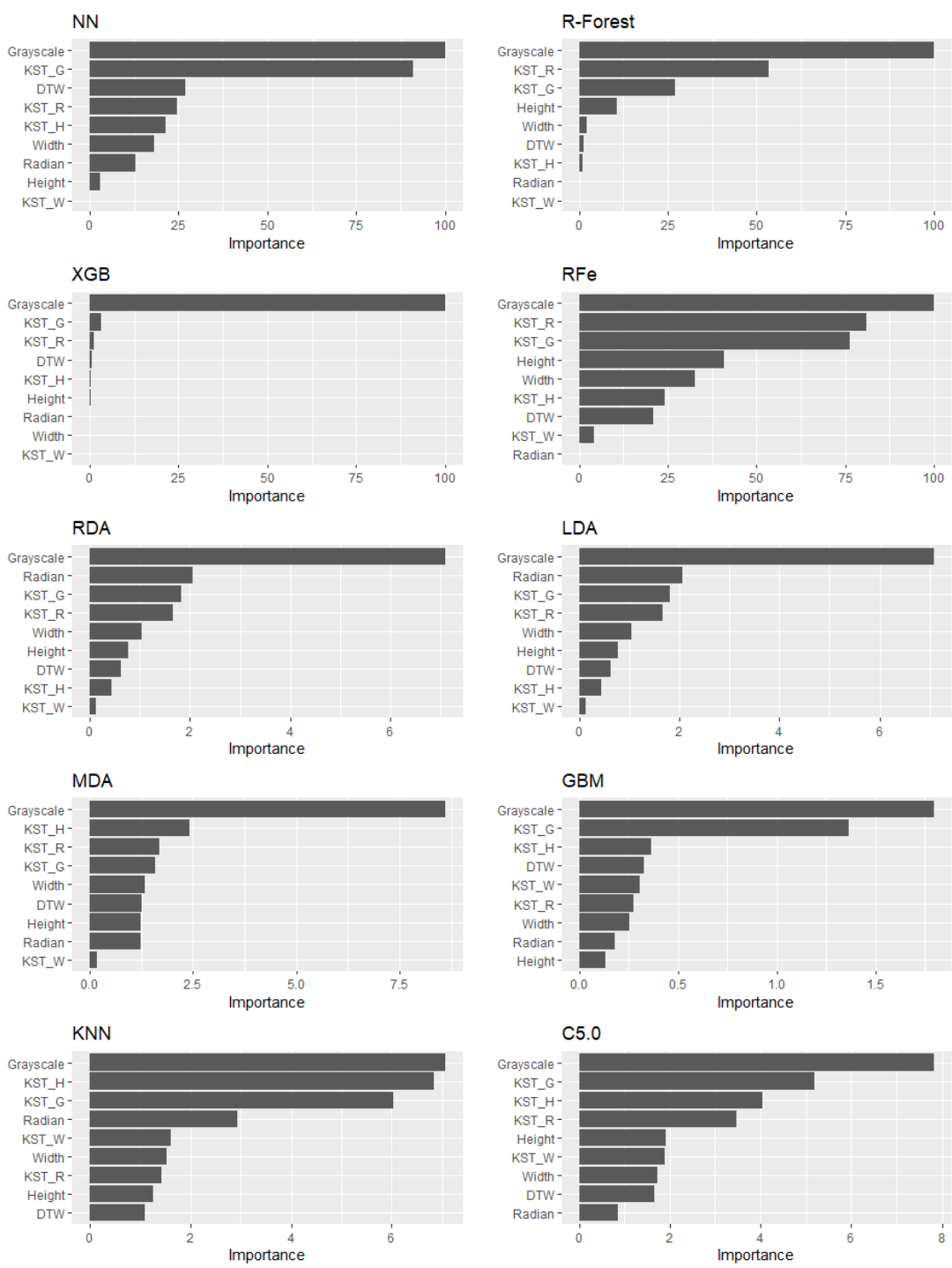
Using a different number of signatures as references for the five ML methods (Hsu & Lin, 2002; Dudani, 1976; Ho, 1995; Breilman, 2001; Lippmann, 1987) led to different results (Table 40). As expected, increasing the number of references led to improved performance. The overall results showed outstanding performance: even if there was only one genuine signature for reference, performance was significant (above 96.50% in precision, 99.31% in recall, and 98.04% in F1 score) in *SVM*, *KNN*, *R-Forest*, and *MLP*. When the references were increased to three, significant improvement was obtained for each method (except for the precision of *MLP*). Random forest showed the best performance (99.75% in precision, 100% in recall, and 99.88% in F1 score). Performance continuously improved when there were five references. Random forest again showed the best performance (99.86% in precision, 100% in recall, and 99.93% in F1 score).

Table 40: Performance evaluation of ML

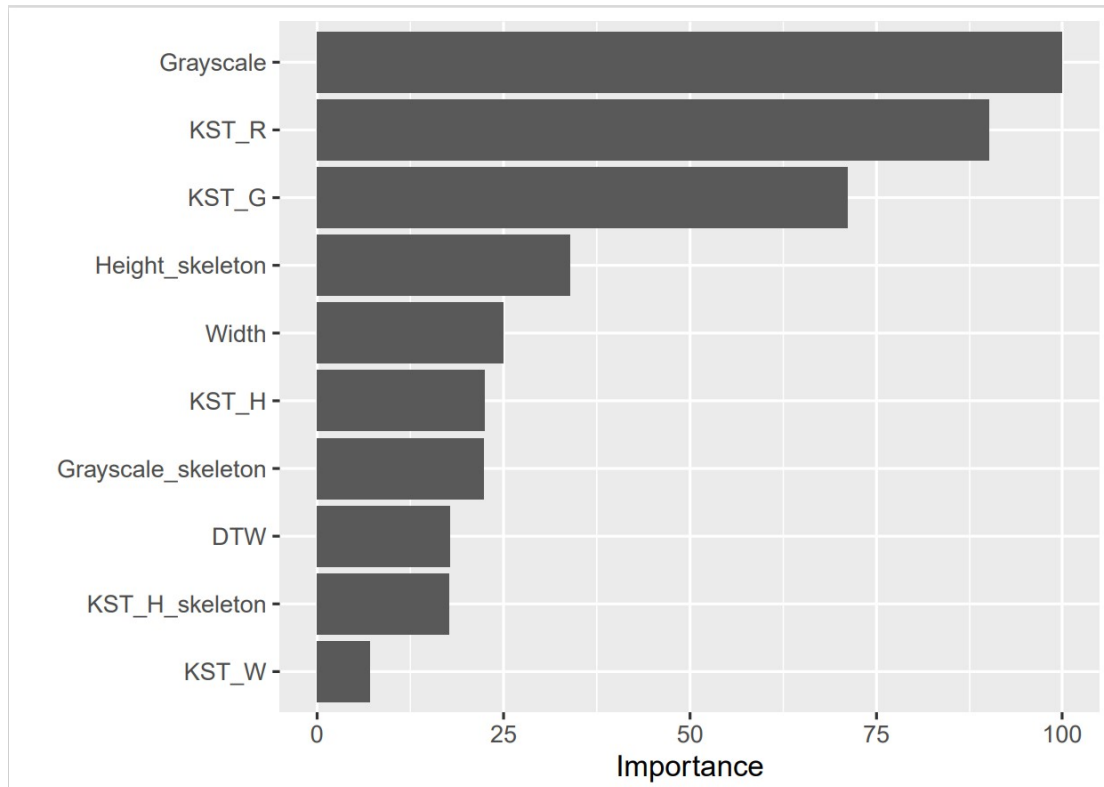
Reference number	ML Algorithm	NB	SVM	KNN	R-Forest	MLP
1	Precision	0.9342	0.9650	0.9784	0.9877	0.9903
	Recall	0.9864	0.9963	0.9931	0.9956	0.9961
	F1 score	0.9596	0.9804	0.9857	0.9916	0.9932
3	Precision	0.9760	0.9877	0.9934	0.9975	0.9897
	Recall	0.9988	1.0000	0.9996	1.0000	0.9996
	F1 score	0.9873	0.9938	0.9965	0.9988	0.9946
5	Precision	0.9843	0.9885	0.9950	0.9986	0.9971
	Recall	1.0000	1.0000	1.0000	1.0000	1.0000
	F1 score	0.9921	0.9942	0.9975	0.9993	0.9986

4.2.10 Variable importance

Variable importance analysis was implemented in *R-Forest*, and *nnet* (see Table 29 and Table 31). To show the variable importance of different ML models, using the *vip* package, eight ML models provide variable importance information (see Figure 32(a)). The *R-Forest* model served as the basis for the variable importance analysis (see Figure 32(b)) for the results of the variable importance analysis.



a) Eight models providing variable importance



b) *R-Forest* model retained on training dataset selected the best parameter.

Figure 32: Variable importance analysis

The variable importance result based on *R-Forest* was used as the weight parameters in the LR calculation using DST.

4.3 Performance evaluation of score-based LR system

The LR's performance evaluation of dataset_2 was published in Chen (2018), which is included in the appendix.

On dataset_3, we have used two methods of calibration, logistic and PAVA, to calibrate the LR system based on MKDE and DST, respectively.

See Table 41 for the method matrix of the performance evaluation.

Table 41: Method matrix of performance evaluation

	LLR calculation	Calibration	Mode	Result
1	MKDE	Logistic	Inner individual	Figure 33
2	MKDE	Logistic	Inter individual	Figure 34
3	MKDE	PAVA	Inner individual	Figure 35
4	MKDE	PAVA	Inter individual	Figure 36
5	DST	Logistic	Inner individual	Figure 39

6	DST	Logistic	Inter individual	Figure 40
7	DST	PAVA	Inner individual	Figure 41
8	DST	PAVA	Inter individual	Figure 42

4.3.1 PAVA calibration of score-based LR using MKDE based on dataset_3

We took advantage of the *comparison* package (Lucy et al., 2020) to perform multivariate LR calculation and evaluation on dataset_3. It calculates the calibrated set of LRs using logistic regression. Additionally, to show the performance based on the two methods of regression, PAVA regression was also used to calculate the calibrated set of LRs.

For inner- and inter-individual scenarios, the ECE plots based on PAVA and logistic regression showed similar performance, as well as RMEP and RMED. The density distribution of same sources and different sources showed different regression results with different density distribution. Logistic regression showed smoother density than PAVA; the former seemed to be more reasonable. In the following validation tests (Section 4.4), calibration based on logistic regression will be used (Figure 33 to Figure 36).

For each LR option in both inner- and inter-individual scenarios, the density distributions for genuine signatures and forgeries were different. Additionally, different types of forgeries (RF, FF, and TF) showed similar density distributions: traced forgeries showed the closest distance to genuine density, and random forgeries showed farthest distance to genuine density (Figure 37).

Skilled forgeries, when selected 30 times by CNN deep learning, showed images similar to genuine signatures. However, the density distribution between skilled and non-skilled forgeries did not show significant differences in calibrated LLR. This was because CNN and this system each depend on different features: 2D images vs 3D and pseudo-dynamic features. The similar images did not necessarily mislead skilled forgeries as genuine signatures. For instance, Figure 37 shows the histogram of calibrated LLR and skilled forgery vs ordinary forgery (five feature options, using MKDE, inner-individual mode). The histogram distribution of skilled forgery showed slightly closer to genuine than ordinary forgery to genuine.

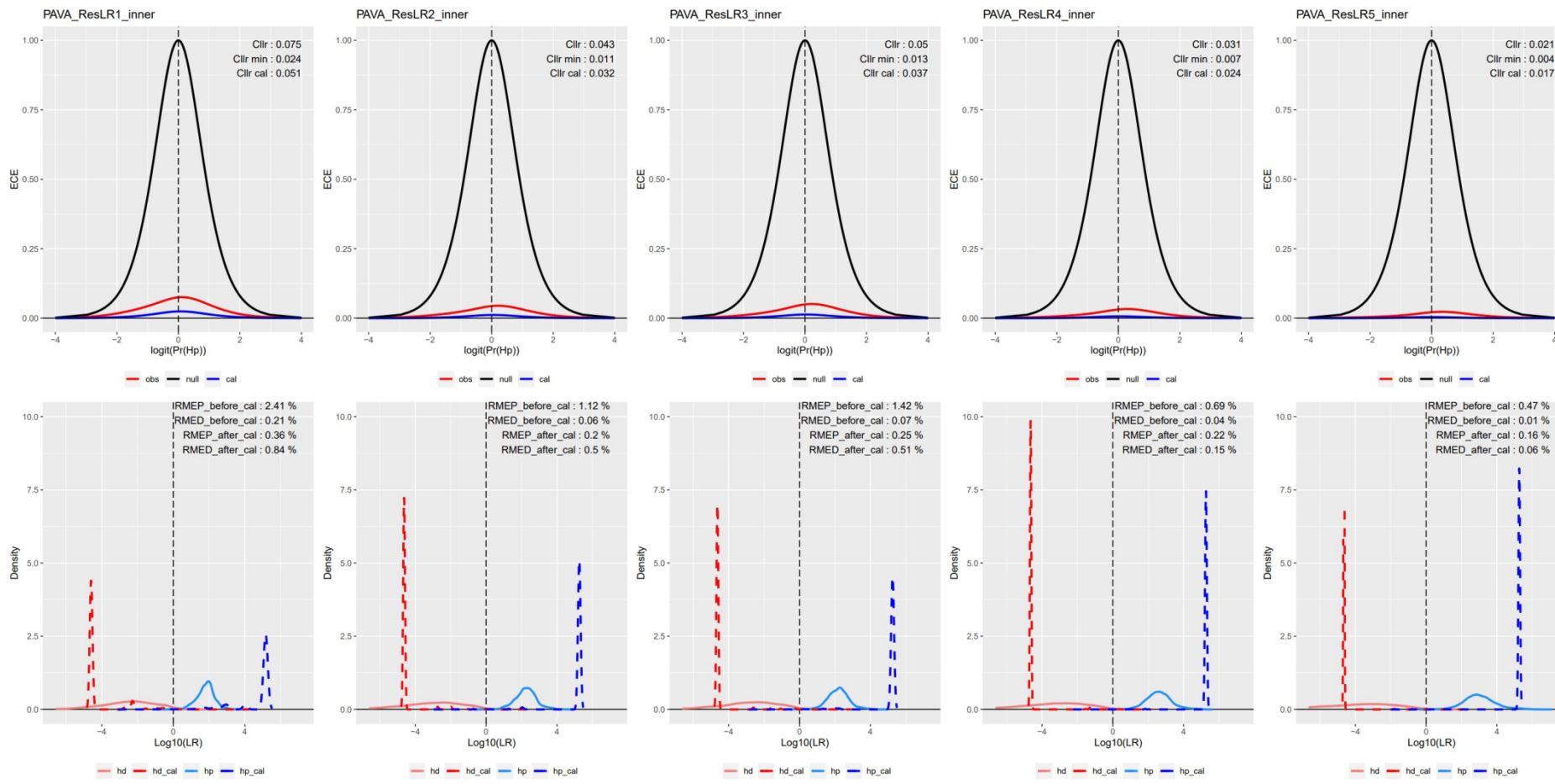


Figure 33: PAVA calibration for the inner-individual mode using MKDE. Upper figures are ECE plots; lower figures are $\log_{10}(\text{LR})$ density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

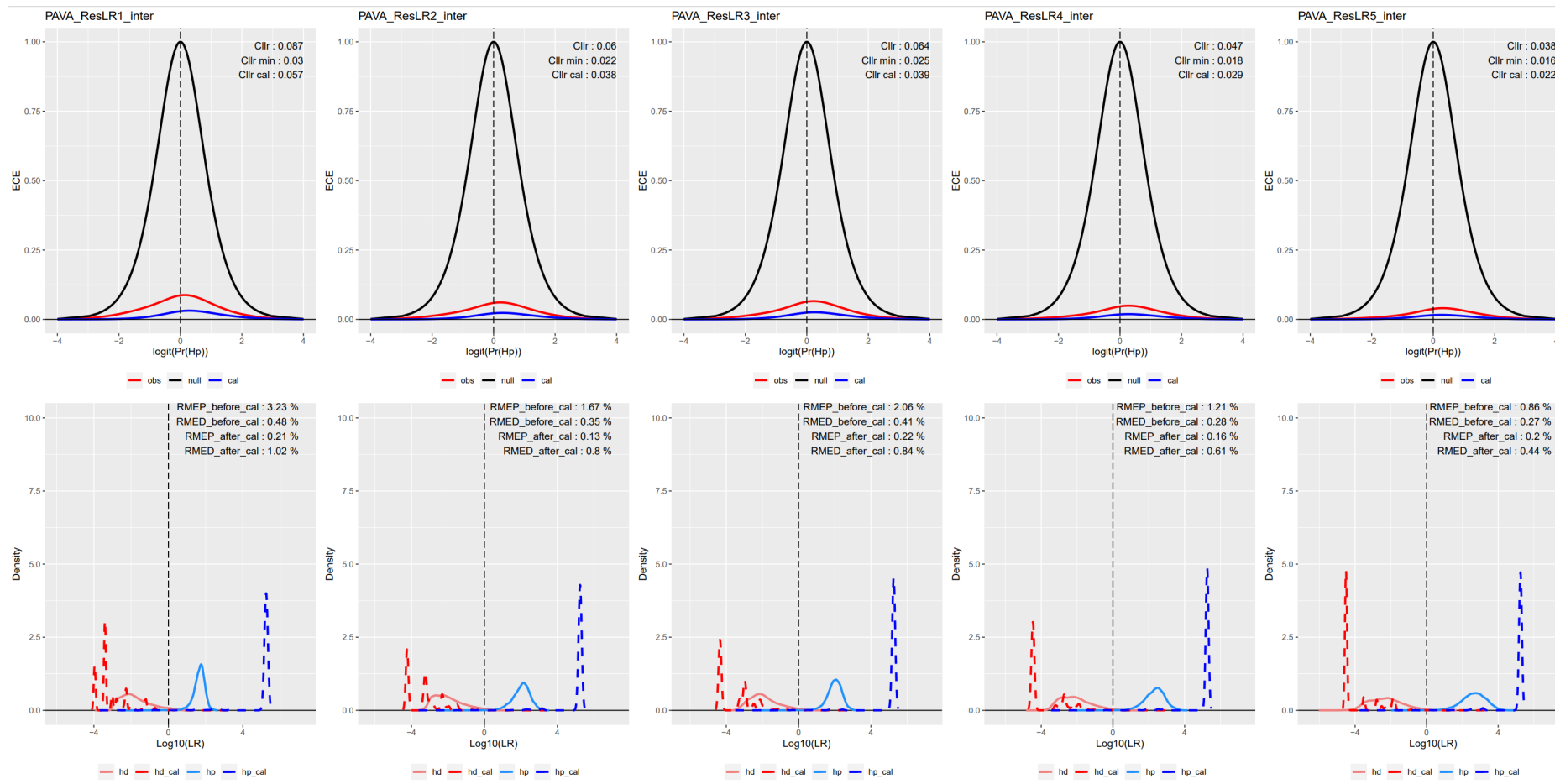


Figure 34: PAVA calibration for the inter-individual mode using MKDE. Upper figures are ECE plots; lower figures are log10(LR) density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

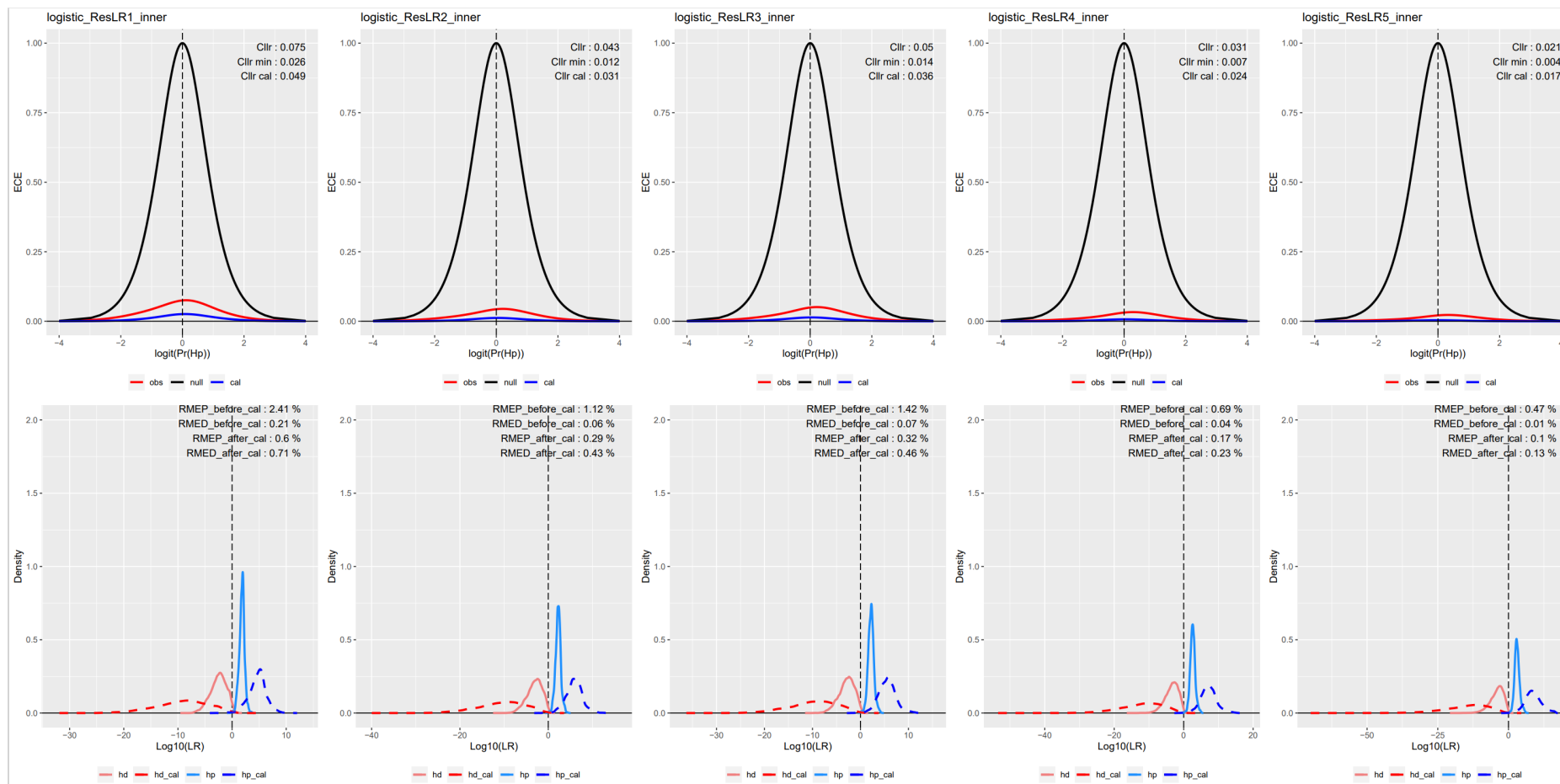


Figure 35: Logistic calibration for the inner-individual mode using MKDE. Upper figures are ECE plots; lower figures are log10(LR) density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

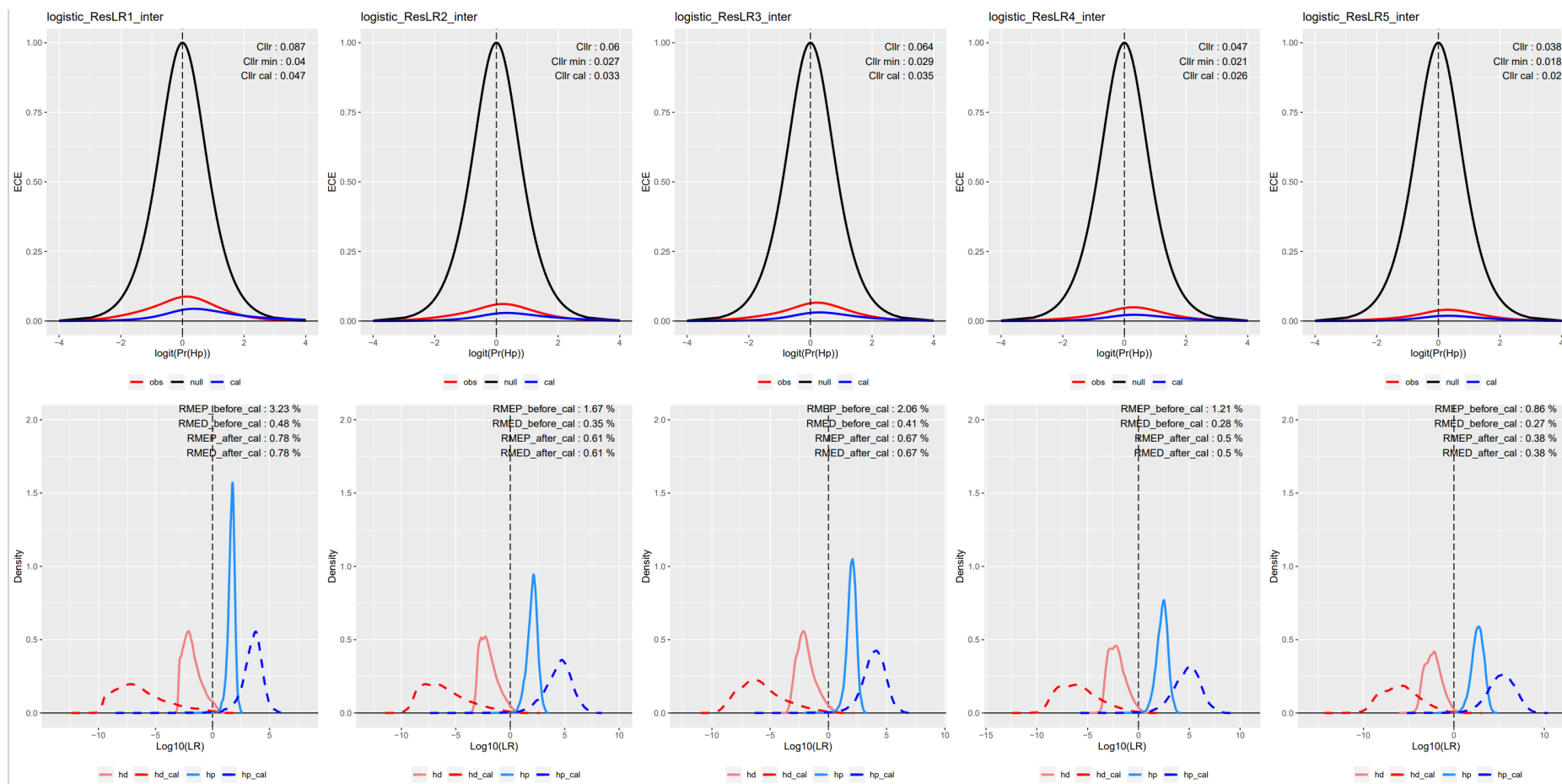
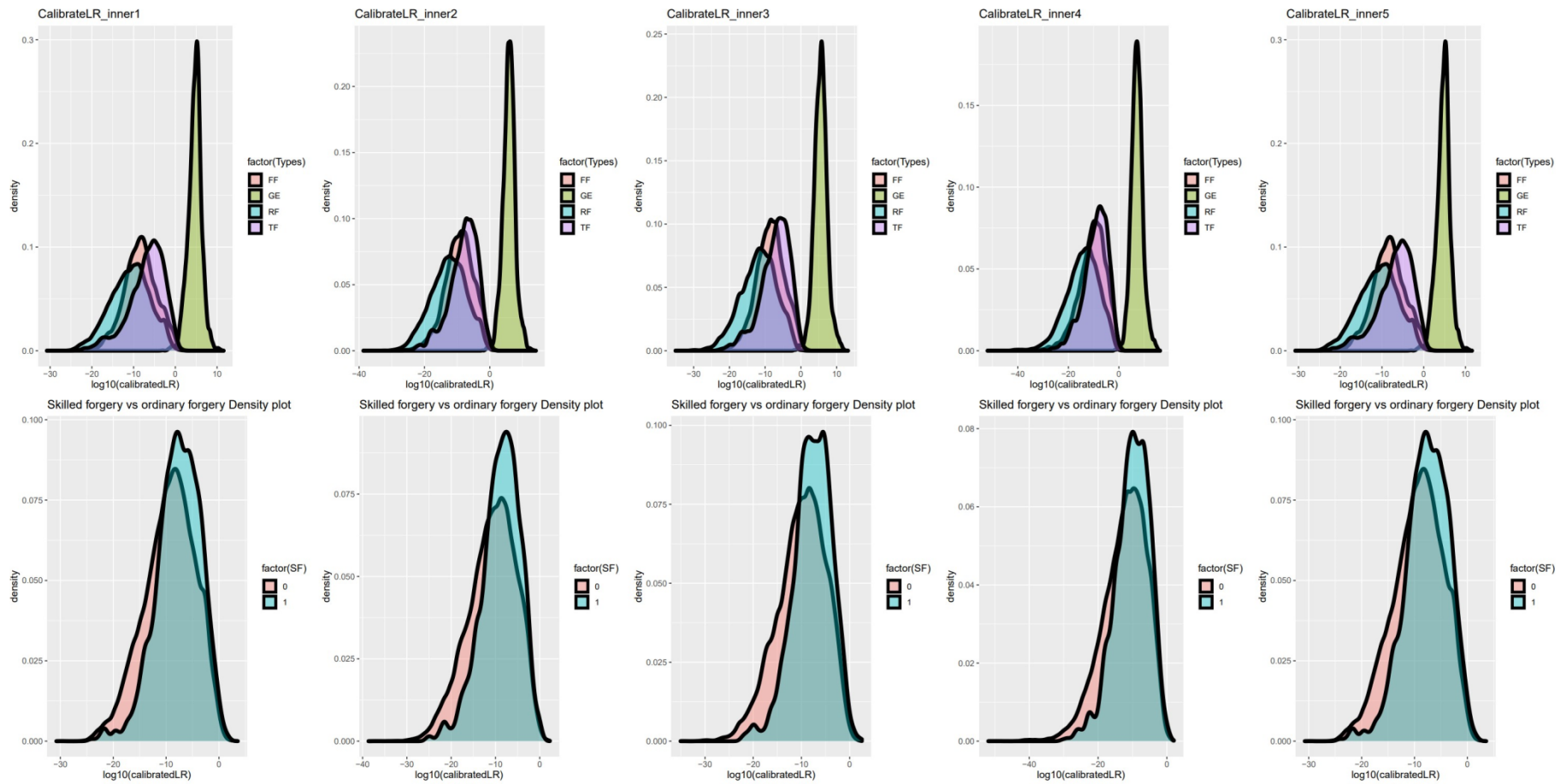


Figure 36: Logistic calibration for the inter-individual mode using MKDE. Upper figures are ECE plots; lower figures are log10(LR) density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.



a) Inner individual mode

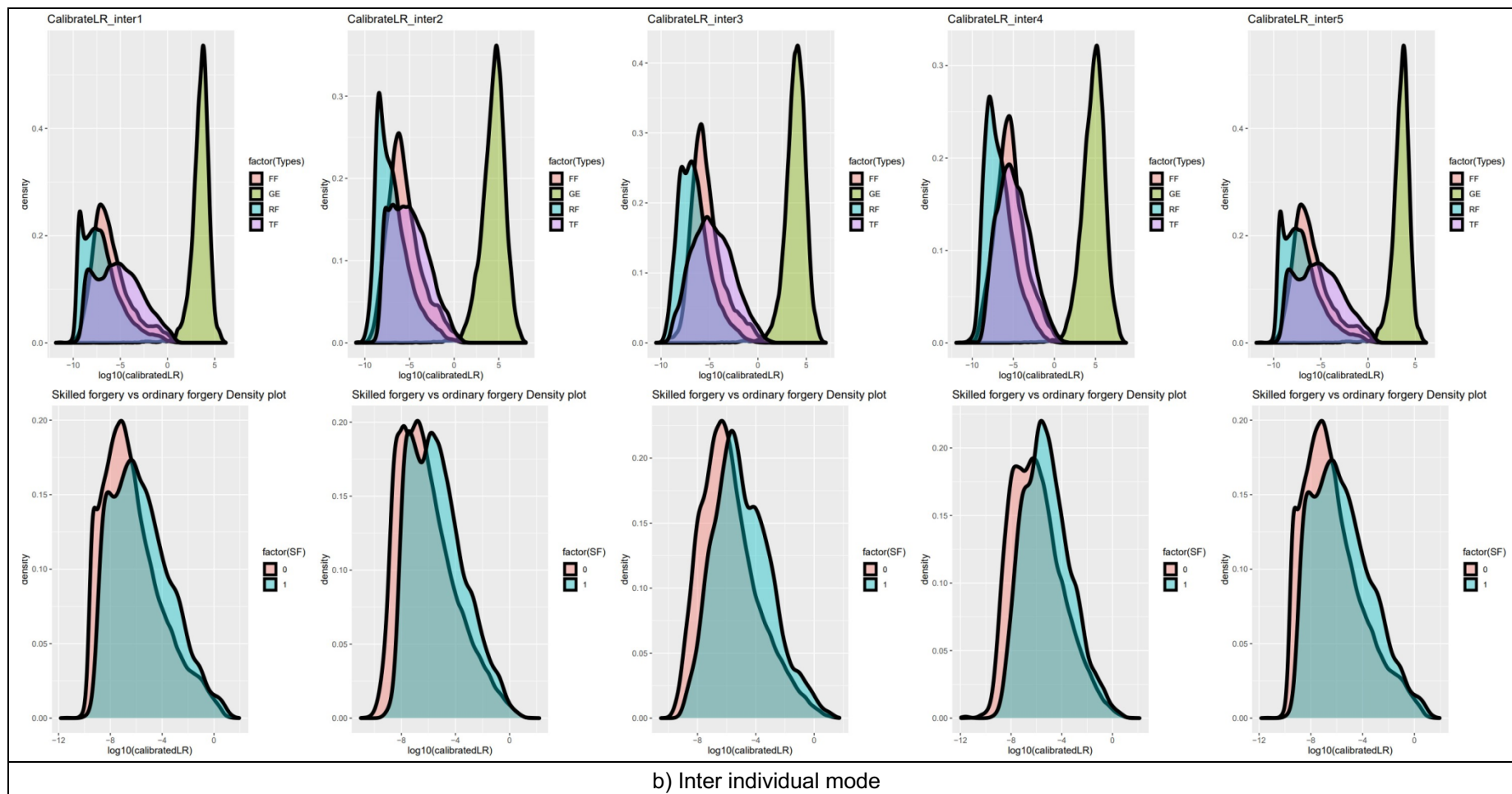
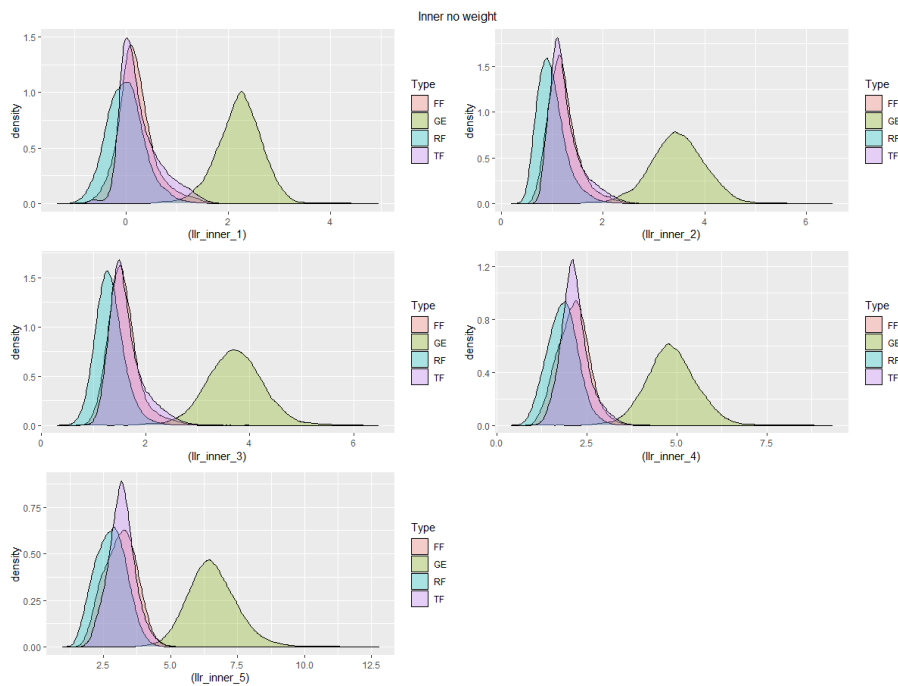


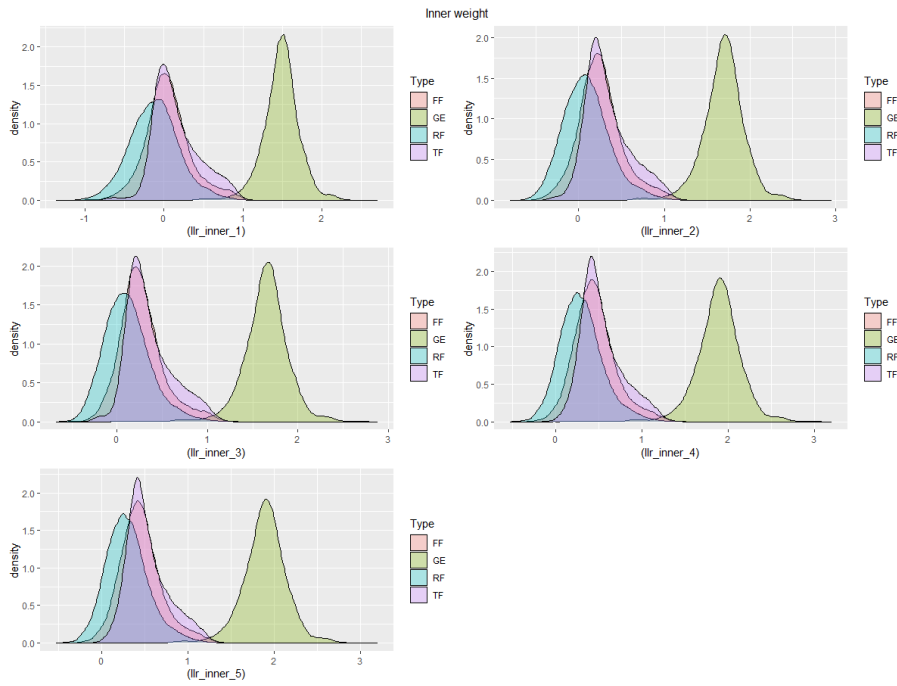
Figure 37: Histogram of calibrated $\log_{10}(\text{LR})$ and skilled forgery vs ordinary forgery (five feature options, using MKDE). In the upper subplot of each figure, red areas represent freehand forgery (FF), green areas represent genuine signature (GE), blue areas represent random forgery (RF), and purple areas represent tracing forgery (TF). In the lower subplot of each figure, green areas represent skilled forgery, and red areas represent non-skilled forgeries.

4.3.2 PAVA calibration of score-based LR using DST based on dataset_3

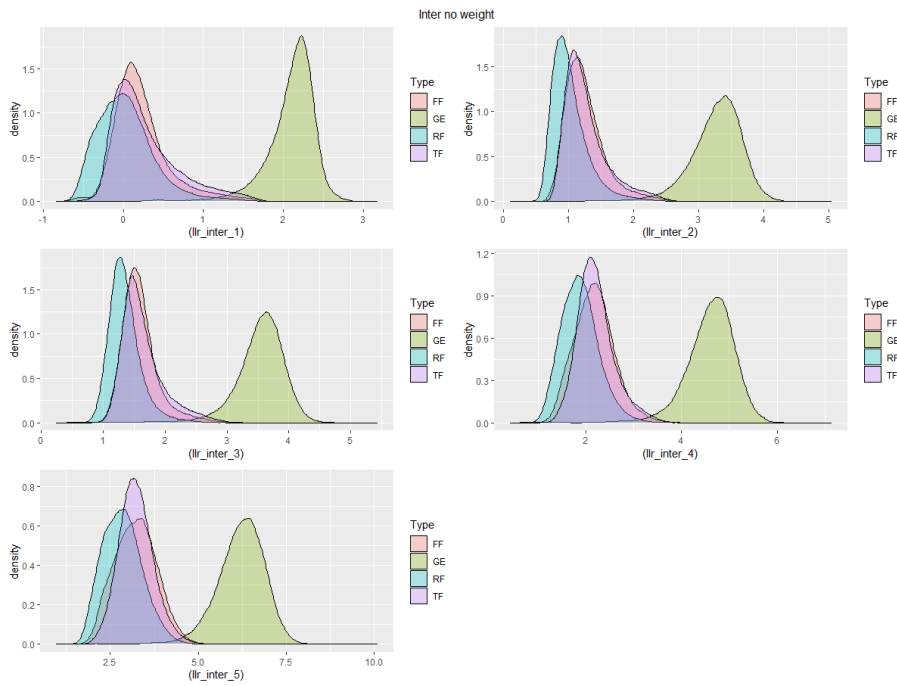
Two options, with or without weight for variables, were tested in the DST calculation. The weight parameters were adapted from variable importance-based *R-Forest* obtained from ML. Figure 38 shows the histogram distribution for GE, RF, FF, and TF. Regarding the overlaps between GE and forgery, DST with weight was significantly better than without weight for the inter- and inter-individual modes. In the following steps, DST with weight was selected.



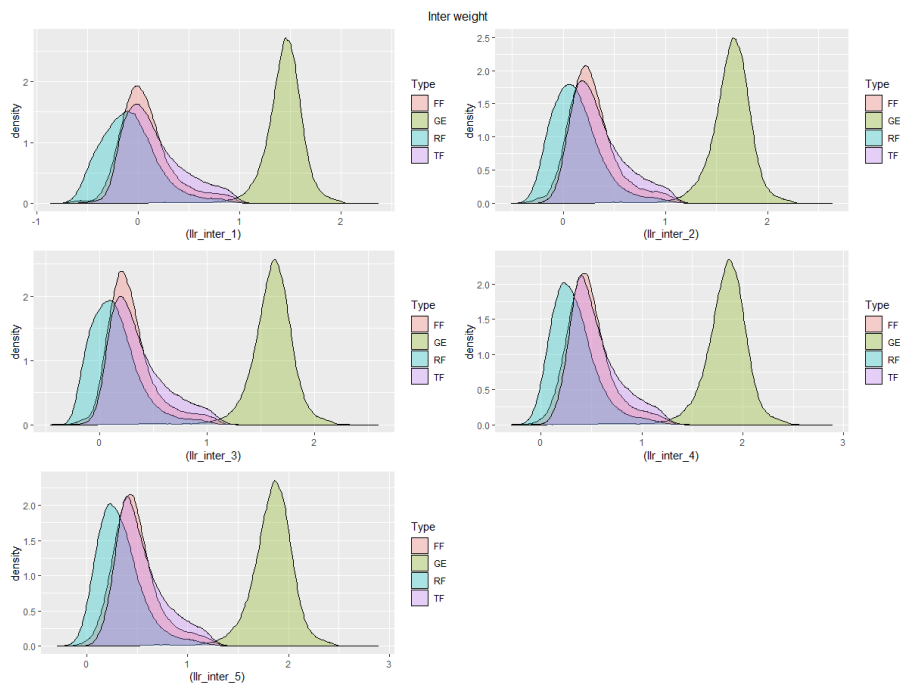
a) Inner-individual without weights



b) Inner-individual with weights



c) Inter-individual without weights



d) Inter-individual with weights

Figure 38: Histogram distribution of the $\log_{10}(\text{LR})$ for GE, RF, FF, and TF, respectively. Red lines and areas represent freehand forgery (FF), green lines and areas represent genuine signature (GE), blue lines and areas represent random forgery (RF), and purple lines and areas represent tracing forgery (TF).

The RMEP in LR before calibration is high. The LR system results using DST showed a significantly lower performance compared to MKDE for calibration (Figure 39–Figure 42). The density distribution between skilled forgeries and ordinary forgeries showed a different distribution density in calibrated Figure 43, for example, shows the histogram of calibrated LR and skilled forgery vs ordinary forgery (five feature options, using DST) under inner- and inter-individual mode. Skilled forgery was slightly closer to genuine than ordinary forgery. DST presents a greater similarity than MKDE.

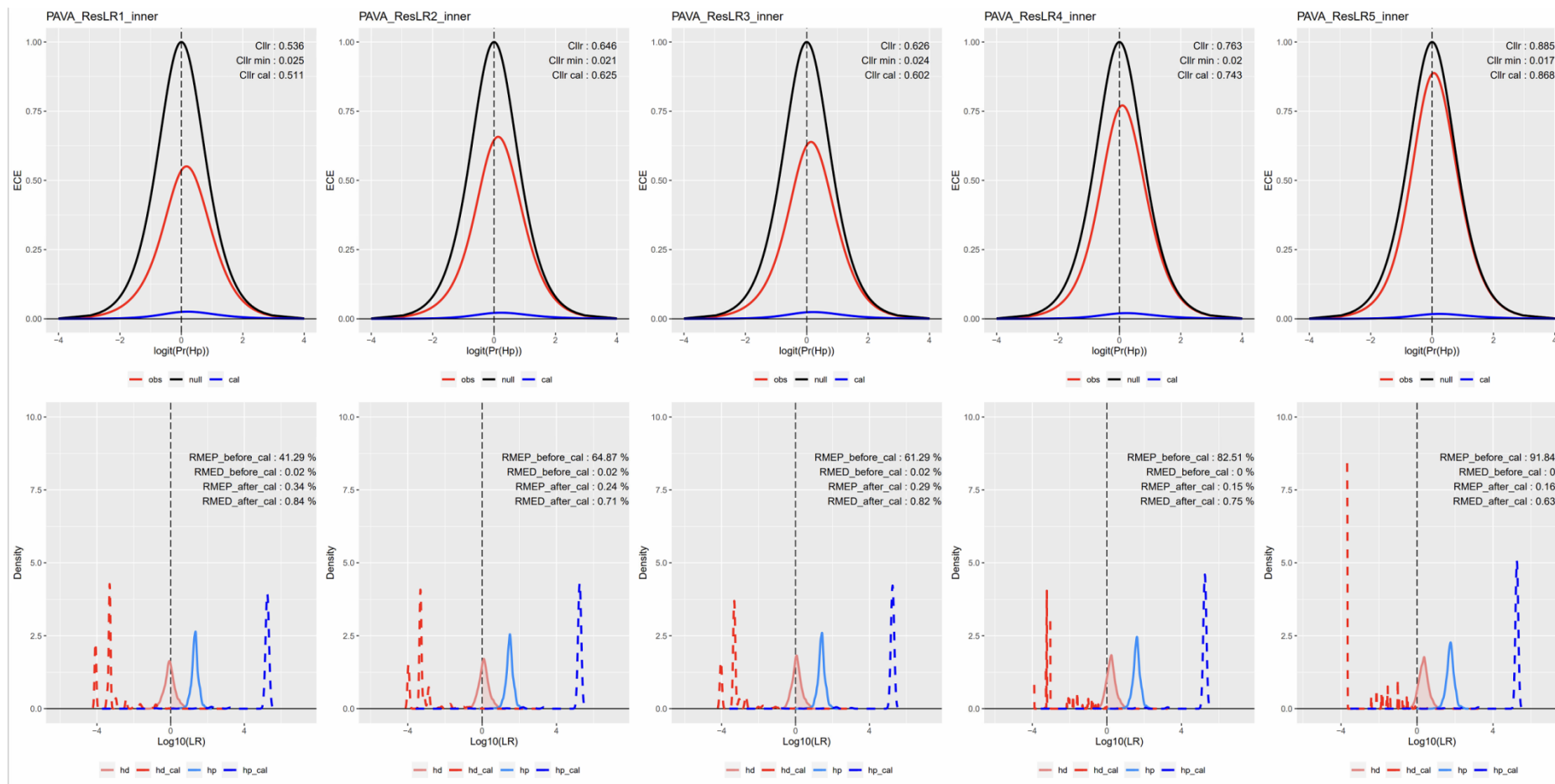


Figure 39: PAVA calibration for the inner-individual mode using DST. Upper figures are ECE plots; lower figures are $\log_{10}(\text{LR})$ density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

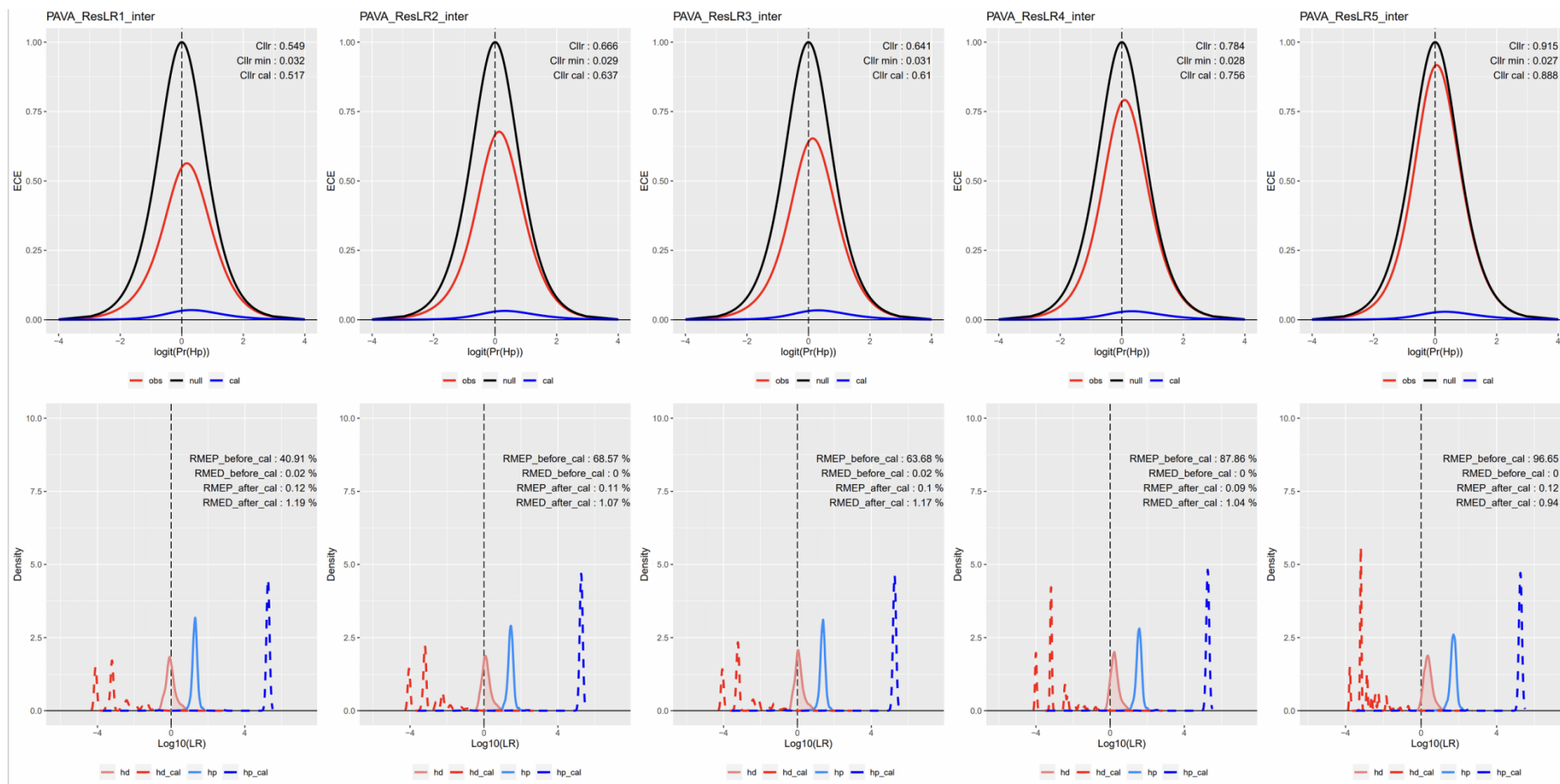


Figure 40: PAVA calibration for the inter-individual mode using DST. Upper figures are ECE plots; lower figures are $\log_{10}(\text{LR})$ density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

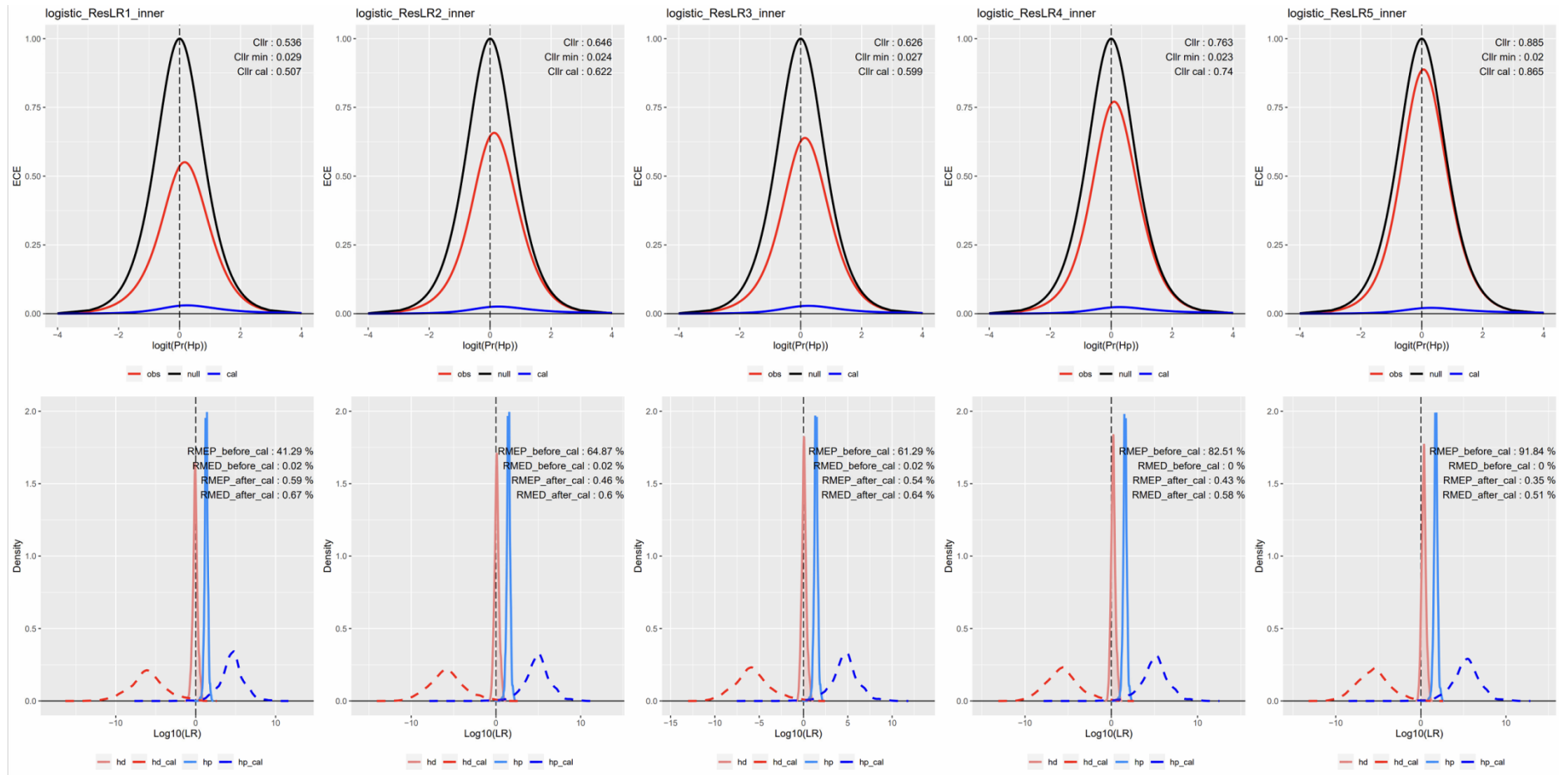


Figure 41: Logistic calibration for the inner- individual mode using DST. Upper figures are ECE plots; lower figures are log10(LR) density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.

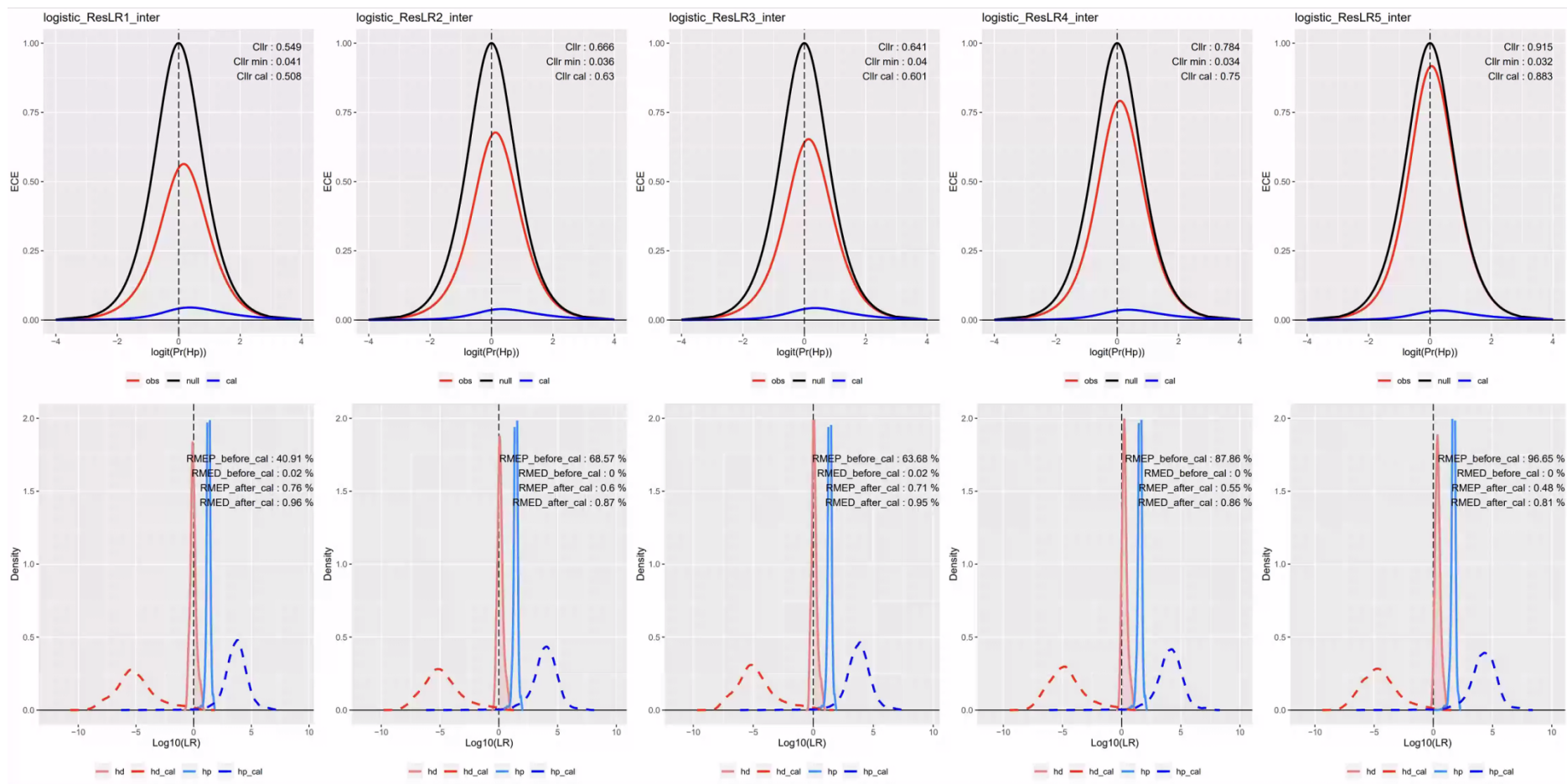
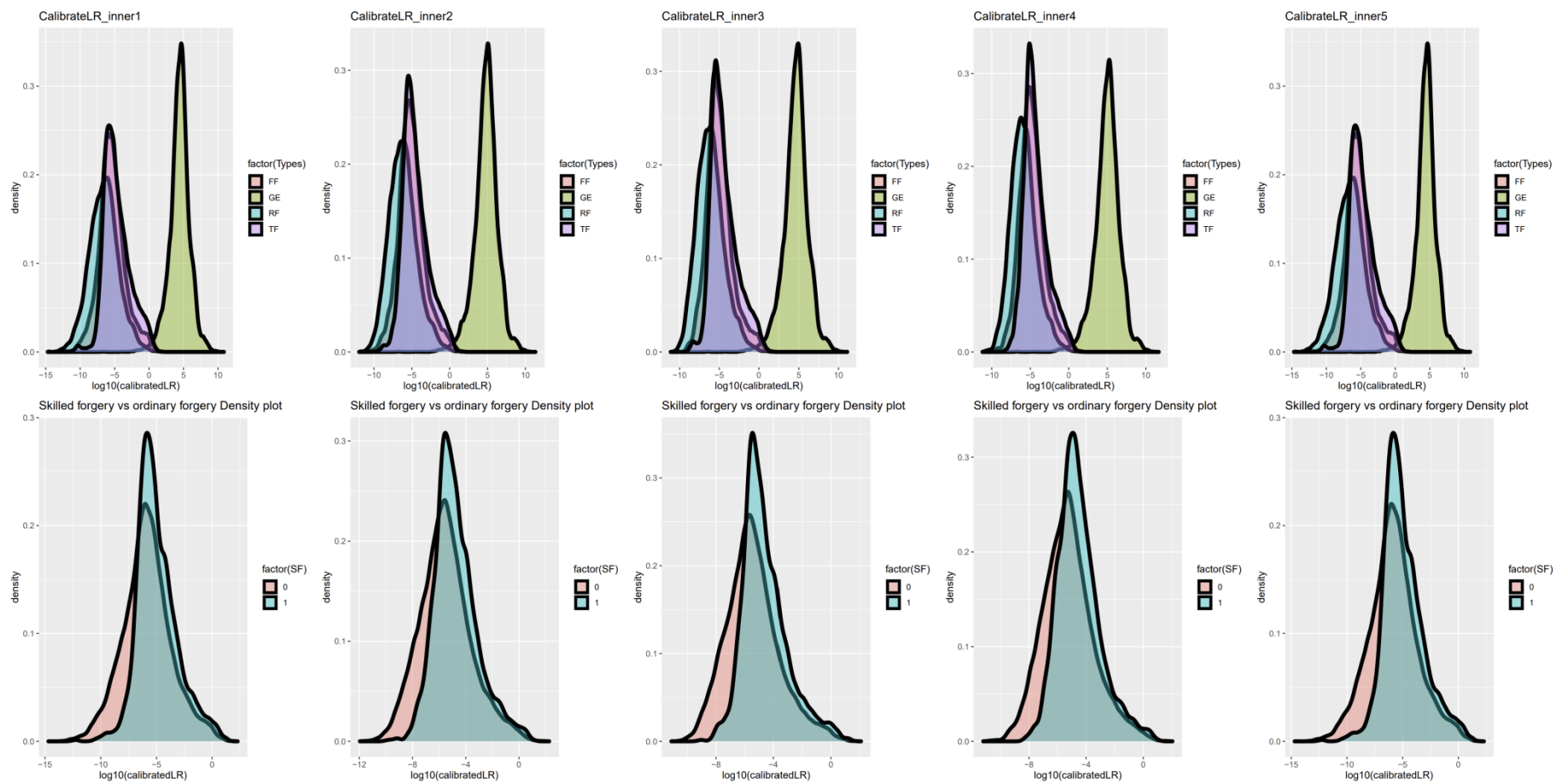


Figure 42: Logistic calibration for the inter-individual mode using DST. Upper figures are ECE plots; lower figures are $\log_{10}(\text{LR})$ density distribution before and after calibration. LR1 to LR5 refer to Table 19: Five options for variable combinations.



a) Inner-individual mode

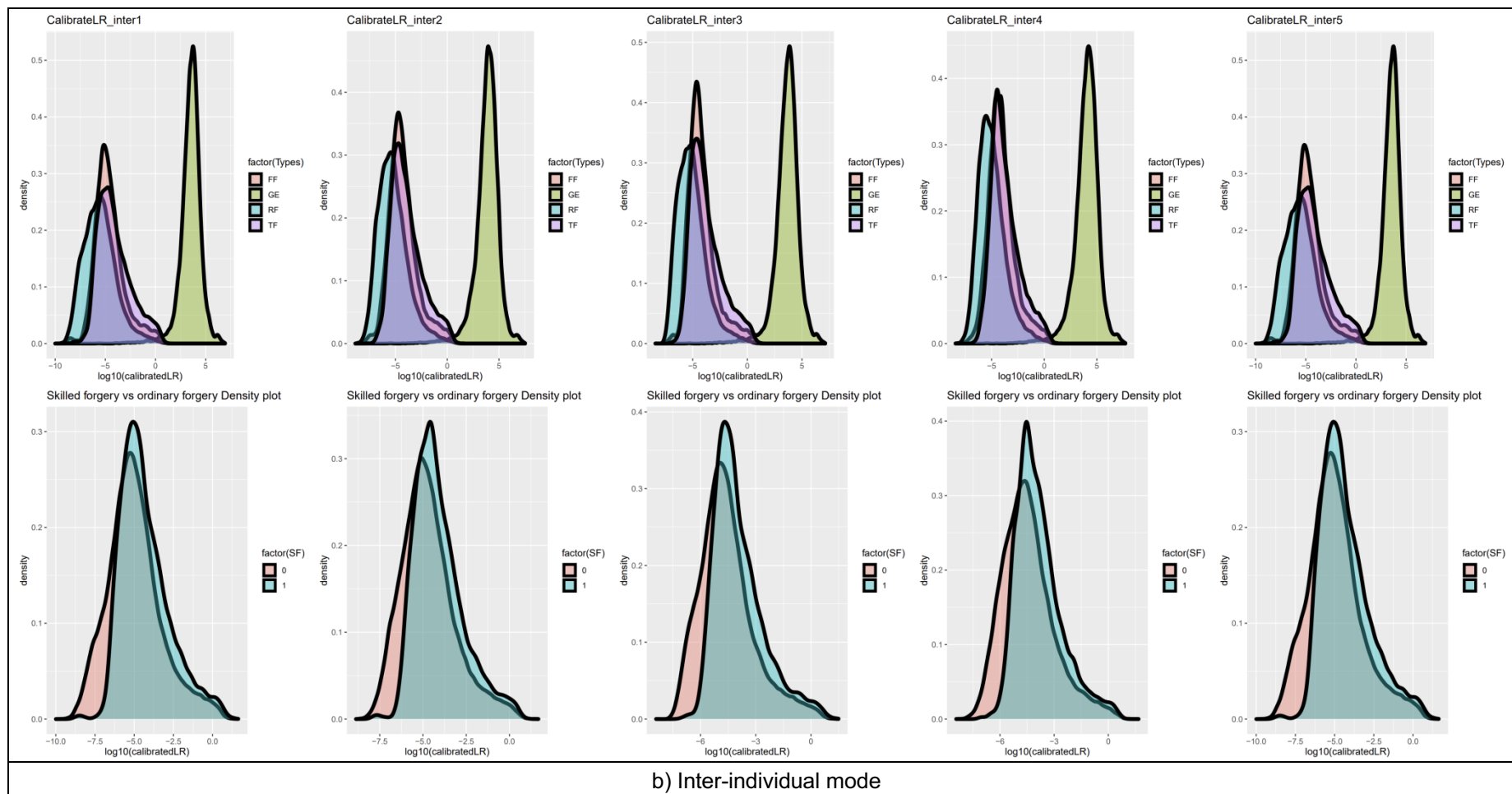


Figure 43: Histogram of calibrated $\log_{10}(\text{LR})$ and skilled forgery vs ordinary forgery (five feature options, using DST). In the upper subplot of each figure, red areas represent freehand forgery (FF), green areas represent genuine signature (GE), blue areas represent random forgery (RF), and purple areas represent tracing forgery (TF). In the lower subplot of each figure, green areas represent skilled forgery, and red areas represent non-skilled forgeries.

Finally, the logistic calibration of LR using MKDE was selected. Figure 44 shows the results of \log_{10} (mean LR) for the inner- and inter-individual modes. Though there was a significant difference between GE and forgery in the inner-individual mode compared to the inter-individual mode, there was a stable and obvious difference between GE and forgery in LR per individual. This shows the feasibility of using the inter-individual data as background information.

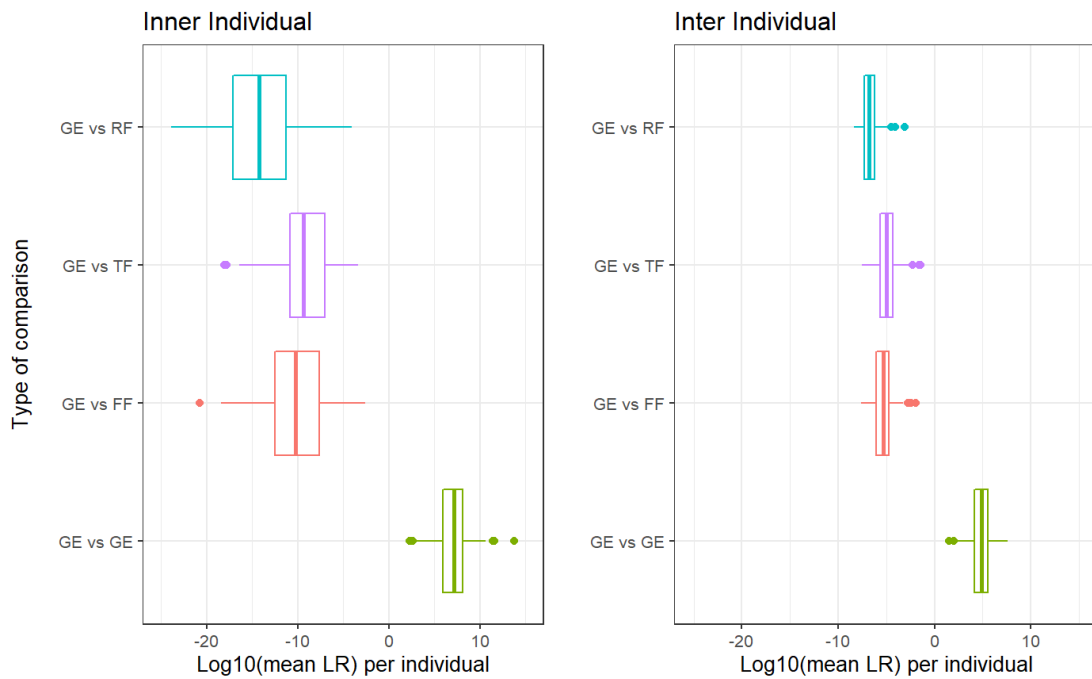


Figure 44: Results of \log_{10} (mean LR) for the inner-individual and inter-individual modes. Red boxplots represent freehand forgery (FF), green boxplots represent genuine signature (GE), blue boxplots represent random forgery (RF), and purple boxplots represent traced forgery (TF).

Figure 45 shows the RMEP and RMED for the inner- and inter-individual modes. There was a lower rate of misleading in the inner-individual (mean value of RMEP = 0.0017, mean value of RMED = 0.0023) than in the inter-individual (mean value of RMEP = 0.0033, mean value of RMED = 0.0048). This is reasonable according to the results of LR distribution. The mean value of RMEP was smaller than that of RMED, both in the inner- and inter-individual modes.

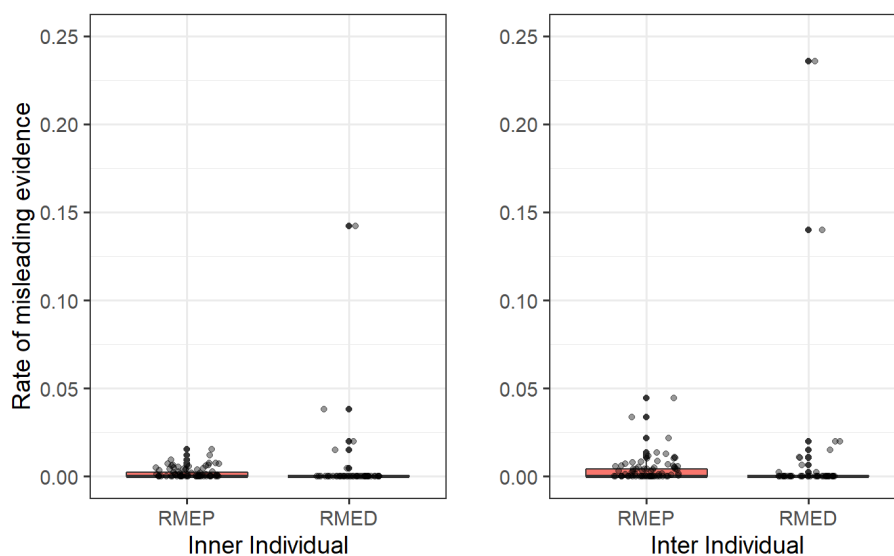
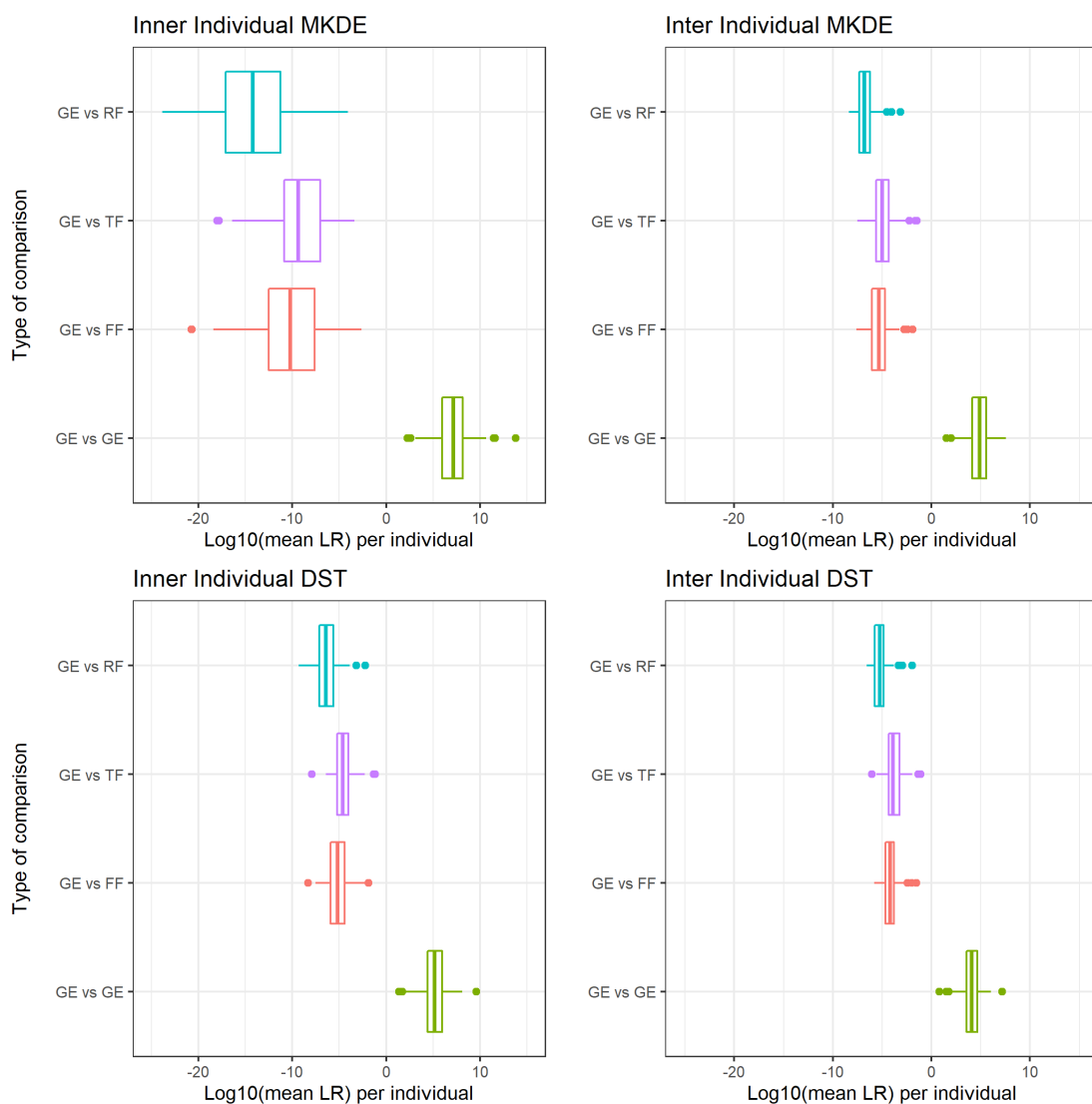


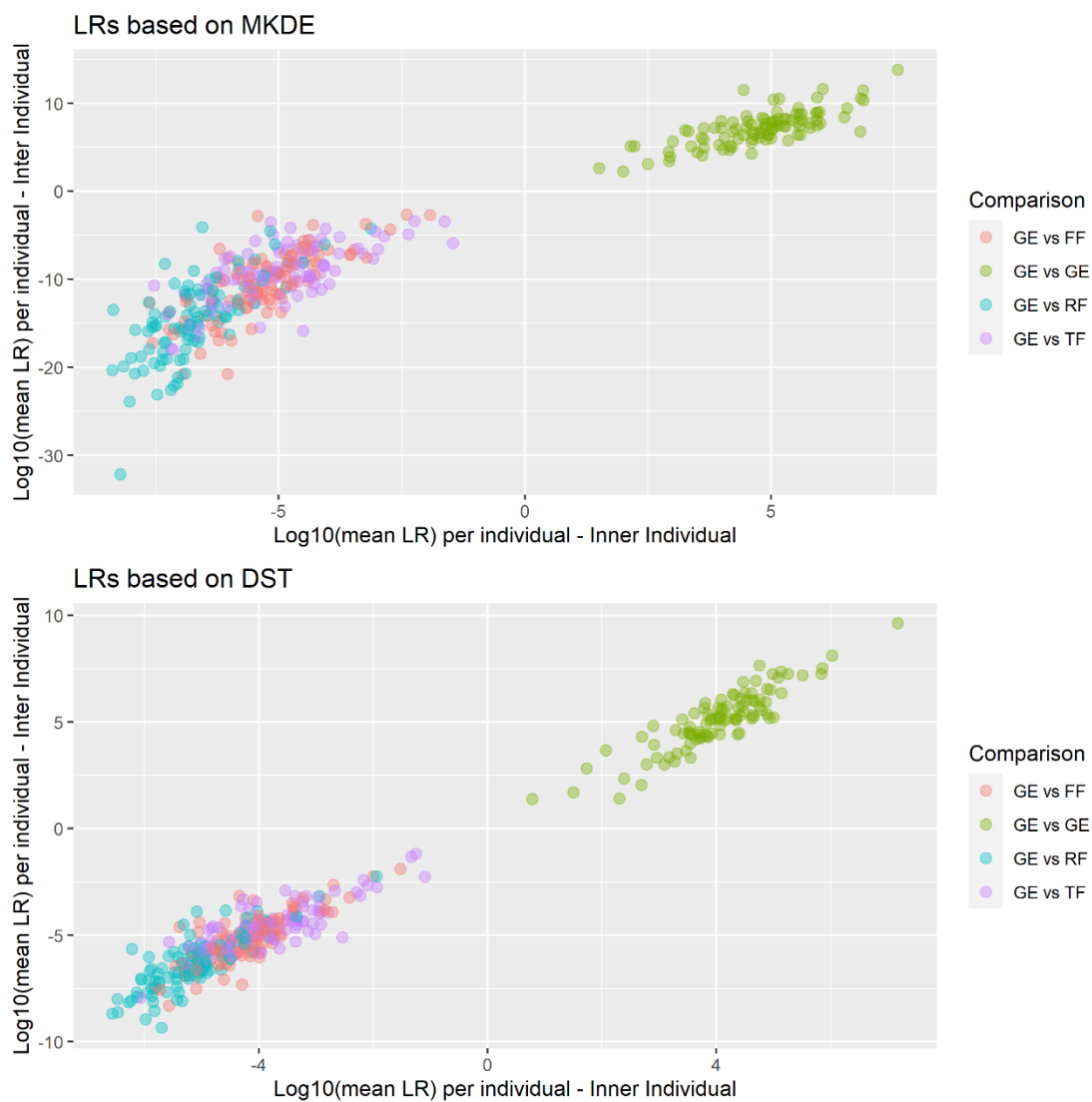
Figure 45: Rate of misleading evidence for inner-individual and inter-individual modes.

4.3.3 Comparison of performance between MKDE and DST

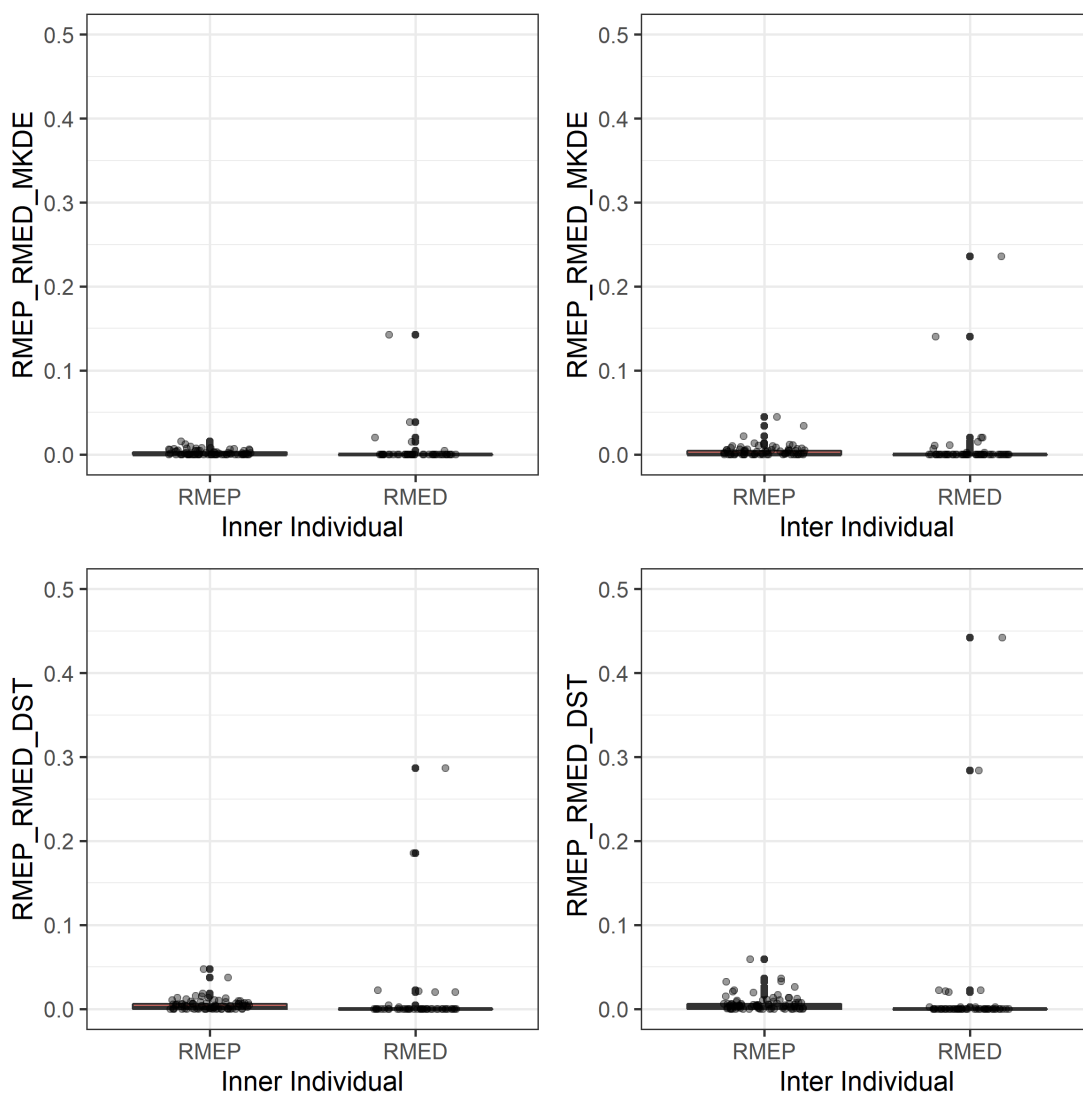
The comparison between MKDE and DST in \log_{10} (mean LR), RMEP, and RMED for the inner-individual and inter-individual modes showed no significant difference, and MKDE was slightly better than DST (see Figure 46).



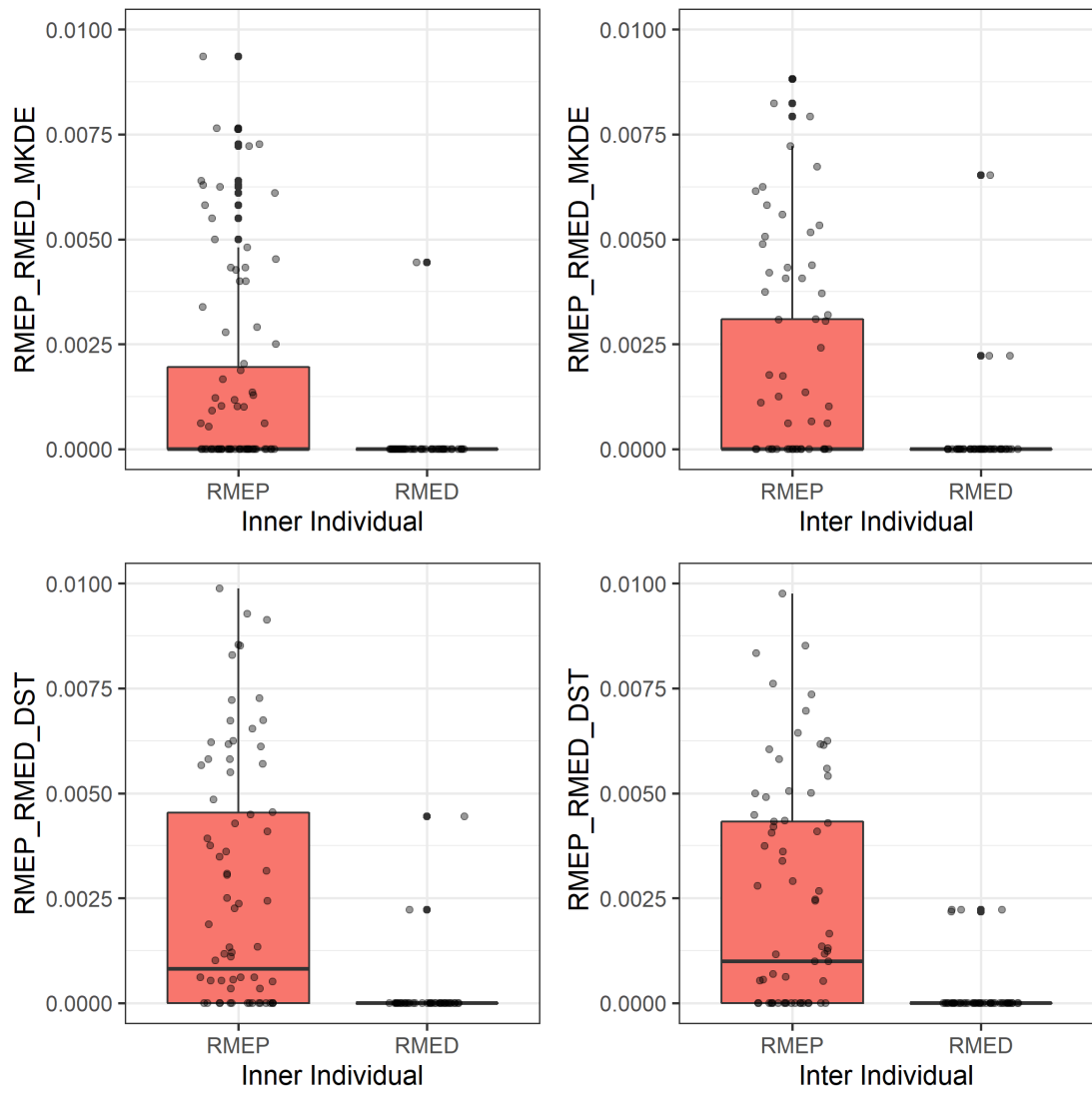
a) Boxplots of calibrated score-based LR for MKDE and DST, respectively. Red boxplots represent freehand forgery (FF), green boxplots represent genuine signature (GE), blue boxplots represent random forgery (RF), and purple boxplots represent traced forgery (TF).



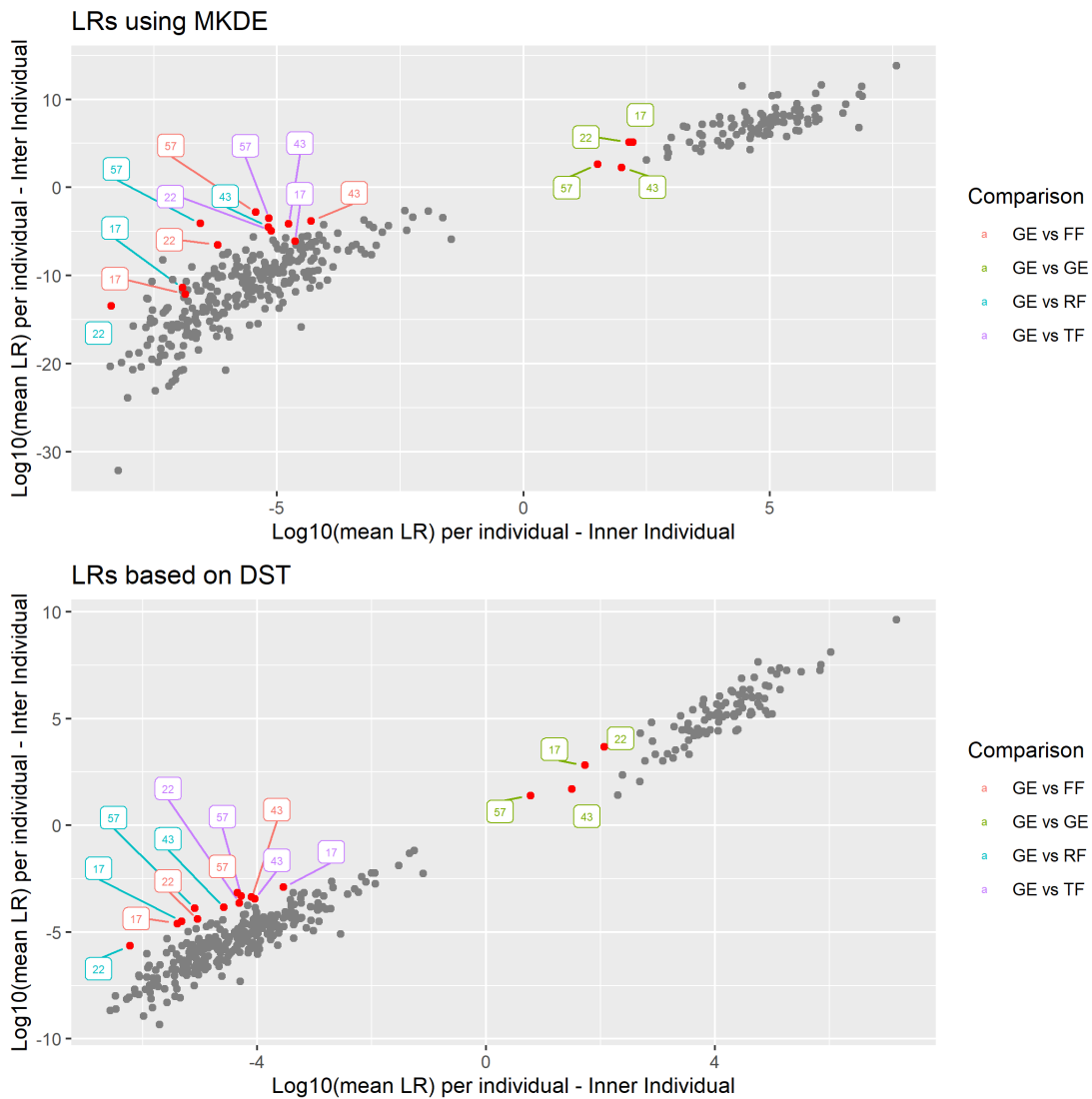
b) Calibrated score-based LR distribution of inner-individual vs inter-individual using MKDE and DST, respectively. Red dots represent freehand forgery (FF), green dots represent genuine signature (GE), blue dots represent random forgery (RF), and purple dots represent traced forgery (TF).



c) Overall view of rate of misleading evidence under inner-individual and inter-individual modes using MKDE and DST.



d) Detailed view of rate of misleading evidence under inner-individual and inter-individual modes using MKDE and DST.



e) Detailed view of LR's distribution for closer individuals under inner-individual and inter-individual modes using MKDE and DST.

Figure 46: Comparison between MKDE and DST

4.4 Validation tests

The test based on Sigcom2011 is presented in Chen et al. (2018) and is included in the appendix.

4.4.1 Proficiency tests (PT)

The PT results were similar to the results obtained for real forensic cases (see below). Interestingly, the system gave reasonable LLRs ($\log_{10}(\text{LR})$) to disguised signatures, such as CNAS_2017mym (Figure 47).

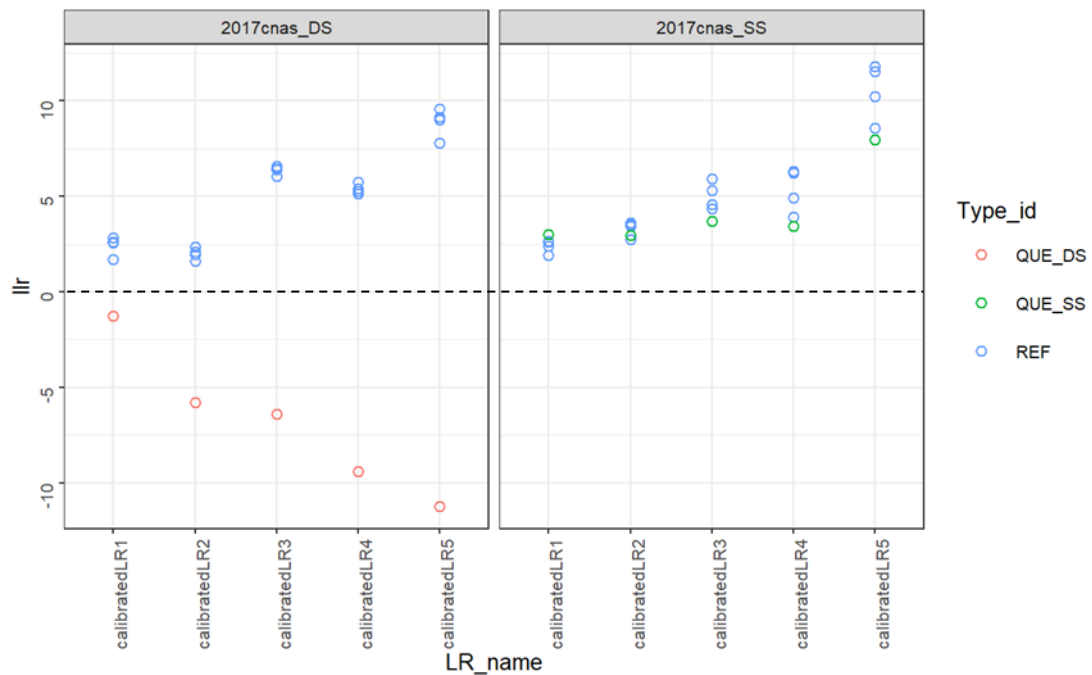


Figure 47: Calibrated log10 score-based LR result for PTs (left: different source; right: same source; REF denotes comparisons involving reference signatures, blue dots; QUE_SS denotes comparison of questioned signatures with references from same source, green dots; and QUE_DS denotes comparison of signatures from different sources compared to the references, red dots).

4.4.2 Test of real forensic cases

Figure 48 shows performance in real cases in which the conclusion reached by the practitioners was an exclusion. The calibrated LLR of the questioned signatures showed good calibration as well.

However, in many cases in which an identification conclusion was reached by the practitioners, the LLR of questioned signatures was lower than zero. Figure 49 show the LLRs of questioned signatures located among those of the references.

These results will be further discussed in chapter 5.

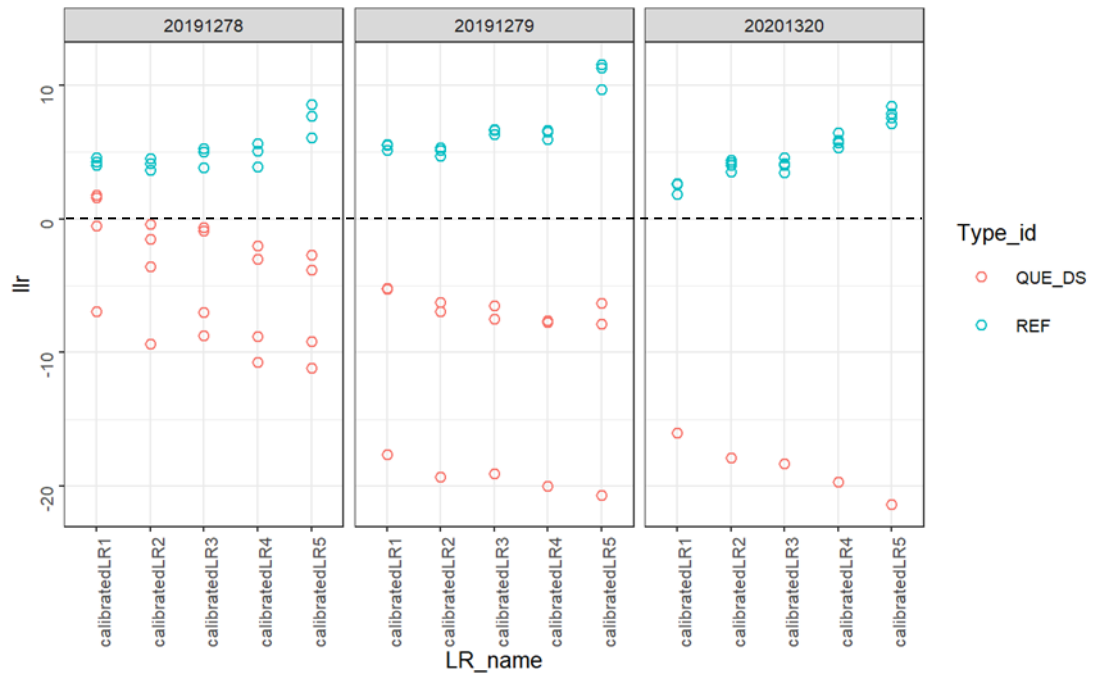


Figure 48: Calibrated log10 score-based LR for cases where practitioners reached exclusion conclusions (REF denotes comparisons among reference signatures; and QUE_DS denotes comparisons of questioned signatures from the same source as the references)

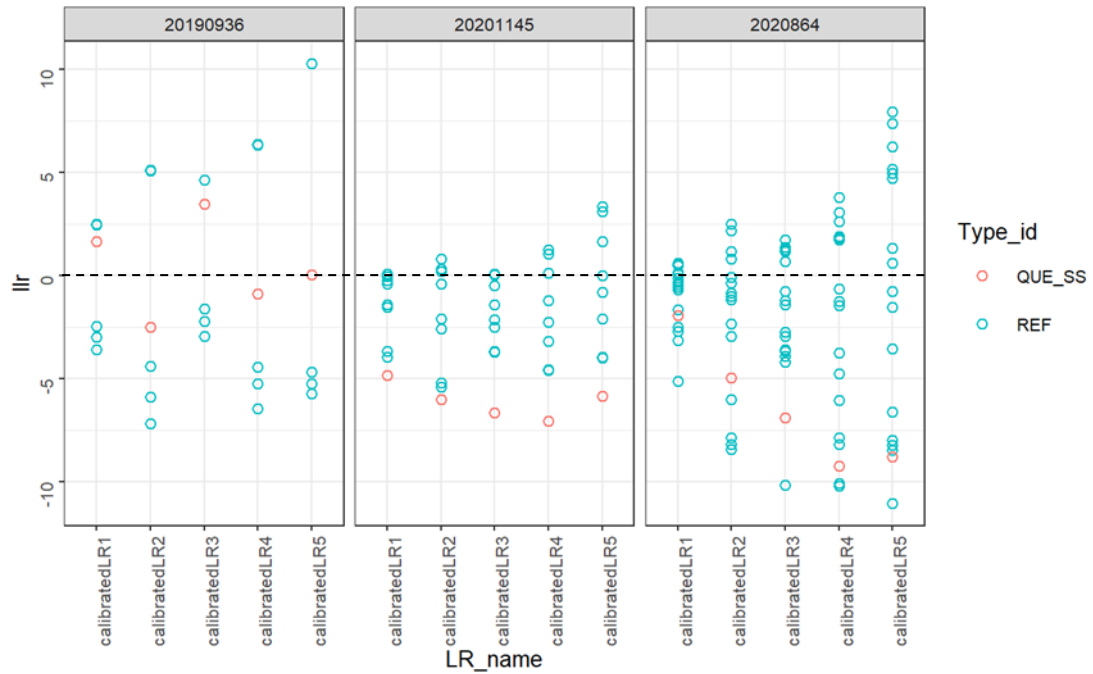


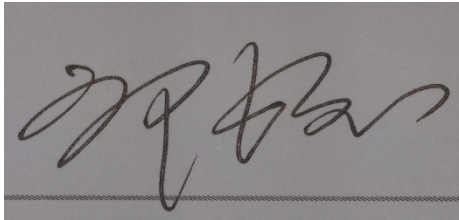
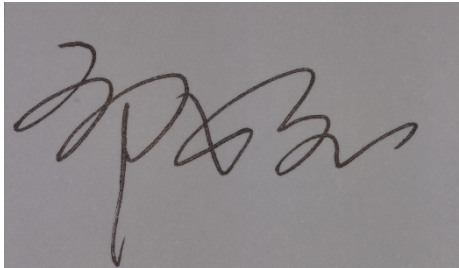
Figure 49: Calibrated log10 score-based LR for cases of where practitioners reached identification conclusions (REF denotes comparisons among reference signatures; QUE_SS denotes comparison of questioned signatures from different sources as the references)

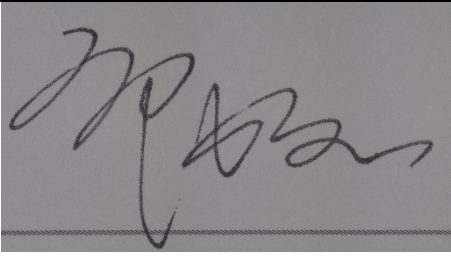
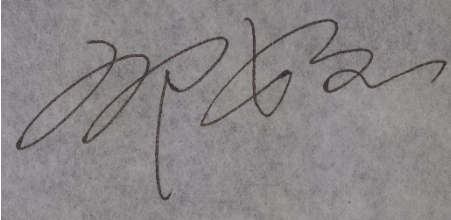
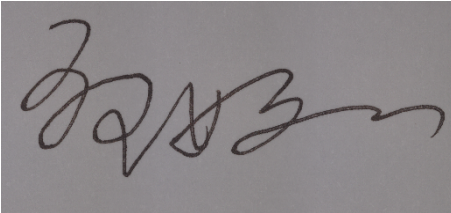
4.4.3 Impact of different writing conditions on signatures

Many conditions contribute to differences in handwriting. These can be divided into two parts: one is the internal writing conditions, which changes according to the writer, and the other is the external writing conditions based on physical circumstances. Internal writing conditions include changes in the health, physical condition, or mental state of the writer, as well as carelessness or negligence. Different writing conditions (place or circumstances) can include writing instruments, surfaces, or positions; these are external writing conditions based on physical circumstances. In practice, it is very common, of course, to write using different writing instruments or on different types of paper.

An additional 20 individuals were organized to collect signatures with different writing instruments or on different types of paper (Table 42: Signature types in dataset_4). In this last dataset (dataset_4), each individual was asked to generate genuine signatures with a ballpoint pen on 120 g paper with or without an underlay (P1 and UN, respectively), with a ballpoint pen on 17 g paper with an underlay (P2), and with a fountain pen on 120 g paper with an underlay (PE). In addition, three individuals were asked to generate freehand simulation forgeries with a ballpoint pen on 120 g paper with an underlay (FF).

Table 42: Signature types in dataset_4

Type ID	Writing instrument	Writing paper	Underlay	Signature image
P1	Ballpoint pen	120 g	Yes	
UN	Ballpoint pen	120 g	No	

PE	Fountain pen	120 g	Yes	
P2	Ballpoint pen	17 g	Yes	
FF	Ballpoint pen	120 g	Yes	

The PCA for the genuine signatures from 17 individuals showed large within-writer variations using different instruments and paper. The PCA for genuine signatures from three individuals showed a similar distribution between freehand simulation forgeries and genuine signatures using different instruments (Figure 50). Different writing instruments and paper types could thus lead to significant differences between genuine signatures. The calibrated LRs show misleading results in genuine signatures under different writing conditions (see

Figure 51). The mixed dots between P1vs FF, P1 vs Ps, and P1 vs PE showed the confusion of forgeries and genuine signatures under different writing conditions.

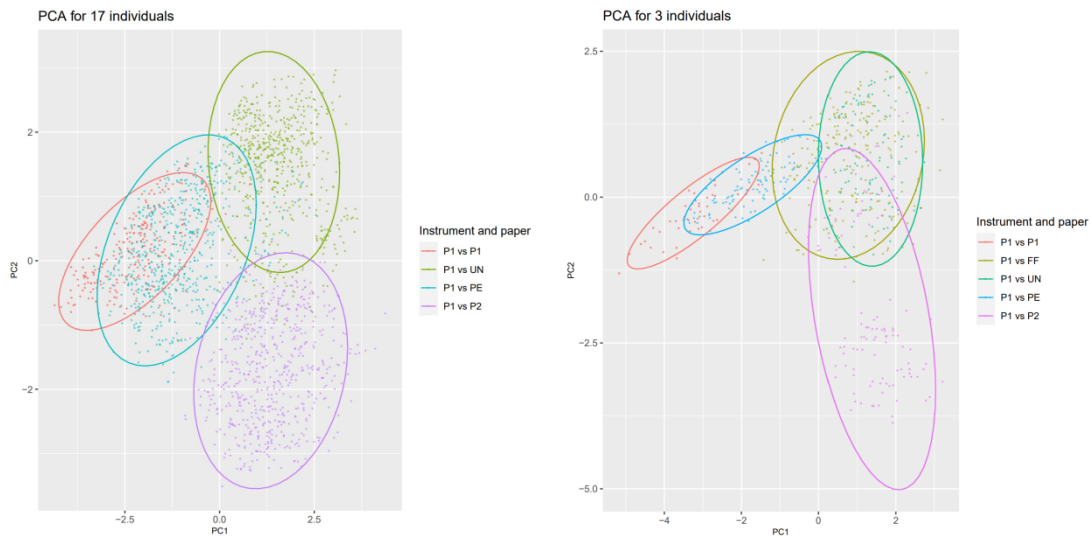
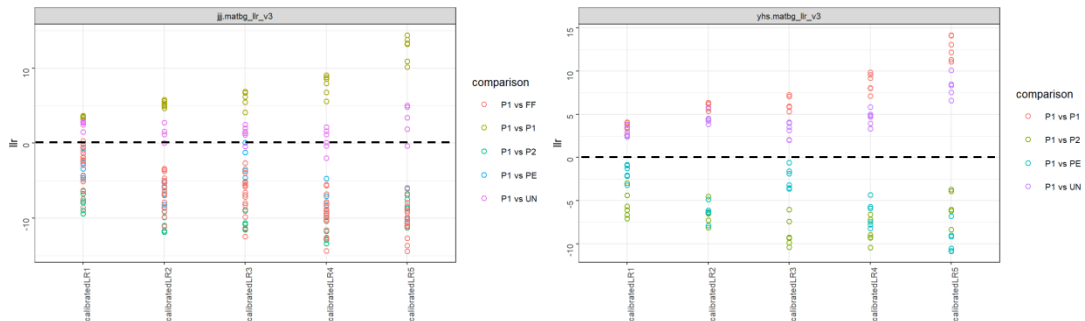


Figure 50: PCA plots for genuine signatures under different writing conditions and freehand simulation forgeries. P1, P2, PE, PS, and UN refer to Table 42: Signature types in dataset_4.



a) Calibrated LR for individuals with FF

b) Calibrated LR for individuals without FF

Figure 51: Calibrated LR (in log10) for individuals under different writing conditions (Dataset_3 as background information). P1, P2, PE, PS, and UN refer to Table 42: Signature types in dataset_4.

These findings call attention to the impact of the writing conditions, such as writing instruments and paper. Maintaining similar writing instruments and paper between references and questioned signatures is the rule for the application of this system. It will be necessary, therefore, to pursue the research to understand and mitigate the impact of different writing conditions.

Chapter 5 Discussion and Future Perspectives

5.1 Scientific basis for handwriting comparison and assessment in this research

When we discuss whether or not forensic handwriting examination is scientific, we mainly need to answer two generic questions: first, does forensic handwriting examination rests on measurable features allowing discrimination between individuals? Second, are the specific techniques and methods used in handwriting examination repeatable and reliable for the task at hand?

Harralson and Miller (2017, p. 67) described the fundamentals of forensic handwriting examination as follows:

Handwriting identification is based on two accepted premises or principles and a corollary to one of them. The first is habituation.

The second premise or principle upon which handwriting identification is founded involves the individuality or heterogeneity of writing.

Corollary—the discriminative reliability of the identification process—pertains to the accuracy of judgments made across samples of writing from different persons, including those that are simulations of another person’s writing, by whatever process of imitation may have been employed.

This description sounds reasonable. Moreover, experts can cite hundreds of examples to demonstrate relative within-writer stability and general between-writer specificity in handwriting. However, the supporting evidence from neurology, brain science, and other disciplines for the mechanism of handwriting examination, such as why the handwriting of different individuals has enough specificity to identify them, is beyond the scope of this research.

Our intention was to use rigorous methods to verify the individual specificity of handwriting in a larger population, rather than exploring or trying to explain the formation mechanism of the individual specificities, such as supports from neurology, brain science, and other disciplines. Although the mechanism behind individual specificity remains controversial or is not clear enough, from the perspective of application, the individual specificity of handwriting is an

important fundamental requirement for forensic handwriting examination.

The research results indicate that random forgeries (RF) are easier to be distinguished than simulation forgeries. As such it is not a surprising result for forensic practitioners, but it is one that now rest on systematic data.

We explored whether skilled forgeries pose a challenge to the individual specificity of handwriting, in particular, collected skilled simulation forgeries. This difficult problem is often encountered in real forensic cases because they are similar to genuine signatures. We tried a variety of methods to realistically simulate genuine signatures and collected these simulation forgeries. The results show that the skilled simulation forgeries are indeed closer to the genuine signatures than other signatures, but the difference between skilled forgeries and genuine entries was significant and sufficient to be recognized.

After examining various intentional or unintentional forgeries similar to genuine signatures, the individual specificity of the handwriting was stable. From an empirical point of view, this study confirmed the individual specificity of handwriting in the population. This conclusion refutes the argument that handwriting identification has no scientific basis, and strongly supports the existing theoretical system of handwriting identification.

Science requires verifiability, reproducibility, and operability. Verifiability means a certain proposition can be tested empirically. Repeatability means a certain proposition must be able to be reproduced by multiple people. Operability means the verification of the proposition can be accomplished through practical and limited technical steps (Popper, 1959). As described in Section 2.1, many technical standards aim to standardize the forensic activity of handwriting identification. No matter which standard is followed, however, the consistency and correctness of the identification conclusion cannot be guaranteed. Each step in the standards or guidelines have a subjective process that depends on experts' experience. Because this is a subjective process, it will vary from expert to expert, although these experts have received extensive professional training and perform better in these tasks than laypeople. These technical standards or guidelines are more principled norms, and they lack practical operability in a stricter sense. They have not fundamentally changed the dependence of forensic handwriting examination on subjective empirical judgments.

This research, like other research dedicated to the quantitative identification of handwriting, firmly adheres to the reproducibility and operability in this strict sense. This research does not intend to compete with the existing forensic handwriting identification. On the contrary, our research is designed to provide

reproducibility and operable detailed steps for each process that may unfortunately vary from person to person under the framework of principled norms.

We suggest combining this system with experts' work. This system can be summarized as follows:

- 1) A series of methods: features acquisition and extraction for examination of questioned signatures and references, respectively, statistical analysis for signatures comparison, score-based LR calculation, and evidence evaluation. All of these methods were designed to offer objective (operator independent) measurements.
- 2) Thanks to quantitative measurement and analysis, all of the results can be rigorously repeated and verified through limited steps.

5.2 An effort to change the operative model of handwriting examination

We next discuss a deeper question: Why is it difficult for forensic handwriting examination to achieve developments and make progress like in other disciplines such as computer, chemistry, or biology? The present author believe that this has something to do with the development model of forensic handwriting examination. Indeed, the main mode of the development of the discipline of handwriting appraisal is based on the inheritance of experience. Just like writing poetry, it is difficult for us to surpass Shakespeare and Pushkin; it is also difficult for a composer to say that his or her level has reached the level of Beethoven or Mozart; but Newton took it. You may have learned the results of decades of his research in half a year. You may know the basics of the geometry written by Euclid in elementary school. This is the difference between science and literature or art. From this example, we should be able to understand that the future development path of this field should rely on the superposition of scientific discoveries, rather than solely on the accumulation of experience. Science relies on objective methods. The research of predecessors can be obtained by future generations without any difference, and it can be further developed on the basis of the former. This idea embodies the internal logic of the rapid development of science. Therefore, use of the scientific method can bring about super-positional improvement.

After seven years of professional education and training, and with 17 years of practical casework experience, the author deeply feels that experience is difficult to spread and inherit. It is particularly prominent in the process of

receiving training and teaching and also in teaching. Although endless examples and observations can be cited, there is no guarantee that experience will be effectively disseminated and accepted. So, it is difficult to develop this field on the basis of previous experience because all of the necessary experience cannot be acquired and disseminated in a rigorous way, and deviations in understanding and specific operation caused by subjective factors are unavoidable. For example, CNAS publishes proficiency tests (PTs) every year. The participating experts are legally qualified and well-trained appraisers, and they follow the same forensic handwriting examination guidelines. Although the difficulty of the PT has varied over years, the feedback results have shown that different or even contradictory conclusions are often reported (see Table 43: Feedback results of PT from CNAS).

Table 43: Feedback results of PT from CNAS

Year	Questioned signature	Writing condition	Conclusion of SS*	Conclusion of DS**	Candidate number
2014	1 st of 2	Different writing condition	36 (50%)	26 (36.6%)	71
	2 nd of 2	Normal condition	64 (90.1%)	4 (5.6%)	71
2015	1	Normal condition	153 (89.4%)	13 (7.6%)	171
2016	1	freehand forgery	14 (6.9%)	186 (92.1%)	202
2017	1 st of 2	Normal condition	6 (3.1%)	186 (95.3%)	193
	2 nd of 2	Disguised in references	17 (9.3%)	45 (23.3%)	193

*SS: Same source; **DS: Different source.

There is an old saying in China: “Stones from other hills may serve to polish jade.” This means that you can use good things from other people to help you develop. It is time to move beyond this concept. The combination of forensic handwriting examination with quantitative measurement, statistical analysis, appropriate help from computer technology (e.g., image processing, deep learning) is in the right direction for the development of forensic handwriting examination.

5.3 Adaptation of the claim of uniqueness of handwriting

In the traditional sense, class characteristics within a handwriting are those writing habits or features that emanate from the published and/or prescribed method (i.e., system) of writing that has been utilized in the learning process. In the more distant past, they have been of two kinds:

(1) *unique characteristics or features that serve to distinguish one method or system from another and (2) common characteristics such as slope, spacing, height, proportions, and letter designs that are shared with other systems.* (Harralson & Miller, 2017, p. 40)

The hypothesis of the uniqueness of handwriting is widely accepted in forensic handwriting circles and we can find numerous instances to support this hypothesis by expert's observation. The results of this research, however, show that the difference between genuine signatures and forgeries varies from individual to individual. Moreover, the similarity between genuine signatures also varies from individual to individual. In other words, for some individuals, their signatures are easier to associate because their handwriting has strong specificity, whereas for other individuals, it is more difficult because their signatures are similar to those of others. Based on this assertion, we argue that the hypothesis of *unique* handwriting needs to be corrected, and handwriting should be qualified as *specific by degree* rather than *unique*.

5.4 Technical contributions of this research

The contributions of this research include multidimensional and dynamic perspectives, an application oriented for forensic science, and multidisciplinary crossover.

The 3D features of handwriting were introduced to forensic handwriting examination. The research has shown that the 3D profile is a valuable feature that has been neglected for a long time due mainly to the lack of ways to easily measure it. The fusion of 2D and 3D features significantly improves the performance of the system.

The pseudo-dynamic features of handwriting are another original aspect of this research. This concept reminds everyone that handwriting is the product of a dynamic writing process. Writing sequence is like a needle and thread, stringing all the features in the signatures into a series of pseudo-dynamic features.

The introduction of ML techniques provided quantitative measures, and a possible use for the system in a more traditional biometric system. The best accuracy of R-Forest reached as high as 99.86%. Other ML systems also may increase performance (if required).

An LR system was developed based on MKDE and logistic calibration, and the metrics used for these choices are RMEP and RME. The comparison between MKDE and DST in log₁₀ (mean LR), RMEP, and RMED for the inner-individual and inter-individual modes showed that MKDE is slightly better than DST.

Two solutions are tailored for forensic handwriting examination: inner-individual and inter-individual modes. The inner-individual mode (mean value of RMEP = 0.0017, mean value of RMED = 0.0023) was significantly better than in the inter-individual mode (mean value of RMEP = 0.0033, mean value of RMED = 0.0048). In real forensic cases, the inner-individual mode is encouraged, which means that samples of genuine signatures and forgeries should be collected as much as possible. Of course, the inter-individual model also can be used in the case of insufficient inner-individual samples and can deliver satisfactory results.

RMEP and RMED are low enough for forensic casework under the conditions established in the database in this research. Both RMEP and RMED are lower than 0.005, This is a satisfactory result compared with the results of the CNAS PT cases by certified forensic handwriting examiners. In a positive and prudent manner, we recommend the mechanisms of interactions between experts and machine (i.e., a system allowing for an assignment of a score-based LR) should be clarified in the future.

The other options tested have shown less efficiency based on the comparison between different the options of LR computation and calibration. The comparison between MKDE and DST in log₁₀ (mean LR), RMEP, and RMED for the inner-individual and inter-individual modes showed no significant difference, and MKDE was slightly better than DST. The density distribution of the same sources and different sources showed different regression results with different density distributions. Logistic regression showed smoother density than PAVA regression, and the former seemed to be more reasonable.

Although the performance of the system on our research datasets is high, we do not believe that this study is perfect and that all problems have been solved. On the contrary, we know that this is just the beginning. We tested the developed model on real cases and PT cases. The purpose was not only to discover and analyse limitations but also to delineate the direction of follow-up research efforts.

In these operational conditions, we have shown that the performance of the system for signatures from different source are relatively satisfactory, but the results for signatures from the same source under different conditions (e.g.,

writing instrument, paper) are not satisfactory. The experimental results based on dataset_4 also confirm that the system has blind spots for the same-source signatures with different writing conditions and the results are misleading. This finding served as a reality check. It does not alter the conclusions reached in cases that match the types of cases acquired in our datasets, but care must be exercised when applying an immediate application of this system to all cases and all conditions.

5.5 Open-ended questions for an optimal DTW algorithm

The technological choices that guided this research effort are not set in stone and could be revisited. For example, we tried two different DTW algorithms: (1) the algorithm used in this research, the DTW search path, is limited, and the code used is from Lee et al. (2005), which is denoted as DTW_1; and (2) the *DTW* function¹⁵ provided in Matlab2018a (Sakoe et al., 1978; Paliwal et al., 1982), with no path restriction, is denoted as DTW_2. The results presented earlier are all based on DTW_1 (see Figure 52: Different algorithms for DTW (DTW_1 aligns data with restriction on the warping path; DTW_2 aligns data without restriction on warping path.)). DTW_2 presents a better alignment than DTW_1. To illustrate how results may vary depending on algorithmic choice, we have tested the replacement of DTW_1 with DTW_2. These two DTW algorithms show some different and interesting results, such as the probability distribution of features, and differences between forgeries (TF, RF, and FF) using DTW_1 are significant, whereas the distributions between forgeries (TF, RF, and FF) using DTW_2 merged together (see Figure 53). Similar behaviour was also presented in pair plots (see Figure 54). The performance between the two algorithms shows that DTW_2 is slight better than DTW_1, with fewer misleading cases, as shown in Figure 56. The LRs of the inter-individual mode using DTW_2 (mean value <-20) are much smaller than that using DTW_1 ($-5 >$ mean value >-10). Moreover, focusing on the distance between genuine signatures and forgeries, when using DTW_1, the LRs of the inter-individual mode is significantly closer than that of the inner-individual mode, whereas the results for DTW_2 are the opposite (see Figure 55 and Figure 57). Finally, the performance of ML showed bigger variation and lower accuracy when using DTW_2 than DTW_1, as shown in Figure 58.

These results exemplify that changes in algorithms or parameters can lead to different results. Optimisation is a never-ending task and validation protocols (based on defined performance metrics) should be systematically applied. This

¹⁵ <https://www.mathworks.com/help/signal/ref/dtw.html>

research provides a building block in that direction.

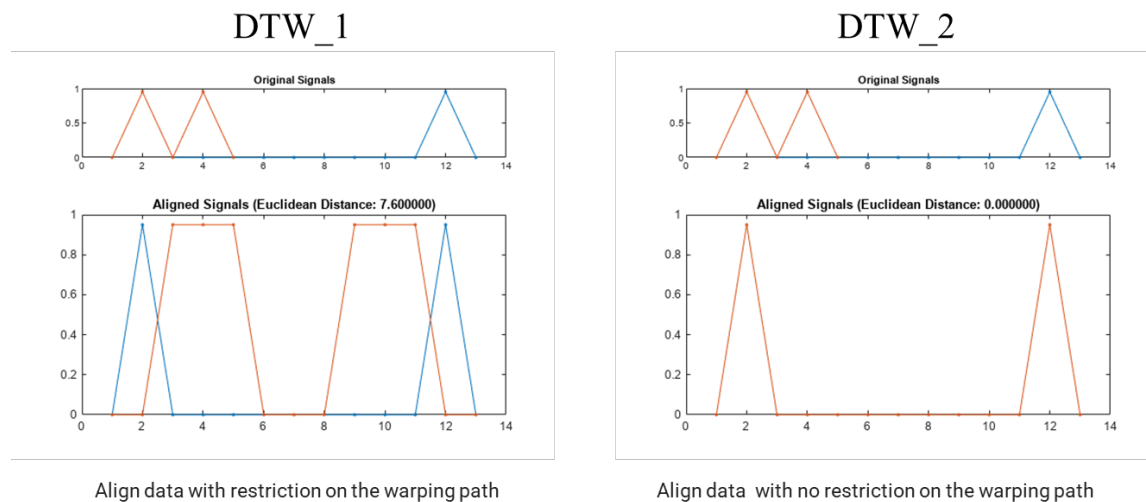


Figure 52: Different algorithms for DTW (DTW_1 aligns data with restriction on the warping path; DTW_2 aligns data without restriction on warping path.)

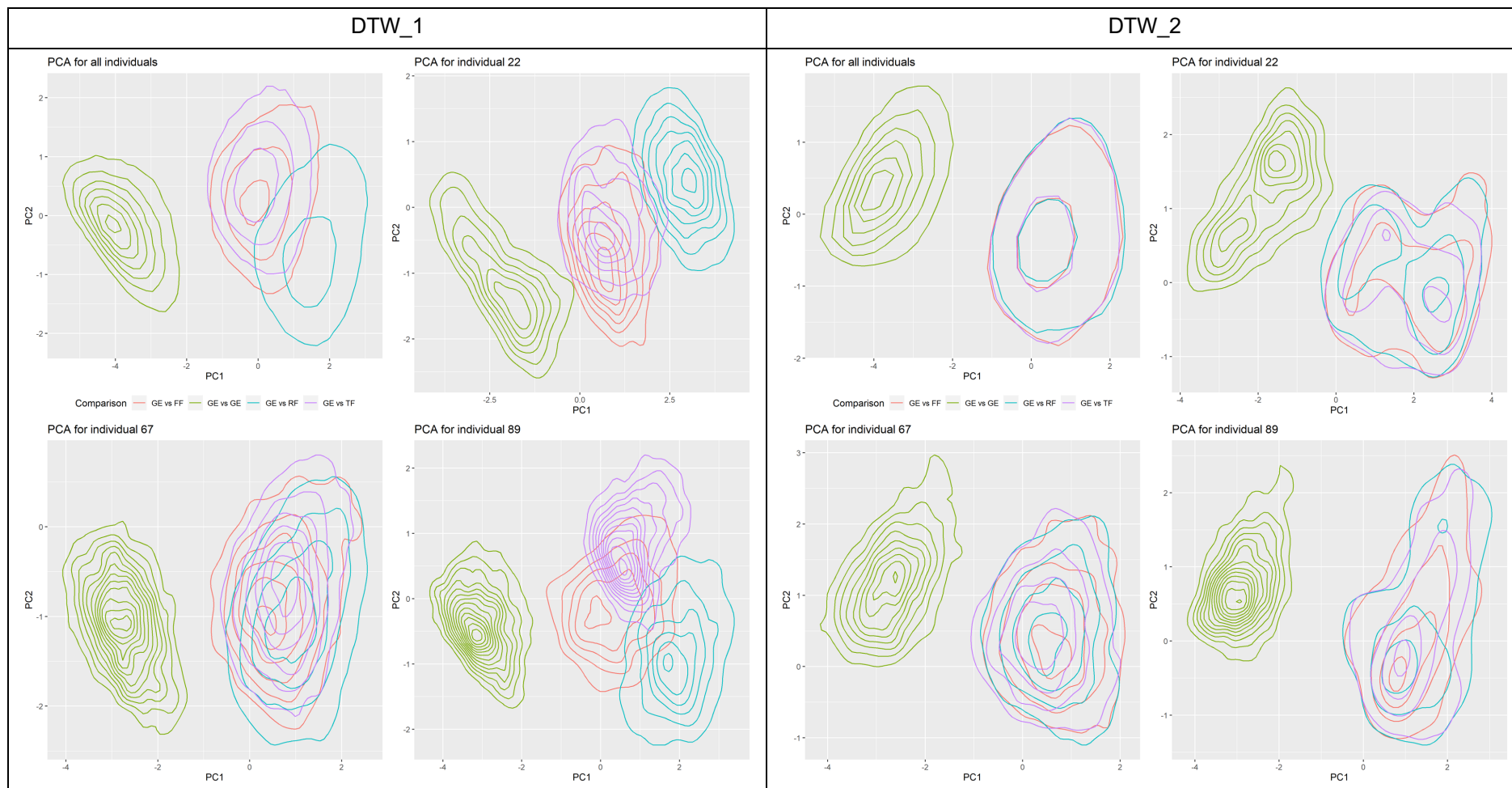
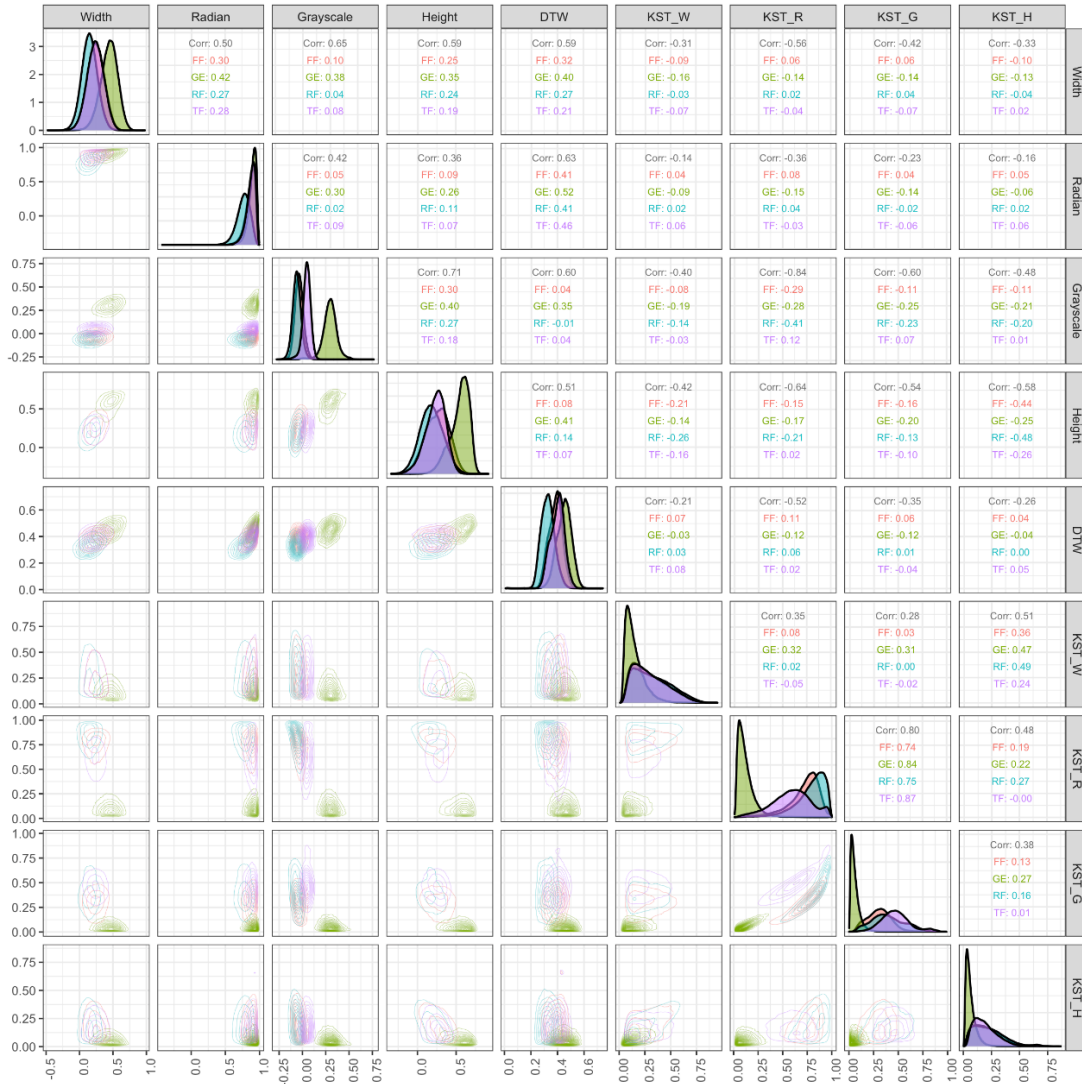
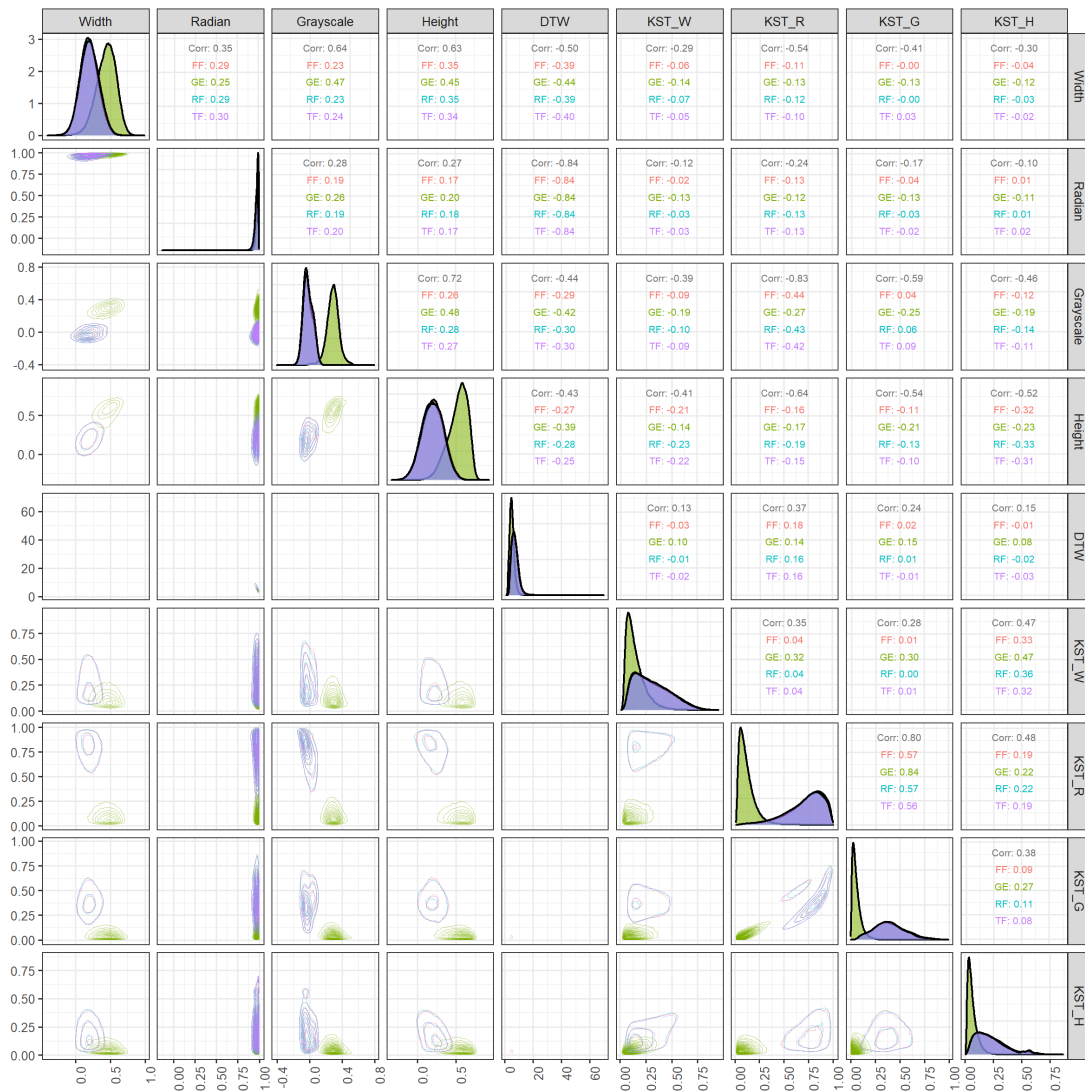


Figure 53: 2D kernel density distribution plot after PCA using two different DTW algorithms. Red lines represent freehand forgery (FF), green lines represent genuine signature (GE), blue lines represent random forgery (RF), and purple lines represent tracing forgery (TF).

Because of the non-restriction warping path of DTW_2, the distribution plot using DTW_2 merges three types of forgeries (RF, FF, and TF), and the difference between genuine signatures and forgeries using DTW_2 is more significant than that using DTW_1.



a) Pair plot using DTW_1. Red lines and areas represent freehand forgery (FF), green lines and areas represent genuine signature (GE), blue lines and areas represent random forgery (RF), and purple lines and areas represent tracing forgery (TF).



b) Pair plot using DTW_2

Figure 54: Pair plots based on comparative measurement from all individuals in Dataset_3 using two different DTW algorithms: red lines and areas represent freehand forgery (FF), green lines and areas represent genuine signature (GE), blue lines and areas represent random forgery (RF), and purple lines and areas represent tracing forgery (TF).

The pair plots show that the radian feature using DTW_2 is much higher than that of DTW_1, and the DTW feature (distance in DTW calculation) using DTW_2 is much lower than that using DTW_1. Similar to the distribution plot, these three types of forgeries do not show any difference between each other, and the difference between genuine signatures and forgeries increases.

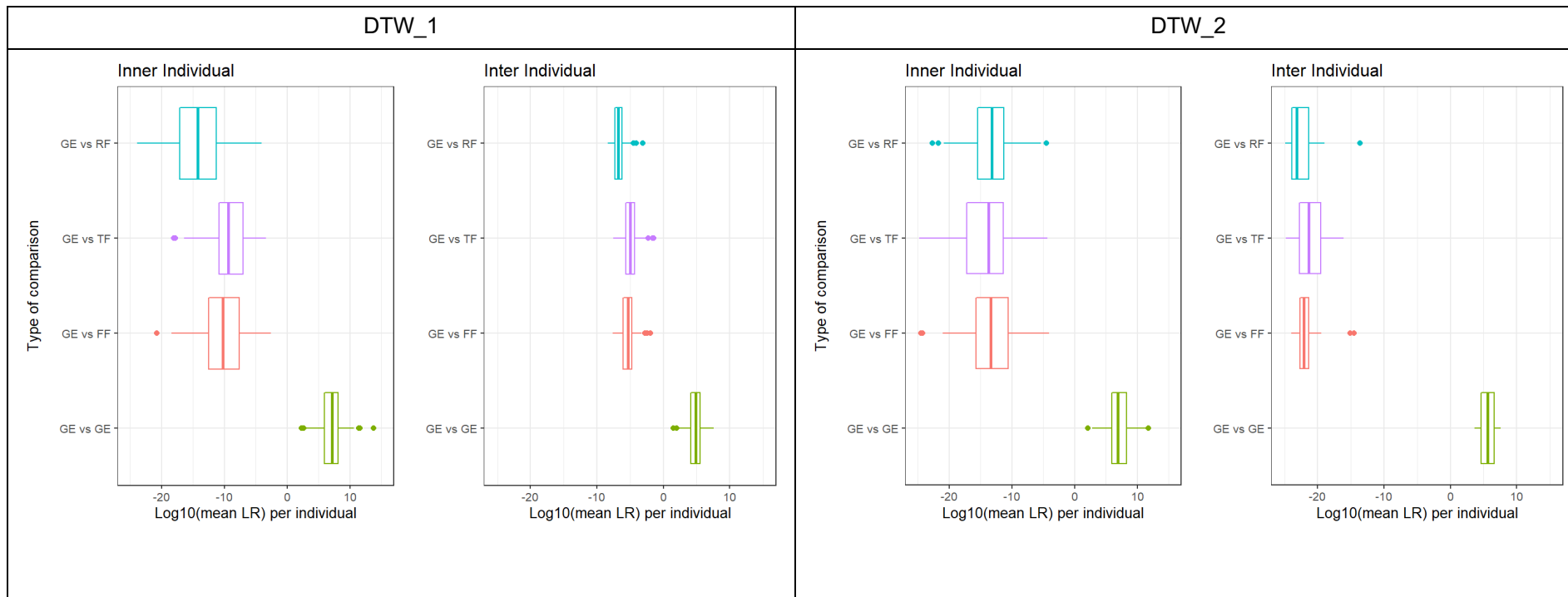


Figure 55: Results of log10 (mean LR) for the inner-individual mode and inter-individual mode respectively using two different DTW algorithms. Red boxplots represent freehand forgery (FF), green boxplots represent genuine signature (GE), blue boxplots represent random forgery (RF), and purple boxplots represent tracing forgery (TF).

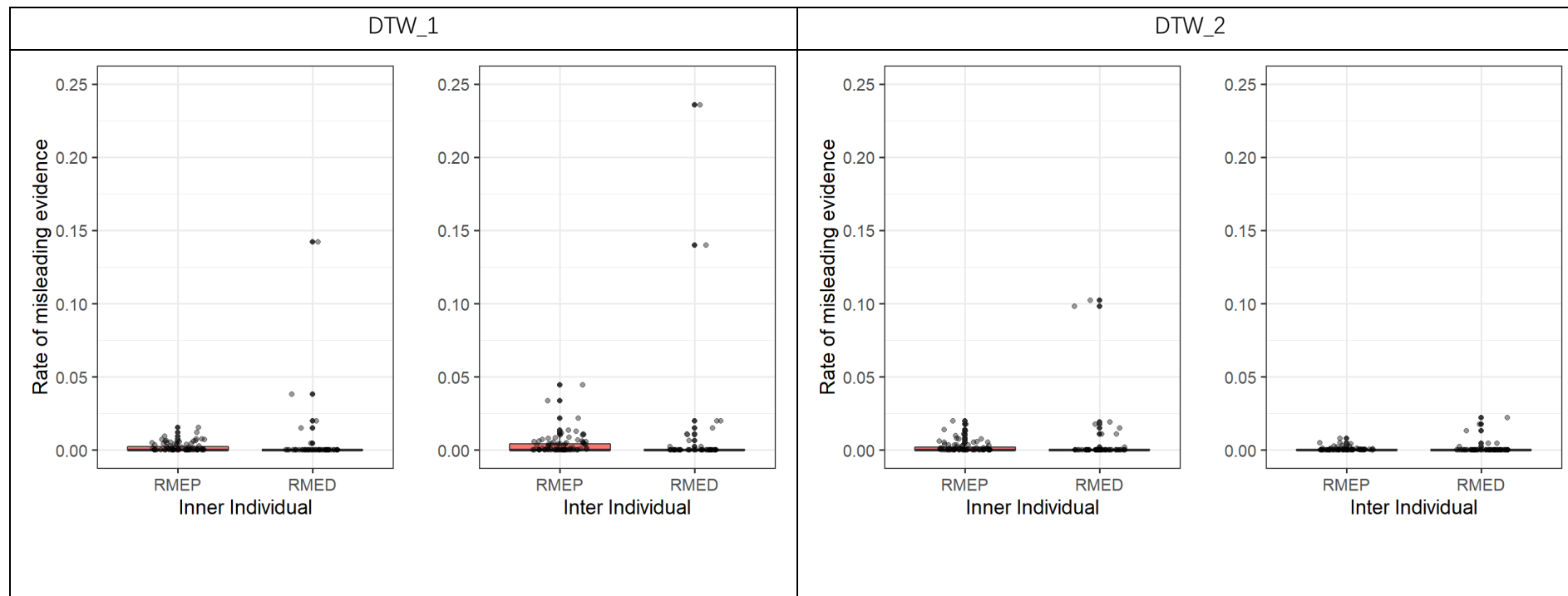


Figure 56: Rate of misleading evidence for inner-individual and inter-individual modes respectively using two different DTW algorithms

The results of \log_{10} for the inner-individual mode using DTW_1 show significantly better performance than those using DTW_2. The results of \log_{10} for the inter-individual mode using DTW_1 show some variation among the three types of forgeries; however, when using DTW_2, no significant variation exists among the three types of forgeries.

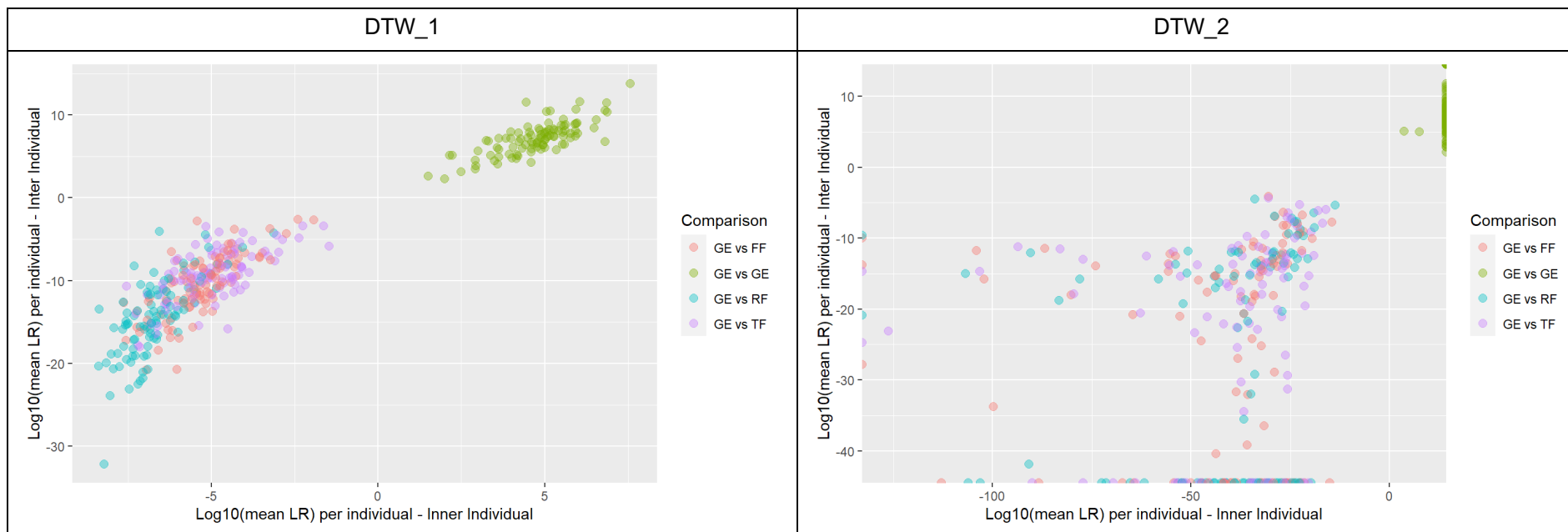


Figure 57: Calibrated score-based LR distribution of inner-individual mode vs inter-individual mode using two different DTW algorithms. Red dots represent freehand forgery (FF), green dots represent genuine signature (GE), blue dots represent random forgery (RF), and purple dots represent tracing forgery (TF).

In the calibrated score-based LR distribution using DTW_2, genuine signatures are pushed to the top-right direction, which shows a bigger variation in the LLR of forgeries. Nevertheless, the calibrated score-based LR distribution using DTW_1 shows a more balanced variation for the LLR in genuine signatures and forgeries.

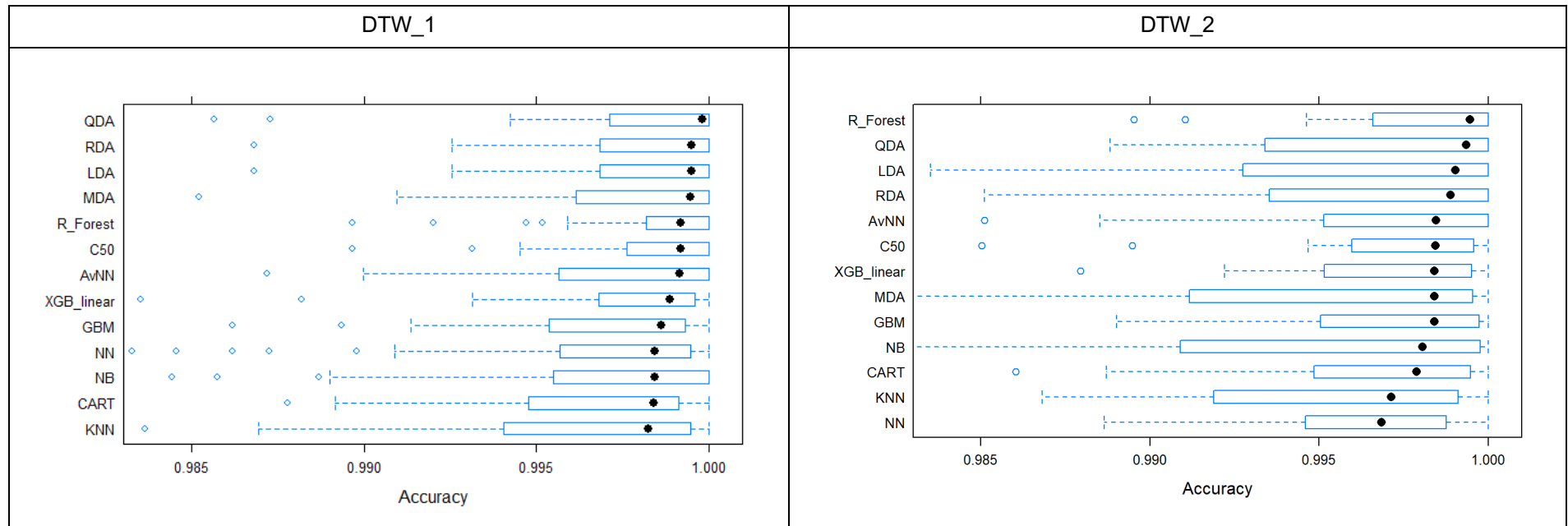
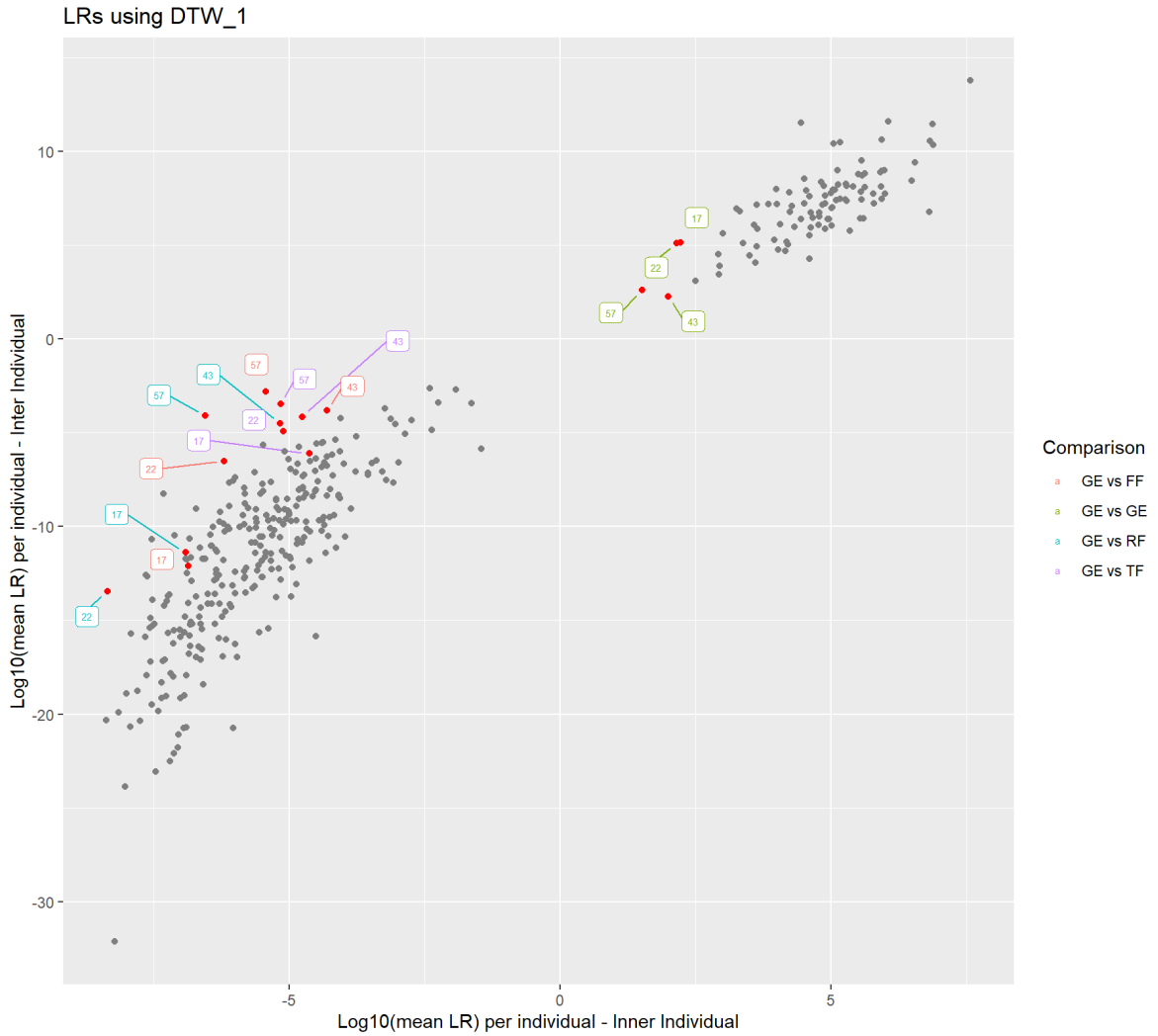


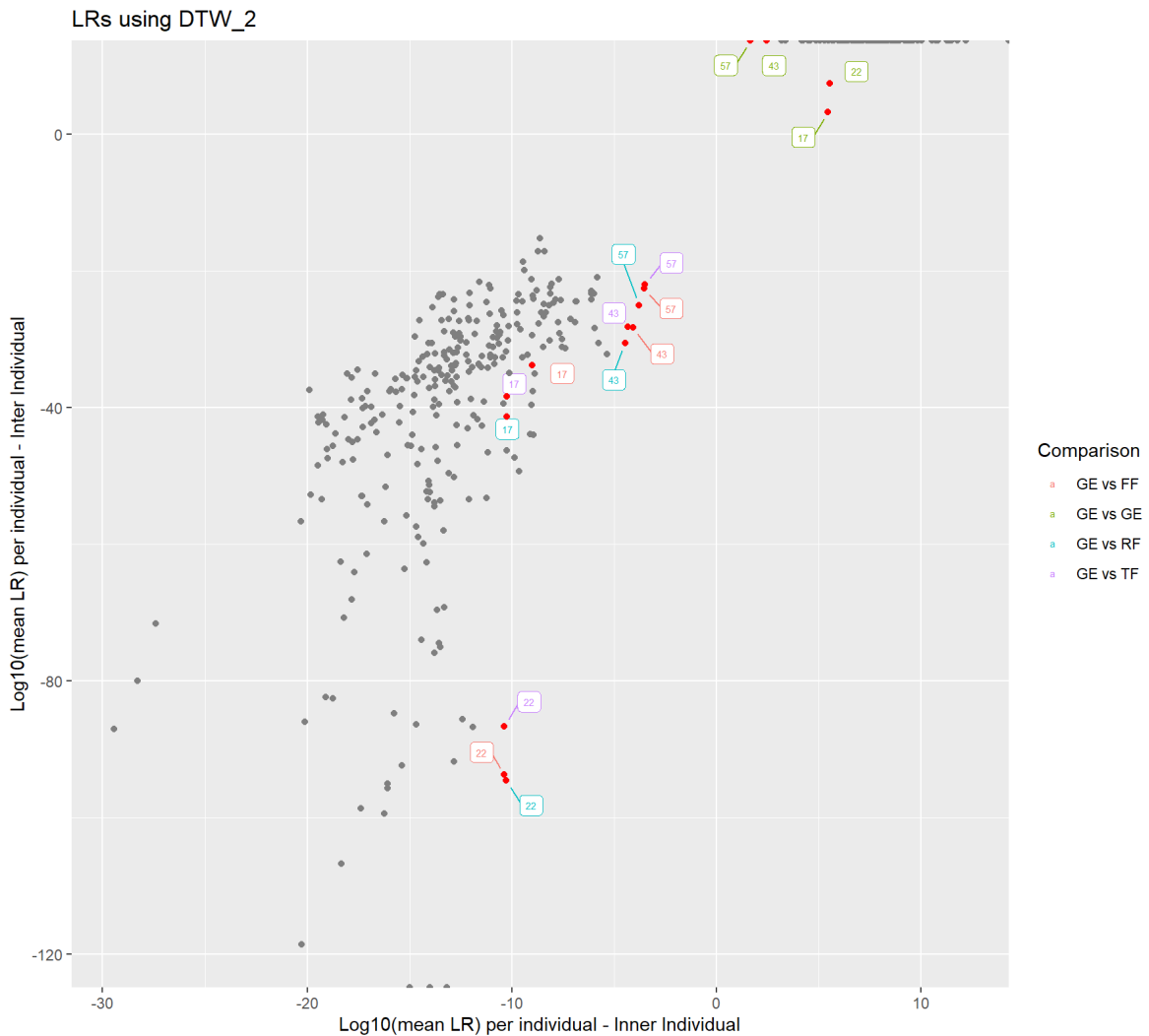
Figure 58: Model accuracy comparison using different DTW algorithms

The DTW_2 algorithm showed better performance than DTW_1, and the LR distribution showed different behaviour toward individuals. For example, four similar individuals (individuals 17, 22, 43, and 57, their genuines are close to their fgeries, respectively) are highlighted in

Figure 59 using DTW_1 and DTW_2.



a) LR distribution using DTW_1 for similar individuals (individuals 17, 22, 43, and 57).



b) LR distribution using DTW_2 for similar individuals (individuals 17, 22, 43, and 57).

Figure 59: LR distribution for similar individuals using different DTW algorithms.

5.5 Future perspectives

This research is just a start. The following paragraphs outline future research efforts.

The proposed set of solutions for the current difficulties faced in forensic handwriting examination offers potential but also has some limitations in current casework. The results have shown that the performance of the system for signatures from different source is relatively satisfactory, but the results for signatures from the same source under different conditions (e.g., writing instrument, paper) are not satisfactory. The experimental results based on

dataset_4 also confirm that the system may produce misleading results in the same-source signatures under different writing conditions. That was an anticipated limitation (see Section 3.1.1), and the results confirmed that the developed system may have a large discrimination power due to the rather constrained within-source variability obtained with the chosen sample production.

To improve this system, the database of signatures ought to be expanded, especially incorporating signatures obtained under different writing conditions and over larger period of time. This means that it could also be necessary to optimize the methods for feature extraction and comparative analysis in response to these more complex situations. The performance metrics will also need to be adapted and reported accordingly.

To use scientific research methods as an enabler to drive the development of handwriting comparison is one thing, but the efforts of researchers should not be limited to development of tools or prototypes residing only in the research sphere. Future work should also include the provision of a user-friendly interface for handwriting experts, and establishing feedback mechanisms so that experts can raise issues and opportunities for improvement. These developments should be “with experts and for experts.” In addition, it will be necessary to disseminate new technologies, new methods, and scientific outputs, for them to be ultimately accepted and recognized by practitioners. Training practitioners on the use of these systems will be a key component.

We hope to provide forensic handwriting experts with a set of software tools to help them use the system provided in this study to perform examinations. Experts will need to import only the corresponding 2D and 3D images to observe the results of statistical analysis and visualization of features. It will be necessary to compare whether or not the experts’ opinions are consistent with the guidance provided by the system. We currently are developing a more user-friendly graphical user interface (GUI) to be used by forensic handwriting experts who, for most of them, do not have a background in statistics or computer science.

We hope that way to advance the application of this research in a gradual and pragmatic way for the following two reasons: (1) we recognize that improvements in the system and improvements in its performance take time; and (2) it also takes time for a new system to be accepted by the community. Therefore, we will consider a human–machine joint venture; in other words, the system can be regarded as another expert, and the results given by the system can be regarded as the conclusion of another expert (i.e., a critical friend).

There is a body of research dealing with human–machine or multi-expert appraisal models. Swofford and Champod (2021) explored why practitioners generally oppose algorithmic intervention and discussed how to overcome their concerns considering issues related to human–algorithm interaction in real-world domains and laboratory research, as well as issues related to algorithmic litigation in the US legal system. With these issues in mind, they proposed a strategy to implement the algorithm in a responsible and practical way, as well as different ways to implement the algorithm. Montani et al. (2019) explored the procedural mechanisms for resolving different conclusions when two experts initially work independently. These experts can be two human experts, or one of them can be a computer-based model. They proposed a resolution process, such as the ACE-V protocol that sets the operating conditions, and described a resolution process based on the principles of transparency and detailed argumentation.

The development of a path based on quantitative measurement, calculation, and analysis is a difficult but needed direction for forensic handwriting examination. We need to gradually introduce some of these technologies or tools into traditional forensic handwriting examination and implement steadily advancing strategies to promote the development of this discipline.

Chapter 6: Conclusion

This research proposes a set of feasible schemes based on objective quantification and scientific statistical analysis to assess the strength to be assigned to forensic signature comparison as carried out by forensic laboratories. Although the object of this research is signature handwriting, the examination of signature handwriting is not essentially different from other handwriting samples and this system should be applicable to other handwriting samples.

This research introduced in the field of handwriting research two new types of features that were termed 3D and pseudo-dynamic. They allowed to add extra information beyond the obvious 2D features in offline handwriting samples. More specifically, pseudo-dynamic features capture the writing sequence in handwriting and 3D features capture the height profile of the handwriting. Compared with 2D features, the results with the addition of 3D features have shown significant improvement in performance.

In most previous studies, when mentioning the specificity of handwriting features, it had to be conditioned on the same signature text, hence they were text-dependent. This research effort transformed direct measurement features into relative measurement features from a comparative perspective. This means that regardless of the text of the signature, a signature dataset can be used as background information for other questioned signatures. This research proposed the addition of 3D pseudo-dynamic properties of handwriting and successfully used these 3D pseudo-dynamic features to achieve the individual recognition of offline handwriting. The system showed excellent performance.

This research offers a quantitative method to measure, compare, and analyse handwriting features. It provided a scientific way to verify the relative stability and general specificity of handwriting. It supports the understanding of between- and within-writer variations in handwriting.

This research is based on extensive and task-relevant datasets composed of more than 300,000 signatures from more than 140 individuals. In addition to the genuine signatures, this research also included several forgeries as could be produced in real cases. Considering the possible similarities in the handwriting of some individuals in the sample, we also asked volunteers to write other people's signatures in their own writing methods (what we named a random forgery). After measuring various intentional or unintentional forgeries against genuine signatures, this study confirmed the individual specificity of handwriting in the population.

This conclusion refutes the argument that handwriting identification has no scientific basis.

This thesis started from the observation that the whole expertise of handwriting examination rests on experts' experience and training that lacks transparency and is not supported by a body of systematic measures and solid statistical underpinning. The proposed quantitative method of signature comparison adheres to the reproducibility and operability requirements. It provides a method to extract features and, thanks to statistical modelling (and appropriate calibration), it allows for an expert-independent assessment of the weight to be attached to the findings among the compared signatures. This weight is based on a score-based likelihood ratio (LR). The method has shown very good forensic performance under controlled conditions with low rates of misleading evidence. These rates are dependent on the individual and show that the specificity of a given signature depends on the individual. They are in the large majority of cases below 0.5%.

When comparing genuine to genuine signatures, the average expected score-based LR is of the order of 7 (on log10, inner-individual mode). When genuine signatures are compared with forgeries, the LR is of the order of -5 on average (in log10, inner-individual mode).

When the system is applied to signatures used in proficiency tests or obtained from real forensic cases, however, we noted some deviations in terms of direction of support from the expected results (as per the known ground truth or as declared by the forensic handwriting examiners (FHEs)). The performance of the system for signatures from different sources was mostly in line with expectations, but the results for signatures from the same source under different conditions (e.g., writing instrument, paper) could be misleading.

This result highlights two important aspects to consider in the future. First, further research should investigate more thoroughly the impact of the writing instrument and paper; and, second, in the future, the mechanisms of interactions between experts and machine (i.e., a system allowing an assignment of a score-based LR) need to be clarified. This research paves the way for a handwriting/signature discipline in which automatic systems will help the examination of disputed signatures. This work will diminish the sole reliance on the expert's judgment and increase transparency. Experience cannot be acquired and disseminated in a rigorous way, and the subjective factor is unavoidable. Systems based on systematic measures and a broad corpus of data indeed will be able to overcome this difficulty. This change will not be successful, however, if it is approached as a standalone endeavour operating in isolation. FHEs should be associated with developments and operational

deployments of these systems. Their boundaries of usage need to be defined. This research has shown that when operating in a robust application, which brings added value to the examiner and transparency to the discipline.

Achievements

Publications

Journals

1. **Chen X.** Extraction and analysis of the width, gray scale and radian in Chinese signature handwriting. *Forensic Sci Int.* 2015;255:123-132. doi:10.1016/j.forsciint.2015.07.008
2. **Chen X, Champod C, Yang X, et al.** Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic Sci Int.* 2018;282:101-110. doi:10.1016/j.forsciint.2017.11.022
3. **Chen X, Yang X, Luo Y, Zhang Q.** Inkjet classification based on a few letters. *Forensic Sci Int.* 2021;325:110869. doi:10.1016/j.forsciint.2021.110869

Book

Xiaohong Chen, Forensic handwriting examination: 2018. Scientific & Technical Publishers, Shanghai.

Patents

1. **Xiaohong Chen et al.**, A vacuum suction stage, China National Intellectual Property Administration, No. ZL 2019 2 0522847.4, authorized data: 20-Mar-2020.
2. **Xiaohong Chen et al.**, Offline handwriting individual recognition system and method based on three-dimensional dynamic features, China National Intellectual Property Administration, No. ZL 2019 1 0780141.2, authorized data: 19-OCT-2021.
3. **Xiaohong Chen et al.**, Offline handwriting individual recognition system and method based on two-dimensional dynamic features, China National Intellectual Property Administration, No. 201910780111.1, receipt data: 22-AUG-2019.
4. **Xiaohong Chen et al.**, Offline handwriting individual recognition system and method based on two-dimensional dynamic features, the United

States Patent and Trademark Office, International Application No. PCT/CN2019/122178, receipt data: 12-AUG-2020.

5. **Xiaohong Chen et al.**, Offline handwriting individual recognition system and method based on three-dimensional dynamic features, the United States Patent and Trademark Office, International Application No. PCT/CN2019/122176, receipt data: 10-AUG-2020.

Support from foundations and research institutions

1. National Natural Science Foundation of China, No. 61605132, **Xiaohong Chen** (PI), 250,000 RMB, Jan. 1, 2017—Dec. 31, 2019.
2. National Natural Science Foundation of China, No. U1736102, **Xiaohong Chen** (PI), 850,000 RMB, Jan. 1, 2018—Dec. 31, 2020.
3. Shanghai Outstanding Academic Leaders Plan, No. 202079, **Xiaohong Chen** (PI), 1600,000 RMB, Jan. 1, 2020—Dec. 31, 2025.
4. Fundamental Research Funds for the Central Scientific Research Institutions, No. 2019-G-1, **Xiaohong Chen** (PI), 400,000 RMB, Jan. 1, 2018—Dec. 31, 2021.

References

- Agius, A., Morelato, M., Moret, S., Chadwick, S., Jones, K., Epple, R., ... & Roux, C.** (2018). Using handwriting to infer a writer's country of origin for forensic intelligence purposes. *Forensic Science International*, 282, 144- 156 <https://doi.org/10.1016/j.forsciint.2017.11.028>
- Aitken, C. G., & Lucy, D.** (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1), 109-122. <https://doi.org/10.1046/j.0035-9254.2003.05271.x>
- Aitken, C., Taroni, F., & Bozza, S.** (2021). *Statistics and the Evaluation of Evidence for Forensic Scientists*. <https://onlinelibrary.wiley.com/doi/book/10.1002/9781119245438>
- ASTM International.** (Withdrawn 2016). *Standard Guide for Examination of Handwritten Items. E2290-07a*, West Conshohocken, PA, 2007, www.astm.org
- Bennour, A., Djeddi, C., Gattal, A., Siddiqi, I., & Mekhaznia, T.** (2019). Handwriting based writer recognition using implicit shape codebook. *Forensic Science International*, 301, 91–100. <https://doi.org/10.1016/j.forsciint.2019.05.014>
- Bhattacharyya, D., Ranjan, R., a, F. A., & Choi, M.** (2009). Biometric Authentication : A Review. *International Journal of Service, Science and Technology*, 2(3), 13–28.
- Bhunia A. K., Alaei A., Roy P. P.** (2019) Signature verification approach using fusion of hybrid texture features. *Neural Computing and Applications* 31:8737– 8748. <https://doi.org/10.1007/s00521-019-04220-x>
- Biau, G., & Scornet, E.** (2016). A random forest guided tour. *Test*, 25(2), 197- 227. <https://doi.org/10.1007/s11749-016-0481-7>
- Biedermann A, Bozza S, Taroni F.** (2018). Analysing and exemplifying forensic conclusion criteria in terms of Bayesian decision theory. *Science & Justice*, 58(2):159-165. [doi:10.1016/j.scijus.2017.07.002](https://doi.org/10.1016/j.scijus.2017.07.002)
- Biedermann, A., & Taroni, F.** (2012). Bayesian networks for evaluating forensic DNA profiling evidence: A review and guide to literature. *Forensic Science International: Genetics*, 6(2), 147–157. <https://doi.org/10.1016/j.fsigen.2011.06.009>
- Biedermann, A., Bozza, S., & Taroni, F.** (2009). Probabilistic evidential assessment of gunshot residue particle evidence (Part I): Likelihood ratio calculation and case pre-assessment using Bayesian networks. *Forensic Science International*, 191(1–3), 24–35. <https://doi.org/10.1016/j.forsciint.2009.06.004>
- Biedermann, A., Bozza, S., & Taroni, F.** (2011). Probabilistic evidential assessment of gunshot residue particle evidence (Part II): Bayesian parameter estimation for

- experimental count data. *Forensic Science International*, 206(1–3), 103–110. <https://doi.org/10.1016/j.forsciint.2010.07.009>
- Biedermann, A., Bozza, S., Taroni, F., and Aitken, C.G.G.** (2016). Reframing the debate: a question of probability, not of likelihood ratio. *Science & Justice* 56: 392–396. <https://doi.org/10.1016/j.scijus.2016.05.008>
- Biedermann, A., Taroni, F., Delemont, O., Semadeni, C., & Davison, A. C.** (2005 a). The evaluation of evidence in the forensic investigation of fire incidents (Part I): An approach using Bayesian networks. *Forensic Science International*, 147(1), 49–57. <https://doi.org/10.1016/j.forsciint.2004.04.014>
- Biedermann, A., Taroni, F., Delemont, O., Semadeni, C., & Davison, A. C.** (2005 b). The evaluation of evidence in the forensic investigation of fire incidents. Part II. Practical examples of the use of Bayesian networks. *Forensic Science International*, 147(1), 59–69. <https://doi.org/10.1016/j.forsciint.2004.04.015>
- Bird, C., Found, B., & Rogers, D.** (2010a). Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors. *Journal of Forensic Sciences*, 55(5), 1291- 1295. <https://doi.org/10.1111/j.1556-4029.2010.01456.x>
- Bird, C., Found, B., Ballantyne, K., & Rogers, D.** (2010 b). Forensic handwriting examiners' opinions on the process of production of disguised and simulated signatures. *Forensic Science International*, 195(1- 3), 103- 107. <https://doi.org/10.1016/j.forsciint.2009.12.001>
- Blankers, V. L., van den Heuvel, C. E., Franke, K. Y., & Vuurpijl, L. G.** (2009, July). Icdar 2009 signature verification competition. In 2009 10th International Conference on Document Analysis and Recognition (pp. 1403- 1407). IEEE. <https://doi.org/10.1109/ICDAR.2009.216>
- Blumenstein, M., Ferrer, M. A., & Vargas, J. F.** (2010, November). The 4NSigComp2010 off-line signature verification competition: Scenario 2. In 2010 12th International Conference on Frontiers in Handwriting Recognition (pp. 721-726). IEEE. <https://doi.org/10.1109/ICFHR.2010.117>
- Bolck, A., Ni, H., and Lopatka, M.** (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data. *Law, Probability and Risk* 14: 243–266. <https://doi.org/10.1093/lpr/mgv009>
- Bozza, S., Taroni, F., Marquis, R., & Schmittbuhl, M.** (2008). Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(3), 329-341. <https://doi.org/10.1111/j.1467-9876.2007.00616.x>
- Breilman, L.** (2001). Random forests. *Machine learning*, 45(1), 5- 32. <https://doi.org/10.1023/A:1010933404324>

- Brümmer, N.** (2010). Measuring, refining and calibrating speaker and language information extracted from speech (Doctoral dissertation, Stellenbosch: University of Stellenbosch). <http://hdl.handle.net/10019.1/5139>
- Brümmer, N., & Du Preez, J.** (2006). Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2- 3), 230- 275. <https://doi.org/10.1016/j.csl.2005.08.001>
- BVLC AlexNet Model.** https://github.com/BVLC/caffe/tree/master/models/bvlc_alexnet
- Castro, D. R.** (2007). Forensic evaluation of the evidence using automatic speaker recognition systems (Doctoral dissertation, Universidad autónoma de Madrid). <https://dialnet.unirioja.es/servlet/dctes?codigo=29105>
- Joshi, C.** (Feb. 21, 2019). Generative adversarial networks (GANs) for synthetic dataset generation with binary classes, Data Science Campus. <https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/>
- Champod, C., & Evett, I. W.** (2009). Evidence Interpretation: A Logical Approach. In Wiley Encyclopedia of Forensic Science. <https://doi.org/10.1002/9780470061589.fsa122>
- Champod, C., & Meuwly, D.** (2000). Inference of identity in forensic speaker recognition. *Speech Communication*, 31(2), 193– 203. [https://doi.org/10.1016/S0167-6393\(99\)00078-3](https://doi.org/10.1016/S0167-6393(99)00078-3)
- Champod, C., Lennard, C., Margot, P., & Stoilovic, M.** (2004). Fingerprints and other ridge skin impressions. In CRC International Forensic Science and Investigation series. <https://doi.org/10.1201/9780203485040>
- Chen, S.** (280). Records of The Three Kingdoms. https://en.wikipedia.org/wiki/Records_of_the_Three_Kingdoms.
- Chen, X.** (2015). Extraction and analysis of the width, gray scale and radian in Chinese signature handwriting. *Forensic Science International*. 255: 123- 132. <https://doi.org/10.1016/j.forsciint.2015.07.008>
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H.** (2015). Xgboost: extreme gradient boosting. R package version 0.4- 2, 1(4), 1- 4. <https://xgboost.readthedocs.io/en/stable/index.html>
- Chen, X.H., Champod, C., Yang, X., Shi, S.P., Luo, Y.W., Wang, N., Wang, Y.C., & Lu, Q.M.** (2018). Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features. *Forensic Science International*, 282: 101- 110. <https://doi.org/10.1016/j.forsciint.2017.11.022>
- Chu, W., Thompson, R. M., Song, J., & Vorburger, T. V.** (2013). Automatic identification of bullet signatures based on consecutive matching striae (CMS) criteria. *Forensic Science International*, 231(1-3), 137-141. <https://doi.org/10.1016/j.forsciint.2013.04.025>

- Corzo, R, Hoffman T, Weis P, Franco-Pedroso J, Ramos D, Almirall J.** (2018). The use of LA-ICP-MS databases to calculate likelihood ratios for the forensic analysis of glass evidence. *Talanta*, 186(January):655-661. <https://doi.org/10.1016/j.talanta.2018.02.027>
- Crawford, A. M., Berry, N. S., & Carriquiry, A. L.** (2021). A clustering method for graphical handwriting components and statistical writership analysis. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(1), 41-60. <https://doi.org/10.1002/sam.11488>
- Curran, J. M., Triggs, C. M., Buckleton, J. S., Walsh, K. A. J., & Hicks, T.** (1998). Assessing transfer probabilities in a Bayesian interpretation of forensic glass evidence. *Science & Justice - Journal of the Forensic Science Society*, 38(1), 15–21. [https://doi.org/10.1016/S1355-0306\(98\)72068-4](https://doi.org/10.1016/S1355-0306(98)72068-4)
- Das, A., Ferrer, M. A., Pal, U., Pal, S., Diaz, M., & Blumenstein, M.** (2016). Multi-script versus single-script scenarios in automatic off-line signature verification. *IET Biometrics*, 5(4), 305–313. <https://doi.org/10.1049/iet-bmt.2016.0010>
- Daubert v. Merrell Dow Pharmaceuticals, Inc.,** (1993). 509 US 579.
- De Baere, T., et al.** (2014). Guidelines for the single laboratory validation of instrumental and human based methods in forensic science. ENSFI. Version 2.0. <http://ensfi.eu/wp-content/uploads/2017/06/Guidelines-for-the-single-laboratory-Validation-of-Instrumental-and-Human-Based-Methods-in-Forensic-Science-2014-version-2.0.pdf>
- Demelle, F,** *Advis pour juger des inscriptions en faux, et comparaison des escritures et signatures.*, René Duelle, Paris, 1604.
- Dempster, A. P.** (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2), 205- 232. <https://doi.org/10.1111/j.2517-6161.1968.tb00722.x>
- Dewhurst, T, Found, B, Rogers, D.** (2008). Are expert penmen better than lay people at producing simulations of a model signature? *Forensic Science International*, 180(1):50-53. [Doi:10.1016/j.forsciint.2008.06.009](https://doi.org/10.1016/j.forsciint.2008.06.009)
- Diaz, M., Chanda, S., Ferrer, M. A., Banerjee, C. K., Majumdar, A., Carmona-Duarte, C., ... & Pal, U.** (2016c, October). Multiple generation of Bengali static signatures. In 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 42-47). IEEE. <https://doi.org/10.1109/ICFHR.2016.0021>
- Diaz, M., Ferrer, M. A., Eskander, G. S., & Sabourin, R.** (2016a). Generation of duplicated off-line signature images for verification systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5), 951- 964. <http://doi.org/10.1109/TPAMI.2016.2560810>
- Diaz, M., Ferrer, M. A., & Sabourin, R.** (2016b, December). Approaching the intra-class variability in multi-script static signature evaluation. In 2016 23rd International

- Conference on Pattern Recognition (ICPR) (pp. 1147-1152). IEEE. 1147-1152. <http://doi.org/10.1109/ICPR.2016.7899791>
- Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T.** (2015). Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition. Verlag für Polizeiwissenschaft, Frankfurt.
- Dudani, S. A.** (1976). The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, (4), 325- 327. <https://doi.org/10.1109/TSMC.1976.5408784>
- Egli, N., Champod, C., & Margot, P.** (2007). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – modelling with finger variability. Forensic Science International 167: 189– 195. <https://doi.org/10.1016/j.forsciint.2013.12.003>
- Ellen, D., Day, S., & Davies, C.** (2018). Scientific Examination of Documents: Methods and Techniques. Fourth Edition. CRC Press, Boca Raton. <https://doi.org/10.4324/9780429491917>
- ENFSI.** Best Practice Manual for the Forensic Examination of Handwriting. BPM-FHX-01 Version 02. 2018 (June). http://enfsi.eu/wp-content/uploads/2016/09/2._forensic_examination_of_handwriting_0.pdf
- Evelt, I. W., Lambert, J. A., & Buckleton, J. S.** (1998). A Bayesian approach to interpreting footwear marks in forensic casework. Science and Justice - Journal of the Forensic Science Society, 38(4), 241–247. [https://doi.org/10.1016/S1355-0306\(98\)72118-5](https://doi.org/10.1016/S1355-0306(98)72118-5)
- Ferrer, M. A., Chanda, S., Diaz, M., Banerjee, C. K., Majumdar, A., Carmona-Duarte, C., ... & Pal, U.** (2017). Static and dynamic synthesis of Bengali and Devanagari signatures. IEEE Transactions on Cybernetics, 48(10), 2896- 2907. <https://doi.org/10.1109/TCYB.2017.2751740>
- Ferrer, M. A., Diaz, M., Carmona-Duarte, C., & Morales, A.** (2016). A behavioral handwriting model for static and dynamic signature synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 10– 11 - 1053. <https://doi.org/10.1109/TPAMI.2016.2582167>
- Ferrer, M. A., Diaz-Cabrera, M., & Morales, A.** (2014). Static signature synthesis: A neuromotor inspired approach for biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3), 667-680. <https://doi.org/10.1109/TPAMI.2014.2343981>
- Ferrer, M. A., Vargas, J. F., Morales, A., & Ordonez, A.** (2012). Robustness of offline signature verification based on gray level features. IEEE Transactions on Information Forensics and Security, 7(3), 966-977. <https://doi.org/10.1109/TIFS.2012.2190281>

- Fierrez, J., & Ortega-Garcia, J.** (2008). On-line signature verification. In Handbook of biometrics (pp. 189-209). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-71041-9_10
- Fierrez, J., Galbally, J., Ortega-Garcia, J., Freire, M. R., Alonso-Fernandez, F., Ramos, D., ... & Gracia-Roche, J. J.** (2010). BiosecurID: a multimodal biometric database. *Pattern Analysis and Applications*, 13(2), 235-246. <https://doi.org/10.1007/s10044-009-0151-4>
- Fierrez-Aguilar, J., Alonso-Hermira, N., Moreno-Marquez, G., & Ortega-Garcia, J.** (2004, May). An off-line signature verification system based on fusion of local and global information. In *International workshop on biometric authentication* (pp. 295-306). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-25976-3_27
- Found, B., & Rogers, D.** (2003). The impact of forger practice on the validity of forensic document practitioners' opinions in Document examination & handwriting analysis, *Forensic Science International*, 136, 69– 87. [https://doi.org/10.1016/S0379-0738\(03\)90009-0](https://doi.org/10.1016/S0379-0738(03)90009-0)
- Found, B., Rogers, D., & Schmittat, R.** (1994). A computer program designed to compare the spatial elements of handwriting. *Forensic Science International*, 68(3), 195-203. [https://doi.org/10.1016/0379-0738\(94\)90358-1](https://doi.org/10.1016/0379-0738(94)90358-1)
- Franke, K. Y., Schomaker, L. R. B., Veenhuis, C., Vuurpijl, L. G., Erp, M. V., & Guyon, I.** (2004). WANDA: A common ground for forensic handwriting examination and writer identification.
- Franke, K., & Srihari, S. N.** (2008, August). Computational forensics: An overview. In *International Workshop on Computational Forensics* (pp. 1-10). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-85303-9_1
- Friedman J. H.** (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pp. 1189–1232. <https://www.jstor.org/stable/2699986>
- Friedman J., Hastie T., Tibshirani R., et al.** (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2), 337–407. <https://doi.org/10.1214/aos/1016218223>
- Friedman, J. H.** (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
- Fuglsby, C, Saunders, C, Ommen, D.M., Buscaglia, J. A, & Caligiuri, M. P.** (2021). Elucidating the relationships between two automated handwriting feature quantification systems for multiple pairwise comparisons. *Journal of Forensic Science*. <https://doi.org/10.1111/1556-4029.14914>

- Gaborini, L., Biedermann, A., & Taroni, F.** (2017). Towards a Bayesian evaluation of features in questioned handwritten signatures. *Science and Justice*, 57(3), 209–220. <https://doi.org/10.1016/j.scijus.2017.01.004>
- Galbally, J., Diaz-Cabrera, M., Ferrer, M. A., Gomez-Barrero, M., Morales, A., & Fierrez, J.** (2015). On-line signature recognition through the combination of real dynamic data and synthetically generated static data. *Pattern Recognition*, 48(9), 2921–2934. <https://doi.org/10.1016/j.patcog.2015.03.019>
- Galbally, J., Plamondon, R., Fierrez, J., & Ortega-Garcia, J.** (2012 a). Synthetic on-line signature generation. Part I: Methodology and algorithms. *Pattern Recognition*, 45(7), 2610-2621. <https://doi.org/10.1016/j.patcog.2011.12.011>
- Galbally, J., Fierrez, J., Ortega-Garcia, J., & Plamondon, R.** (2012 b). Synthetic on-line signature generation. Part II: Experimental validation. *Pattern Recognition*, 45(7), 2622-2632. <https://doi.org/10.1016/j.patcog.2011.12.007>
- Galbally, J., Gonzalez-Dominguez, S., Fierrez, J., & Ortega-Garcia, J.** (2012 c). Biografo: An Integrated Tool for Forensic Writer Identification. In *Computational Forensics* (pp. 200-211). Springer, Cham. https://doi.org/10.1007/978-3-319-20125-2_17
- Ghosh, R.** (2021) A recurrent neural network based deep learning model for offline signature verification and recognition system. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2020.114249>
- Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J.** (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech and Language*, 20(2-3 SPEC. ISS.), 331–335. <https://doi.org/10.1016/j.csl.2005.08.005>
- Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., & Ortega-Garcia, J.** (2005). Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International*, 155(2– 3), 126– 140. <https://doi.org/10.1016/j.forsciint.2004.11.007>
- Gonzalez-Rodriguez, J., Rose, P., Ramos, D., Toledano, D. T., & Ortega-Garcia, J.** (2007). Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), 2104–2115. <https://doi.org/10.1109/TASL.2007.902747>
- GPDS**, GPDS Synthetic Signature Database Website, <https://gpds.ulpgc.es/downloadnew/download.htm>.
- Greenwell, B., & Boehmke, B.** (2020). Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, 12(1), 343–366. <https://doi.org/10.32614/RJ-2020-013>

- Grivas, C. R. , & Komar, D. A. .** (2010). Kumho, daubert, and the nature of scientific inquiry: implications for forensic anthropology. *Journal of Forensic Sciences*, 53(4), 771-776. <https://doi.org/10.1111/j.1556-4029.2008.00771.x>
- Guerbai, Y., Chibani, Y., & Hadjadji, B.** (2015) The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. *Pattern Recognition* 48(1):103–113. <https://doi.org/10.1016/j.patcog.2014.07.016>
- Haack, S.** (2005). Trial and error: The Supreme Court's philosophy of science. *American Journal of Public Health*, 95(S1), S66- S73. <https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2004.044529>
- Hafemann, L. G., Sabourin, R. & Oliveira, L. S.** (2016) Writer-independent feature learning for offline signature verification using deep convolutional neural networks. In: *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp 2576–2583. <https://doi.org/10.1109/IJCNN.2016.7727521>
- Hafemann, L. G., Sabourin, R. & Oliveira, L. S.** (2017). Learning features for offline handwritten signature verification using deep convolutional neural networks. *Pattern Recognition* 70:163–176. <https://doi.org/10.1016/j.patcog.2017.05.012>
- Hafemann, L. G., Sabourin, R. & Oliveira, L. S.** (2018). Fixed-sized representation learning from offline handwritten signatures of different sizes. *International Journal on Document Analysis and Recognition (IJDAR)* 21:219–232. <https://doi.org/10.1007/s10032-018-0301-6>
- Hamadene, A. & Chibani, Y.** (2016). One-class writer-independent off-line signature verification using feature dissimilarity thresholding. *IEEE Transactions on Information Forensics and Security*, 11(6), 1226-1238. <https://doi.org/10.1016/j.patrec.2011.10.009>
- Hanmandlu, M., Hafizuddin, M., Yusof, M. & Krishna V.** (2004) Off-line signature verification and forgery detection using fuzzy modeling. *Pattern Recognition*, 38(3), 341- 356. <https://doi.org/10.1016/j.patcog.2004.05.015>
- Hannad, Y., Siddiqi, I., & El Kettani, M. E. Y.** (2016). Writer identification using texture descriptors of handwritten fragments. *Expert Systems with Applications*, 47, 14-22. <https://doi.org/10.1016/j.eswa.2015.11.002>
- Haraksim, R., Ramos, D., Meuwly, D., & Berger, C. E.** (2015). Measuring coherence of computer-assisted likelihood ratio methods. *Forensic Science International*, 249, 123-132. <https://doi.org/10.1016/j.forsciint.2015.01.033>
- Harralson, H.H., & Miller, L.S.** (2017). *Huber and Headrick's Handwriting Identification: Facts and Fundamentals* (2nd ed.). CRC Press, Boca Raton. <https://doi.org/10.4324/9781315152462>

- Harris, C., & Stephens, M.** (1988, August). A combined corner and edge detector. In *Alvey Vision Conference* (Vol. 15, No. 50, pp. 10-5244).
- Hecker, M.** (1993). Forensic information system for handwriting (FISH). Technical Document from the Kriminaltechnisches Institut, Bundeskriminalamt. Weisaden, Germany
- Heikkinen, V. V., Kassamakov, I., Barbeau, C., Lehto, S., Reinikainen, T., & Hæggeström, E.** (2014). Identifying Diagonal Cutter Marks on Thin Wires Using 3 D Imaging. *Journal of Forensic Sciences*, 59(1), 112-116. <https://doi.org/10.1111/1556-4029.12291>
- Hepler, A. B., Saunders, C. P., Davis, L. J., & Buscaglia, J.** (2012). Score-based likelihood ratios for handwriting evidence. *Forensic Science International*, 219(1-3), 129-140. <https://doi.org/10.1016/j.forsciint.2011.12.009>
- Ho, T. K.** (1995, August). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282). IEEE. <https://doi.org/10.1109/ICDAR.1995.598994>
- Horrocks, M., & Walsh, K. A. J.** (1998). Forensic palynology: Assessing the value of the evidence. *Review of Palaeobotany and Palynology*, 103(1-2), 69-74. [https://doi.org/10.1016/S0034-6667\(98\)00027-X](https://doi.org/10.1016/S0034-6667(98)00027-X)
- Houmani N, Garcia-Salicetti S, Dorizzi B.** (2012) On measuring forgery quality in online signatures. *Pattern Recognition*. 2012;45(3):1004- 1018. <https://doi.org/10.1016/j.patcog.2011.08.019>
- Hsu, C. W., & Lin, C. J.** (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415- 425. <https://doi.org/10.1109/72.991427>
- Huber, R. A.** (1959). Expert witnesses, *Crim. Law Q.* 2, 276-295.
- Huber, R. A.** (1972). The Philosophy of Identification. *RCMP Gazette*, July-August, 9-14.
- Impedovo, D., & Pirlo, G.** (2008). Automatic signature verification: The state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(5), 609-635. <https://doi.org/10.1109/TSMCC.2008.923866>
- Istik, Y. Z., Kanak, A., Bicil, Y., & Dogan, M. U.** (2007, June). A case study for the application of text-independent forensic speaker recognition using Bayesian interpretation. In *2007 IEEE 15th Signal Processing and Communications Applications* (pp. 1-4). IEEE. <https://doi.org/10.1109/siu.2007.4298673>
- Jacquet, M. & Champod, C.** (2020). Automated face recognition in forensic science: review and perspective. *Forensic Science International* 307: 110- 124. <https://doi.org/10.1016/j.forsciint.2019.110124>

- Johnson, M. Q., & Ommen, D. M.** (2021). Handwriting identification using random forests and score-based likelihood ratios. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. <https://doi.org/10.1002/sam.11566>
- Justino, E. J., Bortolozzi, F., & Sabourin, R.** (2001, September). Off-line signature verification using HMM for random, simple and skilled forgeries. In *Proceedings of Sixth International Conference on Document Analysis and Recognition* (pp. 1031-1034). IEEE. <https://doi.org/10.1109/ICDAR.2001.953942>
- Kang, T.-Y., Kim, H., Yook, S., & Lee, J.** (2022). A study on factors that affect error rates in handwriting examinations of Korean characters by forensic document examiners and non-experts. *Forensic Science International*, 334, 111266. <https://doi.org/10.1016/j.foresci.2022.111266>
- Kassambara, A.** (2018). *Machine learning essentials: Practical guide in R*. STHDA <http://www.sthda.com>
- Kramer, O.** (2013). K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors* (pp. 13-23). Springer, Berlin, Heidelberg.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E.** (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.
- Kuhn M.** (2021). caret: Classification and Regression Training. R package version 6.0-90. <https://CRAN.R-project.org/package=caret>
- Kumar, R., Sharma, J. D., & Chanda, B.** (2012). Writer-independent off-line signature verification using surroundedness feature. *Pattern Recognition Letters*, 33(3), 301-308. <https://doi.org/10.1016/j.patrec.2011.10.009>
- Kumho Tire Company, Ltd. v. Carmichael.** (1999). 526 US 137.
- Lee, F.** (1998). Review of 'Handwriting identification evidence in the Post-Daubert world'. *Journal of the American Society of Questioned Document Examiners*, 1, 67-68.
- Lee, H. R., Chen, C., & Jang, J. S. R.** (2005). Approximate lower-bounding functions for the speedup of DTW for melody recognition. *International Workshop on Cellular Neural Networks & Their Applications*. IEEE. <https://doi.org/10.1109/CNNA.2005.1543190>
- Leegwater, A.J., Meuwly, D., Sjerps, M., Vergeer, P., & Alberink, I.** (2017). Performance study of a score-based likelihood ratio system for forensic fingerprint comparison. *Journal of Forensic Sciences* 62: 626–640. <https://doi.org/10.1111/1556-4029.13339>
- Lewis, R. J.** (2000, May). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14).

- Li, B., & Ma, T.** (2018). Research on subjective bias cognition effect in handwriting identification. *Journal of Forensic Science and Medicine*, 4(4), 203-212.
- Linden J, Taroni F, Marquis R, Bozza S.** (2021). Bayesian multivariate models for case assessment in dynamic signature cases. *Forensic Science International*, 318, 110611. <https://doi.org/10.1016/j.forsciint.2020.110611>
- Ling, S.** (2002). A preliminary investigation into handwriting examination by multiple measurements of letters and spacing. *Forensic Science International*, 126(2), 145-149. [https://doi.org/10.1016/S0379-0738\(02\)00048-8](https://doi.org/10.1016/S0379-0738(02)00048-8)
- Lippmann, R.** (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(2), 4-22. <https://doi.org/10.1109/MASSP.1987.1165576>
- Liwicki, M., Malik, M. I., Alewijnse, L., van den Heuvel, E., & Found, B.** (2012, September). ICFHR 2012 competition on automatic forensic signature verification (4NsigComp 2012). In 2012 International Conference on Frontiers in Handwriting Recognition (pp. 823-828). IEEE. <https://doi.org/10.1109/ICFHR.2012.217>
- Liwicki, M., Malik, M. I., Van Den Heuvel, C. E., Chen, X., Berger, C., Stoel, R., ... & Found, B.** (2011, September). Signature verification competition for online and offline skilled forgeries (sigcomp2011). In 2011 International Conference on Document Analysis and Recognition (pp. 1480- 1484). IEEE. <https://doi.org/10.1109/ICDAR.2011.294>
- Liwicki, M., van den Heuvel, C. E., Found, B., & Malik, M. I.** (2010, November). Forensic signature verification competition 4NSigComp2010-detection of simulated and disguised signatures. In 2010 12th International Conference on Frontiers in Handwriting Recognition (pp. 715-720). IEEE. <https://doi.org/10.1109/ICFHR.2010.116>
- Lucena-Molina, J. J., Ramos-Castro, D., & Gonzalez-Rodriguez, J.** (2015). Performance of likelihood ratios considering bounds on the probability of observing misleading evidence. *Law, Probability and Risk*, 14(3), 175- 192. <https://doi.org/10.1093/lpr/mgu022>
- Lucy, D., Curran, J. and Martyna, A.** (2020). comparison: Multivariate Likelihood Ratio Calculation and Evaluation. R package version 1.0- 6. github.com/jmcurran/comparison
- Malik, M. I.** (2015). Automatic signature verification: Bridging the gap between existing pattern recognition methods and forensic science. <https://kluedo.ub.uni-kl.de/frontdoor/index/index/year/2015/docId/4253>
- Malik, M. I., Ahmed, S., Marcelli, A., Pal, U., Blumenstein, M., Alewijns, L., & Liwicki, M.** (2015, August). 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) (pp. 1186-1190). IEEE. <https://doi.org/10.1109/ICDAR.2015.7333948>

- Malik, M. I., Liwicki, M., Alewijnse, L., Ohyama, W., Blumenstein, M., & Found, B.** (2013, August). ICDAR 2013 competitions on signature verification and writer identification for on-and offline skilled forgeries (SigWiComp 2013). In 2013 12th International Conference on Document Analysis and Recognition (pp. 1477-1483). IEEE. <https://doi.org/10.1109/ICDAR.2013.220>
- Marquis, R., Bozza, S., Schmittbuhl, M., & Taroni, F.** (2011a). Handwriting evidence evaluation based on the shape of characters: Application of multivariate likelihood ratios. *Journal of Forensic Sciences*, 56, S238-S242. <https://doi.org/10.1111/j.1556-4029.2010.01602.x>
- Marquis, R., Bozza, S., Schmittbuhl, M., & Taroni, F.** (2011b). Quantitative Assessment of Handwriting Evidence: The Value of the Shape of the Letter "A". *Journal of Forensic Document Examination*, 21, 17-22.
- Marquis, R., Schmittbuhl, M., Mazzella, W. D., & Taroni, F.** (2005). Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of the capital character O. *Forensic Science International*, 150(1), 23-32. <https://doi.org/10.1016/j.forsciint.2004.06.028>
- Marquis, R., Taroni, F., Bozza, S., & Schmittbuhl, M.** (2006). Quantitative characterization of morphological polymorphism of handwritten characters loops. *Forensic Science International*, 164(2-3), 211-220. <https://doi.org/10.1016/j.forsciint.2006.02.008>
- Marsaglia, G., Tsang, W. W., & Wang, J.** (2003). Evaluating Kolmogorov's distribution. *Journal of statistical software*, 8(18), 1-4. <https://www.jstatsoft.org/article/view/v008i18>
- Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M.** (1997). The DET curve in assessment of detection task performance. National Inst of Standards and Technology Gaithersburg MD. <https://apps.dtic.mil/sti/citations/ADA530509>
- Martire, K. A., Grows, B., & Navarro, D. J.** (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review.*, 25(6), 2346-2355. <https://doi.org/10.3758/s13423-018-1448-3>
- Massey Jr, F. J.** (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68- 78. <https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769>
- Meuwly, D.** (2006). Forensic individualisation from biometric data. *Science & Justice*, 46(4), 205-213. [https://doi.org/10.1016/S1355-0306\(06\)71600-8](https://doi.org/10.1016/S1355-0306(06)71600-8)
- Meuwly, D., Ramos, D., & Haraksim, R.** (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276, 142–153. <https://doi.org/10.1016/j.forsciint.2016.03.048>

- Michalska, A., Martyna, A., & Zadora, G.** (2018). Investigation of various factors influencing Raman spectra interpretation with the use of likelihood ratio approach. *Forensic Science International*, 282, 60–73. <https://doi.org/10.1016/j.forsciint.2017.10.034>
- Microsoft and Steve Weston** (2020). foreach: Provides Foreach Looping Construct. R package version 1.5.1. <https://CRAN.R-project.org/package=foreach>
- Microsoft Corporation and Stephen Weston** (2020). doSNOW: Foreach Parallel Adaptor for the 'snow' Package. R package version 1.0.19. <https://CRAN.R-project.org/package=doSNOW>
- Miller, J. J., Patterson, R. B., Gantz, D. T., Saunders, C. P., Walch, M. A., & Buscaglia, J.** (2017). A set of handwriting features for use in automated writer identification. *Journal of Forensic Sciences*, 62(3), 722-734. <https://doi.org/10.1111/1556-4029.13345>
- Miller, L. H.** (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, 51(273), 111- 121. <https://www.tandfonline.com/doi/abs/10.1080/01621459.1956.10501314>
- Moenssens, A. A.** (1997). Handwriting identification evidence in the post-Daubert world. *UMKC Law Review*, 66, 251. <https://heinonline.org/HOL/P?h=hein.journals/umkc66&i=263>.
- Mohammed, R. A., Nabi, R. M., Sardasht, M., Mahmood, R., & Nabi, R. M.** (2015, December). State-of-the-art in handwritten signature verification system. In 2015 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 519-525). IEEE. <https://doi.org/10.1109/CSCI.2015.180>
- Montani, I.** (2015). Exploring transparent approaches to the authentication of signatures on artwork. PhD thesis. Lausanne: The University of Lausanne, School of Criminal Justice, https://www.unil.ch/files/live/sites/esc/files/Fichiers%202019bis/The%CC%80se_Montani.pdf
- Montani, I., Marquis, R., Egli Anthonioz, N., & Champod, C.** (2019). Resolving differing expert opinions. *Science and Justice*, 59(1), 1– 8. <https://doi.org/10.1016/j.scijus.2018.10.003>
- Morrison, G. S.** (2018). 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 Science International*, 283, e1-e7. <https://doi.org/10.1016/j.forsciint.2017.12.024>
- Morrison, G. S.** (2021). In the context of forensic casework, are there meaningful metrics of the degree of calibration?. *Forensic Science International: Synergy*, 100157. <https://doi.org/10.1016/j.fsisyn.2021.100157>
- Morrison, G. S., & Kinoshita, Y.** (2008). Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English/o/formant trajectories. 9th

- Annual Conference of the International Speech Communication Association
<http://publications.aston.ac.uk/id/eprint/37641/>
- Morrison, G.S.** (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic phonetic data multivariate kernel density (MKVD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication* 53: 242–256. <https://doi.org/10.1016/j.specom.2010.09.005>
- Morrison, G.S.** (2013). Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio, *Australian Journal of Forensic Sciences*, 45:2, 173-197, <https://doi.org/10.1080/00450618.2012.733025>
- Morrison, G.S. & Enzinger, E.** (2018). Score-based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality. *Science & Justice* 58: 47–58. <https://doi.org/10.1016/j.scijus.2017.06.005>
- Naika, R.** (2018). An overview of automatic speaker verification system. *Intelligent Computing and Information and Communication*, 603-610. https://doi.org/10.1007/978-981-10-7245-1_59
- National Research Council.** (2009). Strengthening forensic science in the United States: a path forward. National Academies Press. <https://www.ojp.gov/pdffiles1/nij/grants/228091.pdf>
- Nguyen, V., Blumenstein, M., Muthukkumarasamy, V., & Leedham, G.** (2007, September). Off-line signature verification using enhanced modified direction features in conjunction with neural classifiers and support vector machines. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 734-738). IEEE. <https://doi.org/10.1109/ICDAR.2007.4377012>
- Niels, R., Vuurpijl, L., & Schomaker, L.** (2005). Introducing Trigraph-trimodal writer identification. *Proceedings of European Network of Forensic Handwriting Experts*.
- Okawa M.** (2016a) Offline signature verification based on bag-of-visual words model using KAZE features and weighting schemes. In: *Proceedings of 29th IEEE conference on computer vision and pattern recognition workshops*, pp 252–258.
- Pal, S., Blumenstein, M., & Pal, U.** (2011, December). Non-English and non-Latin signature verification systems: a survey. In *CEUR Workshop Proceedings*. <http://hdl.handle.net/10453/120545>
- Paliwal, K. K., Agarwal, A., & Sinha, S. S.** (1982). A modification over Sakoe and Chiba's dynamic time warping algorithm for isolated word recognition. *Signal Processing*, 4(4), 329-333. [https://doi.org/10.1016/0165-1684\(82\)90009-3](https://doi.org/10.1016/0165-1684(82)90009-3)
- Pandya, R., & Pandya, J.** (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16), 18-21. <https://doi.org/10.5120/20639-3318>

- Popper, K. R.** (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, 10(37), 25-42. <https://www.jstor.org/stable/685773>
- R Core Team** (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Ramos-Castro, D. and Gonzalez-Rodriguez, J.** (2013a). Reliable support: measuring calibration of likelihood ratios. *Forensic Science International* 230: 156–169. <https://doi.org/10.1016/j.forsciint.2013.04.014>
- Ramos, D., Gonzalez - Rodriguez, J., Zadora, G., & Aitken, C.** (2013b). Information - theoretical assessment of the performance of likelihood ratio computation methods. *Journal of Forensic Sciences*, 58(6), 1503-1518. <https://doi.org/10.1111/1556-4029.12233>
- Rish, I.** (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* (Vol. 3, No. 22, pp. 41-46). <https://dominoweb.draco.res.ibm.com/db24eb109a77428785256aff005d3df2.html>
- Riva, F. & Champod, C.** (2014). Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases. *Journal of Forensic Sciences* 59: 637–647. <https://doi.org/10.1111/1556-4029.12382>
- Riva, F.** (2011). Etude de la valeur indicielle des traces présentes sur les douilles. PhD thesis. Lausanne: The University of Lausanne, School of Criminal Justice. https://serval.unil.ch/en/notice/serval:BIB_65EEF2A5AAD4
- Riva, F., Hermsen, R., Mattijssen, E., Pieper, P., & Champod, C.** (2017). Objective evaluation of subclass characteristics on breech face marks. *Journal of Forensic Sciences* 62: 417–422. <https://doi.org/10.1111/1556-4029.13274>
- Robertson, B., Vignaux, G. A., & Berger, C. E.** (2016). *Interpreting evidence : evaluating forensic science in the courtroom*. Chichester: John Wiley & Sons. <https://doi.org/10.1002/9781118492475>
- Robin R.** (Aug. 23, 2019). The Prospects and Limitations of Synthetic Data, <https://www.linkedin.com/pulse/prospects-limitations-synthetic-data-robin-r%C3%B6hm>
- Royall, R.** (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association* 95: 760–768
- RStudio Team** (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.
- Sakarya, U., Leloğlu, U. M., & Tunali, E.** (2008). Three-dimensional surface reconstruction for cartridge cases using photometric stereo. *Forensic Science International*, 175(2-3), 209-217. <https://doi.org/10.1016/j.forsciint.2007.07.003>

- Sakoe, H., & Chiba, S.** (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1), 43-49. <https://doi.org/10.1109/TASSP.1978.1163055>
- Saks, M. J.** (1998). Merlin and Solomon: Lessons from the Law's Formative Encounters with Forensic Identification Science. *Hastings Law Journal*, 49(4), 1069-1141.
- Saks, M. J.** (2010). Forensic identification: from a faith-based "Science" to a scientific science. *Forensic Science International*, 201(1- 3), 14- 17. <https://doi.org/10.1016/j.forsciint.2010.03.014>
- Saks, M. J., & Koehler, J. J.** (2005). The coming paradigm shift in forensic identification science. *Science*, 309(5736), 892-895. <https://doi.org/10.1126/science.1111565>
- Sanders, J.** (2001). "Kumho" and How We Know. *Law and Contemporary Problems*, 64(2/3), 373-415. <https://doi.org/10.2307/1192317>
- Sarle, W. S.** (1994). Neural networks and statistical models. Proceedings of the Nineteenth Annual SAS Users Group International Conference, April. SAS Institute Inc., Cary, NC. https://people.orie.cornell.edu/davidr/or474/nn_sas.pdf
- Scholkopf, B., Sung, K. K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., & Vapnik, V.** (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11), 2758-2765. <https://doi.org/10.1109/78.650102>
- Serdouk, Y., Nemmour, H., & Chibani, Y.** (2016). New off-line handwritten signature verification method based on artificial immune recognition system. *Expert Systems with Applications*, 51, 186-194. <https://doi.org/10.1016/j.eswa.2016.01.001>
- Sharif M, Khan M. A., Faisal M., Yasmin M., Fernandes S. L.** (2018). A framework for offline signature verification system: best features selection approach. *Pattern Recognition Letter* 139:50–59. <https://doi.org/10.1016/j.patrec.2018.01.021>
- Skerrett, J., Neumann, C., & Mateos-Garcia, I.** (2011). A Bayesian approach for interpreting shoemark evidence in forensic casework: Accounting for wear features. *Forensic Science International*, 210(1-3), 26-30. <https://doi.org/10.1016/j.forsciint.2011.01.030>
- Soleimani, A., Araabi, B. N., & Fouladi, K.** (2016). Deep multitask metric learning for offline signature verification. *Pattern Recognition Letters*, 80, 84- 90. <https://doi.org/10.1016/j.patrec.2016.05.023>
- Spagnolo, G. S.** (2006). Potentiality of 3D laser profilometry to determine the sequence of homogenous crossing lines on questioned documents. *Forensic Science International*, 164(2-3), 102-109. <https://doi.org/10.1016/j.forsciint.2005.12.004>
- Spagnolo, G. S., Cozzella, L., & Simonetti, C.** (2013). Linear conosopic holography as aid for forensic handwriting expert. *Optik*, 124(15), 2155- 2160. <https://doi.org/10.1016/j.ijleo.2012.06.097>

- Srihari, S. N., Beal, M. J., Bandi, K., Shah, V., & Krishnamurthy, P.** (2005, August). A statistical model for writer verification. In Eighth International Conference on Document Analysis and Recognition (ICDAR'05) (pp. 1105- 1109). IEEE. <https://doi.org/10.1109/ICDAR.2005.33>
- Srihari, S. N., Srinivasan, H., & Desai, K.** (2007). Questioned Document Examination Using CEDAR-FOX. *Journal of Forensic Document Examination*, 18, 1-20.
- Srihari, S. N., Xu, A., & Kalera, M. K.** (2004, October). Learning strategies and classification methods for off-line signature verification. In Ninth International Workshop on Frontiers in Handwriting Recognition (pp. 161-166). IEEE. <https://doi.org/10.1109/IWFHR.2004.61>
- State Administration for Market Regulation & Standardization Administration of The People's Republic China.** (2018). Specification for forensic identification of handwriting. GB/T 37239-2018,
- Sulner, A.** (2018). Critical Issues Affecting the Reliability and Admissibility of Handwriting Identification Opinion Evidence—How They Have Been Addressed (or Not) Since the 2009 NAS Report, and How They Should Be Addressed Going Forward: A Document Examiner Tells All. *Seton Hall Law Review*, 48(3), Article 5.
- Supreme Court of the United States**, Brief Amicus Curiae of Americans for Effective Law Enforcement et al., submitted in *Kumho Tire v. Carmichael*, No. 97-1709, in the Supreme Court of the United States (October Term, 1997)
- SWGDOC**, Standard for Examination of Handwritten Items, SWGDOC, <http://www.swgdoc.org/index.php/standards/published-standards>, 2013.
- Swofford, H. J., Koertner, A. J., Zemp, F., Ausdemore, M., Liu, A., & Salyards, M. J.** (2018). A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation. *Forensic Science International*, 287, 113-126. <https://doi.org/10.1016/j.forsciint.2018.03.043>
- Swofford, H., & Champod, C.** (2021). Implementation of algorithms in pattern & impression evidence: A responsible and practical roadmap. *Forensic Science International: Synergy*, 3, 100142. <https://doi.org/10.1016/j.fsisyn.2021.100142>
- Taherzadeh, G., Karimi, R., Ghobadi, A., & Beh, H. M.** (2011, February). Evaluation of online signature verification features. In 13th International Conference on Advanced Communication Technology (ICACT2011) (pp. 772- 777). IEEE. <https://ieeexplore.ieee.org/document/5745925>
- Tan, G. J., Sulong, G., & Rahim, M. S. M.** (2017). Writer identification: A comparative study across three world major languages. *Forensic Science International*, 279, 41-52. <https://doi.org/10.1016/j.forsciint.2017.07.034>

- Taroni, F., Lambert, J. A., Fereday, L., & Werrett, D. J.** (2002). Evaluation and presentation of forensic DNA evidence in European laboratories. *Science & Justice*, 42(1), 21-28. [https://doi.org/10.1016/S1355-0306\(02\)71793-0](https://doi.org/10.1016/S1355-0306(02)71793-0)
- Taroni, F., Marquis, R., Schmittbuhl, M., Biedermann, A., Thiéry, A., & Bozza, S.** (2012). The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination. *Forensic Science International*, 214(1-3), 189-194. <https://doi.org/10.1016/j.forsciint.2011.08.007>
- Taroni, F., Marquis, R., Schmittbuhl, M., Biedermann, A., Thiéry, A., & Bozza, S.** (2014). Bayes factor for investigative assessment of selected handwriting features. *Forensic Science International*, 242, 266-273. <https://doi.org/10.1016/j.forsciint.2014.07.012>
- Taylor, D.** (2014). Using continuous DNA interpretation methods to revisit likelihood ratio behaviour. *Forensic Science International: Genetics* 11: 144– 153. <https://doi.org/10.1016/j.fsigen.2014.03.008>
- Taylor, M. , Bird, C. , Bishop, B. , Burkes, T. , Caligiuri, M. , Found, B. , Grose, W. , Logan, L. , Melson, K. , Merlino, M. , Miller, L. , Mohammed, L. , Morris, J. , , J. , Osborne, N. , Ostrum, B. , Saunders, C. , Shappell, S. , , H. , Srihari, S. , Stoel, R. , Vastrick, T. , Waltke, H. and Will, E.** (2020). *Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach*, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], <https://doi.org/10.6028/NIST.IR.8282> (Accessed November 9, 2021)
- Van Leeuwen, D. A., & Brümmer, N.** (2007). An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I* (pp. 330-353). Springer, Berlin, Heidelberg.
- Vargas, J. F., Ferrer, M. A., Travieso, C. M., & Alonso, J. B.** (2011). Off-line signature verification based on grey level information using texture features. *Pattern Recognition*, 44(2), 375-385. <https://doi.org/10.1016/j.patcog.2010.07.028>
- Venables, W. N., & Ripley, B. D.** (2000). *Modern Applied Statistics with S-Plus*. New York : Springer Verlag
- Wickham H, Averick M, Bryan J, et al.** (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43): 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H.** (2009). Elegant graphics for data analysis. *Media*, 35(211), 10-1007. <https://doi.org/10.1007/978-0-387-98141-3>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H.** (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

- Williams, M. R., Sigman, M. E., Lewis, J., & Pitan, K. M.** (2012). Combined target factor analysis and Bayesian soft-classification of interference-contaminated samples: Forensic fire debris analysis. *Forensic Science International*, 222(1-3), 373-386. <https://doi.org/10.1016/j.forsciint.2012.07.021>
- Yager N, Dunstone T.** (2010). The biometric menagerie. *IEEE Trans Pattern Anal Mach Intell*;32(2):220–30. <https://doi.org/10.1109/TPAMI.2008.291>
- Yilmaz, M. B.** (2015). Offline signature verification with user-based and global classifiers of local features (Doctoral dissertation). <http://research.sabanciuniv.edu/31181/>
- Ypma, R. J. F., Maaskant-van Wijk, P. A., Gill, R., Sjerps, M., & van den Berge, M.** (2021). Calculating LR for presence of body fluids from mRNA assay data in mixtures. *Forensic Science International: Genetics*, 52, 102455. <https://doi.org/10.1016/j.fsigen.2020.102455>
- Zhang Z.** (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, 4(11), 218. <https://doi.org/10.21037/atm.2016.03.37>
- Zois E. N., Tsourounis D., Theodorakopoulos I., Kesidis A. L., Economou G.** (2019) A comprehensive study of sparse representation techniques for offline signature verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1), 68-81. <https://doi.org/10.1109/TBIOM.2019.2897802>
- Zois, E. N., Alewijnse, L., & Economou, G.** (2016). Offline signature verification and quality characterization using poset-oriented grid features. *Pattern Recognition*, 54, 162-177. <https://doi.org/10.1016/j.patcog.2016.01.009>

Appendix

Publications

Forensic Science International 255 (2015) 123–132



Extraction and analysis of the width, gray scale and radian in Chinese signature handwriting



Xiaohong Chen*

Institute of Forensic Science, Ministry of Justice, 1347, West Guangfu Road, Shanghai 200063, PR China

ARTICLE INFO

Article history:
Available online 13 July 2015

Keywords:
Stroke measurement
Identification
Pictograms
Statistical analysis
Dynamic time warping

ABSTRACT

Forensic handwriting examination is a relevant identification process in forensic science. This research obtained ideas from the process of features detection and analysis in forensic handwriting examination. A Chinese signature database was developed and comprised original signatures, freehand imitation forgeries, random forgeries and tracing imitation forgeries. The features of width, gray scale and radian combined with stroke orders were automatically extracted after image processing. A correlation coefficient was used to precisely characterize and express the similarities between signatures. To validate the differences between writers, a multivariate analysis of the variance was employed. The canonical discriminant analysis was performed between the original and non-original signatures; the cross-validation estimated the discriminating power of the width, gray scale and radian data. It is suggested that the extraction and analysis of these properties in Chinese signatures is reasonable. Meanwhile, forensic handwriting examination using the quantitative feature extraction and statistical analysis methods in this research could be performed with a satisfactory result in the discriminant analysis.

© 2015 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Forensic handwriting examination is a relevant identification process in forensic science, and it is one important part of the tasks performed by forensic document examiners [1,2]. Forensic handwriting examination appears as an empirical discipline. Debates regarding the scientific basis of forensic handwriting examination have arisen in recent years [3,4]. The fundamental law considering the variability and individuality of handwriting were challenged for the lack of objectivity. Various studies have shown that trained forensic document examiners perform significantly better than lay people in the analysis, comparison and identification processes based on handwriting evidence [5]. Much research on the applications of quantitative extraction methods for handwriting features [6–10] and on the statistics methods for forensic handwriting examination assessment [11–16] has been conducted to reduce the subjective aspect in forensic handwriting examination.

Various features are extracted and applied for Chinese handwriting verification [17]. However, this research lacks the

validation of the variability within-writer and the individuality between-writer based on the features that are used in comparison. Additionally, the differences between forensic handwriting examination and computer handwriting verification lead to verification methods that are not practical for forensic application. However, the quantitative feature extraction, statistical analysis and overall assessment are the trend of current forensic handwriting examination [18–20].

This study based on the process of feature detection and analysis in forensic handwriting examination. In addition to spatial features, forensic document examiners should recover the stroke orders and grasp the stroke contours, width and gray scale for comparison. The aim of this paper was to develop new quantitative and objective feature extraction and statistical analysis methods for Chinese signature handwriting examination.

A Chinese signature database was comprised of 981 signatures of 12 groups produced by 12 volunteers; each group consisted of original signatures, freehand imitation forgeries, random forgeries and tracing imitation forgeries. A threshold was applied in image binarization after the signatures were imported into the computer. The skeleton and the signature edges were extracted by image processing. Then, a program for stroke order recovery processed the skeleton of the signatures. The width, gray scale and radian values were automatically extracted in the stroke order.

* Corresponding author.
E-mail address: chenxh@ssfjd.cn

A dynamic time warping (DTW) method was applied to cope with the different writing speeds. Corresponding strokes were aligned for comparative analysis after DTW. In the correlation analysis, the correlation coefficient was used to precisely characterize and express the similarities between signatures.

In the statistical analysis, the mean and standard deviation of the width, gray scale and radian obtained in the correlation analysis were calculated as the data description. A multivariate analysis of the variance was employed to validate the differences between writers based on width, gray scale and radian. A discriminant analysis helped to distinguish different writers with the canonical discriminant functions.

2. Materials and data

2.1. Sampling

Chinese signatures were written by means of a ballpoint pen with black ink, on A4 paper (for original signatures, random forgeries and freehand imitation forgeries) printed with 12 squares and 195 mm × 271 mm highly transparent paper (for tracing imitation forgeries), with the signer in a sitting pose. Twelve volunteers who could produce skilled imitation forgeries were organized. The Chinese signature database was composed of 981 signatures of 12 groups produced by 12 volunteers; every groups contained 20–24 original signatures, 30–36 freehand imitation forgeries, 10–12 random forgeries and 10–12 tracing imitation forgeries.

Original Signature (OR): Each volunteer wrote 20–24 original signatures using their normal writing style in two writing sessions. One original signature was chosen as the model at random and was duplicated into a blank A4 paper for forgery production.

Freehand imitation forgery signature (FF): Three volunteers other than the writer of the original signature were chosen at random. After practicing for 2–4 days until the forgeries were produced reasonably well, each volunteer wrote 10–12 forgeries while attempting to imitate the model signature (FF1, FF2, and FF3, respectively).

Random forgery signature (RF): With the knowledge of only the name of the original writer, one volunteer other than writer of the original signature was chosen at random; this volunteer produced 10–12 signatures based on their own writing habits.

Tracing imitation forgery Signature (TF): One volunteer other than the writer of the original signature was chosen at random. The volunteer was asked to place a piece of highly transparency paper

on top of the original signature model and carefully trace the model signature, attempting to imitate the model signature.

2.2. Image processing procedure and features extraction

The Chinese signatures were imported into the computer by means of an EPSON PERFECTION V700 PHOTO scanner with a resolution of 400 dpi. Matlab 7.0 software was applied for to extract the features. Stroke order or sequence is an important feature in handwriting. A program of stroke order tracing is referred to in the research of Oshiharu Kato and Makoto Yasuhara [21].

Firstly, a threshold was applied in the image binarization. Then, the skeletons and edges of the signatures were extracted by separately skeletonizing and edging the images (Fig. 1). The stroke order tracing was automatically used in the signature skeletons after a begin point was manually provided (Fig. 2). If any error occurred, it was corrected manually. The stroke order or the sequence of signatures (S) were assigned with x coordinates (X) and y coordinates (Y). Finally, the width (W), gray scale (G) and radian (R) values were automatically extracted in the stroke order (Fig. 3). The gray scale values of the points in the skeletons were used as the gray scale data. The width and radian data were calculated as the following functions, as shown in Fig. 3:

$$S = \sum_{i=1}^n (X_i, Y_i); \quad (1)$$

$$W = \sum_{i=1}^n (L_i \cdot \sin \theta_i); \quad (2)$$

$$R = \sum_{i=1}^n \theta_i \left(\frac{\pi}{180} \right); \quad (3)$$

$$F = \sum_{i=1}^n W_i, G_i, R_i; \quad (4)$$

where stroke order (S); x coordinate (X); y coordinate (Y); width (W), gray scale (G), radian (R); tangent line (T), n = length (X) or length (Y); $i = 1, 2, 3, \dots, n$.

2.3. Feature data preparation

2.3.1. Dynamic time warping application

Dynamic time warping is a well-known technique to find an optimal alignment between two time-dependent sequences under

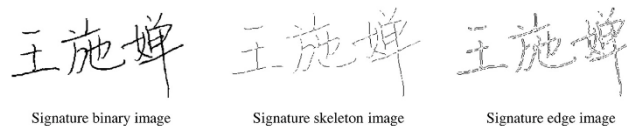


Fig. 1. Image pressing.

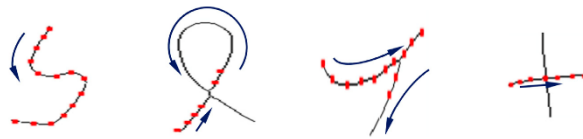


Fig. 2. Basic stroke order recovery processing image.

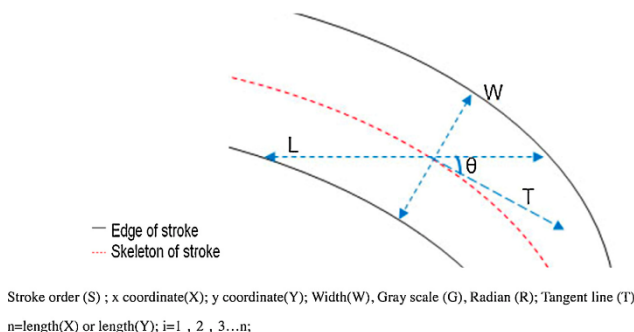


Fig. 3. Width and radian extraction.

certain restrictions. Intuitively, the sequences are warped in a nonlinear fashion to match certain patterns. In fields such as signature statistical analysis, verification and classification [22–25], the DTW method has been successfully applied to automatically cope with time deformations and the different speeds associated with time-dependent data. Matlab 7.0 was used in the DTW calculation.

It is commonly known that writing speed cannot be repeated even when signatures are derived from the same writer (Fig. 4(a and b)). The corresponding stroke should be aligned for comparative analysis. The radian data were assumed to exactly reflect the contour or shape of the stroke. As a result of DTW processing based on the radian data, the corresponding strokes were aligned and matched (Fig. 4(c)).

2.3.2. Correlation analysis

In statistics, the correlation coefficient measures the strength and direction of a linear relationship between two variables. In this research, the correlation coefficient was used to precisely characterize the similarities between signatures. For original signatures (OR), for example, the correlation coefficients of one OR signature were calculated between each possible pair of this OR signature and the other OR signatures. The maximum value was used as the final correlation coefficient of this OR signature. For non-original signatures (Non-OR signatures, including FF, RF and TF signatures), for example, the correlation coefficients of one Non-OR signature were calculated between each possible pair of this Non-OR signatures and all OR signatures. The maximum value was

used as the final correlation coefficient of this Non-OR signature. All of the statistical analyses described below were performed on the correlation coefficient of the width, gray scale and radian data obtained in the correlation analysis. Matlab 7.0 was also used for the correlation analysis.

2.4. Statistical analysis

IBM SPSS statistics 19.0 software was utilized in the statistical analysis.

Initially, the mean and the standard deviation of the width, gray scales and radian data for OR, FF1, FF2, FF3, RF and TF signatures were calculated in each group. Then, to observe the variability of OR, FF, RF and TF signatures in each group, multivariate analysis of variance (MANOVA) was used to validate the differences between OR, FF, RF and TF signatures.

Eventually, the canonical discriminant analysis was performed on the width, gray scale and radian data in two levels: the discriminant analysis between OR, FF, RF and TF signatures, and the discriminant analysis between the OR and Non-OR signatures. The cross-validation was used to express the discriminating power of the width, gray scale and radian data between the OR and Non-OR signatures.

3. Result

The mean and standard deviation of the width, gray scale and radian data in the 12 groups is summarized in Table 1, Table 2 and

Table 1 Summary statistics^a of the correlation coefficient of width in 12 groups.

Signature	Statistics	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	X	0.50	0.40	0.44	0.36	0.53	0.48	0.53	0.35	0.42	0.48	0.37	0.42
	S.D.	0.07	0.05	0.04	0.04	0.07	0.07	0.11	0.06	0.05	0.06	0.07	0.06
FF1	X	0.35	0.26	0.24	0.28	0.42	0.35	0.36	0.25	0.24	0.22	0.23	0.33
	S.D.	0.07	0.07	0.07	0.03	0.04	0.05	0.07	0.08	0.04	0.11	0.06	0.08
FF2	X	0.29	0.22	0.17	0.20	0.38	0.31	0.28	0.21	0.15	0.18	0.24	0.35
	S.D.	0.06	0.07	0.04	0.05	0.06	0.07	0.06	0.03	0.04	0.06	0.06	0.07
FF3	X	0.28	0.28	0.23	0.24	0.38	0.33	0.38	0.28	0.27	0.21	0.24	0.20
	S.D.	0.08	0.04	0.06	0.04	0.07	0.05	0.05	0.06	0.05	0.04	0.08	0.04
RF	X	0.20	0.15	0.12	0.20	0.26	0.34	0.24	0.24	0.21	0.25	0.19	0.17
	S.D.	0.07	0.05	0.08	0.03	0.06	0.07	0.06	0.06	0.05	0.05	0.05	0.04
TF	X	0.34	0.20	0.35	0.28	0.43	0.25	0.25	0.23	0.23	0.36	0.25	0.24
	S.D.	0.04	0.07	0.07	0.04	0.05	0.05	0.07	0.05	0.08	0.05	0.06	0.05

^a X, mean; S.D., standard deviation.

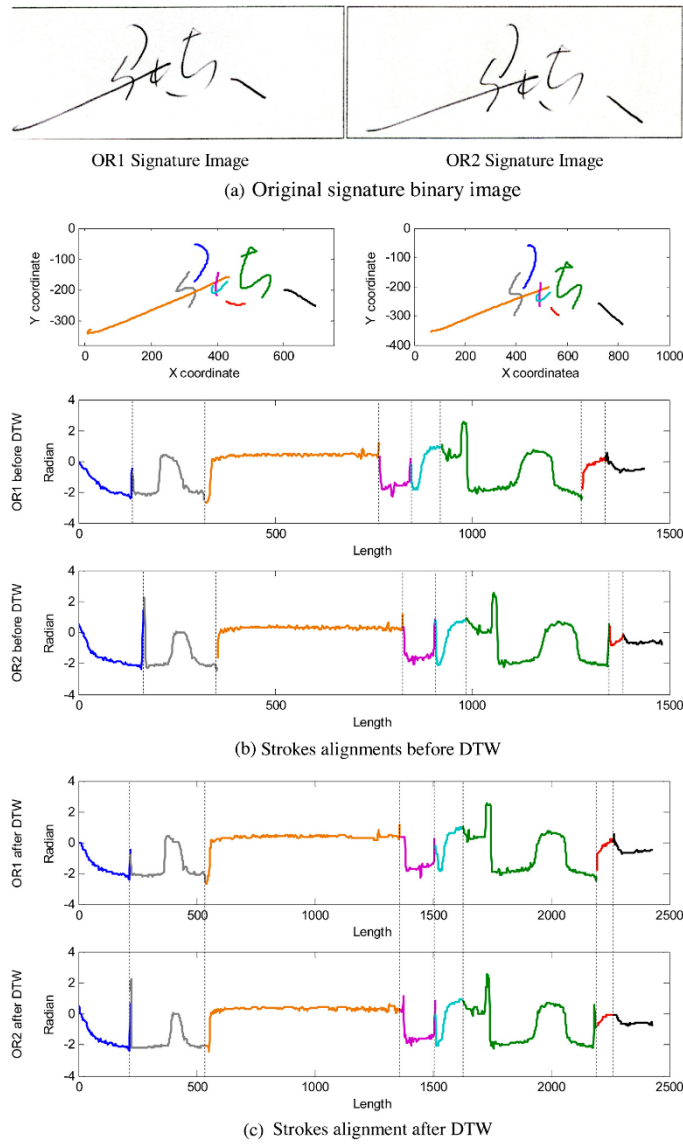


Fig. 4. (a) Original signature binary image; (b) strokes alignments before DTW; (c) strokes alignment after DTW.

Table 3, respectively. The results showed that the overall mean radian values were higher than the width and gray scale values in each group. The mean values of the width, gray scale and radian values were the always highest in the OR signatures in each group

(Tables 1–3). The mean values of the TF signatures were highest in the non-original signatures (Table 3) in respect to the radian data; however, the mean TF values were not always higher than the other non-original signatures in respect to the width and gray scale

Table 2
Summary statistics^a of the correlation coefficient of gray scale in 12 groups.

Signature	Statistics	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	X	0.63	0.64	0.66	0.58	0.69	0.69	0.73	0.59	0.63	0.75	0.65	0.74
	S.D.	0.05	0.04	0.04	0.05	0.06	0.05	0.07	0.03	0.05	0.08	0.07	0.03
FF1	X	0.44	0.52	0.43	0.46	0.41	0.49	0.47	0.42	0.36	0.29	0.46	0.47
	S.D.	0.06	0.04	0.09	0.05	0.09	0.07	0.09	0.08	0.07	0.15	0.06	0.05
FF2	X	0.39	0.33	0.34	0.42	0.43	0.41	0.48	0.35	0.27	0.23	0.46	0.49
	S.D.	0.11	0.09	0.06	0.05	0.05	0.05	0.10	0.09	0.07	0.07	0.05	0.05
FF3	X	0.35	0.46	0.38	0.4	0.48	0.48	0.52	0.37	0.38	0.34	0.46	0.42
	S.D.	0.07	0.05	0.08	0.06	0.09	0.07	0.06	0.08	0.07	0.10	0.04	0.08
RF	X	0.31	0.37	0.27	0.37	0.25	0.45	0.19	0.39	0.28	0.29	0.22	0.34
	S.D.	0.04	0.06	0.06	0.05	0.08	0.05	0.10	0.06	0.10	0.05	0.06	0.06
TF	X	0.34	0.38	0.38	0.38	0.41	0.35	0.31	0.28	0.22	0.45	0.37	0.34
	S.D.	0.05	0.06	0.10	0.07	0.05	0.07	0.09	0.10	0.13	0.08	0.05	0.07

^a X, mean; S.D., standard deviation.

values (Tables 1 and 2). The mean values of the RF signatures were typically the lowest in each group with respect to the radian data (Table 3), and the mean overall values of RF signatures were also lower than other Non-OR signatures with respects to the width and gray scale.

3.1. Result of MANOVA

According to the mean values of the width, gray scale and radian data, the differences between FF1, FF2 and FF3 seem minor. As a result, FF1, FF2 and FF3 were gathered together as FF. A MANOVA was performed on the correlation coefficient of the width, gray scale and radian data obtained from the correlation analysis. The results of the pairwise comparison based on the width, gray scale and radian data are summarized in Table 4, Table 5 and Table 6 respectively.

The mean distances between signatures with width as the dependent variable are shown in Table 4. The minimum mean distance values between OR and Non-OR signature in 12 groups were OR-FF (G2, G6, G7, G8, and G12) and OR-TF (G1, G3, G4, G5, G9, G10 and G11). The result of the MANOVA showed a significant difference between OR and Non-OR signatures based on width. However, significant differences did not always exist between Non-OR signatures. For instance, differences between FF and RF signatures were not significant in three groups (G6, G8 and G9); differences between FF and TF signatures were not significant in five groups (G1, G5, G8, G9 and G11); differences between RF and

TF signatures were not significant in four groups (G2, G7, G8, and G9).

The mean distances between signatures with gray scale as the dependent variable are shown in Table 5. The minimum mean distance values between OR and Non-OR signature in 12 groups were OR-FF (G1, G2, G4, G5, G6, G7, G9, G11 and G12), OR-TF (G3 and G10) and OR-RF (G8). The result of the MANOVA also showed highly significant difference between OR and Non-OR signatures based on gray scale. However, significant differences did not always exist between Non-OR signatures. For instance, differences between FF and RF signatures were not significant in three groups (G6, G9 and G10); differences between FF and TF signatures were not significant in two groups (G3 and G5); differences between RF and TF signatures were not significant in five groups (G1, G2, G4, G9, and G12).

The mean distances between signatures with radian as the dependent variable are shown in Table 6. The minimum distance values between OR and Non-OR signatures were shown as OR-FF (G2 and G11) and OR-TF (G1, G3 to G10 and G12). The result of the MANOVA showed highly significant differences between OR and Non-OR signatures except for the TF signature. Only 7 groups (G1, G2, G6, G8, G9, G11 and G12) showed significant differences between OR and TF signatures. In addition, differences between OR and TF signatures were not significant in 5 groups (G3, G4, G5, G7 and G10). Differences between FF and TF signatures were not significant in 5 groups (G1, G2, G9, G11 and G12).

Table 3
Summary statistics^a of the correlation coefficient of radian in 12 groups.

Signature	Statistics	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	X	0.95	0.94	0.89	0.9	0.97	0.92	0.97	0.91	0.91	0.96	0.91	0.95
	S.D.	0.02	0.02	0.03	0.03	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02
FF1	X	0.9	0.92	0.79	0.89	0.95	0.88	0.96	0.88	0.88	0.58	0.85	0.95
	S.D.	0.03	0.02	0.03	0.03	0.01	0.02	0.02	0.02	0.03	0.2	0.04	0.02
FF2	X	0.89	0.91	0.7	0.78	0.93	0.83	0.94	0.85	0.78	0.5	0.85	0.91
	S.D.	0.02	0.02	0.05	0.04	0.03	0.04	0.02	0.04	0.06	0.07	0.03	0.02
FF3	X	0.87	0.93	0.84	0.86	0.93	0.79	0.86	0.84	0.86	0.83	0.87	0.93
	S.D.	0.03	0.02	0.03	0.03	0.02	0.03	0.03	0.03	0.05	0.17	0.03	0.02
RF	X	0.77	0.8	0.67	0.79	0.76	0.8	0.73	0.78	0.74	0.78	0.68	0.77
	S.D.	0.03	0.03	0.03	0.04	0.04	0.05	0.07	0.03	0.06	0.04	0.05	0.04
TF	X	0.89	0.92	0.88	0.89	0.97	0.88	0.96	0.89	0.85	0.95	0.84	0.94
	S.D.	0.03	0.02	0.03	0.03	0.01	0.03	0.01	0.02	0.03	0.01	0.05	0.02

^a X, mean; S.D., standard deviation.

Table 4
MANOVA: mean distance from I to J in the pairwise comparison with width as the dependent variable.

(I)	(J)	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	FF	0.187 [*]	0.152 [*]	0.224 [*]	0.120 [*]	0.139 [*]	0.144 [*]	0.192 [*]	0.105 [*]	0.200 [*]	0.275 [*]	0.129 [*]	0.134 [*]
	RF	0.293 [*]	0.247 [*]	0.319 [*]	0.164 [*]	0.274 [*]	0.131 [*]	0.294 [*]	0.110 [*]	0.212 [*]	0.224 [*]	0.175 [*]	0.250 [*]
	TF	0.152 [*]	0.211 [*]	0.087 [*]	0.082 [*]	0.105 [*]	0.223 [*]	0.284 [*]	0.119 [*]	0.189 [*]	0.119 [*]	0.118 [*]	0.187 [*]
FF	RF	0.106 [*]	0.095 [*]	0.094 [*]	0.044 [*]	0.134 [*]	-0.013	0.102 [*]	0.005	0.011	-0.052 [*]	0.046 [*]	0.115 [*]
	TF	-0.035	0.060 [*]	-0.137 [*]	-0.039 [*]	-0.034	0.078 [*]	0.092 [*]	0.014	-0.011	-0.157 [*]	-0.011	0.052 [*]
RF	TF	-0.141 [*]	-0.036	-0.231 [*]	-0.083 [*]	-0.169 [*]	0.092 [*]	-0.010	0.009	-0.022	-0.105 [*]	-0.057 [*]	-0.063 [*]

^{*} The mean difference is significant at <0.05 level.

Table 5
MANOVA: mean distance from I to J in the pairwise comparison with gray scale as the dependent variable.

(I)	(J)	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	FF	0.239 [*]	0.207 [*]	0.280 [*]	0.154 [*]	0.246 [*]	0.227 [*]	0.237 [*]	0.205 [*]	0.296 [*]	0.458 [*]	0.189 [*]	0.275 [*]
	RF	0.323 [*]	0.268 [*]	0.389 [*]	0.208 [*]	0.440 [*]	0.234 [*]	0.540 [*]	0.195 [*]	0.350 [*]	0.454 [*]	0.425 [*]	0.401 [*]
	TF	0.293 [*]	0.263 [*]	0.278 [*]	0.202 [*]	0.279 [*]	0.340 [*]	0.416 [*]	0.300 [*]	0.409 [*]	0.293 [*]	0.272 [*]	0.393 [*]
FF	RF	0.084 [*]	0.061 [*]	0.109 [*]	0.054 [*]	0.194 [*]	0.006	0.303 [*]	-0.010 [*]	0.053	-0.004	0.236 [*]	0.126 [*]
	TF	0.054 [*]	0.056 [*]	-0.003	0.049 [*]	0.034	0.113 [*]	0.180 [*]	0.095 [*]	0.113 [*]	-0.165 [*]	0.083 [*]	0.118 [*]
RF	TF	-0.030	-0.005	-0.112 [*]	-0.005	-0.161 [*]	0.107 [*]	-0.124 [*]	0.105 [*]	0.059	-0.161 [*]	-0.153 [*]	-0.008

^{*} The mean difference is significant at <0.05 level.

According to the result of the MANOVA, the discriminant analysis was designed to perform on two levels. Firstly, the discriminant analysis was performed between OR, FF, RF and TF signatures. Then, the discriminant analysis was necessary to performed between OR and Non-OR signatures.

3.2. Discriminant between OR, FF, RF and TF signatures

In the discriminant analysis of OR, FF, RF and TF signatures in each group, three functions were used to distinguish OR, FF, RF and TF signatures. The first two canonical discriminant functions accounted for not less than 95.6% of the total variances. The discriminating power of first two functions was significant according to the Wilks' Lambda test ($p < 0.001$) in each group.

Scatter-plots of first two axes of the canonical discriminant analysis in 12 groups are shown in Fig. 5. In this figure, OR signatures were clearly classified from Non-OR signatures in 8 groups (G1, G3, G5, G6, G7, G8, G9 and G12); misclassification always occurred between FF and TF, but less misclassification occurred between RF and other Non-OR signatures. Four groups (G5, G7, G11 and G12) were classified correctly.

3.3. Discriminant between OR and NON-OR signatures

A cross-validation was applied to estimate how accurately the discriminating functions will perform in practice, excluding one

method. The result of the cross-validation is summarized in Table 7. The best cross-validation value was 98.8% in G12, and the lowest cross-validation value was 90.5% in G11. The mean and standard deviation of the cross-validation was 95.8% and 0.023, respectively.

4. Discussion and perspective

This paper provided a new methodology for the quantitative feature extraction and statistical analysis of Chinese signature examination. The width, gray scale and radian features combined with the stroke order can be automatically extracted in a quantitative and objective way.

The result of the MANOVA confirmed highly significant differences between OR and NON-OR signatures in 12 groups with respect to width and gray scale. Moreover, highly significant differences between OR and FF and between OR and TF were validated with respect to radian data. The mean distances between the observations showed that the imitation signatures, such as the FF and TF signatures were close to the OR signatures in width, gray scale and radian values. The differences between the OR and TF signatures were not significant with regards to radian data. The imitation signatures were more similar to OR signatures than to RF signatures considering their width, gray scale and radian data.

Table 6
MANOVA: mean distance from I to J the pairwise comparison with radian as the dependent variable.

(I)	(J)	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
OR	FF	0.057 [*]	0.025 [*]	0.112 [*]	0.058 [*]	0.035 [*]	0.083 [*]	0.053 [*]	0.061 [*]	0.072 [*]	0.314 [*]	0.056 [*]	0.024 [*]
	RF	0.175 [*]	0.150 [*]	0.218 [*]	0.110 [*]	0.211 [*]	0.121 [*]	0.240 [*]	0.140 [*]	0.173 [*]	0.183 [*]	0.233 [*]	0.183 [*]
	TF	0.054 [*]	0.036 [*]	0.009	0.008	0.001	0.034 [*]	0.011	0.022 [*]	0.054 [*]	0.008	0.071 [*]	0.017 [*]
FF	RF	0.118 [*]	0.124 [*]	0.106 [*]	0.052 [*]	0.176 [*]	0.038 [*]	0.187 [*]	0.079 [*]	0.101 [*]	-0.131 [*]	0.177 [*]	0.159 [*]
	TF	-0.004	0.011	-0.103 [*]	-0.050 [*]	-0.034 [*]	-0.049 [*]	-0.042 [*]	-0.039 [*]	-0.018	-0.306 [*]	0.015	-0.007
RF	TF	-0.121 [*]	-0.114 [*]	-0.209 [*]	-0.103 [*]	-0.209 [*]	-0.086 [*]	-0.229 [*]	-0.117 [*]	-0.119 [*]	-0.175 [*]	-0.162 [*]	-0.166 [*]

^{*} The mean difference is significant at <0.05 level.

The result of the MANOVA indicated that the width, gray scale and radian information are effective features for discriminant analysis.

OR, FF, RF and TF signatures were classified in the discriminant analysis. No overlaps occurred between OR and RF signatures in 12 groups, between OR and Non-OR signatures in eight groups, or between RF and other Non-OR signatures in four groups. Overlaps always existed between FF and TF signatures. However, determining whether one signature was original or not was more important than the manner in which the signature was written in. The discriminant analysis between OR and Non-OR signatures showed high scores in the cross-validation rate with a mean of 95.8%. The

high cross-validation rate suggested that the width, gray scale and radian obtained are useful to distinguish OR from Non-OR signatures.

The established quantitative feature extraction and statistical analysis methods eliminated the subjective aspect, and helped the scientific analysis of forensic handwriting examination. It is suggested that extraction and analysis of the width, gray scale and radian combined with the stroke order in Chinese signatures is reasonable. Meanwhile, forensic handwriting examination using quantitative features extraction and statistical analysis methods in this research could be performed with a satisfactory discriminant analysis result.

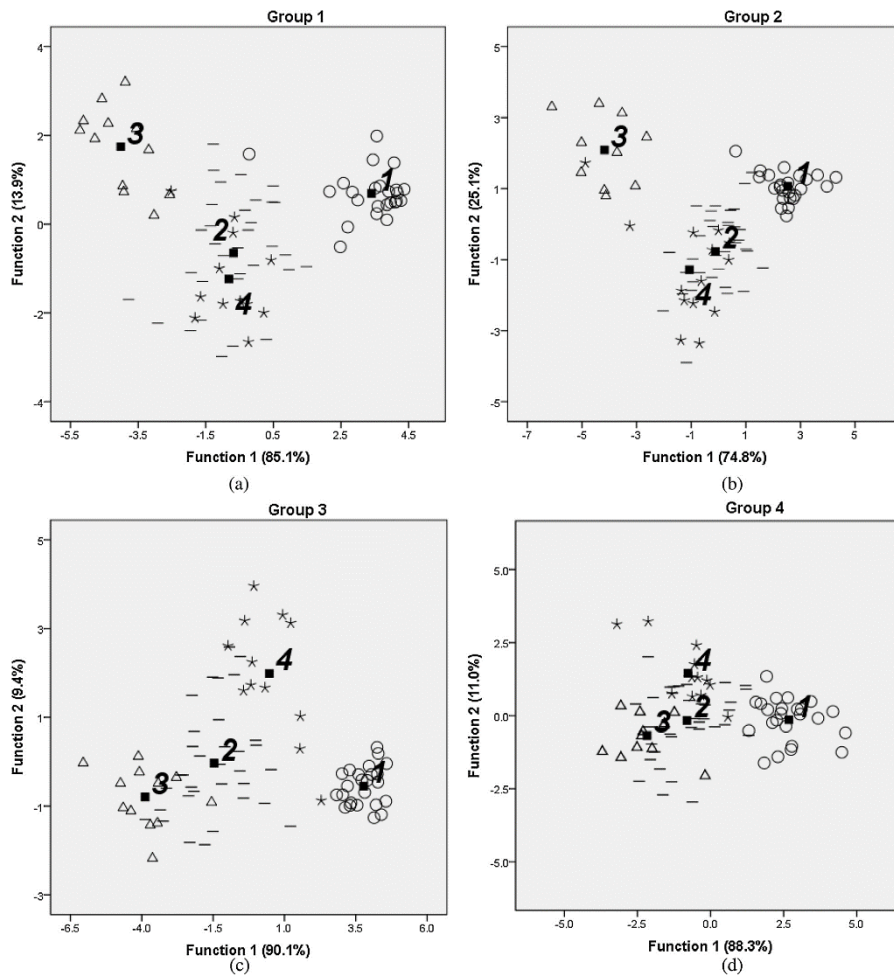


Fig. 5. Results of discriminant analysis performed on width, gray scale and radian data of 12 groups. Percentages given in parenthesis are the rates of explained by the corresponding functions (1:OR, 2:FF, 3:RF, 4:TF).

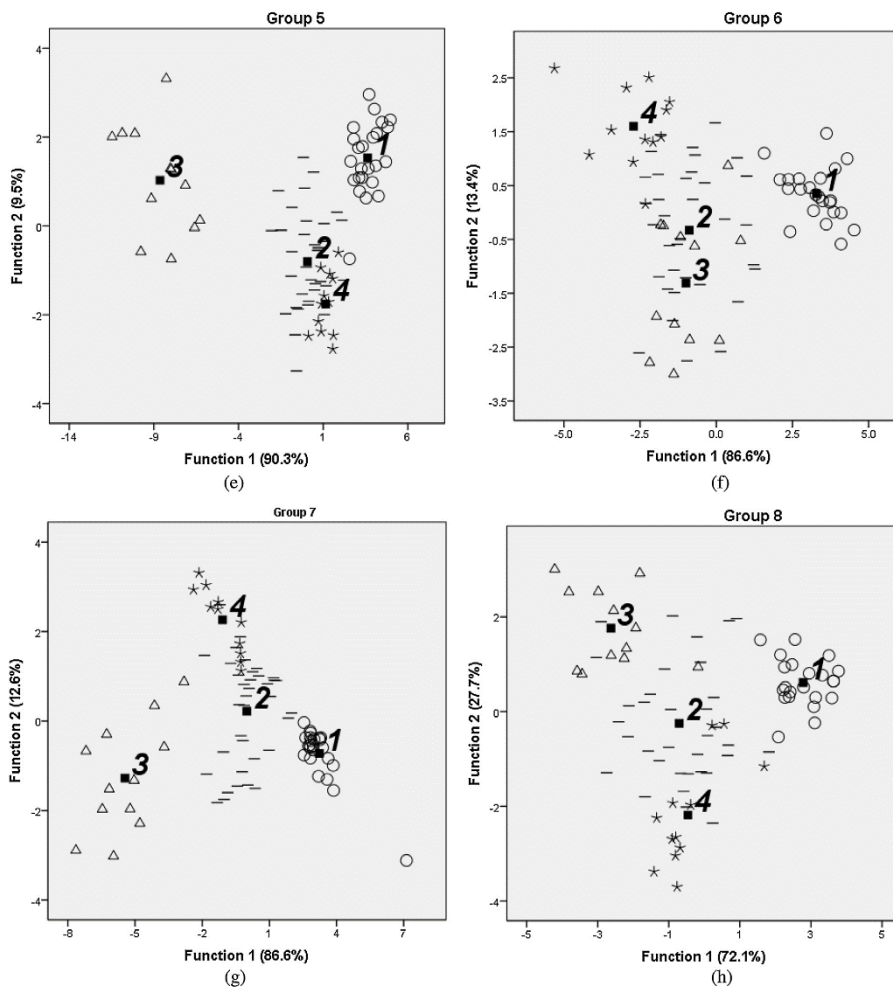


Fig. 5. (Continued).

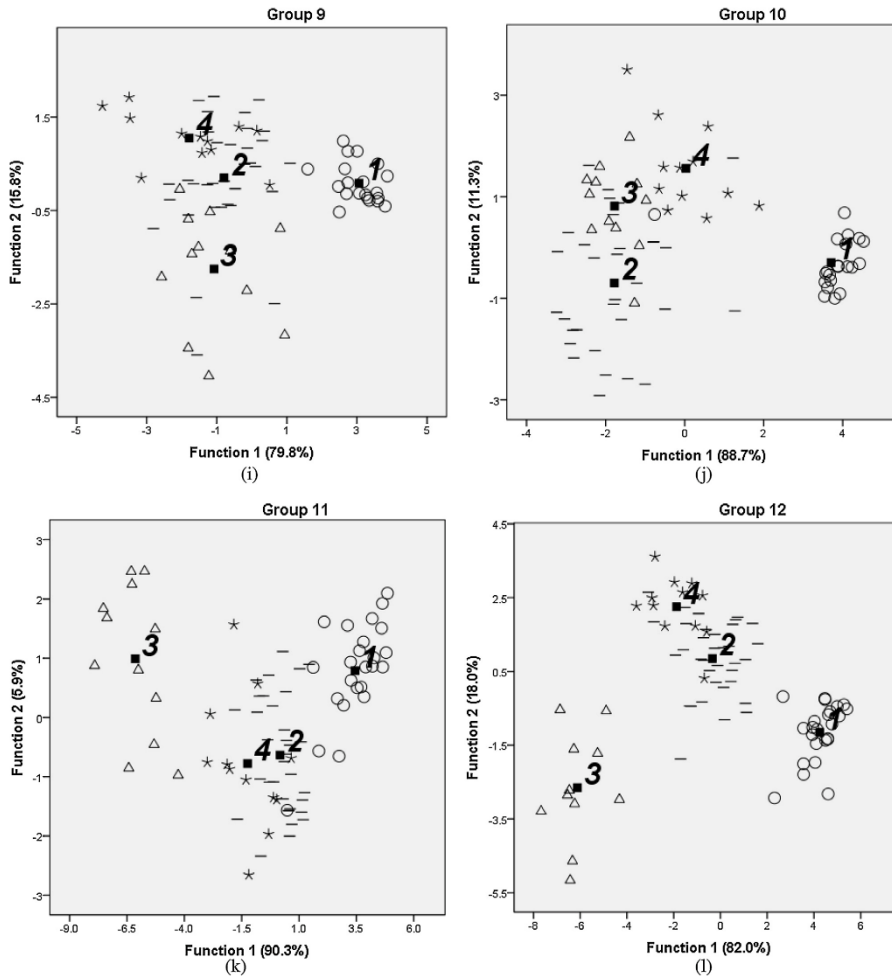


Fig. 5. (Continued).

Table 7

Discriminant analysis based on the width, gray scale and radian data: cross-validation between OR and Non-OR in each group.

Group	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
CV. ^a	97.6%	97.6%	95.1%	96.4%	94%	97.6%	95.2%	95%	97.4%	93.9%	90.5%	98.8%

^a C.V., cross-validation.

It is assumed that quantitative features extractions and statistical analysis methods should be generalized in forensic handwriting examination in addition to signature handwriting. Consequently, it would be interesting to extend the present research to a larger population with more handwriting and statistical analysis methods.

Acknowledgement

This study was supported by grants from the National Key Technology Research & Development Program of the Ministry of Science and Technology of People's Republic of China (2012BAK16B05) and Ministry of Finance, PR China (GY2015G-6).

References

- [1] D. Ellen, *The Scientific Examination of Documents: Methods and Techniques*, Ellis Horwood Limited, West Sussex, 1989.
- [2] R.A. Huber, A.M. Headrick, *Handwriting Identification: Facts and Fundamentals*, CRC Press, Boca Raton, 1999.
- [3] A.A. Moenssens, Handwriting identification evidence in the post-Daubert world, *U.M.K.C.L. Rev.* 66 (1997) 251–310.
- [4] J.J. Koehler, The coming paradigm shift in forensic identification science, *Science* 309 (2005) 892–895.
- [5] Bird Carolyne, Found Bryan, Rogers Doug, Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors, *J. Forensic Sci.* 55 (2010) 1291–1295.
- [6] R. Marquis, M. Schmittbuhl, W. Mazzella, F. Taroni, Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of the capital character o, *Forensic Sci. Int.* 150 (2005) 23–32.
- [7] R. Marquis, F. Taroni, S. Bozza, M. Schmittbuhl, Quantitative characterization of morphological polymorphism of handwritten characters loops, *Forensic Sci. Int.* 164 (2006) 211–220.
- [8] R. Marquis, S. Bozza, M. Schmittbuhl, F. Taroni, Quantitative assessment of handwriting evidence: the value of the shape of the letter A, *J. Forensic Document Exam.* 21 (2011) 17–22.
- [9] A. Stephen Ling, Preliminary investigation into handwriting examination by multiple measurements of letters and spacing, *J. Forensic Sci.* 126 (2002) 145–149.
- [10] B. Found, D. Rogers, R. Schmittat, A computer program to compare the spatial elements of handwriting, *Forensic Sci. Int.* 69 (1994) 195–203.
- [11] Marquis Raymond, Bozza Silvia, Schmittbuhl Matthieu, Taroni Franco, Handwriting evidence evaluation based on the shape of characters: application of multi-variate likelihood ratios, *J. Forensic Sci.* 56 (51) (2011) s238–s242.
- [12] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiery, S. Bozza, Bayes factor for investigative assessment of selected handwriting features, *Forensic Sci. Int.* 242 (2014) 266–273.
- [13] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiery, S. Bozza, The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination, *Forensic Sci. Int.* 214 (2012) 189–194.
- [14] R. Marquis, S. Bozza, M. Schmittbuhl, F. Taroni, Handwriting evidence evaluation based on the shape of characters: application of multi-variate likelihood ratios, *J. Forensic Sci.* 56 (2011) S238–S242.
- [15] Bozza Silvia, Taroni Franco, Marquis Raymond, Schmittbuhl Matthieu, Probabilistic evaluation of handwriting evidence: likelihood ratio for authorship, *Appl. Stat.* 57 (Part 3) (2008) 329–341.
- [16] Amanda B. Hepler, Christopher P. Saunders, Linda J. Davis, JoAnn Buscaglia, Score-based likelihood ratios for handwriting evidence, *Forensic Sci. Int.* 219 (2012) 129–140.
- [17] Srikanta Pal, Michael Blumenstein, Umapada Pal, Non-English and Non-Latin Signature Verification Systems: a survey, *Proc. 1st Int. Workshop Automated Forensic Handwriting Anal. (AFHA)* (2011) 1–5.
- [18] Michael J. Saks, Jonathan J. Koehler, The coming paradigm shift in forensic identification science, *Science* 309 (2005) 892.
- [19] J. Michael, Saks Forensic identification: from a faith-based science to a scientific science, *Forensic Sci. Int.* 201 (2010) 14–17.
- [20] D. Meuwly, Forensic individualisation from biometric data, *Sci. Just.* 46 (2006) 205–213.
- [21] Y. Kato, M. Yasubara, Recovery of drawing order from single-stroke handwriting images, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (9) (2000) 938–949.
- [22] W. Tian, J. Lv, A different approach to off-line signature verification using the optimal DTW algorithm, in: 2012 International Conference on Computer Science and Service System, 2012, 18–21.
- [23] Y. Qiao, X. Wang, C. Xu, Learning Mahalanobis distance for DTW based online signature verification, in: Proceedings of the IEEE International Conference on Information and Automation, Shenzhen, China, June, (2011), pp. 333–338.
- [24] D.-S. Huang, X.-P. Zhang, G.-B. Huang, Improved DTW algorithm for online signature verification based on writing forces, in: Advances in Intelligent Computing, Springer Express, 2005, pp. 631–640.
- [25] A. Thiery, R. Marquis, I. Montani, Statistical evaluation of the influence of writing postures on on-line signatures. Study of the impact of time, *Forensic Sci. Int.* 230 (2013) 107–116.



Assessment of signature handwriting evidence via score-based likelihood ratio based on comparative measurement of relevant dynamic features



Xiao-hong Chen^{a,b,*}, Christophe Champod^b, Xu Yang^a, Shao-pei Shi^a, Yi-wen Luo^a, Nan Wang^a, Ya-chen Wang^a, Qi-meng Lu^a

^a Institute of Forensic Science, Ministry of Justice, 1347, West Guangfu Road, Shanghai 200063, PR China

^b School of Criminal Justice, University of Lausanne, 1015 Lausanne-Dorigny, Switzerland

ARTICLE INFO

Article history:

Received 22 May 2017

Received in revised form 10 November 2017

Accepted 13 November 2017

Available online 21 November 2017

Keywords:

Signature

Comparative measurement

Relevant dynamic feature

Offline handwriting

Evidence evaluation

ABSTRACT

This paper extends on previous research on the extraction and statistical analysis on relevant dynamic features (width, grayscale and radian combined with writing sequence information) in forensic handwriting examinations. In this paper, a larger signature database was gathered, including genuine signatures, freehand imitation signatures, random forgeries and tracing imitation signatures, which are often encountered in casework. After applying Principle Component Analysis (PCA) of the variables describing the proximity between specimens, a two-dimensional kernel density estimation was used to describe the variability of within-genuine comparisons and genuine–forgery comparisons. We show that the overlap between the within-genuine comparisons and the genuine–forgery comparisons depends on the imitated writer and on the forger as well. Then, in order to simulate casework conditions, cases were simulated by random sampling based on the collected signature dataset. Three-dimensional normal density estimation was used to estimate the numerator and denominator probability distribution used to compute a likelihood ratio (LR). The comparisons between the performance of the systems in SigComp2011 (based on static features) and the method presented in this paper (based on relevant dynamic features) showed that relevant dynamic features are better than static features in terms of accuracy, false acceptance rate, false rejection rate and calibration of likelihood ratios.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Despite the advances of digitalization and the move towards paperless offices, forensic handwriting examinations involving signatures are still common in cases. For instance, 1017 forensic handwriting cases were examined in our institute, in 2014: 963 cases referred to signatures, comprising about 95% in all submitted handwriting cases. Such a high percentage of signature handwriting submissions continued in 2015 and 2016.

While comparing questioned and reference handwriting samples, forensic handwriting examiners (FHEs) observe and evaluate similarities and differences. Then they provide an opinion as to the authorship of the questioned handwriting based on their training and experience [1]. The process of

comparison and follow-up assessment of the observations are highly dependent on the experts. In the National Research Council's report to the US Congress, the committee said that the scientific basis for handwriting comparisons and assessment in forensic handwriting examinations should be strengthened [2]. This paper focuses on operator independent techniques. In that area, the weighing of the observed similarities and differences in handwriting examination is not straightforward and has not been submitted to a lot of systematic research.

In order to evaluate the authorship based on the similarities and differences observed on questioned and reference handwriting, previous studies on automatically extracted features [3–5] have already contributed to help FHEs to quantitatively measure the features of handwriting and assess the value of handwriting evidence. The application of a likelihood ratio framework for handwriting evidence evaluation received particular attention [6–10]. In this paper, we will build on this framework. Previous research [3,4] was rather limited in terms of features used (loops), was focused essentially on handwriting and with a limited set of

* Corresponding author at: Institute of Forensic Science, Ministry of Justice, 1347, West Guangfu Road, Shanghai 200063, PR China.
E-mail address: chenxh@ssfjd.cn (X.-h. Chen).

imitations used as genuine forgeries to test the system. This contribution aims at studying signatures, on a large corpus involving genuine and skilled imitated signatures.

Handwriting, including the production of signatures, is the effect of a dynamic behaviour. This behaviour materialises on paper in the form of static traces that are submitted to FHEs. FHEs then reconstruct the dynamic writing sequence based on the analysis of the traced images. It means that the operator-independent features used to characterise the handwriting should capture the dynamic nature of the behaviour and not rest only on static measures such as relative proportions, sizes and shapes of letters. Most of the past research focused on the static features, such as contour, gradient, direction of slopes, etc. The sequence of handwriting was neglected. The writing sequence is a new measured feature in forensic science. This paper will take advantage dynamic time wrapping techniques to capture all features while maintaining the writing sequence.

In this paper, we follow the process of features detection and analysis in forensic handwriting examination described in Ref. [5]. In summary, it takes the following steps: following image capture of signatures, a threshold is applied to the image to obtain binarized images. The skeleton and the signature edges are extracted by digital image processing. The skeletonized signatures are submitted to a programme allowing the extraction of the writing sequence. The width, grayscale and radian were automatically extracted from the writing sequence. Thus, the features of width, grayscale and radian combined with writing sequence are automatically extracted. Next, a dynamic time warping method is applied to cope with the difference writing speeds. The pairwise correlation coefficient was used to characterize and express the similarities between signatures.

The extracted features, namely width, grayscale and radian, are fully described in Ref. [5]. They are measured at every pixel following the skeleton of the whole signature. The skeleton is constructed to reflect the writing sequence of the signature, acquired at 400 dpi.

They are qualified as “dynamic features” because they are extracted accounting for the writing order. They are not extracted at the time of capture, but acquired after the writing act on the images itself. Because these features are different from dynamic features extracted from on-line handwriting, but still reflect the writing sequence, we called them “relevant dynamic features”.

This paper presents three major improvements compared to previous work in [5]: (1) the signature database was enlarged to

twenty groups and 1654 signatures; (2) Probability density distributions were estimated to show the variability of within-genuine comparisons and genuine-forgery comparisons; (3) Likelihood ratios (LRs) were calculated based on the relevant dynamic features.

Finally, the comparisons between the performance of the systems in Signature Verification Competition for Online and Offline Skilled Forgeries [10] (SigComp2011, based on static features) and the system presented in this paper (based on relevant dynamic features) assess the performance of the methodology presented in this paper.

2. Material and method

2.1. Signature database

A signature database (20 groups, 1654 signatures) was acquired based on a previous signature database, including genuine signatures, freehand imitation signatures, tracing imitation signatures and random forgeries without any model.

That was done to reflect situations often encountered in casework. Chinese signatures were written by 20 volunteers using a ballpoint pen with black ink on A4 paper (for genuine signatures, random forgeries and freehand imitation forgeries) printed with 12 squares and 195 mm–271 mm, and highly transparent paper (for tracing imitation forgeries), with the signatories sitting while signing. Twenty volunteers who could produce skilled imitation forgeries were also recruited. The skilled imitations were produced by trained forensic document examiners who have developed skills in producing imitations. The Chinese signature database was composed of 1654 signatures of 20 groups produced by 20 groups of volunteers (each composed of one writer and a set of forgers); every group contained 20–24 genuine signatures (denoted as GE), 30–36 freehand imitation forgeries with a genuine model by three volunteers (denoted as FF), 10–12 random forgeries without any model by one volunteer (denoted as RF) and 10–12 tracing imitation forgeries by one volunteer (denoted as TF). For the production of forgeries, one genuine signature was chosen as the model at random. The freehand imitation forgeries, tracing imitation forgeries and forgeries without any model were all called “forgeries” in this paper.

Our signature database is summarized in Table 1. We have grouped the forgeries in two categories:

Table 1
Chinese signature database.

Group ID	Genuine signature (GE)	Freehand imitation forgery (FF)	Tracing imitation forgery (TF)	Random forgery (RF)
G1	24	35	12	12
G2	24	36	12	12
G3	24	34	12	12
G4	24	36	12	12
G5	24	35	12	12
G6	23	36	12	12
G7	24	36	12	12
G8	22	34	12	12
G9	21	34	12	12
G10	23	35	12	12
G11	24	34	12	12
G12	24	35	12	12
G13	24	36	12	12
G14	24	36	12	12
G15	22	35	12	12
G16	24	36	12	12
G17	24	36	12	12
G18	24	33	12	12
G19	24	36	12	12
G20	24	35	12	12

(1) The forgeries with model consist of the forgeries obtained when a genuine model was available to the forger.

(2) The forgeries without model are the forgeries obtained when the forger tried to produce the signature of a given named person with any model available. In China, most individuals would sign using a legible handwritten-like signature of their name [11]. Hence forgeries without model can be challenging for a handwriting expert.

2.2. Comparison feature extraction

As described in Ref. [5], dynamic time warping was used to make the corresponding stroke aligned and matched for correlation analysis. The correlation coefficient between signatures was calculated. For instance, a number m of genuine signatures was used as the reference signatures; a number n of unknown signatures was used as the questioned signatures. For each genuine signature, $m-1$ measurements of width, grayscale and radian between reference signatures, respectively; for each questioned signature, m measurements of correlation with regards to width, grayscale and radian between questioned signature and reference signatures, respectively as well.

Hence a comparison between one questioned signature (Q) and 3 specimens (S1–S3), we obtain the following matrix:

	Correlation on width	Correlation on grayscale	Correlation on radian
Q – S1	CW _{Q-S1}	CG _{Q-S1}	CR _{Q-S1}
Q – S2	CW _{Q-S2}	CG _{Q-S2}	CR _{Q-S2}
Q – S3	CW _{Q-S3}	CG _{Q-S3}	CR _{Q-S3}

In this paper, the output of a comparison between a questioned signature (genuine or forged) and a set of genuine specimens will be obtained by either using the mean or the maximum of the correlations obtained for each variable. Both mean and maximum values are used respectively.

2.3. Descriptive representations of the within genuine and genuine–forgery variability

For each group, all pairwise comparisons between signatures will be carried out. Hence all comparisons between genuine signatures and all comparisons between genuine and forgeries will be made. To graphically represent these results, a reduction from three dimensions to two using principal component analysis (PCA) has been carried out, followed by a two-dimensional kernel density estimation (KDE). These statistical analyses were carried out in R (version 3.3.0 [12]) using the PCA implementation of MASS package and the 2D KDE function (kde2d) of the same package [13]. Contour plots will represent the variability among genuines and the genuine versus forgery variability.

The purpose of the PCA – 2D KDE process is solely illustrative. It allows a graphical representation per writer of their within-genuine variability and their genuine against forgery variability. The graphical representations are obtained (with the package ggplot2 [14]) for the max values and the mean values on width, grayscale and radian features.

2.4. Score-based likelihood ratio

The measure of the strength of handwriting evidence estimation using subsampling via score-based likelihood ratio was illustrated in Ref. [15]. The total evidence is consisting of three parts:

E_U = Questioned signature.

E_S = Signatures known to have been written by the person of interest (POI) (hence genuine) leading to a template for POI.

E_A = Signatures obtained from writers other than the POI (hence forgeries).

For a given questioned, compared to a set of known specimens (E_S), the findings are represented by a set of correlation (mean or max) scores (Cw, Cg, Cr) denoted as $s(E_U, E_S)$.

The estimated score-based likelihood ratio \hat{SLR} (Eq. (1)) is obtained by the ratio between the probability density observed for $s(E_U, E_S)$ given two alternative propositions H_p and H_d . H_p stands for the proposition that the questioned signature share common authorship with the signatures from the POI. The signature is then genuine. H_d stands for the proposition that the questioned signature is not from the POI's hand but is a forgery. To obtain the relevant probability densities, the distributions of the scores under both propositions are required. It will be done by conducting comparisons between signatures from E_S (under H_p) and signatures from E_A against E_S (under H_d). The detailed process for computing these background densities is described below.

$$\hat{SLR} = \frac{\hat{g}[s(E_U, E_S)|H_p]}{\hat{g}[s(E_U, E_S)|H_d]} \quad (1)$$

In order to simulate operational conditions, simulated cases were generated based on the signatures dataset. For instance for a given group (writer) composed of m genuine signatures and n forgery signatures, we can generate forensic cases by taking one questioned signature (denoted as E_U) and five reference signatures (denoted as E_S) from the genuine signatures, a set of $\{s(E_U, E_S)\}$ is then obtained it represent one forensic case. The number of 5 references is considered to be a minimum number of references that should be available in an operational case.

Two conditions are possible depending on the proposition considered (H_p or H_d):

- (A) If the questioned signature came from the genuine signatures (under H_p), five reference signatures were selected from the remaining $(m-1)$ genuine signatures, given a number of possible combinations of ${}^c_{m-1}^5$, $m \cdot {}^c_{m-1}^5$ simulation cases were then generated. The remaining $(m-6)$ genuine signatures were used to generate the $(m-6) \cdot {}^c_{m-7}^5$ simulations to obtain the scores associated with cases of known same source (genuine), see Eq. (2); the other $m-6$ genuine signatures and n forgery signatures were used to generate $n \cdot {}^c_{m-6}^5$ simulation cases to obtain the scores associated with cases of known different source (forgeries).
- (B) If the questioned signature came from the forgery signatures, five reference signatures were selected from the m genuine signatures, possible combination denoted as ${}^c_m^5 \cdot n \cdot {}^c_m^5$ simulation cases were generated. The remaining $m-5$ genuine signatures were used to generate the $(m-5) \cdot {}^c_{m-6}^5$ simulation cases to get the scores associated with cases of same source (genuine); the rest $m-6$ genuine signatures and n forgery signatures were used to generate $(n-1) \cdot {}^c_{m-5}^5$ simulation cases to get the scores associated with cases originating from different sources (forgery).

For each specific simulation case, the signatures used for the simulation case were taken out from the group. The data of scores associated with genuine sources was denoted as $\hat{g}[\cdot|H_p, I]$, see function (2); the data of scores associated with genuine to forgery comparisons was denoted as $\hat{g}[\cdot|H_d, I]$, see Eq. (3). The estimated score-based likelihood ratio \hat{SLR} for $E = (E_U, E_S, I)$ was evaluated at (E_U, E_S) , see Eq. (4).

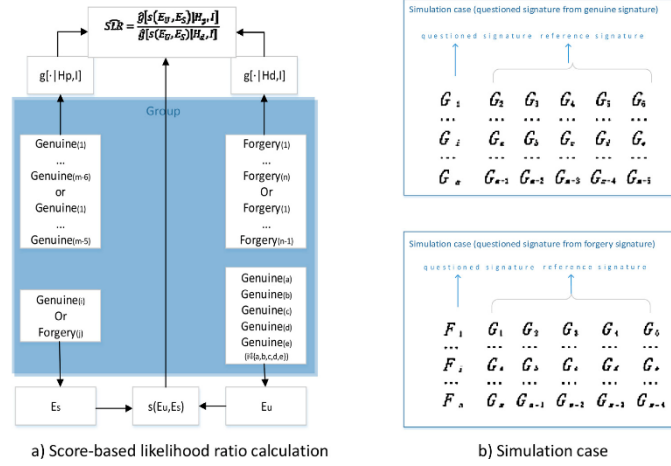


Fig. 1. The process of score-based likelihood ratio calculation. ($i, a, b, c, d, e = 1, 2, \dots, m; i \notin \{a, b, c, d, e\}; j = 1, 2, \dots, n; a, b, c, d, e = 1, 2, \dots, m$).

The probability densities (under H_p and H_d) are modelled using a 3-dimensional normal distribution. The parameters of the distributions are obtained by maximum likelihood estimation from the score data for each group. The process of score-based likelihood ratio calculation see Fig. 1.

$$g[\cdot|H_p, I] = \{S(G_i, \{G_a, G_b, G_c, G_d, G_e\}) : i, a, b, c, d, e = 1, 2, \dots, m; i \notin \{a, b, c, d, e\}\}$$

$$g[\cdot|H_d, I] = \{S(F_j, \{G_a, G_b, G_c, G_d, G_e\}) : j = 1, 2, \dots, n; a, b, c, d, e = 1, 2, \dots, m\}$$

$$SLR = \frac{g[s(E_p, E_s)|H_p, I]}{g[s(E_p, E_s)|H_d, I]} \quad (4)$$

2.5. Performance evaluation

Several measures for performance of LR-based system have been presented in the literature [16–19]. In this research, a validation toolkit (Beta v1.06) available from Ref. [20] has been used to measure the performance and calibration of the present system. Given the dependence on the writer, analysis has been carried out per writer (a group in our data collection comprises of genuine and forged signatures of a given individual). The toolbox calculates the C_{lr} (log likelihood ratio cost), C_{lr}^{ad} , C_{lr}^{min} , EER (equal error rate), RMED (Rate of misleading Evidence in favour of H_d) and RMEP (Rate of misleading Evidence in favour of H_p). The performance graphical representations include Tippett plots, DET (Detection error trade-off) plots, ECE (Empirical cross-entropy) plots and APE (applied probability of error) plots. Accuracy is measured by the C_{lr} and ECE; discriminating power is measured by the C_{lr}^{min} and the EER; calibration is measured by the C_{lr}^{ad} and the difference between ECE and ECE-after-PAV [21].

The above computation involves a very large number of cases and we explored if performance metrics could be estimated with a lower number of cases. In order to reduce the computing time,

three percentages (50%, 30% and 10%) of the data were randomly selected and the metrics were bootstrapped (1000 bootstrap samples).

At last, the present system (based on relevant dynamic features) has been compared to seven systems (based on static features) that took part in the Chinese offline signature competition in Signature Verification Competition for Online and Offline Skilled Forgeries (SigComp2011). These comparisons were performed based on the following performance characteristics: accuracy, false accept rate (FAR), false reject rate (FRR), C_{lr}^{min} and C_{lr} .

3. Results

Two-dimensional contour plot per group after PCA – 2D KDE based on the mean value of comparative measurement led to the observations presented Fig. 2. In Group 16 and Group 19, the overlaps between genuine signatures and forgeries is obvious; in Group 4, Group 11, Group 15 and Group 18, slight overlaps between genuine signatures and forgeries; in the other groups, no overlaps between genuine signatures and forgeries. These observations were made both when using the mean or using the max value. Both of them show the similar results. Fig. 2 shows only the results from the mean values. As expected the forgeries obtained with a model were closer to the genuine signatures than the forgeries obtained without a model. We note that each group has a different behaviour and that translates the fact that signatures of some individual are easier to forge than signatures of others. Indeed, the distances between genuines and forgeries vary significantly between groups. This is not a novel observation as such (see Ref. [22]), but it transpires clearly from the present results.

Hence, for evidence evaluation, and specifically to obtain the background distributions required to estimate the score-based likelihood ratios, the scores should be from the designated writer and not obtained by merging all writers together.

The performance evaluation, based on mean value of comparative measurement results (EER, C_{lr}^{ad} , C_{lr}^{min} , C_{lr} , RMEP and RMED) for all the twenty groups, are presented in Table 2. In line with the overlaps observed on the PCA plots, the performance evaluation based on mean value of comparative measurement for Group 12

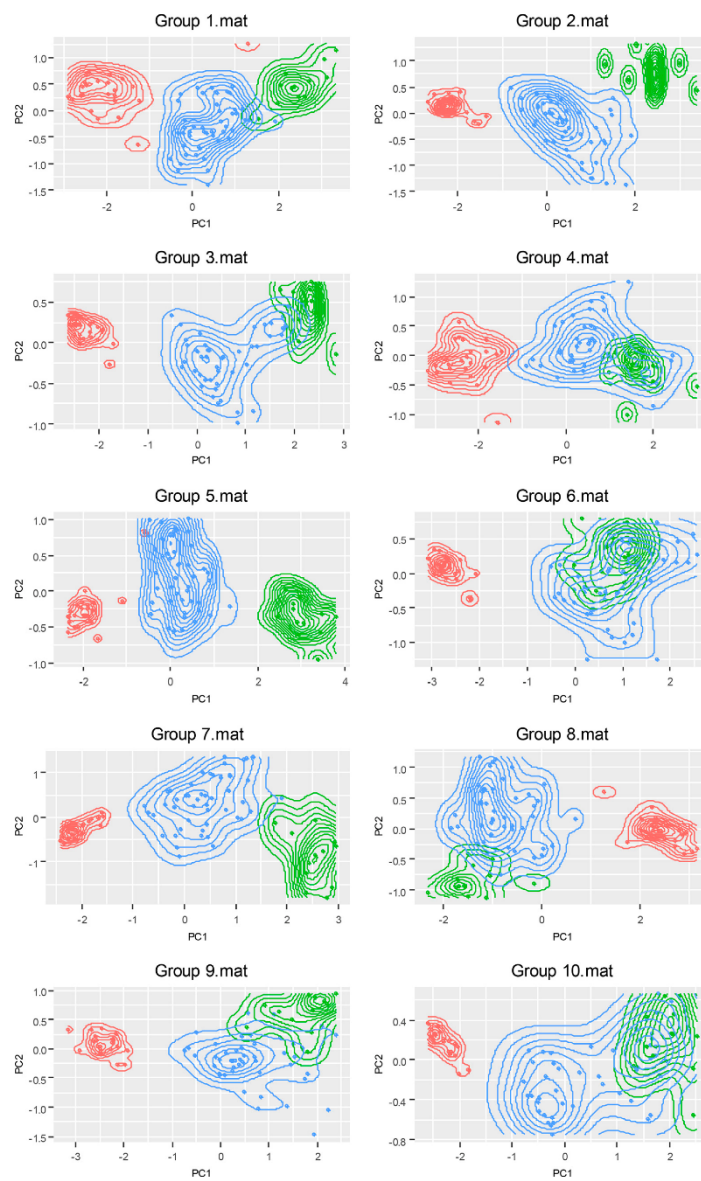


Fig. 2. Two-dimensional density distribution plot after Principle Component Analysis (PCA) for twenty groups based on the mean value of comparative measurement of relevant dynamic features: red lines and dots represent data for genuine signatures; green lines and dots represent data for random forgeries (without model); blue lines and dots represent data for imitation forgeries (with model).

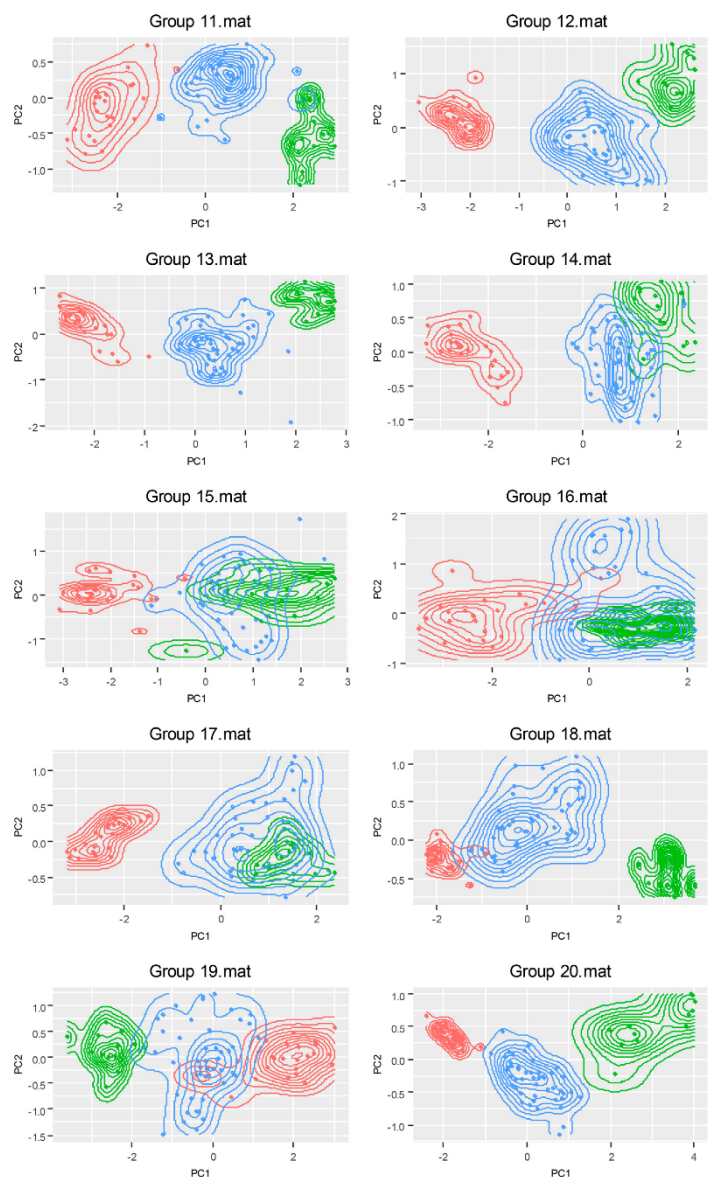


Fig. 2. (Continued)

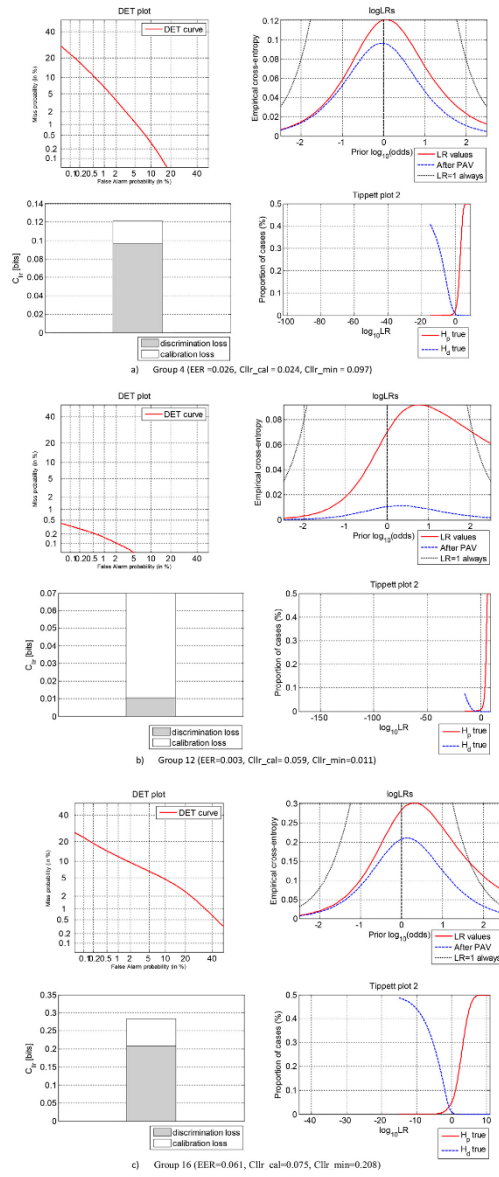


Fig. 3. Performance evaluation for Group 12 (no overlap), Group 4 (small overlap) and Group 16 (large overlap): DET plot, ECE plot, C_{fr} plot and Tippett plot.

Table 2
Results of performance evaluation (based on the mean value of comparative measurement) for signatures of twenty groups.

Group	EER	C_{lr_cal}	C_{lr}^{min}	C_{lr}	RMED	RMEP
1	0.087	0.389	0.307	0.696	0.110	0.046
2	0.010	0.067	0.036	0.103	0.040	0.000
3	0.006	0.135	0.023	0.158	0.037	0.000
4	0.026	0.025	0.094	0.120	0.043	0.016
5	0.044	0.592	0.114	0.706	0.066	0.000
6	0.002	0.051	0.007	0.058	0.023	0.000
7	0.009	0.053	0.032	0.084	0.033	0.002
8	0.003	0.015	0.012	0.027	0.010	0.001
9	0.008	0.040	0.028	0.068	0.021	0.001
10	0.002	0.276	0.007	0.283	0.054	0.000
11	0.053	0.111	0.193	0.304	0.099	0.019
12	0.002	0.056	0.009	0.065	0.013	0.000
13	0.019	0.154	0.070	0.223	0.067	0.001
14	0.002	0.013	0.009	0.022	0.010	0.000
15	0.055	0.111	0.181	0.292	0.092	0.024
16	0.061	0.076	0.211	0.287	0.092	0.024
17	0.006	0.016	0.022	0.038	0.014	0.002
18	0.032	0.050	0.117	0.168	0.053	0.021
19	0.102	0.185	0.328	0.513	0.132	0.062
20	0.013	0.041	0.046	0.087	0.032	0.003

(no overlap), Group 4 (small overlap) and Group 16 (large overlap) are presented in Fig. 3. Given the variability between groups, it was required to conduct the performance evaluation per group.

In order to show the performance based on the max value and mean value of comparative measurement respectively, the EER and C_{lr}^{cal} were used (Fig. 4). There is not a marked difference in EER between the mean value (mean=0.027) and the max value

Result of different percentage of data for bootstrap

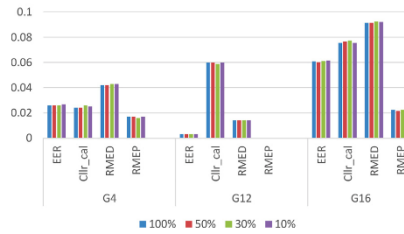


Fig. 5. Result of different percentage of data for bootstrap simulation based on the mean value of comparative measurement.

(mean = 0.028), but the max value (mean = 0.062) was better than the mean value (mean = 0.123) in the calibration metric C_{lr}^{cal} .

The mean value of EER, C_{lr}^{cal} , C_{lr}^{min} , C_{lr} , RMEP and RMED for when 100% and then only 50%, 30% or 10% of the data are used are reported in Fig. 5. The results show that adequate estimates can be obtained when the percentage of data for computing is significantly reduced.

Factors, such as a bad choice of databases, of statistical models, or limited quantity or quality of the signatures can lead to misleading likelihood ratios (meaning that they may provide support for the wrong propositions). Calibration is a way to mitigate this problem and ensure that the likelihood ratios can be

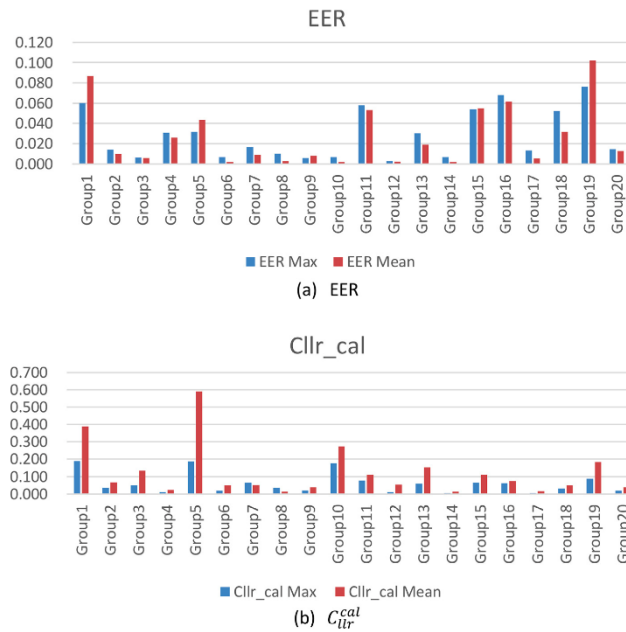


Fig. 4. Comparison between the results based on the max value and mean value of comparative measurement.

Table 3
Summary statistics of results calibrated SLR, RMED and RMEP.

Group	Log10 (Calibrated SLR) (H_p)			Log10 (Calibrated SLR) (H_d)			RMED	RMEP
	Min	Max	Mean	Min	Max	Mean		
G1	-26.613	9.868	2.892	-28.450	3.323	-3.046	0.110	0.046
G2	-8.141	4.599	2.553	-200.063	1.003	-24.152	0.040	0.000
G3	-12.854	7.815	4.371	-187.041	-0.466	-34.749	0.037	0.000
G4	-6.474	7.409	2.794	-85.601	3.243	-9.462	0.042	0.016
G5	-27.092	7.521	2.835	-323.306	1.151	-52.431	0.066	0.000
G6	-8.037	9.075	5.058	-153.999	1.230	-33.737	0.023	0.000
G7	-8.961	5.536	3.069	-323.306	2.367	-60.093	0.033	0.002
G8	-6.342	13.349	6.707	-179.221	4.379	-22.837	0.010	0.001
G9	-9.874	6.061	3.635	-100.954	1.212	-21.708	0.021	0.001
G10	-16.821	7.234	3.799	-323.306	-4.094	-127.283	0.054	0.000
G11	-10.386	8.076	2.217	-60.805	2.737	-9.678	0.099	0.019
G12	-19.103	7.098	4.837	-150.589	-2.039	-30.212	0.013	0.000
G13	-18.085	8.773	2.755	-320.284	1.405	-36.827	0.068	0.001
G14	-5.412	13.171	5.289	-81.183	0.868	-13.299	0.010	0.000
G15	-14.013	7.287	2.529	-81.058	3.488	-10.246	0.092	0.024
G16	-10.861	10.151	2.801	-31.769	2.556	-4.990	0.021	0.024
G17	-7.266	8.089	4.144	-119.133	4.104	-20.535	0.014	0.002
G18	-8.124	3.026	1.789	-265.024	2.595	-30.520	0.053	0.022
G19	-17.012	5.539	1.040	-93.679	1.709	-7.671	0.132	0.062
G20	-7.137	4.850	2.798	-323.306	2.092	-41.136	0.032	0.003

Table 4
Comparisons between the seven systems used in SigComp2011 (from ID = 1 to 7) and the method presented in this paper (ID = 8).

ID	Accuracy (%)	FRR	FAR	C_{fr}	C_{fr}^{min}
1	80.04	21.01	19.62	0.7577	0.6933
2	73.10	27.05	26.70	3.0627	0.7650
3	72.90	27.50	26.98	1.1252	0.7899
4	56.06	45.00	43.60	1.2605	0.8907
5	51.95	50.00	47.41	3.2225	0.9513
6	62.01	37.50	38.15	1.5736	0.9266
7	61.81	38.33	38.15	6.2270	0.9185
8	96.17	3.83	3.83	0.1629	0.1433

probabilistically interpreted as representing the evidential value of the comparison in a Bayesian evaluation framework [17,20]. The calibrated likelihood ratios obtained in this study are presented in Table 3 (in $\log_{10}(\text{LR})$). It can be seen, as expected following the descriptive analysis in Section 2.3, that the expected LRs for one individual (a group) can be higher than for another. Overall for cases when H_p is true, the overall mean strength associated with features analysed are $\log_{10}(\text{LR})$ between 1 and 7. Cases of forgeries (H_d) lead to stronger evidence towards that proposition. However, we have to balance these LRs against the rates of misleading evidence (especially here RMED) that can be high, up to 13%. So, the system is guiding with strength on the appropriate propositions but with an appreciable probability to mislead. Such a system should be used with an appropriate account and declaration of its limitations. The rates of misleading evidence (RMEP and RMED) should then be reported alongside any computed likelihood ratio.

Compared to the performance of the systems used for Sigcom2011, the performance of the solution provided in this paper is much more efficient as shown in Table 4. It shows the improvements offered by the presented method in terms of EER and calibration. Compared with the performance obtained with the systems in SigComp2011 (based on static features), the method presented in this paper (based on relevant dynamic features), is much more discriminative and suffers from low rates of misleading evidence.

4. Conclusion

This research focused only on features that can be automatically extracted from signatures left on paper and their ability to

statistically guide towards their status as being a genuine or a forgery. We have shown that relevant dynamic features (width, grayscale and radian features measured as a function of the writing sequence derived from static images) allow to discriminate between genuine and forged signatures. These features are important additions to traditional features measured statistically on the images. Our data confirmed that some signatures are easier to forge than others, but reasonable discrimination can be achieved.

The LR-based approach adopted in this paper, alongside with measured indicators of performance (through calibration), proved to be efficient and robust at the task of detecting genuine from forgery. We have reported on the forensic error rates (expressed by the rates of misleading evidence RMED and RMEP). As expected, these rates were not equal to zero. As it stands, the system is more efficient at guiding towards forgeries than at helping confirming genuine signatures. However, the proposed system outperforms the other systems that entered the Sigcom2011 competition.

This research is a further step in strengthening the scientific basis of signature comparison and assessing signature examinations based on operator-independent features. In the future, the signature database will be enlarged; additional features will be added with the hope to improve the forensic performance of the system and especially reducing the rates of misleading evidence.

Acknowledgement

This study was supported by grants from the National Natural Science Foundation of China (61605132), Ministry of Finance, PR China (GY2015G-6) and the National Key Research and Development Program of China (2016YFC0800705).

References

- [1] F. Lee, Review of handwriting identification evidence in the post-Daubert world, *J. Am. Soc. Quest. Doc. Exam.* 1 (1998) 67–68.
- [2] National Research Council of the National Academies, *Strengthening Forensic Science in the United States: A Path Forward*, The National Academies Press, Washington, DC, 2009.
- [3] R. Marquis, M. Schmittbuhl, W. Mazzella, F. Taroni, Quantification of the shape of handwritten characters: a step to objective discrimination between writers based on the study of the capital character o, *Forensic Sci. Int.* 150 (2005) 23–32.

- [4] R. Marquis, F. Taroni, S. Bozza, M. Schmittbuhl, Quantitative characterization of morphological polymorphism of handwritten characters loops, *Forensic Sci. Int.* 164 (2006) 211–220.
- [5] X. Chen, Extraction and analysis of the width, gray scale and radian in Chinese signature handwriting, *Forensic Sci. Int.* 255 (2015) 123–132.
- [6] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiéry, S. Bozza, Bayes factor for investigative assessment of selected handwriting features, *Forensic Sci. Int.* 242 (2014) 266–273.
- [7] F. Taroni, R. Marquis, M. Schmittbuhl, A. Biedermann, A. Thiéry, S. Bozza, The use of the likelihood ratio for evaluative and investigative purposes in comparative forensic handwriting examination, *Forensic Sci. Int.* 214 (2012) 189–194.
- [8] A.B. Hepler, C.P. Saunders, L.J. Davis, J. Buscaglia, Score-based likelihood ratios for handwriting evidence, *Forensic Sci. Int.* 219 (2012) 129–140.
- [9] Y. Tang, S.N. Srihari, Likelihood ratio estimation in forensic identification using similarity and rarity, *Pattern Recognit.* 47 (2014) 945–958.
- [10] M. Liwicki, M. Imran 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), 2011 International Conference on Document Analysis and Recognition (2011) 1480–1484.
- [11] Jia Xiaoguang, Comparative research on English and Chinese signature handwriting examination, *J. Chin. People's Pub. Secur. Univ. (Sci. Technol.)* 12 (3) (2006) 13–17.
- [12] A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017. <http://www.R-project.org/>.
- [13] W.N. Venables, B.D. Ripley, *Modern Applied Statistics with S*, 4th edition, Springer, New York, 2002 ISBN 0-387-95457-0.
- [14] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York, 2009.
- [15] 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 (2012) 146–157.
- [16] N. Brünner, J. du-Preez, Application independent evaluation of speaker detection, *Comput. Speech Lang.* 20 (2006) 230–275.
- [17] J. Lucena-Molina, D. Ramos, J. Gonzalez-Rodriguez, Performance of likelihood ratios considering bounds on the probability of observing misleading evidence, *Law Probab. Risk* 14 (2015) 175–192.
- [18] D. Ramos, J. Gonzalez-Rodriguez, Reliable support: measuring calibration of likelihood ratios, *Forensic Sci. Int.* 230 (2013) 156–169.
- [19] A.F. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, Rhodes, Greece, September, Proc. Eurospeech '97, vol. 41997, pp. 1899–1903.
- [20] <https://sites.google.com/site/validationtoolbox/home>.
- [21] R. Haraksim, D. Ramos, D. Meuwly, C.E.H. Berger, Measuring coherence of computer-assisted likelihood ratio methods, *Forensic Sci. Int.* 249 (2015) 123–132.
- [22] T.N. Dewhurst, B. Found, D. Rogers, Are expert penmen better than lay people at producing simulations of a model signature? *Forensic Sci. Int.* 180 (2008) 50–53.