Noname manuscript No.
(will be inserted by the editor)

Demographic-Adapted ROC Curve for Assessing Automated Matching of Latent Fingerprints

Emanuela Marasco*[1] · Mengling He [2] · Larry Tang [2] · Sumanth Sriram [1]

accepted: 02/25/2022

Abstract Although the diagnostic ability of a binary classifier system has been effectively assessed using a receiver operating characteristic (ROC) curve, the presence of covariates can affect the discriminatory capacity. This research investigates how automated tools used in forensics introduce demographic biases and discusses performance unfairness mitigation strategies. In our previous work, we evaluated the impact of demographic differentials in automatic matching of latent fingerprints and incorporated these covariates in the ROC curve. The resulting adjusted ROC curve provided error rates that account for an individual's demographic information, which is a better measure of the discriminatory capacity compared to the pooled ROC curve. Our ROC regression model was also able to handle continuous covariates such as age as well as discrete covariates such as gender and ethnicity. In this paper, we extend the preliminary study carried out on right index latent fingerprints to right thumb instances. We investigate: *i*) until which extent demographic differential vary depending on properties specific to the finger instance (*e.g.*, size of the fingertip); *ii*) the effectiveness of the proposed demographic adjusted-ROC to handle unfairness.

Emanuela Marasco Tel.: +1 (703) 993-5831
orcid: 0000-0003-3373-074X E-mail: emarasco@gmu.edu ·
Sriram Sai Sumanth Tel.: +1 (571) 286-9655
orcid: 0000-0002-4616-1112 E-mail: ssriram2@gmu.edu
Center for Secure Information Systems
Volgenau School of Engineering, George Mason University
4400 University Drive, Fairfax, Virginia 22030, USA

Mengling He Tel.: +1 (321) 347-9095
orcid: 0000-0002-3743-8765 E-mail: menglinghe@knights.ucf.edu ·
Larry Tang Tel.: +1 (407) 823-0638
orcid: 0000-0002-7276-155X E-mail: ltang1@ucf.edu
Department of Statistics and National Center for Forensic Science
4000 Central Florida Blvd, Orlando, FL 32816

Keywords Forensics · Demographic Differentials · Latent fingerprints · unfairness

1 Introduction

The problem of understanding and solving algorithmic bias is extremely important. Unfairness in automated decision making algorithms might reinforce societal discrimination and injustices. Scientists are currently confronting complex research questions about what it means to make an algorithm fair and they are debating about how to best measure algorithmic fairness [19,10]. Several measures for assessing algorithmic fairness have been recently discussed. The two most prominent types of measures focus on whether the scores generated by an algorithm should be equally accurate for members of legally protected groups instead on whether the error rates produced are equal [8]. The problem of mitigating unfairness of machine learning algorithms has been approached by introducing fairness before, during and after the training process [1]. In this regard, common countermeasures to discovered biases consist of exposing machines to more fresh data, feature engineering, algorithm selection, hyper-parameter optimization and retraining the machine to reduce or eliminate the biased outcome [15,20]. Some solutions focused on producing data that are less biased than the original dataset via generative models or data transformation [2]. Artificially generated data may not provide realistic examples for training.

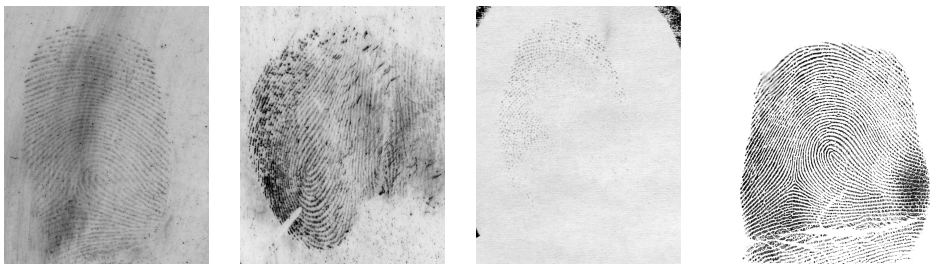
In biometric systems, biases can compromise the invariance of the features representing the identity of an individual [13,17]. In facial recognition systems, examining images from male subjects does not yield the same error rates as examining images from female subjects, and it is unclear whether the pooled error rates lead to the same error rates in subgroups of subjects with different demographics. Recent studies found that also fingerprint decision scores vary with covariates such as subjects' demographic information (e.g., age and gender). In this regard, several approaches for gender estimation from fingerprints have been explored [16,18,17]. Logistic regression models have been applied for this task and found that image quality features, such as ridge valley uniformity, minimum foreground pixels and minutiae counts, are significant covariates associated with gender [18].

In forensic science and criminal investigations, fingerprints are one of the most important sources of evidence left at a crime scene. It generally consists of latent prints that are lifted from surfaces of objects that are inadvertently touched or handled by a person, see Fig. 1 and Fig. 2. Fig. 1 shows sample images of the right index instance while Fig. 2 shows sample images of right thumb, both from same subject. The examination of a latent print involves a comparison of the latent print to a known (or exemplar) print. The FBI's Criminal Justice Information Services (CJIS) Division developed and incrementally integrated a new system the Next Generation Identification (NGI) to improve the effectiveness of latent fingerprint searches [14]. NGI provides

the criminal justice community with the world's largest and most efficient electronic repository of biometric and criminal history information. An examiner usually must do a full comparison while a demographic indicator may refine the search and narrow down the huge pool of candidates.

(a) (b) (c) (d)

Fig. 1 Sample images pertaining to the FBI Biometric Collection (BioCoP) Next Generation Identification Phase 1 (2008 - 2009) collected at West Virginia University: (a), (b) and (c) are latent fingerprint impressions of the right index with different quality, (d) is an exemplar used as an identity reference.



(a) (b) (c) (d)

Fig. 2 Sample images pertaining to the FBI Biometric Collection (BioCoP) Next Generation Identification Phase 1 (2008 - 2009) collected at West Virginia University: (a), (b) and (c) are latent fingerprint impressions of the right thumb with different quality, (d) is an exemplar used as an identity reference.

Latent fingerprints obtained from crime scenes have served as crucial evidence in forensic identification for more than a century [12]. Although accurate, the technology is still probabilistic [3]. Furthermore, when interpreting evidence in a forensic casework, results can be affected by factors such as the experience or expertise of the forensic examiner. Forensic examiners with different training and demographics may not result in the same error rates. More experienced examiners may tend to have higher confidence and lower individualization error rates than newly recruited examiners.

The Receiver Operating Characteristic (ROC) curve is a popular way to evaluate and compare the accuracy of classification markers when the out-

puts are continuous. A matching score is obtained by comparing two copies of the same biometric sample. The accuracy of any given threshold value can be measured by the probability of a true positive (sensitivity) and the probability of a true negative (specificity). The ROC curve is a plot of the sensitivity (Se(c)) versus 1-specificity (1-Sp(c)) over all possible threshold values (c) of the marker. Although ROC curves are very useful to assess fingerprint matching algorithms, the ROC estimation has rarely accounted for covariates from the known source. Whether a specific demographic covariate is associated with matching status, the pooled ROC curve needs to incorporate the proportion of discrimination attributable to the covariate. As covariates for source demographics vary with different known sources, accounting for them provides a way to utilize the background information which may yield more accurate evidence interpretation as more information is integrated.

This paper extends the validation experiments of ROC regression models incorporating covariates into the ROC curve on right thumb latent fingerprints. The resulting covariate-specific ROC curves provide the interpretation of error rates specific to the demographics of source subjects, and error rates which account for them. *The importance of the ROC regression is that it allows the estimation of the ROC curve for continuous covariates.* These ROC curves are compared with the pooled ROC curves commonly studied in the scientific literature. The significance of the focus on latent fingerprints pertains to generating a decision score by an algorithm automatically. Such a score is related to the subject and not to the examiner. The paper is organized as follows: Section 2 reviews research conducted on the analysis of demographics in fingerprint images, Section 3 describes the proposed approach, Section 4 presents the experimental results, Section 5 draws our conclusions and discusses future work.

2 Related Works

An important study on an operational fingerprint database by Yoon and Jain also found that fingerprint decision scores vary with subjects‘ demographic covariates [27]. Variations pertaining to fingerprint matching features over time were analyzed. They experimentally confirmed that, for a given individual, genuine match scores decrease over time, impostor scores do not significantly vary, and that the accuracy remains stable. Image quality was considered as being the best covariate to explain the changes in the genuine match scores. The data used in the study was collected from records of the Michigan State Police with TenPrint cards as acquisition type. The focus was on observing fingerprint matching features changes over time and their experiments are carried out to demonstrate this aspect. Their model considers genuine and impostors separately and focuses on regression only. Our investigation starts from finding out that gender, age and ethnicity impact automatic tools designed for identifying finger marks. Our model provides a more general approach applicable to any relevant demographic information available for each individual.

We first execute a linear regression based on match scores, next we analyse the covariates age, gender and ethnicity from these scores, and then we model them directly in the ROC curve producing covariates-specific ROC curves.

A dependence of fingerprint features on age group (individuals born at a similar time), gender (physical characteristics that distinguish males from females) and ethnicity (common culture and origin) has been reported [11, 7, 16]. Methods for demographics estimation from fingerprints have searched for gender clues in the ridge density structure that can be encoded by the local texture [16, 23]. Marasco *et al.* [18] recently applied logistic regression models to data acquired from four optical sensors. Gender was the response variable and NFIQ2 score the independent variable as well as four minutiae features including ridge valley uniformity, image contrast, minutiae count, and minimum foreground pixels. For fingerprints from L1 Identify Solutions, subjects with higher ridge valley uniformity ($p<.0001$), higher Image Contrast ($p=.023$) and higher foreground pixels ($p<.0001$) exhibited a higher chance of coming from female subjects. Similar results were observed for Cross Match Guardian while for the i3 digID, subjects with higher ridge valley uniformity ($p=.013$) and higher foreground pixels ($p<.0001$) had a higher chance coming from female subjects. Lower minutiae count ($p<.001$) was also associated with a higher chance of female subjects for all the sensors.

To the best of our knowledge, the proposed work is the first research effort that studies the impact of demographics on match scores generated by an automatic matcher designed for latent fingerprints. In 2018, Cao et al. presented the first fully automated latent identification system build by training a convolutional Autoencoder to detect the minutiae points and by a CNN-based minutiae descriptor extractor. The latent-to-reference comparison is then done using two algorithms, I.e., minutiae template comparison and texture template comparison [3]. The report CMC curves on 449 latents on the same WVU database used in our study by highlighting that the latents used for the experiments are dry with broken ridges that affects the enhancement model of their system. Furthermore, our experimental comparison does include standard ROCs for the scenarios under study.

3 The Proposed Approach

Source identification problems aim to determine the link between the known evidence (e.g., suspect) and an unknown evidence (e.g., evidence from the crime scene). The common-source problem relates to whether two evidences are from the same source. Errors in evidence interpretation are mainly related to individualization and exclusion decisions [24]. The term individualization means that, among all the possible suspects, the analyst confidently can state that the defendant is the source of the latent print. Exclusion indicates the determination by an examiner that two friction ridge impressions did not originate from the same source. When the decision scores from examiners or computer algorithms are ordinal or continuous, the accuracy can be assessed with

the ROC curve. The associated error rates are related to the decision on the following two propositions for how the evidence has arisen:

H_0: The unknown and known evidence are from different sources;
H_1: The unknown and known evidence are both from a common source.

Since the presentation by forensic scientists in courts has influence over decisions made by the judge and the jury, as consequences of these errors, an innocent person could be wrongly accused and a criminal could be mistakenly claimed innocent [5]. These error rates and the ROC curve are obtained by pooling all the decisions from examiners or computer algorithms with same-source or different-source prints. These measures report the average error rates across a population of examiners for evidence sources. Ideally, the error rates tend to provide guidance for the error rates for interpreting a given evidence source if error rates are consistent for sources with various aspects and for examiners with various background. Further research is necessary to identify the attributes of prints associated with false positive or false negative errors. ROC regression methodology offers the opportunity to investigate how factors such as characteristics of study subjects influence test accuracy [21]. A method for applying generalized ordinal regression models to categorical rating data to estimate and analyze ROC curves is presented. This model permit adjustment of ROC curve parameters for relevant covariates through two regression equations that correspond to location and scale [25]. The proposed model accounts for both sets of covariates such as source subjects' covariate information including their demographics and/or source images' attributes and quality, and potentially examiners' covariate information such as their training background and demographics.

We recall the definition of the ROC curve without covariates. Let T denote the real-valued random variable associated with a continuous biometric measurement (score) for assessing the above two propositions, and let further D be a $\{0,1\}$-valued status variable, with $D = 1$ indicating a genuine pair (H_1 is true), and $D = 0$ indicating an imposter pair (H_0 is true). Let

$$F_0(t) = P(T \leq t|D = 0), \tag{1}$$

and

$$F_1(t) = P(T \leq t|D = 1). \tag{2}$$

, denote the cumulative distribution functions of T conditional on $D = 0$ and $D = 1$, respectively, $t \in \mathbb{R}$. And let

$$S_j(t) = 1 - F_j(t),\ j \in \{0,1\}, \tag{3}$$

and

$$F_j^{-1}(u) := \inf\{t \in \mathbb{R} : F_j(t) \geq u\},\ j \in \{0,1\}, \tag{4}$$

denote the corresponding survivor and quantile functions, respectively.

Mathematically, without the covariates, the ROC curve can be defined by:

$$\mathrm{ROC}(u) = 1 - F_1(F_0^{-1}(1-u)), u \in (0,1) \tag{5}$$

where F_i denotes the cumulative distribution function (CDF) conditioning on the value of label, ROC(u) is the sensitivity at u, and u is the false positive rate (FPR).

3.1 Covariate-Specific ROC curve

Covariate-specific ROC curves model the covariate effects on the ROC curves and are commonly used in medical diagnostics [28]. Although rarely used in forensics, it provides a useful tool for the interpretation of error rates specific to the demographics of source subjects and covariates of examiners. Using the covariate information, the population of examiners are no longer considered as a homogeneous group with only random variation in their error rates. Instead, the error rates will account for examiners‘ training background and their demographics. Second, important covariates related to the evidence that impact the corresponding error rates will be identified and explored. Third, the covariate specific model provides an intuitive tool to incorporate background information on the known source. The weight of evidence with covariate information essentially account for more information from a case than the model without. Finally, quantification of the uncertainty will present forensic scientists with better understanding on the trade-off between variation and sample sizes involved in the error rate assessment.

Accounting for a set of covariates $X = (X_1, \ldots, X_p)^T$ that may represent subject demographics, information on quality of the measurement on the process, etc., it is often of interest to examine how the ROC curve varies conditional on observed covariates $X = \mathbf{x}$. We define the covariate-specific ROC curve by:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - F_{1,\mathbf{x}}(F_{0,\mathbf{x}}^{-1}(1-u)), \quad u \in (0,1) \tag{6}$$

where $F_{1,\mathbf{x}}(t) = P(T \leq t | D = 1, X = \mathbf{x})$, is the distribution of scores in the genuine group conditional on the covariates, and $F_{0,\mathbf{x}}^{-1}(u) \coloneqq \inf\{t \in \mathbb{R} : F_{0,\mathbf{x}}(t) \geq u\}$ is the quantile function of the imposter group conditional on the covariates[28].

In order to model the conditional distributions $F_{j,\mathbf{x}}(t)$, and $F_{j,\mathbf{x}}^{-1}(u)$, it is common to use linear regression models on genuine and imposter match scores as follows:

$$\mathrm{T}_j = \mu_j(\mathbf{x}) + \sigma_j(\mathbf{x})\epsilon_j, j \in \{0,1\}, \tag{7}$$

where the conditional mean and the conditional variance of T are $\mu_j(\mathbf{x}) = E(T|D = j, X = \mathbf{x})$ and $\sigma_j^2(\mathbf{x}) = var(T|D = j, X = \mathbf{x})$ given observed covariates $\mathrm{X} = \mathbf{x}$, respectively. And the error term ϵ_j is independent of $\mathbf{x}$. Then, for a given covariate $\mathbf{x}$, the covariate-specific ROC curve can be expressed as:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\mu_1(\mathbf{x}) - \mu_2(\mathbf{x})}{\sigma_1(\mathbf{x})}], \tag{8}$$

where $G_i(z) = P(z \leq \epsilon_j)$ is the distribution function of the the regression

error term which is independent of covariates for i $\in \{0, 1\}$. The derivation for this expression can be found in [21] and [25].

In this paper, the unknown source (i.e., the latent fingerprint evidence) is referred to as Query Q while the one pertaining to the known source as Reference R. Subsequently, the demographic covariate pertaining to them will adopt a similar nomenclature. For instance, the age variable pertaining to the unknown source is referred to as Age_Q while the one pertaining to the known source as Age_R. This scheme is also used for the gender and ethnicity covariates.

The match scores are modeled according to linear regression as follows:

$$\begin{aligned} Score =& \beta_0 + \beta_L * label + \beta_{11} * Age_Q + \beta_{12} * Age_R \\ &+ \beta_{21} * Gender_Q + \beta_{22} * Gender_R \\ &+ \beta_{L11} * Age_Q * label + \beta_{L21} * Gender_Q * label, \end{aligned} \quad \text{(Model A)}$$

where the variable *label* equals to '1' if the two images are from the same subjects, while it equals to '0' otherwise. Age_Q and Age_R indicate the age category of the two images being compared, while $Gender_Q$ and $Gender_R$ represent their gender category. Specifically, when the subject is male then gender equals to '1'.

In the regression model, we only consider the interaction term Age_Q*$label$ and $Gender_Q$*$label$ given that the case *label* is '1' indicates that the two images pertain to the same identity, subsequently, Age_Q equals Age_R as well as $Gender_Q$ equals $Gender_R$. Whether *label* corresponds to '0', the two interaction terms are also equal to '0'. Therefore, it is not meaningful to include four interaction terms in the model.

After the model above is estimated, the regression results can be used to compose the covariate-specific ROC curve in Eqn. 8, then the accuracy of the examiners or computer algorithms can be estimated.

The ROC curve based on Model A can be expressed as follows:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\beta_L + \beta_{L11} * Age_Q + \beta_{L21} * Gender_Q}{\sigma_1(\mathbf{x})}]. \quad (9)$$

We build models based only on one covariate. The model with the only variable, age, is given below:

$$Score = \beta_0 + \beta_L * label + \beta_{11} * Age_Q + \beta_{12} * Age_R + \beta_{L11} * Age_Q * label. \quad \text{(Model B)}$$

The model with only variable gender is defined below:

$$Score = \beta_0 + \beta_L * label + \beta_{21} * Gender_Q + \beta_{22} * Gender_R + \beta_{L21} * Gender_Q * label. \quad \text{(Model C)}$$

The corresponding ROC curves based on Model B and Model C can respectively be expressed as:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\beta_L + \beta_{L11} * Age_Q}{\sigma_1(\mathbf{x})}], \tag{10}$$

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\beta_L + \beta_{L21} * Gender_Q}{\sigma_1(\mathbf{x})}]. \tag{11}$$

We discuss the impact on the error rate assessment of the demographic ethnicity.

The model based only on the covariate ethnicity is expressed as follows [28]:

$$\begin{aligned} Score =& \beta_0 + \beta_L * label + \beta_{31} * Ethnicity_Q \\ &+ \beta_{32} * Ethnicity_R + \beta_{L31} * Ethnicity_Q * label. \end{aligned} \tag{Model D}$$

The ROC curve model can then be expressed as:

$$\mathrm{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\beta_L + \beta_{L31} * Ethnicity_Q}{\sigma_1(\mathbf{x})}]. \tag{12}$$

Finally, all the three demographics analysed in this study are incorporated in Model E as below:

$$\begin{aligned} Score =& \beta_0 + \beta_L * label + \beta_{11} * Age_Q + \beta_{12} * Age_R + \beta_{21} * Gender_Q \\ &+ \beta_{22} * Gender_R + \beta_{31} * Ethnicity_Q \\ &+ \beta_{32} * Ethnicity_R + \beta_{L11} * Age_Q * label \\ &+ \beta_{L21} * Gender_Q * label + \beta_{L31} * Ethnicity_Q * label. \end{aligned} \tag{Model E}$$

And the corresponding ROC curve becomes:

$$\begin{aligned} \mathrm{ROC}_{\mathbf{x}}(u) =& 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - (\beta_L + \beta_{L11} * Age_Q + \beta_{L21} * Gender_Q \\ &+ \beta_{L31} * Ethnicity_Q)/\sigma_1(\mathbf{x})]. \end{aligned} \tag{13}$$

4 Experimental Results

4.1 Dataset

The dataset used in this study is a subset of the FBI Biometric Collection of People (BioCoP) Next Generation Identification Phase 1 (2008 - 2009) [9]. The data collection involved the acquisition of latent-deposited fingerprints on common materials as well as standard ink and paper methods. The ink and paper data was used as an exemplar set for both electronic capture performed using BioCOP and the latent substrate capture. Each scanned image is saved as a grayscale type image with a resolution of 1000 ppi.

Subset	Subjects	Samples per Subjects
Latent F01	456	456 x 456
Latent F02	213	213 x 213
Rolled F01	456	456 x 456
Rolled F02	213	213 x 213

Table 1 The total number of subjects and samples from the latent and rolled, right index and right thumb.

These fingerprints were obtained at West Virginia University (WVU) and belong to a total of 1504 people. For this study, a subset of 1135 subjects was selected, of which 219 subjects have latent right index fingerprints, and 456 subjects have latent right thumb fingerprints. There are a total of 12 outliers, 6 outliers each for the latent right index and right thumb fingerprints, respectively. Coming to the rolled fingerprints, 2 sets of images (SET1, SET2) are collected from each of the 1135 subjects, but only the rolled fingerprint images from SET1 are used in this research. There was a nearly equal amount of male to female participants with 52% to 48% ratio. Also, among the participants, the age group between 18-29 was highest accounting for 74% percent of people, 8% between 30-39 years old, 7% between 40-49 years old and 11% above 50 years old. Among the ethnicity, Caucasians accounted for 79% of the people, only 6.2% Asian, 3.8% Asian Indian, 3.7% African American, 2.4% African, 2.1% Hispanic. The number of scores generated from the subjects are as follows: for the right index, since there are 213 subjects, the number of genuine scores generated is 213, while the number of impostor scores is 45,156 (=213 x 212), and for the right thumb, since there are 456 subjects, the number of genuine scores will be 456 and the number of impostors will be 2,07,480 (=456 x 455). Because the number of impostors is so high, the data is divided into five groups, with the first four groups each containing 100 subjects and the fifth group containing 56 subjects, resulting in a total of 39,600 (=4 x 100 x 99) impostors from the four groups and 3,080 (=56 x 55) impostors from the fifth group, resulting in a total of 42,680 impostors and 456 genuine scores from 456 subjects.

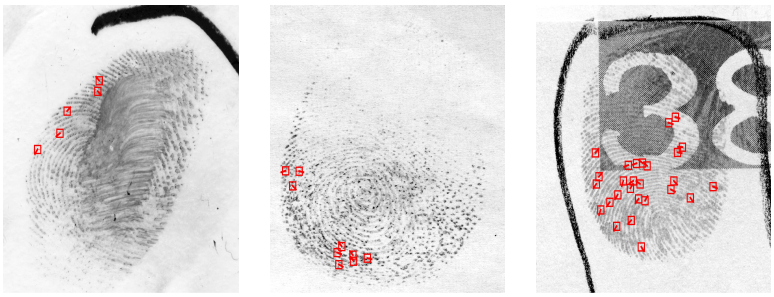
The experiments carried out in this paper consider right index as well as right thumb instances. The number of genuine and the number of impostors for the right thumb and right index are shown in Table4.1.

Finger	Number of Genuine	Number of Impostors	
Right Index	213	45,156 (213 x 212)	
Right Thumb	456	Group 1	9,900 (100 x 99)
		Group 2	9,900 (100 x 99)
		Group 3	9,900 (100 x 99)
		Group 4	9,900 (100 x 99)
		Group 5	3,080 (56 x 55)
		Total	42,680

Table 2 The total number of genuine and impostor images of the Right Index and Thumb fingers.

The latent fingerprints collection was carried out by gloving the subject‘s hands with nitrile gloves that induce sweating required for the development of the first latent fingerprints. Three quality sets were needed, good, bad and ugly, so that three whole or partial impressions for each finger were made on each of the substrates. Three different substrates were used: paper, plastic, and glass/porcelain. The items were separated based on substrate type and processed in one of three ways: *i)* chemical (ninhydrin) processing, *ii)* cyanoacrylate processing, *iii)* lift cards (processed with black fingerprint powder at the collection site). All fingerprints processed with cyanoacrylate were digitally photographed, while all ninhydrin and black powder fingerprints were scanned.

The FBI 2008 Biocop dataset includes partial, occluded, and distorted latent impressions. The latent fingerprints generated in this data collection are realistic and reflect very well the challenges pertaining to the evidence found in the crime scene, thus the proposed assessment as well.



(a) (b) (c)

Fig. 3 Sample images pertaining to the FBI Biometric Collection (BioCoP) Next Generation Identification Phase 1 (2008 - 2009) collected at West Virginia University: (a) distorted (b) partial and (c) occluded latent fingerprint impressions.

4.2 Latent-to-reference print comparison

The match score obtained by comparing two images pertaining to the same source is referred to as genuine score; while, the score generated by comparing images pertaining to two different sources is referred to as impostor score. Latent print studies mainly analyze examiners‘ binary decisions and discuss error rates including false positive rate (FPR) indicating the probability of incorrect individualization on impostor pairs and false negative rates (FNR) corresponding to the probability of incorrect exclusion from the same source [26]. In 2018, Cao *et al.* presented the first fully automated latent identification system build by training a convolutional autoencoder to detect the minutiae points and by a CNN-based minutiae descriptor extractor. The match scores were obtained using this tool, an end-to-end latent fingerprint search system recently made

publicly available by Anil Jain's research group at Michigan State University [3]. This algorithm executes automated region of interest (ROI) cropping, latent image pre-processing, feature extraction, feature comparison, and outputs a candidate list. Two separate minutiae extraction models provide complementary minutiae templates. Furthermore, to compensate for poor quality latents, an additional texture template is also generated. Each reference fingerprint template consists of one minutiae template and one texture template. Two matchers, i.e., minutiae template matcher and texture template matcher are used for comparison between the query latent and reference prints. Three latent minutiae templates are compared to one reference minutiae template, and the latent texture template is compared to the reference texture template. Four comparison scores are fused to generate the final comparison score.

Cao's paper reports CMC curves on 449 latents from the same WVU database used in our study, pointing out that the latents used in the experiments are dry with broken ridges, which affects their system's enhancement model based on training an autoencoder [3]. Fig.4 shows sample images dry latent prints. As a result, the majority of dry latents are failures of the automatic fingerprint identification system.

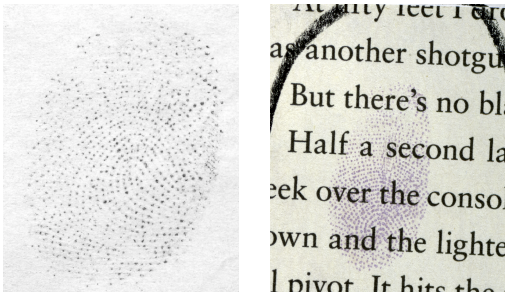


Fig. 4 Examples of dry latent fingerprints featured by broken ridges from the WVU database.

4.2.1 Baseline

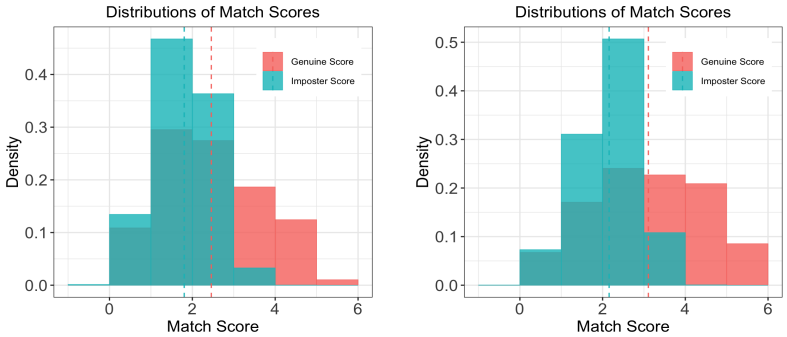
In this paper, we did a further cleaning of the data by removing the match scores generated from samples with no valid minutiae points. The total number of observations deleted from the right index and right thumb fingerprints are shown in Table4.2.1

Fig. 5 (a) and (b) illustrate the distributions of the genuine and impostors match scores for the instances right index and right thumb, respectively. For a specific threshold, the FPR is the percentage of the impostor scores greater than the threshold in the non-genuine pairs, and the FNR is the percentage of genuine scores less than or equal to the threshold in the genuine pairs. As the threshold increase, the FPR decreases while the FNR increases. Error rates

Exclusion	Right Index		Right Thumb	
	imposter	genuine	imposter	genuine
negative and zero score	4522	18	76	1
NA	0	0	1449	4
outliers	101	2	37	6
unknown demographics	37	0	0	0

Table 3 Details on the Cleaning of the data.

such as the FPR and the FNR are calculated for all the possible thresholds and the accuracy is assessed by the receiver operating characteristic (ROC) curve reported in Eqn. 6 which can also be summarized using the area under the ROC curve (AUC)[21].



(a) (b)

Fig. 5 Distributions of the Genuine and Impostor Match Scores: (a) Right Index and (b) Right Thumb.

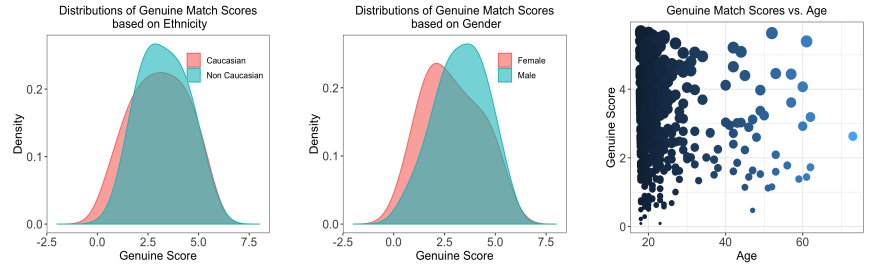
4.3 Results

Fig.6 (a) shows how ethnicity affects the genuine match scores for the right index instance. The class Non-Caucasian appears not to be able to reach the highest values of genuine scores as it happens for the Caucasian population. Fig.7 (a) illustrates the impact of ethnicity on the genuine match scores for the right thumb instance. We can notice that genuine match scores are generally higher for the Non-Caucasian population.

Fig.6 (b) illustrates the distributions of the genuine match scores with respect to gender for the right index instance. Male subjects exhibit higher values of genuine match scores compared to female subjects. Fig.7 (b) illustrates the distributions of the genuine match scores with respect to gender for the right thumb instance. The gender differential is prominent, probably due to a bigger size of the fingertip in thumbs that positively impacts the quality

(a) (b) (c)

Fig. 6 Distributions of Genuine Match Scores of the right index finger with respect to the demographic covariates considered in this study: (a) Density Plot of the Genuine Match Scores by Ethnicity, (b) Density Plot of the Genuine Match Scores by Gender, and (c) Scatter Plot of the Genuine Scores vs. Age.



(a) (b) (c)

Fig. 7 Distributions of Genuine Match Scores of the right thumb finger with Respect to the Demographic Covariates Considered in this Study: (a) Density Plot of the Genuine Match Scores by Ethnicity, (b) Density Plot of the Genuine Match Scores by Gender, and (c) Scatter Plot of the Genuine Scores vs. Age.

of the information processed by the automated matcher such as texture and minutiae.

Fig.6 (c) shows the scatter plot of the genuine match scores with respect to age values for the right index instance. We can see that younger individuals exhibit higher values of genuine match scores. Fig.7 (c) illustrates the scatter plot of the genuine match scores with respect to age values for the right thumb instance. For the 20 years old individuals, the trend obtained with right thumb looks similar to right index; instead, for the subjects in the range of 20-30 years old using right thumb leads to higher genuine scores than those generated by comparing images of right index.

4.3.1 ROC curves

We examine how the ROC curve varies conditioning on observed covariates by reporting the regression results for the three covariate-specific ROC models described in Section 3. Various classifiers are capable of predicting the likelihood of a sample belonging to a class. A probabilistic classifier is implemented

by setting a threshold which divides the entire data into different classes. The results shown in Table 4 and 5 were used to compute the Sensitivities of the right index and the right thumb fingers.

	Model A	**Model B**	**Model C**	**Model D**	**Model E**
Intercept	2.189	1.778	1.738	1.734	1.623
label	0.805	0.589	0.494	0.498	0.206
Age_Q	-0.006	-0.001	-	-	-0.0008
Age_R	0.002	0.002	-	-	0.002
$Gender_Q$	0.060	-	0.008	-	0.015
$Gender_R$	0.084	-	0.099	-	0.100
$Ethnicity_Q$	-	-	-	0.078	0.079
$Ethnicity_R$	-	-	-	0.011	0.020
Age_Q*label	-0.0007	0.0026	-	-	0.004
$Gender_Q$*label	0.296	-	30.259	-	0.280
$Ethnicity_Q$*label	-	-	-	0.201	0.235

Table 4 Result of the designed covariate-specific ROC Models with respective right index fingers.

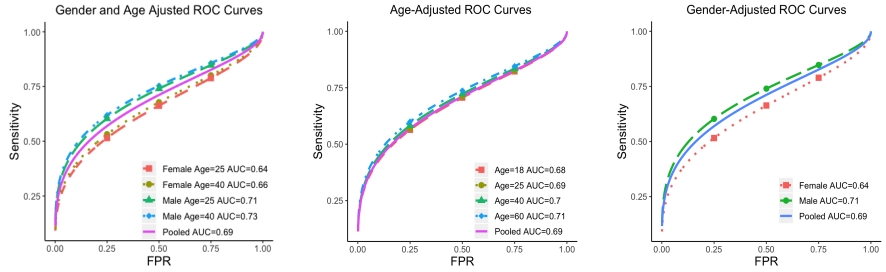
	Model A	**Model B**	**Model C**	**Model D**	**Model E**
Intercept	1.710	2.295	2.071	2.231	2.266
label	0.410	1.015	0.787	1.033	0.880
Age_Q	-0.001	-0.007	-	-	-0.007
Age_R	0.002	0.001	-	-	0.002
$Gender_Q$	0.008	-	0.077	-	0.052
$Gender_R$	0.099	-	0.080	-	0.085
$Ethnicity_Q$	-	-	-	-0.091	-0.092
$Ethnicity_R$	-	-	-	-0.006	0.004
Age_Q*label	0.003	-0.003	-	-	-0.001
$Gender_Q$*label	0.262	-	0.297	-	0.291
$Ethnicity_Q$*label	-	-	-	-0.111	-0.085

Table 5 Result of the designed covariate-specific ROC Models with respective right thumb finger.

Fig.8 (a) compares ROC curves of Model A considering both gender and age covariates for the right index instance. Fig.9 (a) compares ROC curves of Model A considering both gender and age covariates for the right thumb instance.

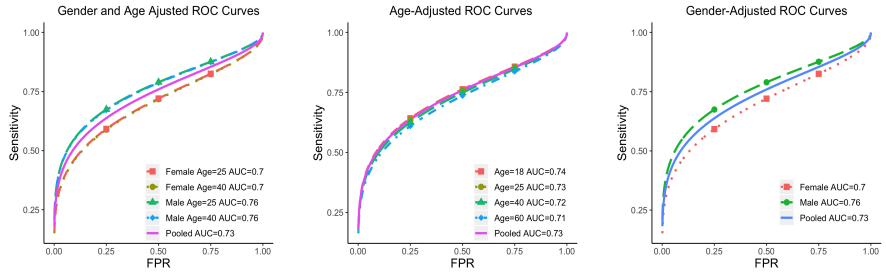
Fig.8 (b) compares ROC curves of Model B with only the covariate age for the right index instance. We can observe that as age increases, the model's identifying ability increases at a modest trend. Fig.9 (b) compares ROC curves of Model B with only the covariate age for the right thumb instance. In contrast to the right index finger, as age increases, the model's identifying ability decreases for the right thumb.

Fig.8 (c) compares ROC curves of Model C with only the covariate gender for the right index instance and Fig.9 (c) compares ROC curves of Model



(a) (b) (c)

Fig. 8 Comparison between the proposed covariate-specific ROC curves conditioned on right index demographic covariates: (a) Conditional ROC Curve corresponding to Model A, (b) Conditional ROC Curve corresponding to Model B, and (c) Conditional ROC Curve corresponding to Model C.



(a) (b) (c)

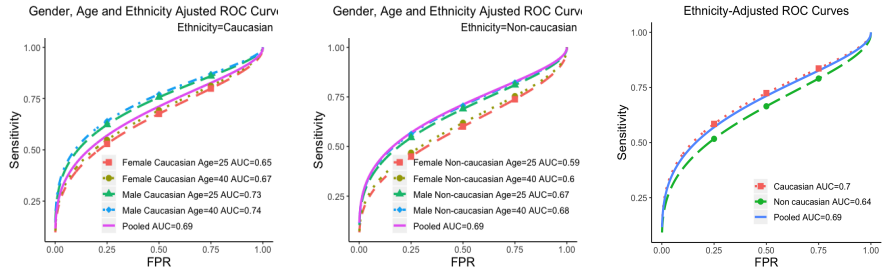
Fig. 9 Comparison between the proposed covariate-specific ROC curves conditioned on right thumb demographic covariates: (a) Conditional ROC Curve corresponding to Model A, (b) Conditional ROC Curve corresponding to Model B, and (c) Conditional ROC Curve corresponding to Model C.

C with only the covariate gender for the right thumb instance. These curves show that the models perform significantly better for male subjects for both instances.

It is observed that, the results obtained from the integration of both age and gender covariates are consistent with their corresponding univariate models for the right index and the right thumb instances. For male subjects, the model performs significantly better in source identification problems, and for elder subjects, the performance seems to be further slightly improved for the right index instance, on the other hand for the right thumb instance, the performance seems to be the same for the elder subjects. Nevertheless, pooling the data regardless of the values of the covariate yields a ROC curve that is below the specific ROC curves for each of the populations determined by a given demographic covariate.

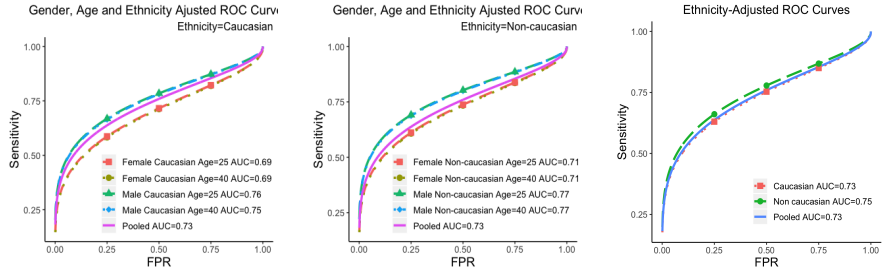
Similarly, Fig.10 (c) compares ROC curves conditioned only on the ethnicity covariate for the right index instance. Fig.11 (c) compares ROC curves conditioned only on the ethnicity covariate for the right thumb instance . It is observed that, the performance for Non-Caucasian subjects is worse than the

Caucasian ones in the right index instance, where the performance of the Caucasians is slightly less than the Non-Caucasian subjects for the right thumb instance. The same trends can also be seen in Fig.10 (a) and Fig.11 (a) that compares ROC curves considering both gender and age covariates for the Caucasian population as well as in Fig.10 (b) and Fig.11 (b) that compares them for the Non-Caucasian population.



(a) (b) (c)

Fig. 10 Comparison between the proposed covariate-specific ROC curves conditioned on the right index demographic covariates observed in this study, including ethnicity.



(a) (b) (c)

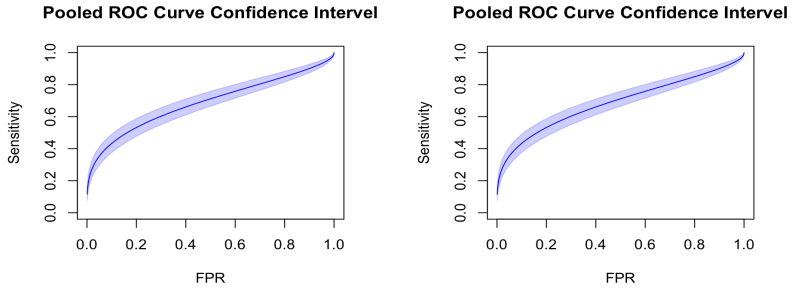
Fig. 11 Comparison between the proposed covariate-specific ROC curves conditioned on the right thumb demographic covariates observed in this study, including ethnicity.

4.3.2 Variability of ROC curves

In order to investigate the sampling variabilities of the estimated ROC curve and its AUC, we calculate the variances and the confidence intervals (CIs) of the ROC curves.

The Delong's method provides the explicit formula for the variance of the AUC for the pooled ROC curve without the consideration of covariates [6]. The results can be obtained from the R Shiny application developed by the authors, which is available upon request [22]. Based on the results of the Shiny

application, the CIs of the AUC for the pooled ROC curve are (0.6108, 0.7055) with its variance = 0.0006, and (0.6827,0.7473) with its variance = 0.0003 for the right index and right thumb, respectively. Figure 12 shows the pooled ROC curves and its confidence intervals at various FPRs.



(a) (b)

Fig. 12 Pooled ROC Curve Confidence Interval: (a) Right Index and (b) Right Thumb.

The bootstrap method is used to estimate the variance of covariate-specific ROC curves. The regression models are run for 1000 bootstrap samples drawn from the original data set, and then the variance of TPR at each FPR point is calculated based on the parameter estimates from each model. For each model with different covariate values, the confidence intervals of AUCs are shown in Table 6. For each model with specified covariates, the confidence interval (CI) for the ROC curve can also be obtained using the bootstrap method. Figure 13 shows for Model C that the confidence interval of the ROC curve is wider for female than male for both right index and right thumb data, while the algorithm performs much better in right thumb data with a larger AUC value and smaller variance.

Model	Covariate	Right Index		Right Thumb	
		Variance	Confidence Interval	Variance	Confidence Interval
Model A	Age=25 Gender=Female	0.0013	(0.5705, 0.7095)	0.0006	(0.6535, 0.7465)
	Age=40 Gender=Female	0.0026	(0.5600, 0.7600)	0.0010	(0.6377, 0.7623)
	Age=25 Gender=Male	0.0007	(0.6574, 0.7626)	0.0003	(0.7272, 0.7928)
	Age=40 Gender=Male	0.0024	(0.6331,0.8269)	0.0009	(0.7017,0.8183)
Model B	Age= 18	0.0007	(0.6276, 0.7324)	0.0003	(0.7050, 0.7750)
	Age= 25	0.0005	(0.6476, 0.7324)	0.0002	(0.7009, 0.7591)
	Age= 40	0.0022	(0.6074, 0.7926)	0.0008	(0.6657, 0.7743)
	Age= 60	0.0088	(0.5265, 0.8935)	0.0032	(0.5992, 0.8208)
Model C	Gender= Female	0.0012	(0.6573, 0.7627)	0.0006	(0.7278,0.7922)
	Gender=Male	0.0007	(0.5718, 0.70822)	0.0003	(0.6538, 0.7462)
Model D	Ethnicity= Caucasian	0.0006	(0.6512, 0.7488)	0.0003	(0.6959, 0.7641)
	Ethnicity= Non-caucasian	0.0017	(0.5596, 0.7204)	0.0006	(0.7037, 0.7963)
Model E	Age=25 Gender=Female Ethnicity= Caucasian	0.0013	(0.5361, 0.7039)	0.0006	(0.6284, 0.7716)
	Age=40 Gender=Female Ethnicity= Caucasian	0.0028	(0.5669, 0.7731)	0.0011	(0.6238, 0.7562)
	Age=25 Gender=Male Ethnicity= Caucasian	0.0009	(0.6698, 0.7902)	0.0004	(0.7217, 0.7983)
	Age=40 Gender=Male Ethnicity= Caucasian	0.0027	(0.6381, 0.8419)	0.0011	(0.6862, 0.8138)
	Age=25 Gender=Female Ethnicity= Non-caucasian	0.0034	(0.4761, 0.7039)	0.001	(0.6484, 0.7716)
	Age=40 Gender=Female Ethnicity= Non-caucasian	0.0043	(0.4709, 0.7291)	0.0013	(0.6387, 0.7813)
	Age=25 Gender=Male Ethnicity= Non-caucasian	0.0017	(0.5884, 0.7516)	0.0005	(0.7252, 0.8148)
	Age=40 Gender=Male Ethnicity= Non-caucasian	0.0032	(0.5696, 0.7904)	0.0010	(0.7069, 0.8331)

Table 6 The CIs of AUCs for each model with different covariates.

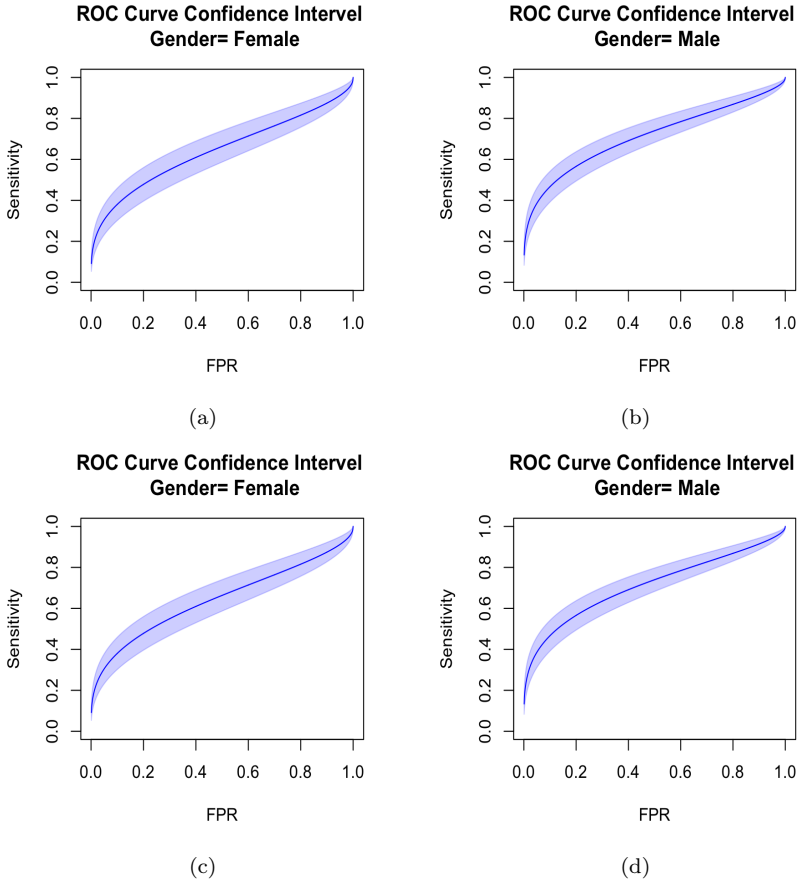


Fig. 13 Gender-specified ROC curve Confidence Interval: (a&b) Right Index and (c&d) Right Thumb.

The study conducted in this paper and the related findings pertain to the specific matcher used for the extraction of the similarity scores. As also previously pointed out by its designers, the identification system used in this work may not be robust to dry fingerprints which results in several missed minutiae points.

5 Conclusions

In this paper, we investigate the impact of demographic information on automatic tools for matching latent fingerprints to the corresponding references with a specific focus on the right index and right thumb fingers. We apply ROC regression techniques to genuine and impostor scores to take into account demographic covariates in the assessment of the identification technology for

a more accurate and less biased estimation of the error rates. The proposed methodology is general and can be extended to sensor-based fingerprint matching as well as to other biometric data. An inherent limitation of our approach consists in assuming the availability of the covariates for each subject. Whether the age and gender are unknown, the model cannot be applied in a real latent fingerprint matching scenario. To conclude, we found that in both univariate and multivariate models, the models perform better on male subjects for both the instances, whereas differing results were observed when the covariates age and ethnicity are considered.

Future research efforts will: *i)* Investigate the causes related to the demographic differentials analysed in this paper; *ii)* Implement score-level fusion strategies to combine confidence scores output by the individual instances right index and right thumb for performance enhancement and for further studying the overall fairness achieved; *iii)* Integrated quality measures extracted using LFIQ to maximize the accuracy of the multi-instances decision system; *iv)* Explore the use of different existing matching algorithms designed for latent fingerprints that may be more robust to the dryness of the fingerprint although the matcher may not be fully automated; *v)* Investigate Fuzzy Logic ROC curves [4].

6 Acknowledgments

The authors thank Dr. Anil Jain at the Michigan State University for the latent fingerprints matcher. This work was funded by the NIJ grant #2019-DU-BX-0011. The contribution of S. Sriram to this work was related to the extraction of match scores using the MSU tool.

7 Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Amini, A., Soleimany, A.P., Schwarting, W., Bhatia, S.N., Rus, D.: Uncovering and mitigating algorithmic bias through learned latent structure. AAAI/ACM Conference on AI, Ethics, and Society pp. 289–295 (2019)
2. Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K., Varshney, K.: Optimized preprocessing for discrimination prevention. Advances in Neural Information Processing Systems pp. 3992–4001 (2017)
3. Cao, K., Nguyen, D., Tymoszek, C., Jain, A.: End-to-End Latent Fingerprint Search. arXiv preprint arXiv:1812.10213 (2018)
4. Castanho, M.J.P., Barros, L.C., Yamakami, A., Vendite, L.L.: Fuzzy receiver operating characteristic curve: An option to evaluate diagnostic tests. IEEE Transactions on Information Technology in Biomedicine **11**(3), 244–250 (2007). https://doi.org/10.1109/TITB.2006.879593

5. Cole, S.A.: More than Zero: Accounting for Error in Latent Fingerprint Identification. J. Crim. l. & Criminology **95**, 985 (2004)
6. Elizabeth R. DeLong, D.M.D., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. International Biometric Society **44**(3) (1988)
7. Gnanasivam, P., Muttan, D.S.: Estimation of age through fingerprints using wavelet transform and singular value decomposition. International Journal of Biometrics and Bioinformatics (IJBB) **6**(2), 58–67 (2012)
8. Hellman, D.: Measuring Algorithmic Fairness. Va. L. Rev. **106**, 811 (2020)
9. Hornak, L., LaRue, W., Cukic, B., Ross, A., Morris, K., Dawson, J., Crihalmeanu, S., Kalka, N., Kayal, N.: FBI Biometric Collection of People (BioCoP): Next Generation Identification Phase 1 (2008 - 2009). 2008 Biometric Collection Project 08-06-2008 to 12-31-2009 FINAL REPORT (2009)
10. Hutchinson, B., Mitchell, M.: 50 Years of Test (un) fairness: Lessons for Machine Learning. Proceedings of the Conference on Fairness, Accountability, and Transparency pp. 49–58 (2019)
11. Huynh, C., Brunelle, E., Halamkova, L., Agudelo, J., Halamek, J.: Forensic identification of gender from fingerprints. Analytical chemistry **87**(22), 11531–11536 (2015)
12. Jain, A.K., Feng, J.: Latent Fingerprint Matching. IEEE Transactions on pattern analysis and machine intelligence **33**(1), 88–100 (2010)
13. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology **14**(1), 4–20 (2004)
14. Lipowicz, A.: Fbi's new fingerprint id system is faster and more accurate, agency says – gcn. Government Computer News (2011)
15. Lu, Y., Guo, H., Feldkamp, L.: Robust neural learning from unbalanced data samples. International Joint Conference on Neural Networks **3**, 1816–1821 (1998)
16. Marasco, E., Lugini, L., Cukic, B.: Exploiting Quality and Texture Features to Estimate Age and Gender from Fingerprints. SPIE Defense and Security (2014)
17. Marasco, E.: Biases in fingerprint recognition systems: Where are we at? IEEE Biometrics: Theory, Applications and Systems - Special Session on Generalizability and Adaptability in Biometrics (BTAS-SS GAPinB)
18. Marasco, E., Cando, S., Tang, L., Tabassi, E.: Cross-sensor evaluation of textural descriptors for gender prediction from fingerprints. 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) pp. 55–62 (2019)
19. Mehrabi, N., Morstatterd, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
20. Miller, C.: When algorithms discriminate. The New York Times **9**, 2015 (2015)
21. Pepe, M.: An Interpretation for the ROC Curve and Inference using GLM Procedures. Biometrics **56**(2), 352–359 (2000)
22. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013), `https://menglingheshiny.shinyapps.io/orderROCweb/`
23. Rattani, A., Chen, C., Ross, A.: Evaluation of texture descriptors for automated gender estimation from fingerprints pp. 764–777 (2014)
24. Ray, E., Dechant, P.: Sufficiency and Standards for Exclusion Decisions. Journal of Forensic Identification **63**(6) (2013)
25. Tosteson, A., Begg, C.: A General Regression Methodology for ROC Curve Estimation. Medical Decision Making **8**(3), 204–215 (1988)
26. Ulery, B., Hicklin, R., Buscaglia, J., Roberts, M.: Accuracy and Reliability of Forensic Latent fingerprint decisions. Proceedings of the National Academy of Sciences **108**(19), 7733–7738 (2011)
27. Yoona, S., Jain, A.: Longitudinal study of fingerprint recognition. Proceedings of the National Academy of Sciences **112**(28), 8555–8560 (2015)
28. Zhou, X., McClish, D., Obuchowski, N.: Statistical Methods in Diagnostic Medicine. John Wiley & Sons (2009)