Quality- and Demographic-Specific ROC Curves derived from Regression Models of Match Scores involving Latent Prints

Anonymous WACV submission

Paper ID ****

Abstract

Recent research has focused on gaining crucial insights about demographic differentials and their sources by analyzing how they influence matching. Previous work analyzes biases in matching algorithms without considering the impact of quality estimators that might also be biased due to their learning-based nature. In the proposed study, the research question we address is: "How do demographics affect matching of latent fingerprints of same image quality?".

We extend our previous approach that uses ROC curves derived from regression models by integrating quality measures extracted from latent prints. The contribution focuses on considering quality in the predictive model and not only demographics.

The evaluation is carried out on the FBI WVU BioCop 2008 database that contains 469 right-thumb and 219 right-index latent fingerprint images with associated demographics. Experimental results provide comprehensive empirical support related to the quality-based demographic differentials' impact on the performance of automatic latent fingerprint matching.

1. Introduction

Uncovering and clarifying biases in biometric technologies has recently gained increasing interest [6, 14]. A lot of attention has focused on how predictive models may be biased and how statistical and machine learning models' performance differs along social axes such as gender, age, and ethnicity. In this paper, we analyze fairness of prediction decision-making in forensics through discovering and mitigation of biases in automated algorithms operating on latent fingerprint images. This data often represents an incomplete or distorted impression of a finger obtained when body's natural oils and sweat on the skin are left on a surface. Thus, processing latent fingerprints usually requires the involvement of experts in determining the value of the print as forensic evidence and only in the past few years scientists have released fully automated approaches. An interesting study by Yoon and Jain found that fingerprint matching scores vary with covariates in terms of demographic differentials [25]. Later, Marasco *et al.* utilized the idea of ROC regression techniques to incorporate demographic differentials in the ROC curve. The resulting covariate-specific ROC curves were able to successfully provide the interpretation of demographic bias by taking them into account when assessing the system's performance [12].

Although image quality has been extensively used as a predictor of matching performance, automated quality estimators have not been investigated from a fairness perspective. The importance of quality has been extensively highlighted for sensor-based fingerprints, its impact is even more significant in the presence of latent fingerprint evidence. For latent prints, quality is also an indicator of which type of evidence we are dealing with, which can help to use a more accurate model for that specific case, where standard ROC would not capture the variations in the sample. In the literature, it is well-known that matching performance is highly impacted by image quality [19]. Incorporating quality measures for latent fingerprints enables a more objective assessment and mitigation of algorithmic unfairness. The quality of a biometric signal expresses its utility to an automated system intended as its suitability to further process by the biometric system. Poor quality fingerprint images do not have a clear pattern of ridge and valleys and may result in spurious features and loss of identifiable information. Most fingerprint matchers are based on minutiae points (e.g., ridge ending and ridge bifurcation). Image quality for query latent prints is defined through the extraction of Latent Fingerprint Image Quality (LFIQ) scores. Factors such as adverse skin conditions (e.g., dry, creased) can lower image quality and degrade performance.

ROC curves are a standard way to evaluate the ability of a continuous marker to predict a binary outcome. This work is inspired by recent studies in which ROC curves are used to evaluate the incremental effect of an additional marker in predicting a binary event. The proposed approach uses

ROC curves derived from regression models, where demographics and quality are both considered in the predictive model [18, 22, 11, 2]. The fitted values from the regression model are used to construct the ROC curve and to compare it with the ROC curve derived from the regression excluding quality. The contribution of this paper is three-fold: *i)* Investigate how demographic differentials affect the latent fingerprint image quality (LFIQ) algorithm, *ii)* Study how demographics impact automatic matching of latent fingerprints of same quality, and *iii)* Discuss the use of covariate-specific ROC regression incorporating not only demographics but also LFIQ measures for a more fair assessment compared to the standard ROC.

The rest of the paper is organized as follows: Section 2 reviews research conducted on latent fingerprint image quality assessment, Section 3 describes the proposed quality-based mitigation approach, Section 4 presents the experimental results, Section 5 draws our conclusions and discusses future work.

2. Literature Review

Fingerprint features have been related to age group (individuals born at a similar time), gender (physical characteristics that distinguish males from females) and ethnicity (common culture and origin) [8, 4, 13]. Methods for demographics estimation from fingerprints have searched for gender clues in the ridge density structure that can be encoded by the local texture [13, 17].

An important study on an operational fingerprint database also revealed that fingerprint decision scores vary with subjects' demographic covariates [25]. The data used in Yoon's study were collected from records of the Michigan State Police with TenPrint cards as acquisition type. 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. Since Yoon's model only considers genuine and impostors separately and focuses on regression, our proposed framework could provide a more general approach applicable to adopt demographic information, as well as the quality of the latent prints.

A Latent Fingerprint Quality assessment was tasked by the FBI to assess the quality of friction ridge images for use by latent print examiners. Latent print examiners were tasked with providing an assessment of overall image usefulness, pattern classification, and anticipated difficulty for over 1,000 exemplar fingerprint images[3]. However, some unexplained variability by minutia could be due to the lack of reproducibility of determinations among examiners. It would be beneficial to develop quality and quantity metrics in order to better analyze relationships and comparison determinations. As a result, Yoon *et al.* proposed a method of defining quality measure for latent fingerprints, namely Latent Fingerprint Image Quality (LFIQ), which can be used as a predictor of latent matching performance. Based on the minutiae from latent image estimation, the earlier LFIQ method defines a latent quality measure by combining a qualitative quality feature (*i.e.*, the average ridge clarity) and a quantitative quality feature (*i.e.*, the number of minutiae) to estimate the objective target quality[24]. However, the earlier method that features minutiae count is not a good measurement for latent quality estimation in the presence of unreliable minutiae. Later on, Yoon further incorporated *i)* the connectivity of good ridge quality regions at global level, *ii)* reliability of minutiae, and *iii)* finger position estimation into the latent quality measurement (LFIQ) [23]. The experimental results show that the modified LFIQ has a high correlation with latent matching accuracy, and the model can be effectively used to reveal the quality measurement of latent fingerprint image.

3. Covariate-Specific ROC curve

The Receiver Operating Characteristic (ROC) curve is a popular way to evaluate and compare the accuracy of classification markers when the outputs are continuous. While the pooled ROC curve doesn't take any covariates into consideration, the covariate-specific ROC curve models the covariate effects on the ROC curves and are commonly used in medical diagnostics [26]. It provides error rates specific to the demographics of source subjects. 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., the ROC curve varies conditional on observed covariates $X = \mathbf{x}$.

We define the covariate-specific ROC curve as

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

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 $F_{0,\mathbf{x}}^{-1}(u) := \inf\{t \in \mathbb{R} : F_{0,\mathbf{x}}(t) \geq u\}$ is the quantile function of the imposter group conditional on the covariates[26].

The conditional distributions $F_{j,\mathbf{x}}(t)$ and $F_{j,\mathbf{x}}^{-1}(u)$, for j $\in \{0,1\}$, are estimated from linear regression models on genuine and imposter match scores as follows:

$$\mathrm{T}_j = \mu_j(\mathbf{x}) + \sigma_j(\mathbf{x})\epsilon_j, j \in \{0,1\}, \qquad (2)$$

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 X = **x**, respectively. And the error term ϵ_j is independent of **x**. Then, for a given covariate **x**, the covariate-specific ROC curve

can be expressed as:

$$\text{ROC}_{\mathbf{x}}(u) = 1 - G_1[G_0^{-1}(1-u)\frac{\sigma_0(\mathbf{x})}{\sigma_1(\mathbf{x})} - \frac{\mu_1(\mathbf{x}) - \mu_2(\mathbf{x})}{\sigma_1(\mathbf{x})}], \tag{3}$$

Here $G_i(z) = P(z \leq \epsilon_j)$ is the distribution function of the the regression error term which is independent of covariates for i $\in \{0, 1\}$. In this paper, we assume that $G(z)$ is the normal CDF, and the ROC curve in (3) is so-called binormal ROC curve. The derivation for this expression can be found in [16] and [21].

We refer to the unknown source (i.e., the latent fingerprint evidence) as Query Q while the one pertaining to the known source as Reference R. The demographic covariates for the source subjects are denoted similary. 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. The same scheme is used for the gender and ethnicity covariates. We used two regression models to account for the effect of LFIQ on matching scores. The unique feature of the regression models for the ROC curve is the inclusion of the label term to indicate whether a matching score is genuine or imposter, and also the inclusion of the interaction between covariates and the label[26]. The interaction terms ensure that genuine and imposter groups have different population means.

The first model accounting for the effect of only LFIQ score is given by

$$\begin{aligned} Score \sim & \beta_0 + \beta_L * label + \beta_1 * LFIQ \\ & + \beta_{L1} * LFIQ * label, \end{aligned} \tag{Model A}$$

where the variable $label$ equals to '1' if the two images are from the same subjects, while it equals to '0' otherwise. To account for the demographics, the second model includes age, gender and ethnicity as follows:

$$\begin{aligned} Score \sim & \beta_0 + \beta_L * label + \beta_1 * LFIQ + \beta_2 * Age_Q \\ & + \beta_3 * Age_R + \beta_4 * Gender_Q + \beta_5 * Gender_R \\ & + \beta_6 * Ethnicity_Q + \beta_7 * Ethnicity_R \\ & + \beta_{L1} * LFIQ * label + \beta_{L2} * Age_Q * label \\ & + \beta_{L3} * Gender_Q * label \\ & + \beta_{L4} * Ethnicity_Q * label. \end{aligned} \tag{Model B}$$

In the model, 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 (male or female), and $Ethnicity_Q$ and $Ethnicity_R$ represent their Ethnicity category (caucasian or non-caucasian). Specifically, when the subject is male then gender equals to '1', when the subject is caucasian then ethnicity equals to '1'. We only consider the interaction term of label and the demographics of the unknown source (e.g.Age_Q) because if label=1, the Score is from comparing the fingerprints of the same person, and hence the demographics of the unknown source and the reference are the same (e.g. $Age_Q = Age_R$). When label=0, then the interaction term is also equal to 0. Therefore, it is not meaningful to include both 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. (3), then the accuracy of computer algorithms can be estimated by the curves and can also be summarized using the area under the ROC curve (AUC).

3.1. Image Quality from Latent Prints

A deep network-based minutiae extractor, referred to as MinutiaeNet, is used to obtain the minutia points from the input latent fingerprint image. LFIQ is then applied to the minutiae map extracted in the previous step to estimate image quality. The algorithms can only be applied to latent prints with more than 5 minutiae points. LFIQ is computed by three components, i.e., ridge quality expressed as local ridge continuity, minutiae reliability and finger position. A diagram illustrating the information flow of the proposed approach can be seen in Fig. 1.

3.1.1 MinutiaeNet

Fingerprint comparison is primarily based on minutiae points comparison. Several hand-crafted approaches have been used to augment the minutiae with their attributes to improve the recognition accuracy [9]. However, a robust automatic fingerprint minutiae extraction that's suitable for noisy fingerprint images, continues to be a bottleneck in fingerprint recognition technique. With rapid developments in computer technology, deep learning approach has been used by other researchers for minutiae extraction. Typically, minutiae extraction and matching involves pre-processing stages such as ridge extraction and ridge thinning mentioned above, followed by minutiae extraction , and some heuristics to define minutiae attributes [9]. While such an approach performs well for high quality images, its performance does degrade for poor quality rolled/plain prints in particular for latent prints.

Extracting minutiae automatically from latent prints is a challenging task. Tang *et al.* utilized the idea of object detection (non-maximum suppression) to detect candidate minutiae location and orientation, but it suffers from two major weaknesses such as hard threshold to delete the candidate patches, and the use of plain stacked CNN that suffers from vanishing gradient [20].

For this paper, a fully automatic minutiae extractor called MinutiaeNet is used for minutiae extraction; specifically,

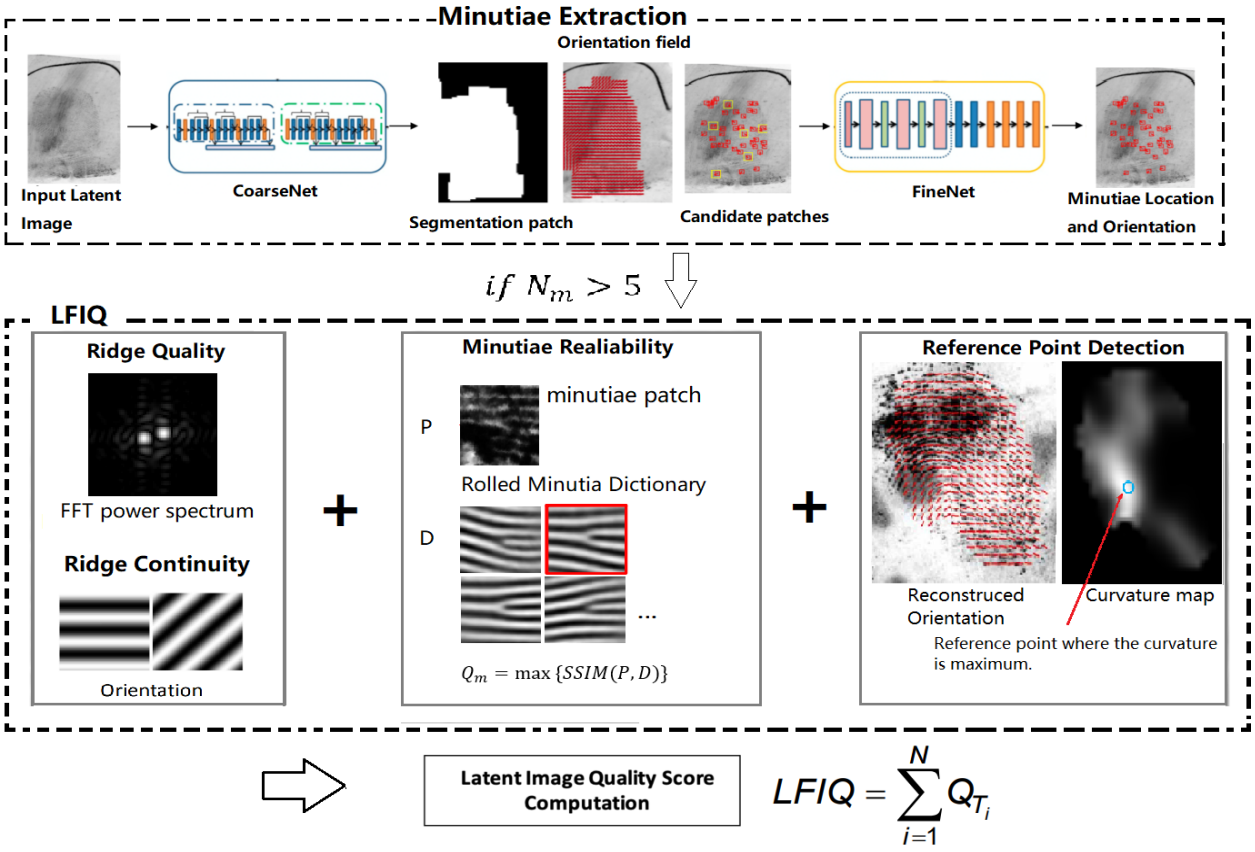


Figure 1. The LFIQ extraction process. Portions adapted from slides of [10].

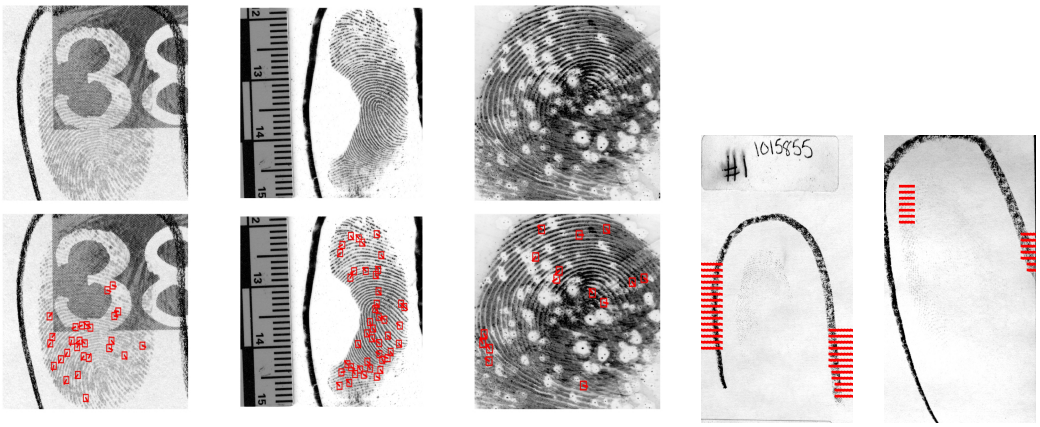
we use the MinutiaeNet introduced by Cao and Nguyen *et al.* [1]. This tool consists of a robust patch-based minutiae classifier that significantly boosts the precision and recall of candidate patches. This approach could provide reliable minutiae location and orientation without using hard threshold or fine-tuning. This method uses residual net instead of just plain stacked convolutional layers to make the classifier more precise [5]. The experimental results show that the MinutiaeNet is robust and has superior performance in terms of precision, recall and F1 values over published state-of-the-art on both benchmark datasets, namely FVC 2004 and NIST SD27 [15]. The architecture is based on two deep neural networks called CoarseNet and FineNet. CoarseNet uses a residual learning based convolutional neural network with fingerprint domain knowledge to predict the minutiae score map and minutiae orientation. This minutiae score map is generated using latent finger print as a primary input and the corresponding enhanced image, segmentation map, and orientation field as secondary input. FineNet, on the other hand, is a minutiae classifier based on an inception residual network that processes each candidate patch to improve the minutiae score map and approximate minutiae orientation using regression [15].

In cases of small amount of candidate minutiae given, an adaptive threshold was applied in minutiae classifier - FineNet for determining final minutiae:

$$\text{Threshold} = \begin{cases} 0.45, & \text{if } N_m > 20 \\ 0.45 - n, & \text{if } N_m \leq 20 \\ n\epsilon\{0.05, 0.1, 0.15 \ldots 0.40\} \end{cases}$$

Where N_m is number of candidates minutiae for each latent fingerprint. This adaptive threshold will obtain all candidate points over 45% threshold, plus additional best ranking candidates until at least 20 minutiae points are provided.

The result is shown in Fig. 2, minutiae points extracted from right index and right thumb with different background noise are illustrated. The algorithm works well in difficult situations such as Fig.2(a) distorted, Fig.2(b) occluded, and Fig.2(c) partial latent fingerprints which demonstrated its robustness to noisy background. Orientation field of two latent impressions marked in Fig.2(d) and 2(e) show the CoarseNet mapping suffers poor performance from dry latent that are extremely low contrast in grey scale (compared to background), due to the discontinuity of broken ridges. The algorithm failed to detect the correct center of a latent at the beginning, this is caused by the weak continuity of true ridges has been further coarsened into background



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

Figure 2. Sample images result from MinutiaeNet on right thumb finger from the FBI Biometric Collection (BioCoP) Next Generation Identification Phase 1 (2008 - 2009) dataset[7]: *(a)* occluded *(b)* partial and *(c)* distorted latent fingerprint impressions; *(d)* and *(e)* are wrongful orientation field (red area) obtained from the low contrast dry-latent print (unsurprisingly LFIQ = 0).

noise while each level of residual net is fused to get the final minutiae score map. In fact, these phenomena commonly happen to latent fingerprints that belong to the no value (NV) category due the texture of the ridges becomes indistinguishable even by a professional examiner's eye.

3.1.2 Latent Fingerprint Image Quality (LFIQ)

In which, LFIQ score can be an independent latent fingerprint's quality indicator for evaluating the matching performance of any existing Automated Fingerprint Identification System (AFIS). This proposed the model can be effectively used to *i)* automatize quality measurement of latent fingerprint image and *ii)* assist latent examiners in their value determination [23].

The objective quality metric for latent fingerprints is the Latent Fingerprint Image Quality (LFIQ). It can be utilized to successfully distinguish high-quality latent fingerprints that don't require human intervention, as well as compensate for the subjective aspect of value determination by latent examiners.

The LFIQ is determined by three parameters that affect the quality of latent fingerprints: (i) ridge quality, (ii) minutiae reliability, and (iii) finger position. Local ridge clarity and the friction ridge regions with high ridge clarity determine ridge quality. The reliability of minutia is determined by its likelihood of being a genuine minutia. The position of a finger is determined by detecting the reference point (e.g., core point(s) or maximum curvature point for arch-type fingerprints) and assigning high weights to minutiae in the central regions of the finger in the LFIQ computation [23]

3.1.3 Latent Quality Score Computation

For each triangle T_i

$$\mathrm{Q}_{T_i} = \mathrm{Q}_{r_i} \sum_{j=1}^{3} \mathrm{Q}_{m_ij} \mathrm{W}_{m_ij},$$

Where Q_{r_i} is the average ridge quality in T_i
Q_{m_ij} is the Reliability of j-th minutia of T_i
W_{m_ij} is the weight based on the finger position
Quality score of a latent is computed as follows:

$$\mathrm{LFIQ} = \sum_{i=1}^{N} \mathrm{Q}_{T_i}$$

N is the number of triangles in the latent print.
LFIQ requires the least amount of 5 candidate minutiae to achieve any meaningful result (LFIQ > 0).

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) [7]. 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. These fingerprints pertain to a total of 1504 subjects and were collected at West Virginia University. 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. For the experiments of this paper, we use right index and right thumb.

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.

4.2. Latent-to-reference print comparison

The match scores in this paper were obtained using the end-to-end latent fingerprint search system recently published by Cao *et al.*. The algorithm does include automated ridge structure cropping, latent image pre-processing, feature extraction, feature comparison, and outputs a candidate list. The model is robust to poor quality latents by generated set of virtual minutiae to construct a texture template. This fully automated latent search system was evaluated on NIST SD27 (258 latents); MSP (1,200 latents), WVU (449 latents) and N2N (10,000 latents) against a background set of 100K rolled prints, which includes the true rolled mates of the latents with rank-1 retrieval rates of 65.7%, 69.4%, 65.5%, and 7.6% respectively[1].

4.3. Results

Fig. 3 (a) and Fig. 4 (a) show the distributions of match scores generated using the MSU identification system for both right index and right thumb.

Fig. 3 also illustrates how demographics impact LFIQ measures for right index. Fig. 3 (b) points out instead that LFIQ differs across different age groups with younger population achieving the highest scores (LFIQ $>$ 60) while the subjects greater that 41 years old exhibit an upper bound of only LFIQ = 20. Fig. 3 (c) suggests that there is no trend with respect to gender which may be due to lack of textural information capture by the LFIQ algorithm. From the literature, gender estimation from fingerprints does exploit textural differences between males and females. Similarly, Fig. 3 (d) highlights that higher LFIQ scores are more likely to be achieved by the Non-Caucasian population. Fig. 4 il-

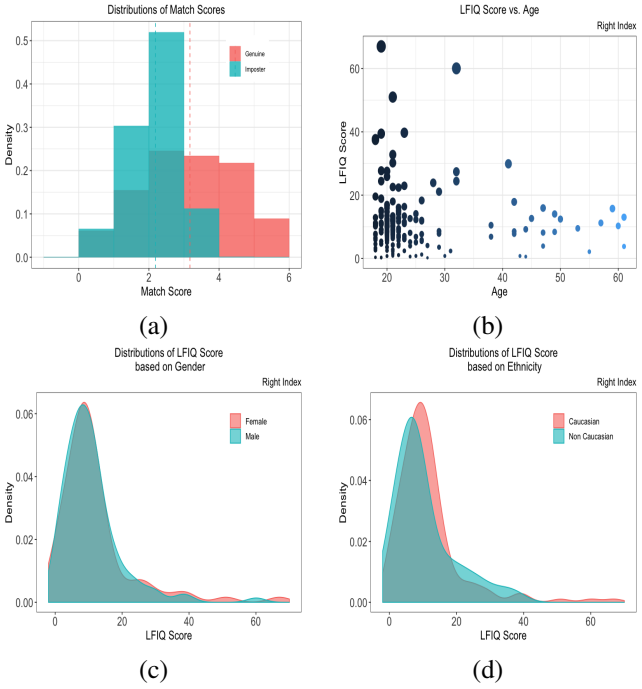


Figure 3. Distributions of the Right Index Finger: (a) Distributions of the Genuine and Impostor Match Scores, (b) Scatter Plot of the LFIQ Scores vs. Age, (c) Density Plot of the LFIQ Scores by Gender and (d) Density Plot of the LFIQ Scores by Ethnicity

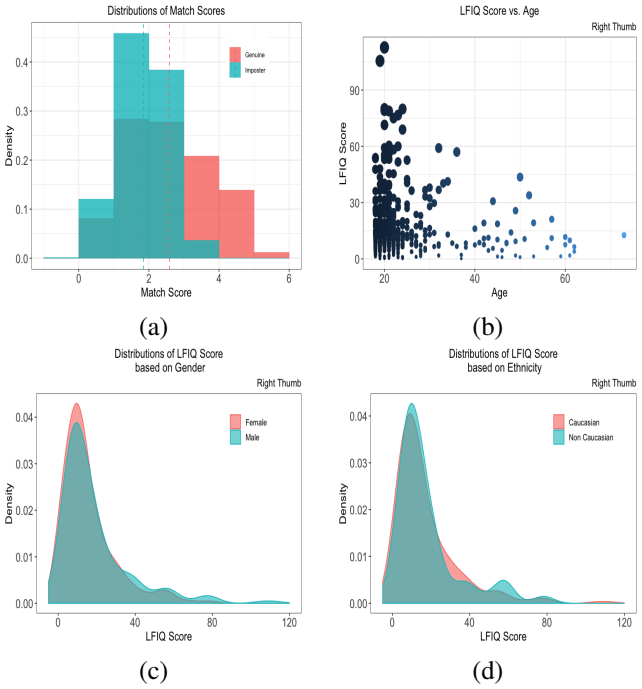


Figure 4. Distributions of the Right Thumb Finger: (a) Distributions of the Genuine and Impostor Match Scores, (b) Scatter Plot of the LFIQ Scores vs. Age, (c) Density Plot of the LFIQ Scores by Gender and (d) Density Plot of the LFIQ Scores by Ethnicity

lustrates how demographics impact LFIQ measures for right thumb which shows same behavior as right index.

In Fig.5, the scatter plot of match scores versus LFIQ measures pertaining to right thumb shows a clear improvement of the separability between genuine and impostors for LFIQ>60. A similar trend has been observed for right index.

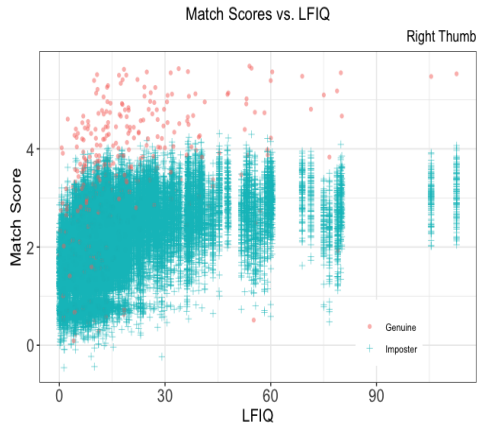


Figure 5. Scatter plot of Match Scores vs. LFIQ scores for Right Thumb.

We examine how the ROC curve varies conditioning on observed covariates by reporting the regression results for the two 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 1 was used to compute the sensitivities of the right index and the right thumb fingers. The QQ plots of the model were used to check the model diagnostics, Fig.6 shows the plot of model B for right index which indicates that the model fits the data very well. The similar results have been observed in right thumb data and also in Model A. The significant interaction terms indicate mean difference between genuine and imposter scores changes with covariates. For example, Model B for the right index fingers has significant interaction terms, LFIQ*label, $Gender_Q$*label and $Ethnicity_Q$*label. This indicates that the difference in mean scores between genuine and imposter groups change when LFIQ scores or age change. The difference in mean matching scores between genuine and imposter groups also change with ethnicity levels. For the right thumb, significant interaction terms indicate the difference in mean matching scores between genuine and imposter groups change as LFIQ scores and for different gender change. The mean difference is a main component in the ROC curve expression. Significant interactions imply when the covariates change, the ROC curve adjusting for the covariates tends to have significant changes.

	Right Index		Right Thumb	
Model	(A)	(B)	(A)	(B)
Intercept	1.5756*	1.3017*	1.8731*	1.8833*
Label	0.4419*	-0.3022	0.6196*	0.4667*
LFIQ	0.0227*	0.0229*	0.0177*	0.0175*
Age_Q	-	0.0015*	-	-0.0025*
Age_R	-	0.0021*	-	0.0020*
$Gender_Q$	-	0.0797*	-	0.0167*
$Gender_R$	-	0.1057*	-	0.0930*
$Ethnicity_Q$	-	0.0624*	-	-0.0827*
$Ethnicity_R$	-	0.0251*	-	0.0117
LFIQ*label	0.0257*	0.0266*	0.0216*	0.0210*
Age_Q*label	-	0.0091	-	0.0034
$Gender_Q$*label	-	0.4077 *	-	0.2535*
$Ethnicity_Q$*label	-	0.3453*	-	-0.0742

* p-value<0.05

Table 1. Result of the designed covariate-specific ROC Models

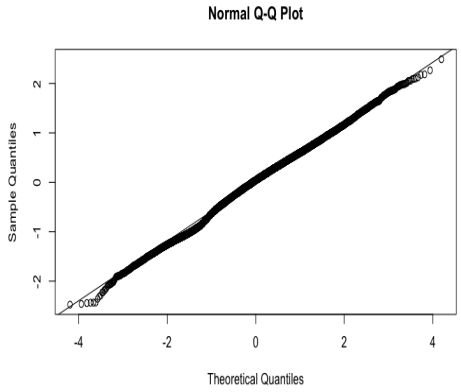


Figure 6. QQ plot of Model B for Right Index.

Fig.7 shows the ROC curves based on Model A only considering LFIQ score for right index, and the corresponding AUC values are also reported in the figure. Here, the 25^{th}, 50^{th}, and 75^{th} percentiles of LFIQ score were chosen to compute the curves. The purple curve is the pooled ROC curve which is used for comparison. We can observe that as LFIQ score increases, the model's identifying ability increases. A similar behaviour has been found also for right thumb with a slightly lower performance compared to right index. Fig.8 and Fig.9 show the ROC curves based on Model B considering demographics and also LFIQ score for right index and right thumb, respectively. We can see that when the demographics are constant, the increase in LFIQ will increase the model's identifying ability, which is consistent with the univariate Model A.

In Fig.8 we can also observe that for the same LFIQ score and demographics, as age increases the model has a much better identifying ability for right index, while we do not see much increase in identifying ability for right thumb in Fig.9 . Fig.8 also shows that when LFIQ scores and other demographics are adjusted, male subjects performs better in

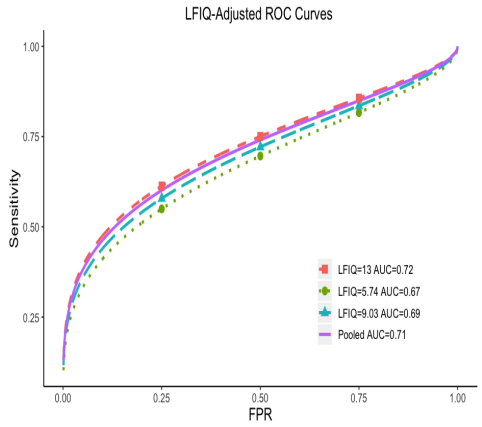


Figure 7. Covariate-specific ROC curves conditioned on LFIQ scores for Right Index.

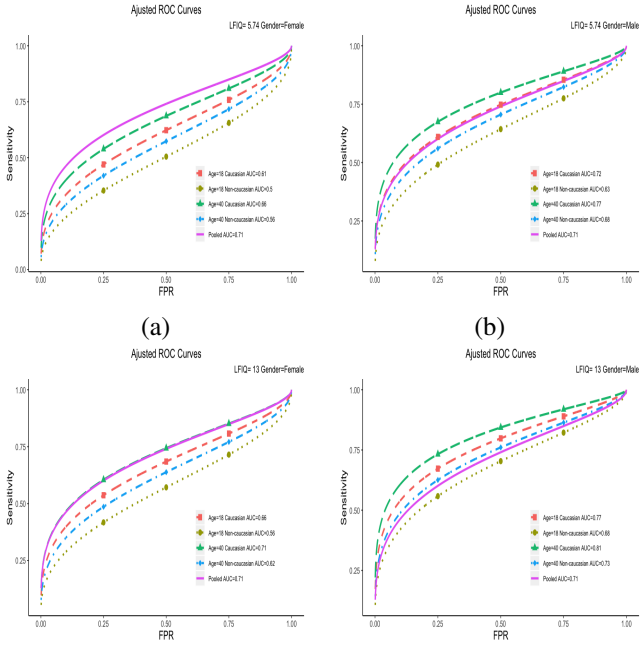


Figure 8. Covariate-specific ROC curves conditioned on LFIQ scores and demographics for right index.

identification than female; we can find the similar behaviour in Fig.9. In Fig.8 it is observed that when LFIQ score and other demographics are same, the caucasian group has a better identifying ability compared with the non-caucasian group for right index. While for right thumb, as we can see in Fig.9, the non-caucasian group performs better in identification.

In Fig.8 (c) and (d), we can see that although LFIQ is not able provide insights about gender differentials, when incorporated as an additional covariate into the proposed demographic-adjusted ROC regression, it can contribute to improving the performance that was not obtained in previous research based only on demographic covariates. Specifically, from Fig.8 (d) we can see that for males we have a better performance using the quality-based adjusted-ROC with higher LFIQ values.

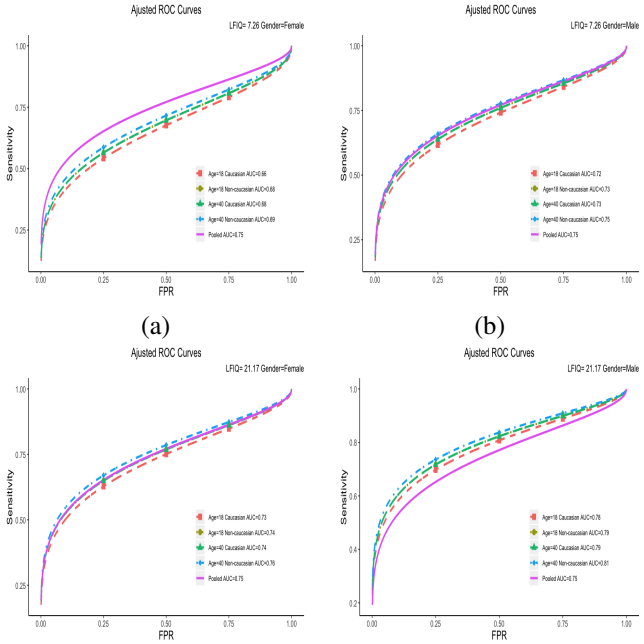


Figure 9. Covariate-specific ROC curves conditioned on LFIQ scores and demographics for right thumb.

5. Conclusions

The identification ability of an automatic matching algorithm is expected to increase in the presence of a high quality image, and it could be further improved if the demographics of each subject can be obtained in some circumstances, such as when the subject is male. Our findings show that the proposed covariate-adjusted assessment scheme conditioned on image quality and demographics is more informative than the traditional ROC curve. It shows that demographics including age, gender and ethnicity tend to have impacts on the discriminating abilities of the matching algorithm.

In this paper, the match scores were extracted using Cao's 2018 algorithm based on comparing minutiae and texture templates. A more "tailored" analysis of the quality covariate can be obtained by accessing to the minutiae template using by the matching algorithm. Furthermore, the MinutiaeNet CNN-based minutiae extractor used in this work was trained on the FVC 2002 dataset that does not contain examples of dry fingers. Thus, the minutiae extracted from dry fingers in the WVU database are often not enough for being processed through LFIQ.

In future work, we will: *i)* Enhance the robustness of the

CNN-based minutiae extractor to dry fingers by fine-tuning it on the WVU database; *ii)* Access to the minutiae templates extracted by the MSU identification system and apply LFIQ to those; and *iii)* Enhance the LFIQ algorithm by fusing multi-layer of minutiae maps from additional extractors for increased robustness in the presence of poor image quality.

References

[1] K. Cao, D. Nguyen, C. Tymoszek, and A. Jain. End-to-End Latent Fingerprint Search. *arXiv preprint arXiv:1812.10213*, 2018. 4, 6

[2] Nancy R Cook. Statistical evaluation of prognostic versus diagnostic models: beyond the roc curve. *Clin Chem.*, Jan;54(1):17–23, 2008. 2

[3] Hicklin et al. Latent fingerprint quality: A survey of examiners. *Journal of Forensic Identification*, 61(4):385–419, 2011. 2

[4] P Gnanasivam and Dr S Muttan. Estimation of age through fingerprints using wavelet transform and singular value decomposition. *International Journal of Biometrics and Bioinformatics (IJBB)*, 6(2):58–67, 2012. 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE CVPR*, pages 770–778, 2016. 4

[6] Deborah Hellman. Measuring Algorithmic Fairness. *Va. L. Rev.*, 106:811, 2020. 1

[7] L. Hornak, W. LaRue, B. Cukic, A. Ross, K. Morris, J. Dawson, S. Crihalmeanu, N. Kalka, and N. Kayal. 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. 5

[8] Crystal Huynh, Erica Brunelle, Lenka Halamkova, Juliana Agudelo, and Jan Halamek. Forensic identification of gender from fingerprints. *Analytical chemistry*, 87(22):11531–11536, 2015. 2

[9] A. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Trans*, PAMI,19(4):302–314, 1997. 3

[10] Anil Jain, Soweon Yoon, Kai Cao, and Eryun Liu. Latent fingerprint quality(lfiq). Lecture PowerPoint - biometrics.cse.msu.edu. 4

[11] Michael W. Kattan. Judging new markers by their ability to improve predictive accuracy. *PubMed.*, May 7;95(9):634–5., 2003. 2

[12] E. Marasco, Mengling He, L. Tang, and Sumanth Sriram. Accounting for demographic differentials in forensic error rate assessment of latent prints via covariate-specific roc regression. *CIVP 2020*, CCIS 1376:338–350, 2020. 1

[13] E. Marasco, L. Lugini, and B. Cukic. Exploiting Quality and Texture Features to Estimate Age and Gender from Fingerprints. *SPIE Defense and Security*, 2014. 2

[14] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–63, 2020. 1

[15] Dinh-Luan Nguyen, Kai Cao, and Anil K. Jain. Robust minutiae extractor: Integrating deep networks and fingerprint domain knowledge. *CoRR*, abs/1712.09401, 2017. 4

[16] M. Pepe. An Interpretation for the ROC Curve and Inference using GLM Procedures. *Biometrics*, 56(2):352–359, 2000. 3

[17] Ajita Rattani, Cunjian Chen, and Arun Ross. Evaluation of texture descriptors for automated gender estimation from fingerprints. pages 764–777, 2014. 2

[18] Venkatraman E. Seshan, Mithat Gönen, and Colin B. Begg. Comparing roc curves derived from regression models. *Stat Med.*, April 30; 32(9)::1483–1493, 2013. 2

[19] Tabassi, Elham, and Grotherand Patrick. *Fingerprint Image Quality*, pages 482–490. Springer US, Boston, MA, 2009. 1

[20] Y. Tang, F. Gao, and J. Feng. Latent fingerprint minutia extraction using fully convolutional network. *IEEE IJCB*, 2017. 3

[21] A. Tosteson and C. Begg. A General Regression Methodology for ROC Curve Estimation. *Medical Decision Making*, 8(3):204–215, 1988. 3

[22] A. N. Tosteson, M. C. Weinstein, J. Wittenberg, and C. B. Begg. Roc curve regression analysis: the use of ordinal regression models for diagnostic test assessment. *PubMed.*, 102 Suppl 8(Suppl 8):73–80, 1994. 2

[23] Soweon Yoon, Kai Cao, Eryun Liu, and Anil K. Jain. Lfiq: Latent fingerprint image quality. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013. 2, 5

[24] Soweon Yoon, Eryun Liu, and Anil K Jain. On latent fingerprint image quality. *Computational Forensics*, pages 67–82, 2012. 2

[25] S. Yoona and A. Jain. Longitudinal study of fingerprint recognition. *Proceedings of the National Academy of Sciences*, 112(28):8555–8560, 2015. 1, 2

[26] X. Zhou, D. McClish, and N. Obuchowski. *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, 2009. 2, 3