Received: 19 July 2021 | Revised: 9 November 2021 | Accepted: 17 November 2021

DOI: 10.1002/sam.11566

WILEY

RESEARCH ARTICLE

Handwriting identification using random forests and score-based likelihood ratios

Madeline Quinn Johnson | **Danica M. Ommen**

Center for Statistics and Applications in Forensic Evidence, Iowa State University, Ames, Iowa, USA

Correspondence
Danica M. Ommen, Department of Statistics, Iowa State University, Ames, IA, USA.
Email: dmommen@iastate.edu

Funding information
This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Abstract

Handwriting analysis is conducted by forensic document examiners who are able to visually recognize characteristics of writing to evaluate the evidence of writership. Recently, there have been incentives to investigate how to quantify the similarity between two written documents to support the conclusions drawn by experts. We use an automatic algorithm within the "handwriter" package in R, to decompose a handwritten sample into small graphical units of writing. These graphs are sorted into 40 exemplar groups or clusters. We hypothesize that the frequency with which a person contributes graphs to each cluster is characteristic of their handwriting. Given two questioned handwritten documents, we can then use the vectors of cluster frequencies to quantify the similarity between the two documents. We extract features from the difference between the vectors and combine them using a random forest. The output from the random forest is used as the similarity score to compare documents. We estimate the distributions of the similarity scores computed from multiple pairs of documents known to have been written by the same and by different persons, and use these estimated densities to obtain score-based likelihood ratios (SLRs) that rely on different assumptions. We find that the SLRs are able to indicate whether the similarity observed between two documents is more or less likely depending on writership.

KEYWORDS

handwriting analysis, machine learning, SLR

1 | BACKGROUND

Forensic handwriting analysis has traditionally been conducted by trained forensic examiners who rely on visual inspection to compare writing samples. The assumption that underlies forensic handwriting examination is that each person has a unique writing style that is developed over years and that may depend on cultural, demographic, and physiological factors [1]. After reviewing the evidence, forensic document examiners summarize their findings using categorical conclusions such as "identification," "strong probability," "probable," "indicate," and "indeterminable," among others [2]. The 2009 National Research Council's report "Strengthening Forensic Science in the United States" includes an outline of current practices and concludes that the scientific

basis of handwriting analysis must be strengthened [3]. In recent years, methods to quantify the similarity between two handwritten samples have been proposed, and software to implement those methods is now available. FLASH ID is a software tool developed by Sciometrics that uses the topology and "geometric features" in handwriting samples from a closed set of writers. FLASH ID then provides a ranked list of the writers in that set who are most likely to have written the questioned document [4]. Another approach proposed by Hepler et al. numerically evaluates the similarity in handwriting with score-based likelihood ratios (SLRs) [5].

Earlier work performed by researchers in the Center for Statistics and Applications in Forensic Evidence (CSAFE) focused on estimating the probability that two documents have the same source [6]. This is known as forensic identification of a common source. In this context, the goal is to determine whether the two documents were written by the same person, even if the identity of that person is unknown [7]. Furthermore, Crawford and collaborators considered only the closed set scenario, which requires all of the potential sources of the handwriting to be known [8].

We extend the scope of the research to the open set case, where the writer of the documents can be anyone in a defined population subgroup. More specifically, the objective is to estimate the probability that two samples were authored by the same or by a different writer without the requirement of knowing all of the potential sources. The SLR approach provides an open-set solution to the question of whether the similarity observed between a questioned and a known document supports the proposition that the documents were written by the same person. To compute the SLR, several decisions regarding the distance/scoring function and the method for estimating score densities, for example, need to be made. We explore both common source and specific source SLR approaches using simple distance measures between two documents as inputs for a random forest which outputs a score between 0 and 1 which can be loosely interpreted as the empirical probability of same source. We then use a kernel density framework to estimate the densities of the similarity scores among pairs of documents known to have been written by the same or by different persons. These approaches are applied to a handwriting data set collected by CSAFE.

2 | DATA

Handwriting samples were collected in a study conducted by CSAFE at Iowa State University [9]. Participants were asked to copy three prompts in their natural handwriting with a ballpoint pen on unlined paper supplied by CSAFE. The prompts include "The London Letter," an excerpt from the book *The Wizard of Oz*, and the short phrase: "The early bird may get the worm, but the second mouse gets the cheese." The longest of the prompts is "The London Letter" which has been used in other studies because it contains the numbers 0 through 9 and all of the letters in the alphabet in both uppercase and lowercase. Here, we use the samples obtained from 90 participants, most of who provided three replicates of each prompt.

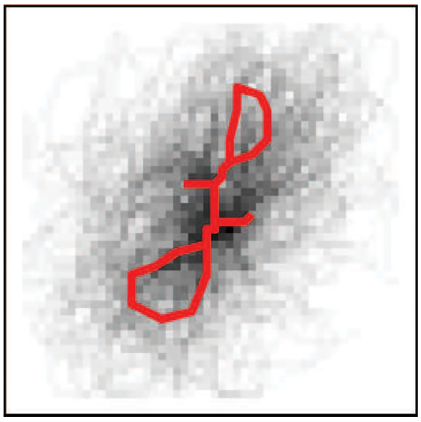

FIGURE 1 Example of a cluster exemplar (red) and all graphs in this cluster (gray)

In an earlier project, researchers in CSAFE developed an R package to extract features from gray-scale images of the writing samples [10]. The algorithm decomposes writing into small graphical units that roughly (but not always) correspond to characters. These graphs were then sorted into 40 clusters using a k-means approach. Figure 1 shows one of the clusters, as illustration [6]. Each cluster is characterized by an *exemplar graph*, shown in red in the figure.

Crawford et al. [6] proposed that the frequency with which a writer contributes graphs to each cluster is characteristic. They assumed that the observed 40-dimensional vectors of frequencies were drawn from a multinomial distribution with parameter vector $\boldsymbol{\pi}_w$, unique for each writer. Crawford et al. considered the case where the writer of a document is in a closed set of writers. Here we relax the closed set assumption, and instead focus on the problem of comparing any two documents to determine whether they were written by the same person. Even though we will also use the 40-dimensional vectors of frequencies as the response variable, we transform these vectors to account for the fact that the total number of graphs in the two documents may vary. See Section 3 for further details.

3 | METHODS

3.1 | Forensic identification of source

Suppose that a crime is committed, and the evidence consists of two handwritten documents E_x and E_y. We distinguish between two different scenarios which can be summarized as follows [7]:

- *Common source scenario*, which can be represented by two propositions, colloquially referred to as the prosecution and defense, where E_x and E_y are questioned documents with unknown writer(s):

  H_p: Two questioned documents, E_x and E_y, were written by the same, unspecified person.
  H_d: Two questioned documents, E_x and E_y, were written by two different unspecified people.

- *Specific source scenario*, where E_x is a questioned document and E_y is now a reference sample obtained from the defendant, and the prosecution and defense propositions are:

  H_p: The questioned document, E_x, was written by the defendant who wrote the known document, E_y.
  H_d: The questioned document, E_x, was written by some person other than the defendant who wrote the known document, E_y.

In both scenarios, a useful approach to quantify the weight of the evidence in favor of the prosecution's (or the defense's) propositions is the *likelihood ratio*. If x, y are features measured from E_x and E_y, respectively, and if $f\left(x, y|H_p\right)$ and $f\left(x, y|H_d\right)$ are probability models that describe the joint distribution of those features under the two propositions, then the likelihood ratio is computed as

$$\mathrm{LR} = \frac{f\left(x, y|H_p\right)}{f\left(x, y|H_d\right)},$$

and quantifies the odds of observing the evidence under the two competing propositions. When the feature vectors of the evidence are highly dimensional and complex, as is the case with handwriting features, it is often very difficult to define reasonable statistical models for these features.

An alternative is to use SLRs to approximate the weight of evidence [11]. SLRs are generically defined as:

$$\mathrm{SLR} = \frac{g\left(\Delta(x, y)|H_p\right)}{g\left(\Delta(x, y)|H_d\right)},$$

where H_p and H_d are as previously defined, $\Delta(x, y)$ is a (dis)similarity score for a comparison between x and y, $g\left(\cdot|H_p\right)$ is the probability density of comparison scores when the prosecution proposition is assumed to be true, and $g\left(\cdot|H_d\right)$ is the probability density of comparison scores when the defense proposition is assumed to be true. Practically, the densities, g, are estimated using a set of comparison scores reflecting what you would expect to see when the relevant proposition is true. In the remainder, we focus on SLRs and their calculation and interpretation in the two different scenarios introduced above.

In the common source scenario, the formulation of the SLR is relatively straightforward. For illustration, refer to the top panel in Figure 2. With three writers and three samples per writer, there are 36 different paired comparisons that are possible, and those are indicated in the colored squares in the figure. In this case, the numerator of the SLR is indicated in red and marked with an "N," and the denominator is indicated in blue and marked with a "D," and correspond to comparisons between sample pairs that are used to estimate $g\left(\cdot|H_p\right)$ and $g\left(\cdot|H_d\right)$, respectively. The numerator scores represent all comparisons between pairs of documents from the same writer, whereas the denominator scores represent all comparisons between pairs of documents from different writers.

In the specific source scenario, the document pairs used to estimate $g\left(\cdot|H_p\right)$ in the numerator are colored in red and marked with an "N" in the bottom panel of Figure 2. The numerator scores represent all comparisons between pairs of documents written by the defendant (denoted "Source" in the figure). Next, the pairs of documents that are used to estimate $g\left(\cdot|H_d\right)$ in the denominator depend on the precise formulation of H_d. We consider three different approaches for the specification of the pairs of documents used to calculate the denominator in the SLR proposed by Hepler et al. [5]. The three approaches are denoted "trace-anchored," "source-anchored," and "general-match," and the pairs of documents to construct the denominator of the SLR in each case are shown in different colors in the bottom panel of Figure 2.

For the trace-anchored approach, only the comparisons between the questioned document (or trace, previously denoted E_x) and a collection of writers different from the specific source (often called the background population or database) are used. These sample pairs are shown in green and marked with a "T" in the bottom panel of Figure 2. For the source-anchored approach, only the comparisons between the writing from the specific source (in this case, the defendant) and the other writers in the background population are used. These document pairs are shown in orange and marked with an "S" in the bottom panel of Figure 2. Finally, for the general-match approach, only the comparisons between samples from different writers in the background population are used. These are the same comparison pairs used in the estimation of the

Common Source Score-based Likelihood Ratios

		QD 1	QD 2	Writer A			Writer B			Writer C		
				1	2	3	1	2	3	1	2	3
QD 1												
QD 2												
Writer A	1											
	2			N								
	3			N	N							
Writer B	1			D	D	D						
	2			D	D	D	N					
	3			D	D	D	N	N				
Writer C	1			D	D	D	D	D	D			
	2			D	D	D	D	D	D	N		
	3			D	D	D	D	D	D	N	N	

Same source numerator Different source denominator

Handwriting Specific Score-based Likelihood Ratios

			Source			Writer A			Writer B			Writer C		
		QD	1	2	3	1	2	3	1	2	3	1	2	3
QD														
Source	1													
	2		N											
	3		N	N										
Writer A	1	T	S	S	S									
	2	T	S	S	S									
	3	T	S	S	S									
Writer B	1	T	S	S	S	G	G	G						
	2	T	S	S	S	G	G	G						
	3	T	S	S	S	G	G	G						
Writer C	1	T	S	S	S	G	G	G	G	G	G			
	2	T	S	S	S	G	G	G	G	G	G			
	3	T	S	S	S	G	G	G	G	G	G			

Numerator Trace-anchored Source-anchored General Match

FIGURE 2 Document pairs used to create the numerators and denominators in common source, trace-anchored, source-anchored, and general-match score-based likelihood ratios

denominator of the common source SLR, and are shown in blue and marked with a "G" in the bottom panel of Figure 2.

3.1.1 | Scores for comparison of handwritten samples

Following Crawford [8], we rely on the frequency of graphs contributed by a writer to each of 40 clusters. An important difference, however, is that in our case, the two documents to be compared can differ in length and consequently, in the total number of graphs extracted from it. Therefore, the multinomial probability model of Crawford et al. [6] is no longer appropriate. Instead, we consider the *proportions* of graphs allocated to each of the clusters. As in Reference 6, we hypothesize that the proportion of graphs allocated to each cluster is typical of the writer. To compare two vectors of observed proportions, denoted by x and y, we considered two different distance metrics. The first is the element-wise absolute differences in the observed feature vectors for both documents, and is given by

$$d_{\mathrm{A}}(x, y) = |x - y|$$

where the difference between the vectors is performed element-wise. The second is the Euclidean distance in the observed feature vector for both documents, and is given by

$$d_{\mathrm{E}}(x, y) = \sqrt{\sum_{i=1}^{K} \left(x_{[i]} - y_{[i]}\right)^2}$$

where the subscript $[i]$ denotes the ith element of each vector and K is the vector length (number of clusters). Concatenating these distances into one single vector results in a 41-dimensional vector of distances between documents in a pair.

To summarize the feature vectors, a random forest algorithm is trained to classify these distance vectors as being a "same-writer" or "different-writers" comparison between two documents. Let $rf_m(D_{\text{KM}}, D_{\text{KNM}})$ be used to denote the training step of the random forest algorithm for method m (which will correspond to the type of SLR) which takes as input labeled training data consisting of vector of distances between known-matching pairs, D_{KM}, and vectors of distances between known-non-matching pairs, D_{KNM}. For example, the random forest that computes the scores needed to construct the common source SLR would be trained on a set of known-match distances computed from sample pairs corresponding to the red "N" in the top panel of Figure 2 and a set of known-non-match distances computed from sample pairs corresponding to the blue "D," denoted by D_{CS} and D_{GM}, respectively. Thus, rf_{CS} is trained on the labeled data $(D_{\text{CS}}, D_{\text{GM}})$. Similarly, the random forests that compute the scores needed to construct the three specific source SLRs would be trained on a set of known-match distances computed from sample pairs corresponding to the red "N" in the bottom panel of Figure 2, denoted D_{SS}, and three different sets of known-non-match distances computed from sample pairs corresponding to the green "T," orange "S," and blue "G," denoted by D_{TA}, D_{SA}, and D_{GM}, respectively. So, the random forest for the trace-anchored approach, rf_{TA}, is trained on the labeled data $(D_{\text{SS}}, D_{\text{TA}})$, the random forest for the source-anchored approach, rf_{SA}, is trained on the labeled data $(D_{\text{SS}}, D_{\text{SA}})$, and the random forest for the general-match approach, rf_{GM} is trained on the labeled data $(D_{\text{SS}}, D_{\text{GM}})$. Once the rf_m is trained, then the prediction step of the random forest algorithm takes as input a vector of distances from a new comparison, and outputs the resulting proportion of decision trees that classified the distance measures as originating from a "same-writer" comparison, denoted in functional form as $\widehat{rf}_m : \left(\mathbb{R}^+\right)^{K+1} \mapsto [0,1]$. These outputs will be referred to as "similarity scores" because a larger score is more indicative of a "same-writer" comparison whereas a smaller score is more indicative of a "different-writer" comparison. In this case, the function for producing scores, Δ_m, can be considered a composition of functions such that $\Delta_m = \widehat{rf}_m \circ (d_{\text{A}}, d_{\text{E}})$. Note that this implies that the random forest algorithms used for the specific source SLRs need to be trained specifically to each defendant. In order to save the computational burden required to retrain each specific source random forest, we will explore using $\widehat{rf}_{\text{CS}}$ for generating the scores, see Section 3.3.2 for further details.

In some circumstances, such as the illustrative example given in Reference 5 with normally distributed features and a squared-difference distance function, the distributions of the scores can be derived analytically from the features of the data. However, this is not the case for this handwriting application. Because we have decided to use these features as input variables for the "black-box" random forest algorithm to generate our scores, it is unclear in this case what the distribution of the scores should be under our two propositions. Therefore, we have chosen a non-parametric kernel density estimation (KDE) approach to modeling the densities of the scores under H_p and H_d, denoted by $\widehat{g}(\cdot|\delta)$ where δ denotes a set of scores used to create the density estimate. For our purposes, the KDE is implemented using the density function in R [12] assuming a Gaussian kernel (with the default bandwidth selection) within the bounds $[0, 1]$.

3.1.2 | Score-based likelihood ratios

Recalling that the score densities defining the common source SLR are estimated by performing all same-writer and different-writers pairwise comparison scores according to the top panel in Figure 2, let the random forest similarity scores used to estimate $g\left(\cdot|H_p\right)$ and $g(\cdot|H_d)$ be denoted by δ_{CS} and δ_{GM}, respectively. Therefore, the common source SLR is defined as

$$\text{SLR}_{\text{CS}}(x,y) = \frac{\widehat{g}\left(\Delta_{\text{CS}}(x,y)|\delta_{\text{CS}}, H_p\right)}{\widehat{g}\left(\Delta_{\text{CS}}(x,y)|\delta_{\text{GM}}, H_d\right)}. \tag{1}$$

Similarly, the score densities defining the three specific source SLRs are estimated by performing same-writer and different-writers pairwise comparison scores according to the bottom panel in Figure 2. The numerators for each of the three SLRs from Reference 5 are estimated using similarity scores in handwriting samples from the same, fixed writer (in this case, the fixed writer will be the defendant). However, with small numbers of samples from the defendant, there are not enough scores available for estimating the density. Following the experiment by Hepler et al. [5], multiple "pseudo-documents" were created from the available reference samples from the defendant. First, all of the reference samples are combined to create one large "template" for the specific source. In each of the templates, there is a given total number of graphs and a random number is drawn to split the template into two pseudo-documents. After splitting the graphs, the proportions in each cluster are recorded in each pseudo-document pair, denoted by y_1^* and y_2^*. A random forest similarity score is calculated between the two feature vectors, $\Delta_m\left(y_1^*, y_2^*\right) = \widehat{rf}_m \circ (d_{\text{A}}, d_{\text{E}})\left(y_1^*, y_2^*\right)$. This process is repeated 1000 times for each template to create 1000 scores used to estimate $g\left(\cdot|H_p\right)$ in the numerator of each specific source SLR, denoted by δ_{SS}. A diagram of these steps is illustrated in Figure 3.

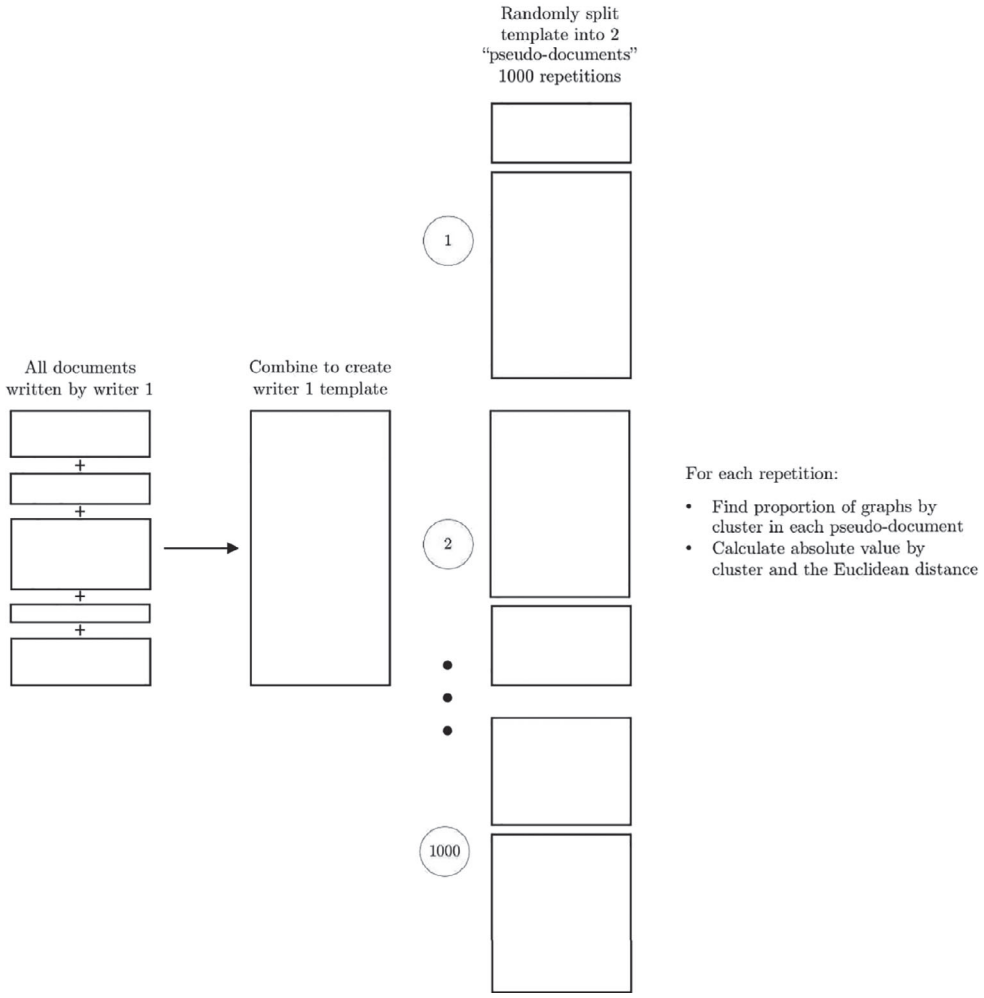

FIGURE 3 Diagram of the procedure for creating the collection of pseudo-documents from all of a writer's documents

Next, the comparison scores for estimating the three denominator densities are computed based on the relevant population. The trace-anchored denominator is defined as "the evidence score arises from the distribution of scores obtained by pairing the questioned document with a template written by a random individual" [5]. So, the denominator is the comparison between this questioned document and all other documents that are not written by the defendant. For the trace-anchored approach, let the random forest similarity scores used to estimate $g(\cdot|H_d)$ in the denominator be denoted by δ_{TA}. Thus, the trace-anchored SLR is defined by

$$\text{SLR}_{\text{TA}}(x,y) = \frac{\hat{g}\left(\Delta_{\text{TA}}(x,y)|\delta_{\text{SS}}, H_p\right)}{\hat{g}\left(\Delta_{\text{TA}}(x,y)|\delta_{\text{TA}}, H_d\right)}. \tag{2}$$

The source-anchored denominator is defined as: "the evidence score arises from the distribution of scores obtained by pairing a questioned document written by a random individual with the template written by the suspect" [5]. To do this, we compare the documents written by the defendant to documents written by all other writers in the relevant population. For the source-anchored approach, let the random forest similarity scores used to estimate $g(\cdot|H_d)$ in the denominator be denoted by δ_{SA}. So, the source-anchored SLR is defined by

$$\text{SLR}_{\text{SA}}(x,y) = \frac{\hat{g}\left(\Delta_{\text{SA}}(x,y)|\delta_{\text{SS}}, H_p\right)}{\hat{g}\left(\Delta_{\text{SA}}(x,y)|\delta_{\text{SA}}, H_d\right)}. \tag{3}$$

Last, the scores used to define the general match denominator are the same as the ones used in the denominator of the common source SLRs. We compare two documents written by two different people in the relevant population. For the general-match approach, let the random forest similarity scores used to estimate $g(\cdot|H_d)$ in the denominator be denoted by δ_{GM}. Consequently, the general-match SLR is defined by

$$\text{SLR}_{\text{GM}}(x,y) = \frac{\hat{g}\left(\Delta_{\text{GM}}(x,y)|\delta_{\text{SS}}, H_p\right)}{\hat{g}\left(\Delta_{\text{GM}}(x,y)|\delta_{\text{GM}}, H_d\right)}. \tag{4}$$

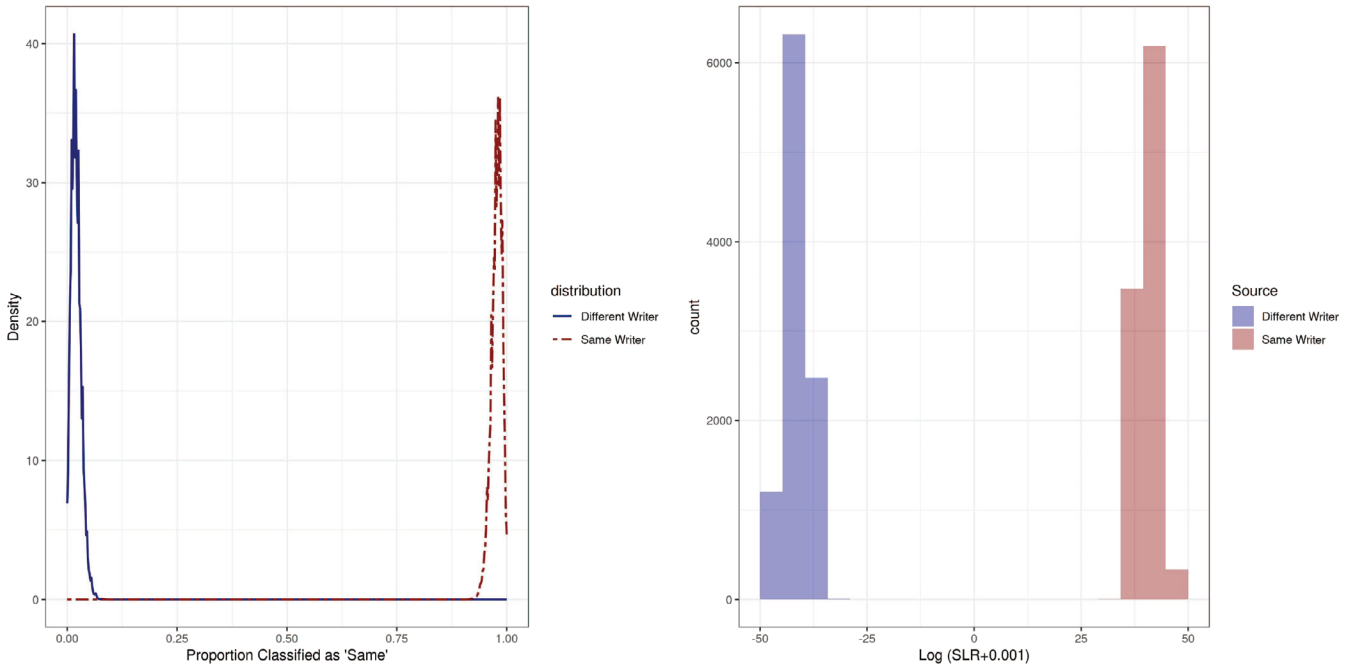

FIGURE 4 Kernel density estimation of random forest similarity scores (left panel) and common source score-based likelihood ratios (right panel) for same (red) and different (blue) writer comparisons of simulated Dirichlet data

The general match denominator excludes all known different-source comparisons that involve either the questioned document or the documents from the defendant. For the source-anchored denominator, known different-source comparisons involving the questioned document are excluded. In contrast, the trace-anchored denominator excludes all known different-source comparisons involving the defendant's documents. Both the trace-anchored and source-anchored denominators exclude known different-source comparisons involving two different writers from the relevant population. Refer back to Figure 2 for an illustrative example of which scores are used in each SLR construction.

3.2 | Simulation study

Because the methods proposed above have not been applied to handwriting analysis before, the feasibility of this SLR approach is explored using simulated data for the common source scenario. To do this, the feature vectors for documents from a given writer are simulated from a Dirichlet distribution. The elements of the parameter vector of this Dirichlet distribution, $\alpha_1, \ldots, \alpha_{40}$, correspond to the rate at which a specific writer emits graphs to each cluster, and for this simulation are generated from a uniform distribution. In order to test the common source SLR approach, scores representing a comparison of writing samples are needed within the same writer and between different writers. Therefore, three sets of 10,000 random variables were simulated. Two sets were drawn from a Dirichlet with the same parameter vector and the third set from a Dirichlet with another parameter vector. The element-wise absolute differences and Euclidean distances between samples in the first and second sets represent known "same-writer" document pairs whereas comparisons between samples in the first and third sets will represent known "different-writers" document pairs. This data constituted the training set.

Next, a random forest trained on this data set will classify new distance measures as deriving from a "same-writer" or "different-writers" comparison. A second testing data set was simulated with the same procedure as the training data set to create more known same-writer and known different-writers pairs. The test set was fed into the trained random forest, and the output gives us a set of similarity scores that are used to estimated score densities for the numerator and denominator of the common source SLR. The resulting KDEs are show in the left panel of Figure 4.

A third validation data set was simulated in the same way as the previous two and the distance metrics computed for each. The similarity scores for each pair were determined using the trained random forest algorithm. Leaving one resulting comparison out at a time to represent the current questioned documents, the density function value for the questioned comparison is estimated from both the prosecution ("same-writer") and defense

("different-writers") estimated distributions plotted in the left panel of Figure 4. SLRs are then calculated by taking the ratio of the two density values. It is expected that SLR value would be small for those scores corresponding to known different-writers comparisons and large for known same-writer comparisons. The distribution of SLR values from the simulated validation data is show in the right panel of Figure 4. For the purposes of visualization, SLRs calculated to be infinite are replaced by the largest finite SLR in the data set. The results from the simulation study suggest that the proposed method may be a reasonable approach for inferring writership for real handwriting data.

3.3 | Application to handwriting data

3.3.1 | Common source SLRs

To evaluate the performance of the common source SLR method, the first session of the CSAFE data set is split into a training set and a testing set. The training set was created by taking a random sample of 80% of the writers. The remaining 20% of the writers are reserved for the testing set. In addition, the training and testing sets are split into four different subsets. The first three subsets correspond to separating the writing by the three different prompts. The fourth subset corresponds to separating the writing by repetition, then all three prompts within the same repetition are concatenated to create one large, combined document. Because there were three replicates of each prompt, each data set will consist of three samples for each writer. Therefore, there will be four different training and testing sets, one corresponding to each of the three prompts and one for the combined document. For all samples in the CSAFE data set, the `handwriter` package was used to extract the features (the proportions of graphs in each of $K = 40$ clusters). So, for sample E_{ws} corresponding to prompt, $p = L, W, P, C$ (for the London Letter, Wizard of Oz, short phrase, and combined prompts, respectively) denote the corresponding features as $z_{ws}^{(p)}$. In the training set $w = 1, 2, \ldots, 72$, in the testing set $w = 1, 2, \ldots, 18$, and in both $s = 1,2,3$.

To create a group of known same-source comparison scores to estimate the density for scores under the prosecution proposition, all pairwise comparison scores within the same writer were taken among all of the documents for a given prompt written by the training-set writers. Let these comparison scores be denoted by $\delta_{\text{CS}}^{(p)} = \left\{ \Delta\left(z_{ws}^{(p)}, z_{w's'}^{(p)}\right) : w = w' \right\}$. To create a group of known different-source comparison scores to estimate the density for scores under the defense proposition, all pairwise comparisons between documents for a given prompt written by two different training-set writers were taken. Let these comparison scores be denoted by $\delta_{\text{GM}}^{(p)} = \left\{ \Delta\left(z_{ws}^{(p)}, z_{w's'}^{(p)}\right) : w \neq w' \right\}$. Since there are many more comparisons between documents from different writers than same writer ($|\delta_{\text{CS}}^{(p)}| = 72\binom{3}{2} = 216$ compared to $|\delta_{\text{GM}}^{(p)}| = \binom{72.3}{2} - 216 = 23{,}004$), the number of document comparisons between different writers was downsampled (randomly) to be equal to the number from same writers. An example of the KDE for the known same-source and different-sources scores from the training set is shown in the left panel of Figure 5 (the KDE plots for the remaining prompts are given in the Appendix). While the training data set consisted of an equal number of distance measures from documents written between same writer and different writers, the testing data set did not. Each comparison score in the testing data set served the role of $\Delta(x, y)$ and the corresponding common source SLR values were calculated using Equation 1. An example of the distribution of common source SLR values resulting from the testing set is provided in the right panel of Figure 5 (the SLR histograms for the remaining prompts are given in the Appendix). Receiver operating characteristic (ROC) curves, associated area under the curve (AUC), and Tippett plots were created to assess the performance of the common source SLRs; see Section 4 for details.

3.3.2 | Specific source SLRs

In the analyses so far, the SLR has been defined with respect to the common source propositions. It is reasonable to expect that repeating the procedure to get specific source SLRs for the same data would result in similar conclusions. To evaluate the performance of the specific source SLR methods, the first session of the CSAFE data set is again used. One difference between this application and the common source applications is that the specific source SLRs include information about the suspected source of the writing and vary for each defendant considered. To compare the behavior of these SLRs to the common source, three writers (sources/defendants) were selected to include one writer that is typical of the relevant population (handwriting features are similar to a large number of other writers), one that is rare in the population (handwriting features are different from a large number of other writers), and one in-between. Within the data set representing the relevant population (the 90 writers in the CSAFE data set), the Euclidean distance between feature vectors was computed for each possible pairwise comparison. After ordering the sum of Euclidean distances by writer, the writer associated

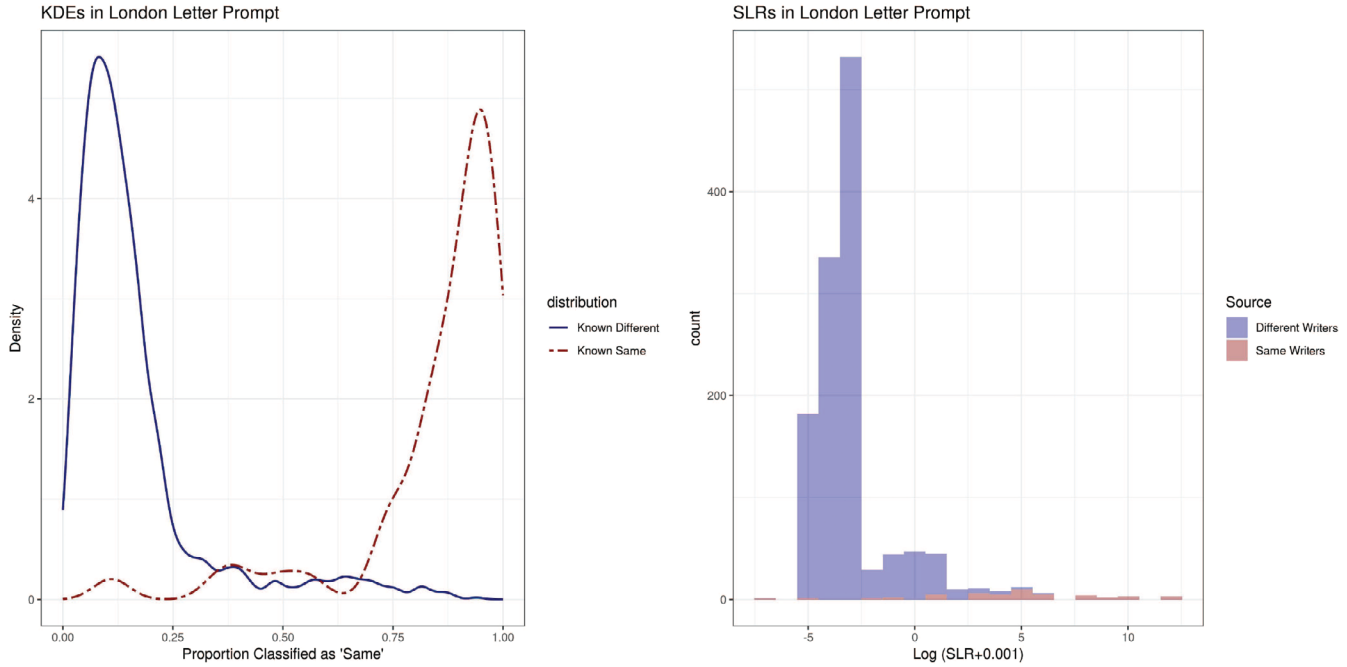

FIGURE 5 Kernel density estimation of random forest similarity scores (left panel) and common source score-based likelihood ratios (right panel) for same (red) and different (blue) writer comparisons for the CSAFE data London Letter prompt. CSAFE, Center for Statistics and Applications in Forensic Evidence

with the minimum, median, and maximum distances were selected (writer 0145, writer 0063, and writer 0004, respectively). We investigate two methods for determining which random forest is employed to compute the similarity scores. The first method uses the random forest trained with all of the combined documents for addressing the common source problem (pretrained random forest). The second trains random forests according to the SLR of interest.

Begin by considering the unmodified specific source SLR approach using the SLR-specific random forest scores. For this approach, we only consider the combined prompts from the 90 CSAFE writers (three samples each). Further, let $w = j$ be the index corresponding to the writer playing the role of the defendant (j is the index corresponding to writer 0145, 0063, or 0004), and so the jth writer wrote the reference document E_y. For the numerator of each of the three specific source SLRs, there are three reference samples from writer j. We combine all three samples to create one large template, and then follow the procedure to create 1000 pseudo-document pairs and extract their corresponding features, denoted by $\left(y_1^{*(s)}, y_2^{*(s)}\right)$ where s indexed the split iteration. Then, use the appropriate random forest to calculate the similarity scores, $\Delta_m\left(y_1^{*(s)}, y_2^{*(s)}\right)$. Finally, randomly assign two-thirds ($n = 667$) of these scores to the training data set and the remaining one-third ($n = 333$) to the testing set. Thus, for a given writer j there is a different set of similarity scores that are used to estimate the numerator density for each of the three different SLR methods: 1) for the trace-anchored approach $\delta_{SS} = \left\{\Delta_{TA}\left(y_1^{*(s)}, y_2^{*(s)}\right) : s = 1, \ldots, 667\right\}$ in Equation (2), 2) for the source-anchored approach $\delta_{SS} = \left\{\Delta_{SA}\left(y_1^{*(s)}, y_2^{*(s)}\right) : s = 1, \ldots, 667\right\}$ in Equation (3), and 3) for the general-match approach $\delta_{SS} = \left\{\Delta_{GM}\left(y_1^{*(s)}, y_2^{*(s)}\right) : s = 1, \ldots, 667\right\}$ in Equation (4).

Next, we need to create the similarity scores for the denominator of each specific source SLR, for which there are 89 available writers (excluding j) and three samples from each writer. So, for the sth sample from the wth available writer, E_{ws}, denote the corresponding features as z_{ws}. For the trace-anchored approach, compute all similarity scores between the features of the questioned document, x, and all samples from the available alternative writers, $\delta^*_{TA} = \{\Delta_{TA}(x, z_{ws}) : w \neq j\}$. For the CSAFE data, there will be $|\delta^*_{TA}| = (w-1)s = 267$ comparisons. Then, randomly assign two-thirds to the training set and the remaining one-third to the testing set. Thus, $\delta_{TA} = \left\{\text{random selection of } n = 178 \text{ scores from } \delta^*_{TA}\right\}$ are the similarity scores used to estimate the denominator density in Equation (2). For the source-anchored approach, compute all similarity scores between the features of the defendant's template (all three prompts combined), y^*, and all samples from the available alternative writers, $\delta^*_{SA} = \{\Delta_{SA}(y^*, z_{ws}) : w \neq j\}$. Like the trace-anchored approach, there are $|\delta^*_{SA}| = (w-1)s = 267$ comparisons for the CSAFE data set. Again,

randomly assign two-thirds to the training set and the remaining one-third to the testing set. Thus, $\delta_{SA} = \{\text{random selection of } n = 178 \text{ scores from } \delta^*_{SA}\}$ are the similarity scores used to estimate the denominator density in Equation (3). For the general-match approach, among the available writers, compute all similarity scores between two samples from different writers, $\delta^*_{GM} = \{\Delta_{GM}(z_{ws}, z_{w's'}) : w \neq w', w \neq j, w' \neq j\}$. With 89 available writers and three samples each, there are $|\delta^*_{GM}| = \binom{(w-1)s}{2} - w\binom{s}{2} = 35{,}244$ comparisons for the CSAFE data. Randomly assigning two-thirds to the training set and the remaining one-third to the testing set, then $\delta_{GM} = \{\text{random selection of } n = 23{,}496 \text{ scores from } \delta^*_{GM}\}$ are the similarity scores used to estimate the denominator density in Equation (4).

Now that the training and testing sets have been defined, each comparison score in the testing data set served the role of $\Delta_m(x, y)$ and the corresponding specific source SLR values were calculated using Equations (2)–(4). ROC curves summarizing the performance of each specific source SLR method are provided in Section 4. See the Appendix for additional details of the results.

Because training the random forest for each defendant can be computationally intensive for utilizing this approach in practice, we explored whether a single, pretrained random forest could be used to get the similarity scores for the specific source SLRs. For this experiment, we chose to use the common source random forest, $\widehat{rf}_{CS}$, trained on the combined prompts from the 72 CSAFE writers in the common source training set. This random forest was used to create the scores for the specific source training sets described above, so essentially there are four changes: 1) the numerator densities for the specific source SLRs are all estimated using the set of similarity scores $\delta_{SS} = \Delta_{CS}\left(y_1^{*(s)}, y_2^{*(s)}\right) : s = 1, \ldots, 667\}$, 2) the denominator density of the SLR for the trace-anchored approach is estimated using the set of similarity scores δ_{TA} corresponding to a random two-thirds selection of the scores in $\delta^*_{TA} = \{\Delta_{CS}(x, z_{ws}) : w \neq j\}$, 3) the denominator density of the SLR for the source-anchored approach is estimated using the set of similarity scores δ_{SA} corresponding to a random two-thirds selection of the scores in $\delta^*_{SA} = \{\Delta_{CS}(y^*, z_{ws}) : w \neq j\}$, and 4) the denominator density of the SLR for the general-match approach is estimated using the set of similarity scores δ_{GM} corresponding to a random two-thirds selection of the scores in $\delta^*_{GM} = \{\Delta_{CS}(z_{ws}, z_{w's'}) : w \neq w', w \neq j, w' \neq j\}$. In addition, the common source random forest was used to compute the similarity scores for the testing set as well. Thus, $\Delta_{CS}(x, y)$ was used to compute the comparisons scores in the testing set, and replaces the approach-specific score $\Delta_m(x, y)$ in Equations (2)–(4). ROC curves summarizing the performance of each specific source SLR method utilizing the pretrained (common source) random forest scoring method are provided in Section 4. See the Appendix for additional details of the results.

4 | RESULTS

The results of applying the trained random forest to the CSAFE data showed promising discrimination for the common source question using the similarity scores as the length of the prompt increased. As the length of the writing sample increases, the same-writer and different-writers KDEs become more separated. However, the KDEs for the shortest prompt are not well separated. The degree of separation between the two KDEs is directly related to the performance of an SLR. When the comparison is known to come from different writers, a well-behaved SLR system would result in a small log-SLR value (less than 0), whereas a comparison known to come from the same writer should result in a large log-SLR value (greater than 0). For a closer inspection of the performance of the common source SLR system to the CSAFE data, Tippett plots were utilized (Figure 6). Tippett plots assess the performance of the SLR system as a whole by splitting the data into known same-writer and known different-writer comparisons. Then, the empirical cumulative distribution function (ECDF) within each group of comparisons is plotted (black for the known same-writer and green for the known different-writers comparisons). Ideally, the ECDF for the known different-writers (green) would be to the left of the vertical line at zero and the ECDF for the known same-writer (black) would be to the right of the vertical line at zero. This would indicate that the SLR system made no errors on the CSAFE data set. We can see that the green line is to the left of zero only for the London Letter and combined prompts, whereas none of the black lines are completely to the right of zero. This indicates that it is more difficult for the SLR system to properly support the prosecution proposition when it is true than it is to properly support the defense proposition when it is true. The area below the black line and to the left of the vertical line at zero indicates the rate of SLR values that support the defense proposition when the prosecution proposition is actually true (Type 1 error). The results in Figure 6 show the largest Type 1 error rate for the combined prompt. Similarly, the area above the green curve and to the right of the vertical line at zero indicates the rate of SLR values that support the prosecution proposition when the defense proposition is actually true (Type 2 error). The results in

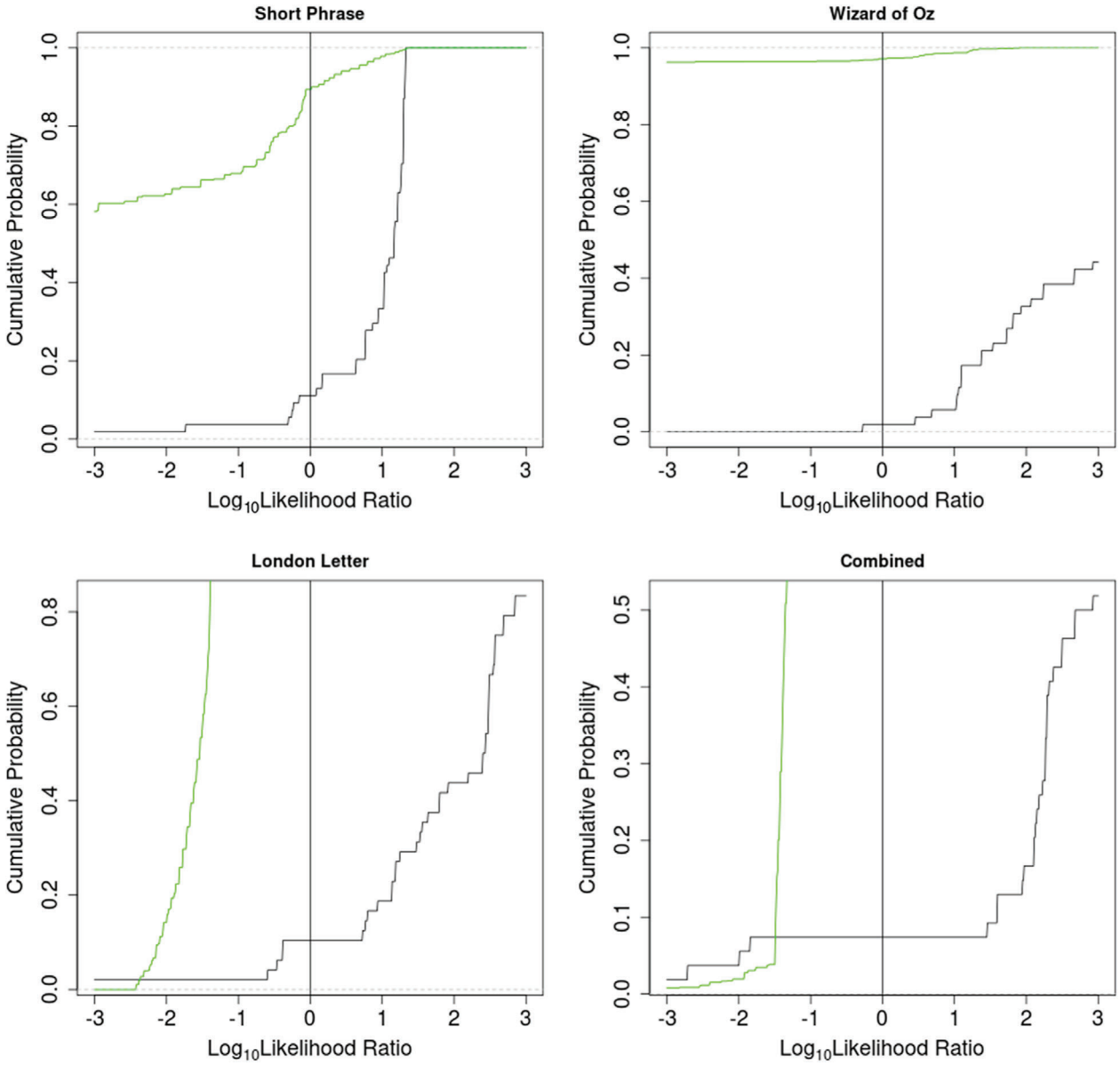

FIGURE 6 Tippett plots for the common source SLR applied to the CSAFE data by handwriting prompt where the green line indicates the known different-writers comparisons and the black line indicates the known same-writer comparisons. CSAFE, Center for Statistics and Applications in Forensic Evidence; SLRs, score-based likelihood ratios

Figure 6 show the largest Type 2 error rate for the short phrase prompt.

For the application of the common source SLR to the CSAFE data set, as the length of the documents increase the SLR system showed better performance, with respect to the threshold on the log-SLR at zero. However, if an SLR system is poorly calibrated, this threshold may actually occur at a different location. The performance of the system (using the Type 1 and Type 2 errors) across a variety of thresholds is typically visualized using a ROC curve. Then, the performance of the SLR system across all possible threshold values can be quantified using the area under the ROC curve (AUC). ROC curves should have a larger AUC (or have curves closer to the upper left corner) when the SLR system has better performance, whereas the curve will be closer to the 45° diagonal line when the SLR system has poorer performance. For this application of the common source SLR, the resulting AUC values are given in Table 1 and are better for the longer prompts (like London letter and combined) and worse for shorter prompts (Wizard of Oz and short phrase). Therefore, we still expect the SLR to perform better for longer prompts using different thresholds on the log-SLR than for shorter prompts. This is not unexpected because the longer documents contain more writing which we would expect to provide more information about a writer's writing style and be the

TABLE 1 Summary of the area under the ROC curve for all SLR methods

Method	AUC			
Common source	Short Phrase	Wizard of Oz	London Letter	Combined
	0.6447	0.7293	0.9692	0.9610
		Writer 0004	Writer 0063	Writer 0145
Specific source	Trace-anchored	0.8815	0.9658	0.9627
Pretrained	Source-anchored	0.8913	0.9742	0.9536
Random Forest	General-match	0.9117	0.9675	0.9635
		Writer 0004	Writer 0063	Writer 0145
Specific source	Trace-anchored	1	1	0.9980
Writer-specific	Source-anchored	1	1	1
Random Forest	General-match	1	1	1

Abbreviations: AUC, area under the curve; ROC, receiver operating characteristic; SLR, score-based likelihood ratio.

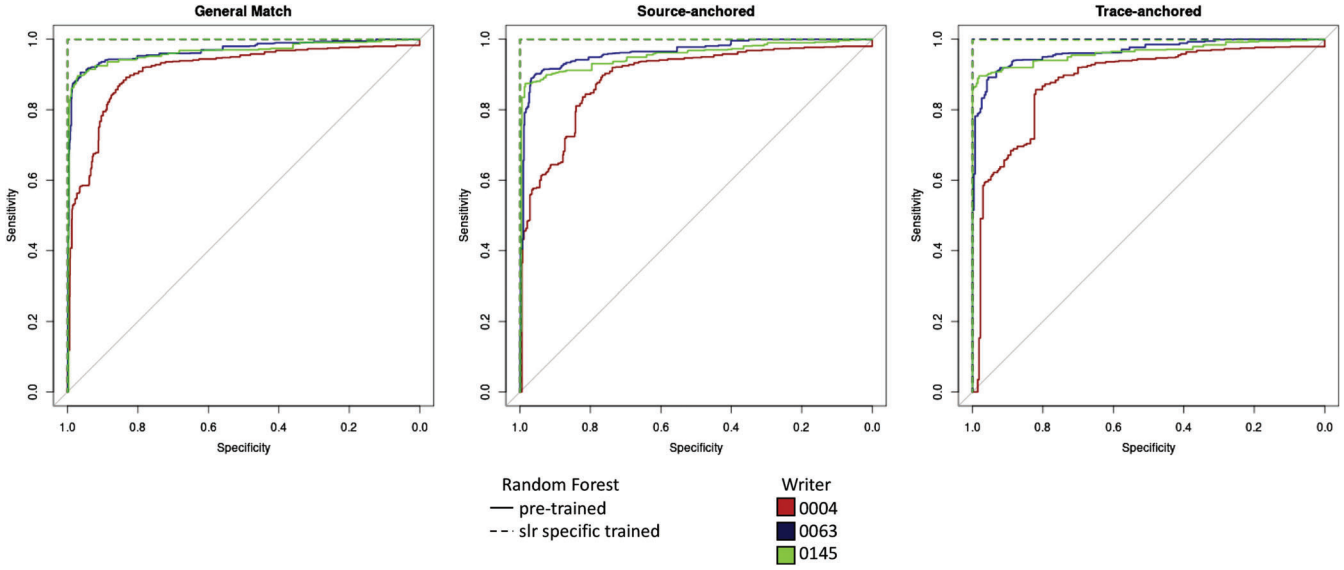

FIGURE 7 ROC curves of SLRs from pretrained and SLR specific random forests across writers 0004, 0063, and 0145. ROC, receiver operating characteristic; SLRs, score-based likelihood ratios

most representative of the proportion of graphs per cluster compared to the shorter documents.

For the specific source problem, it is necessary to train a new random forest for every fixed, specific source to obtain the comparison scores. In an attempt to save on computational time, the pretrained, common source random forest was used to generate the comparison scores and then the specific source approach was used to define the SLRs. To compare the performance of these specific source SLRs (using the common source random forest) to the common source SLRs, the ROC curves were plotted with a solid line in Figure 7 and color corresponding to each of the three specific writers. Furthermore, the AUC was computed and summarized in Table 1. In every case except one (the source-anchored approach for writer 0063), the best ROC curve for the common source SLR (AUC of 0.9629) had better performance than using the same scoring mechanism with the specific source SLR. This behavior is likely due to the fact that there is a mismatch between the method of defining the SLR (specific source) to the method of training the random forest (common source). Furthermore, the specific source SLRs for writer 0004 (corresponding to the solid red lines) performed less well compared to the other two writers. This behavior is expected because this is the writer that was typical of writers in the data set, meaning that there would be several other writers that have similar handwriting making it harder to discriminate writer 0004 from others.

Because the specific source SLRs did not perform as well using the pretrained random forest scores, the fully specific source method using a random forest trained specifically to each writer was investigated next. The results in Figure 7 show that the SLRs from the writer-specific trained random forests (dotted lines) classify between same-writer and different-writer comparisons better than the pretrained version (solid lines). This indicates that the best method of computing the SLR is to match the method of training the random forest to the method of defining the SLR.

5 | CONCLUSIONS

Because two samples of handwriting created by the same source tend to have similar writing patterns, we assume this will result in similar proportions of graphs per cluster. To measure this similarity, we compute two different distance metrics. Theoretically, neither the distribution of the Euclidean distance nor the absolute difference was clearly a known probability density function. Therefore, these distance metrics were included as features in a random forest designed to classify these features as same-source or different-sources. Among the 41 features included to create the random forest, not all have an equal importance based on the Gini importance which is calculated as the Gini index at each node split [13]. Relative to the other variables, the Euclidean distance had the greatest importance. Additionally, when comparing the absolute difference by cluster, some clusters appear to be more influential across prompts, such as clusters 31 and 34 (see Figures A1 and A2 in the Appendix for further details). Cluster 34 contains many complex graphs with a greater range of variability and cluster 31 contains many simplistic graphs with a lower range of variability. We are unsure why these clusters are useful in the training of the random forest, but it is an interesting result worth considering for future directions of the research project.

For this process, more similar proportions of graphs per cluster will result in the smaller random forest scores. So, smaller random forest scores would be observed among documents written by the same writer and larger random forest scores would be observed between documents written by different writers. Then, we used KDE to model the score distributions for the known same-writer and known different-writers distributions. These score distributions are the basis for computing the SLR. Similar to results from Crawford [8], the results here show that the SLRs do not perform as well for short prompts containing fewer numbers of graphs in each document. One possible explanation for this poor performance is a mismatch of the data to the number of clusters used. The data used in these analyses were classified into 40 clusters; however, it is possible to change this number in the clustering algorithm. Future work could look into what number of clusters is optimal based on the length and type of documents being compared and how much, if any, influence this has on the performance of SLRs.

This work demonstrates the common source SLRs in Park and Carriquiry [11] and specific source SLRs in Hepler et al. [5] all show promise applied to this collection of handwriting data. While each of the SLRs perform well, deciding which one is appropriate depends on the question of interest and the evidence available. Common source SLRs should be applied when there are two questioned documents, and the goal is to quantify the probability of them being from the same writer against the probability of them being from different writers, without specifying which writer this is. This SLR incorporates all pairwise comparisons among writing in a relevant population and therefore makes the most efficient use of the data available to the investigator. This is different from the three specific source SLRs. The specific source SLRs should be applied when there is one questioned document, a known suspect who wrote a control document(s), and a database of handwriting from some writers in a relevant population. In all three specific source approaches, the goal is to quantify the probability of the questioned document being written by the suspect against the probability that the questioned document was written by someone else. To evaluate the second probability, the trace-anchored SLR is based on one questioned document and does not account for potential variability of the suspect's handwriting. The source-anchored SLR does include variability among documents written by the suspect, but excludes the question document itself during training. This approach will not account for information that could be present in the questioned document itself. The general match SLR excludes all of the documents written by the suspect and does not include the questioned document, either. Therefore, each of these methods will make less efficient use of the information available to the investigator, but can focus on more relevant data and information depending on the approach.

Another aspect to consider is that the common source approach has a more straight-forward method of evaluating the first probability, whereas it is much more difficult to evaluate the probability that the suspect wrote the questioned document for the specific source. This is because the specific source approach involves the complicated process of creating pseudo-documents when the number of available control documents written by the suspect is too few. Thus, if a suitable approach for creating pseudo-documents does not exist in practice, the common source approach is preferred over a specific source approach. On the other hand, if a suitable approach for

generating pseudo-documents exists (or if there are a large number of control documents from the suspect), a specific source approach with corresponding specifically-trained random forest scores is preferred because it performs better (according to AUC). Finally, a specific source approach more directly answers the question of interest in most forensic settings: whether the suspect wrote the document (rather than the less informative common source approach of determining same, but unknown, source).

In conclusion, this research project demonstrates an approach that uses a machine learning algorithm to assess the similarity between two handwritten documents, resulting in a low-dimensional score. This score is used in conjunction with two distributions of scores under both the common source and specific source propositions. The computed SLR provides information regarding the strength of forensic handwriting evidence when the alternative sources for the questioned document are not completely known. Thus, our approach improves upon the previous closed-set solution of writer identification for the CSAFE data.

ACKNOWLEDGMENTS

The authors thank Dr. Alicia Carriquiry for her support and guidance throughout this project. Her comments and suggestions helped us greatly improve the manuscript. This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreements 70NANB15H176 and 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College, and University of Nebraska, Lincoln. Open access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Iowa State University Digital Repository at https://doi.org/10.25380/iastate.10062203.

ORCID

Madeline Quinn Johnson https://orcid.org/0000-0001-7524-2846
Danica M. Ommen https://orcid.org/0000-0001-9955-3817

REFERENCES

1. R. Huber and A. Headrick, *Handwriting identification: Facts and fundamentals*, CRC Press, Boca Raton, FL, 1999.
2. Expert Working Group for Human Factors in Handwriting Examination, *Forensic handwriting examination and human factors: Improving the practice through a systems approach*. U.S. Department of Commerce, National Institute of Standards and Technology. NISTIR 8282, 2020.
3. National Research Council Committee on Identifying the Needs of the Forensic Sciences Community, *Strengthening forensic science in the United States: A path forward*. The National Academies Press, Washington, DC. https://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward, 2009.
4. J. J. Miller, R. B. Patterson, D. T. Gantz, C. P. Saunders, M. A. Walch, and J. Buscaglia, *A set of handwriting features for use in automated writer identification*, J. Forensic Sci. 62 (2017), no. 3, 722–734. https://doi.org/10.1111/1556-4029.13345
5. A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, *Score-based likelihood ratios for handwriting evidence*, Forensic Sci. Int. 219 (2012), no. 1, 129–140. https://doi.org/10.1016/j.forsciint.2011.12.009
6. A. M. Crawford, N. S. Berry, and A. L. Carriquiry, *A clustering method for graphical handwriting components and statistical writership analysis*, Stat Anal Data Min 14 (2021), no. 1, 41–60. https://doi.org/10.1002/sam.11488
7. D. M. Ommen and C. P. Saunders, *A problem in forensic science highlighting the differences between the Bayes factor and likelihood ratio*, Stat. Sci. 36 (2021), no. 3, 344–359. https://doi.org/10.1214/20-STS805
8. A. Crawford, *Bayesian hierarchical modeling for the forensic evaluation of handwritten documents*, Ph.D. thesis, Iowa State University, 2020.
9. A. Crawford, A. Ray, A. Carriquiry, J. Kruse, and M. Peterson, *CSAFE handwriting database*. https://iastate.figshare.com/articles/dataset/CSAFE_Handwriting_Database/10062203, 2019
10. N. Berry, J. Taylor, and F. Baez-Santiago, *Handwriter: handwriting analysis in R*, R package version 1.0.1. https://CRAN.R-project.org/package=handwriter, 2019.
11. S. Park and A. Carriquiry, *Learning algorithms to evaluate forensic glass evidence*, Ann. Appl. Stat. 13 (2019), no. 2, 1068–1102. https://doi.org/10.1214/18-AOAS1211
12. R Core Team, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2018.
13. L. Breiman, *Random forests*, Mach. Learn. 45 (2001), no. 1, 5–32. https://doi.org/10.1023/A:1010933404324

AUTHOR BIOGRAPHIES

Madeline Quinn Johnson is currently a biostatistician in Minnesota. She graduated with a B.A. in Mathematics from St. Olaf College in 2019 and earned her M.S. in Statistics in Iowa State University in 2021.

Dr. Danica M. Ommen is currently an Assistant Professor in the Department of Statistics at Iowa State University where she collaborates with the NIST-funded Center for Statistics and Applications in Forensic

Evidence (CSAFE) Center of Excellence. In 2017, she received her Ph.D. in Computational Science and Statistics from South Dakota State University. Her Ph.D. research was focused on finding statistically rigorous approximations to the value of evidence for use in the forensic identification of source problems. She previously received her B.S. in Mathematics in 2012 and her M.S. in Mathematics with an emphasis in Statistics in 2014, also from South Dakota State University. Prior to graduating, she participated in the ORISE Visiting Scientist Program with the Counterterrorism and Forensic Science Research Unit at the FBI Laboratory where she learned about the science behind trace element analysis and analysis of impression and pattern evidence. This program furthered her interested in the statistical analysis of trace elements from materials such as copper wire and aluminum powder, as well as the statistical analysis of handwriting. Dr. Ommen also served as a visiting researcher with the statistics group at the Netherlands Forensic Institute in 2016 where she worked to advance her knowledge of the relationship between the different statistical paradigms for forensic evidence interpretation. She has presented her research work at many national and international conferences and is the first recipient of the Stephen E. Fienberg CSAFE Young Investigator Award.

How to cite this article: M. Q. Johnson, and D. M. Ommen, *Handwriting identification using random forests and score-based likelihood ratios*, Stat. Anal. Data Min.: ASA Data Sci. J. (2021), 1–15. https://doi.org/10.1002/sam.11566

APPENDIX: ADDITIONAL FIGURES & RESULTS

This appendix contains additional figures and results, not central to the main body of the article.

Figure A1 shows the Gini importance for each of the distance measures used to train the common source random forest for each of the four different training sets. The figure shows that the Euclidean distance is the most important, followed by the absolute differences in proportions for various clusters. The figure shows that the absolute difference in proportions for clusters 31 and 34 are consistently among the highest rated Gini indices.

Figure A2 shows the cluster exemplar in red along with the cluster member graphs in gray for two clusters with high importance for scoring the difference between documents using the random forest. Cluster 34 contains many complex graphs with a greater range of variability and cluster 31 contains many simplistic graphs with a lower range of variability.

Figure A3 shows the estimated densities of similarity scores for the common source score-based likelihood ratios (SLRs) for each of the four training sets. The figure shows that the three longest prompts (the Wizard of Oz, London Letter, and combined) show promising discrimination between known same-writer and known different-writers comparisons. However, the separation of the two score densities for the short phrase is not adequate for computing effective common source SLRs. We believe this is because the number of clusters $K = 40$ is too large for such a short document.

Figure A4 shows the histogram of common source SLR values for the data in the four different testing sets. Similar to the conclusions from the previous figure, this figure shows that there is promising performance of the common source SLR to the three longer prompts. However, the performance on the short phrase is poor.

Figure A5 shows the distribution of three different specific source SLRs applying the pretrained random forest to get the scores for three different writers playing the role of the defendant. The figure shows poor performance with lots of overlap between the known same-writer and known different-writer values, corresponding to high rates of misleading evidence. Ideally, we would see the red distribution completely above 0 and the blue distribution completely below 0. This is almost the case for the trace-anchored SLR for writer 1045, which shows the best performance among all nine specific source SLRs using the pretrained random forest similarity scores.

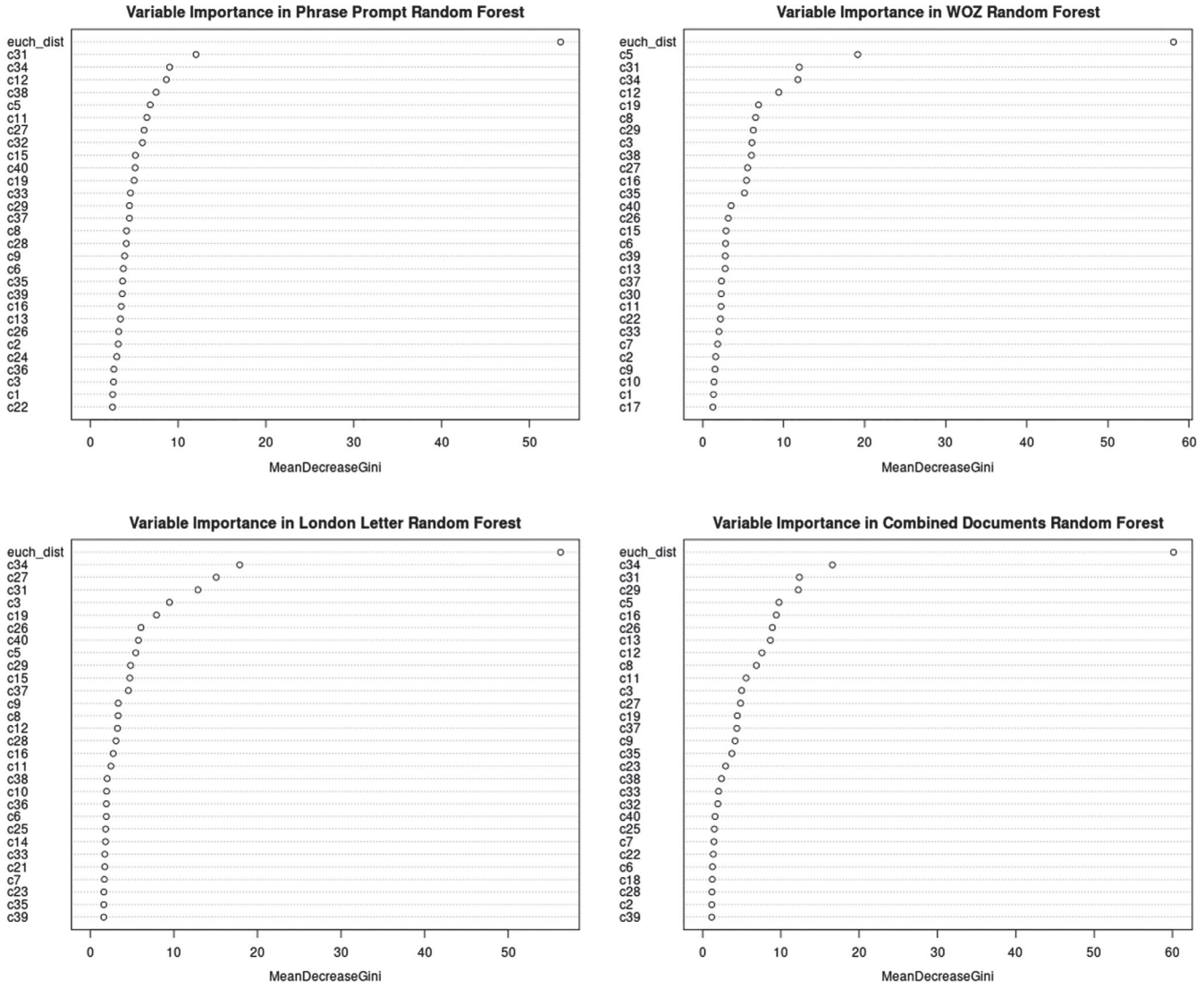

FIGURE A1 Gini variable importance in random forests by CSAFE handwriting data prompt. CSAFE, Center for Statistics and Applications in Forensic Evidence

FIGURE A2 Clusters 31 (left) and 34 (right) with high importance in classification via random forests

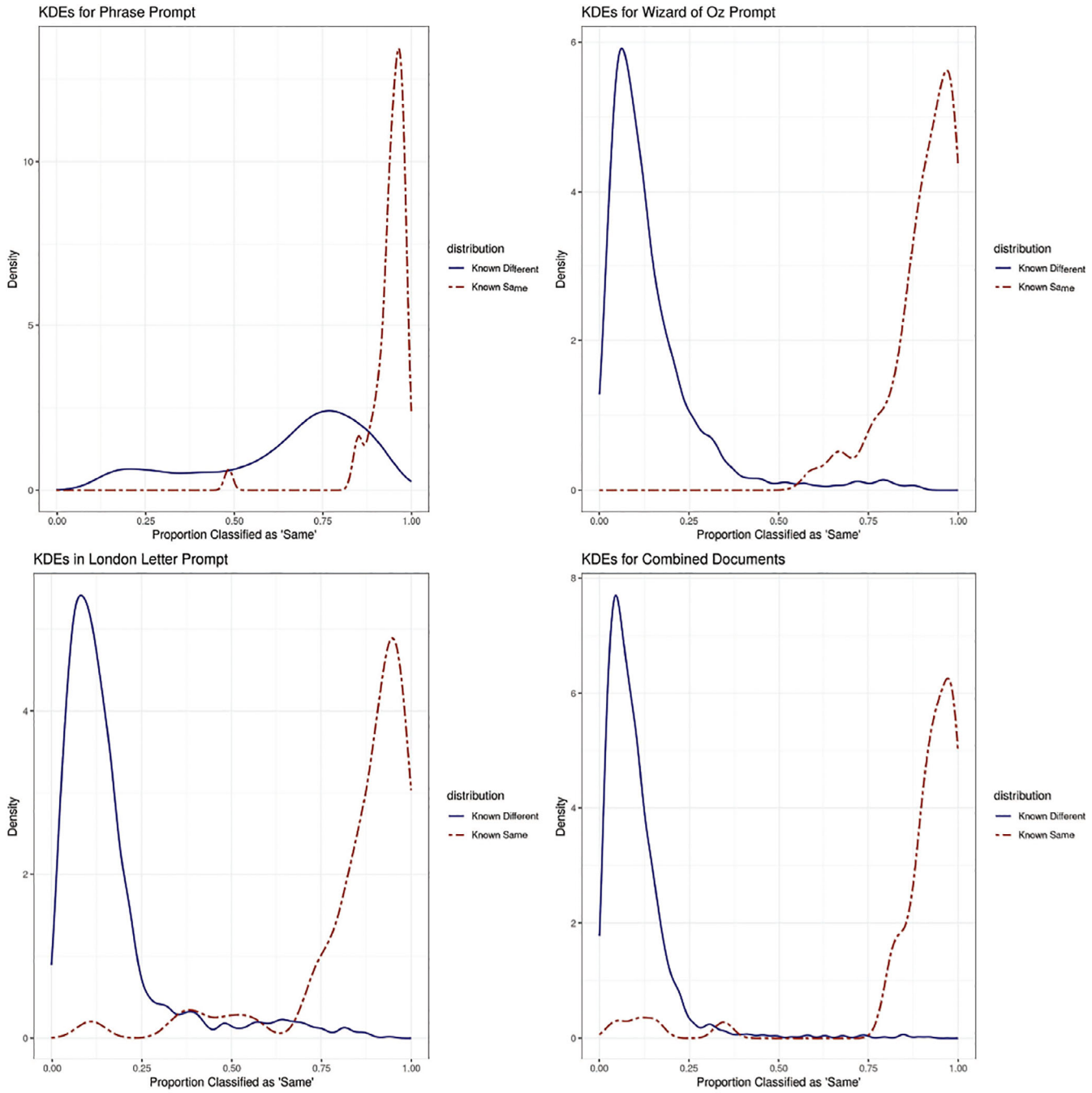

FIGURE A3 Kernel density estimated curves for common source SLR by CSAFE handwriting data prompt. CSAFE, Center for Statistics and Applications in Forensic Evidence; SLR, score-based likelihood ratio

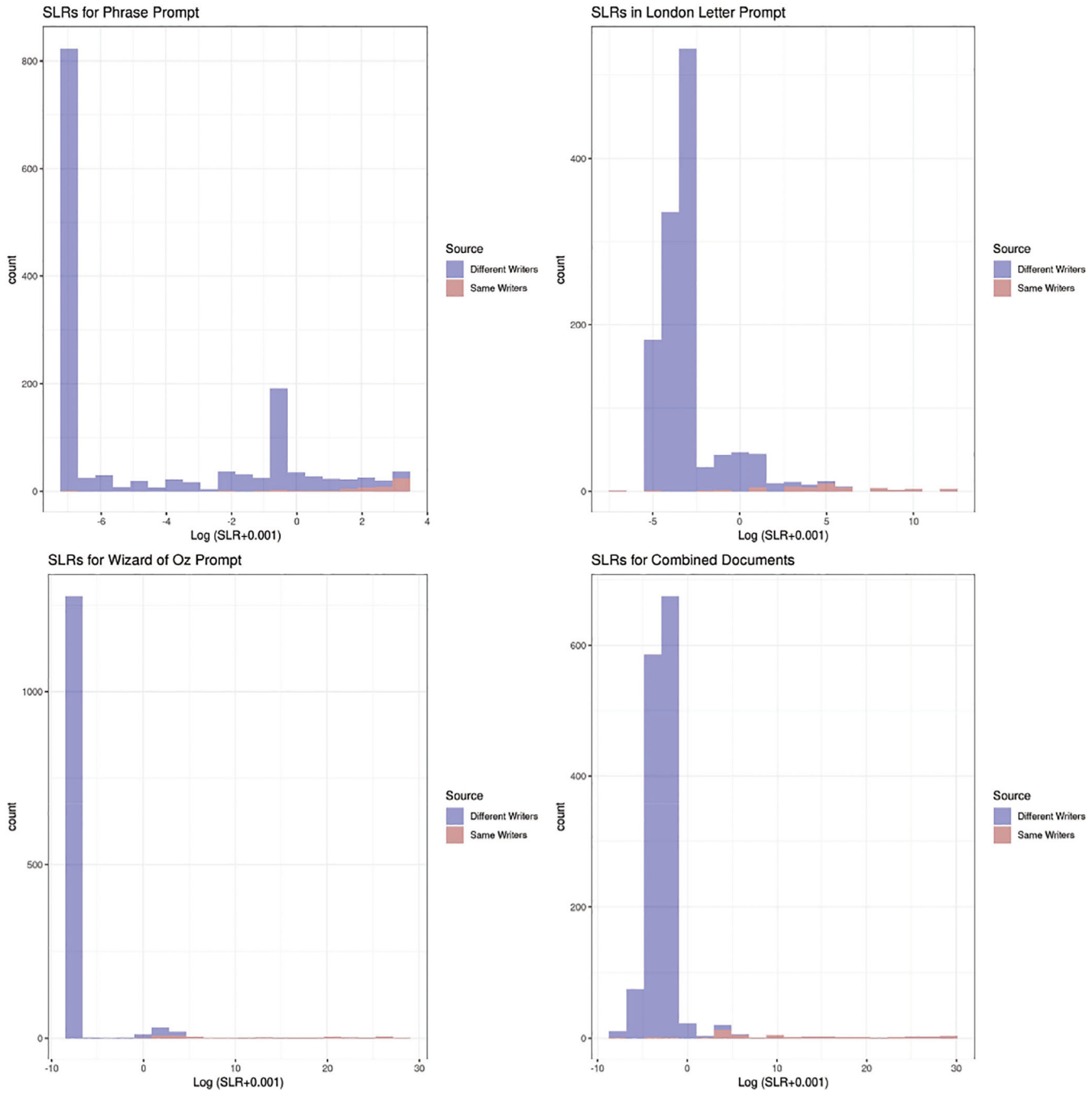

FIGURE A4 Log score-based likelihood ratios for common source by CSAFE handwriting data prompt. CSAFE, Center for Statistics and Applications in Forensic Evidence

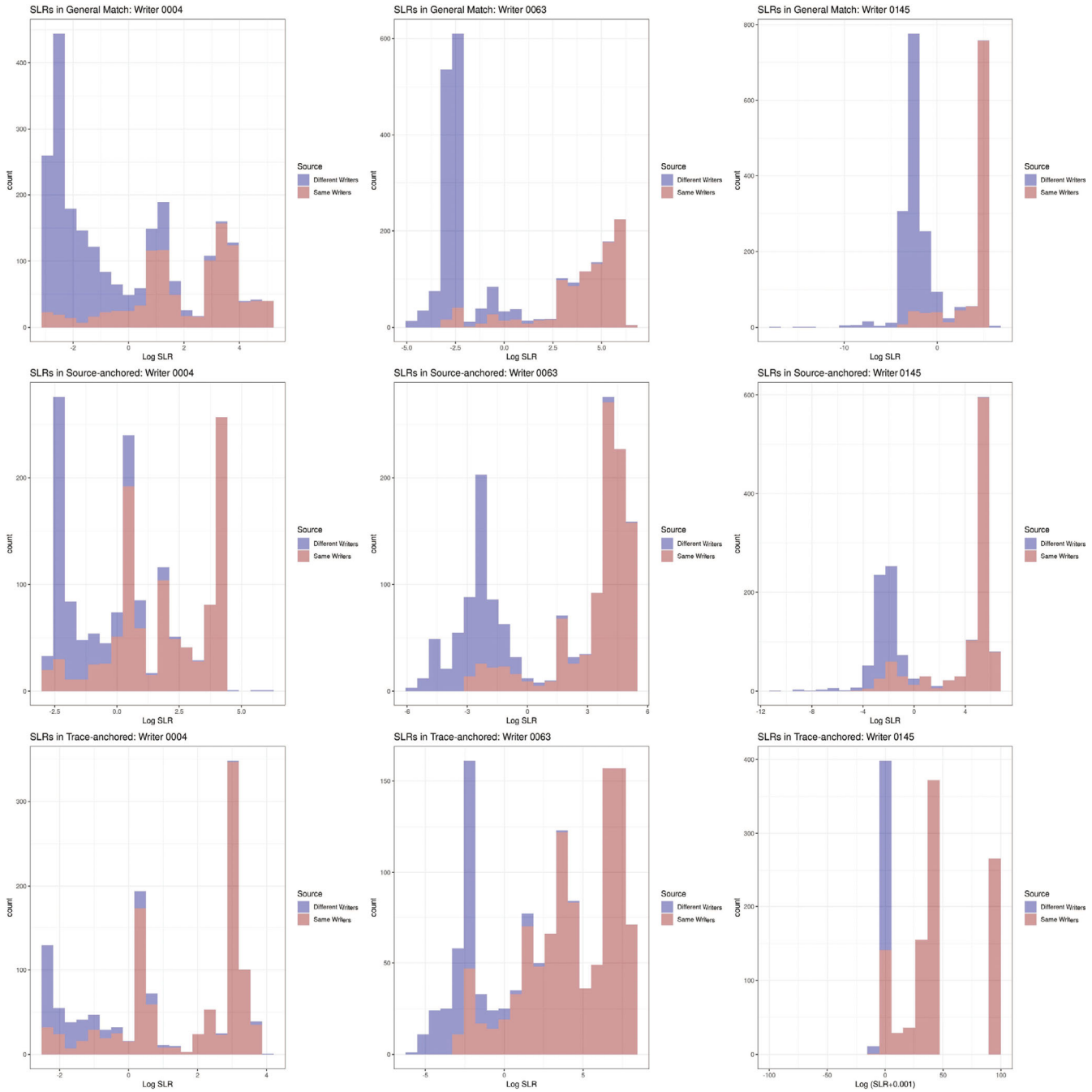

FIGURE A5 General-match, trace-anchored, and source-anchored specific source SLRs with pretrained random forest. SLRs, score-based likelihood ratios