2011

Quantifying the Reliability of Latent Fingerprint Matching via Signal Detection Theory

Eric J. Widman
West Virginia University

Quantifying the Reliability of Latent Fingerprint Matching via Signal Detection Theory

Eric J. Widman

Thesis submitted to the Eberly College of Arts and Sciences
at West Virginia University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Forensic and Investigative Science

Keith B. Morris, Ph.D., Chair
Afzel Noore, Ph.D.
Jon Stimac

Morgantown, West Virginia
2011

Keywords: fingerprint, forensics, AFIS, statistics, ROC curve, signal detection

Abstract

Quantifying the Reliability of Latent Fingerprint Matching via Signal Detection Theory

Eric J. Widman

The definition of standards in fingerprint identification is currently an issue of discussion in the field. Quantitative standards have been used in the past to provide justifications for conclusions; however, a scientific basis for relying on numerical standards alone currently does not exist. The tradeoff for this combined approach is that conclusions are based on a conclusion that is left to the judgment of the examiner and may not be repeatable. To test the implementation of thresholds for conclusion, this research studied the effects of only considering concrete data in quantitative form. In this case, signal detection theory is applied to latent fingerprint matching by using automated fingerprint identification systems from two different program vendors. By searching a test set of fingerprints multiple times with a wide range of detail entered, values for the number of system-matched minutiae and computed match scores can be studied to determine threshold limits based on the amount of the search returns. This in turn allows for the generation of receiver operating characteristic curves that directly measure the reliability of the system. The results show that the ability of the system to distinguish matches and non-matches properly is partly based on the method by which the searches are evaluated. Furthermore, the searched area of the fingerprint and the size of the database play roles in determining how well the system is able to discriminate between states. Through future comparison against results submitted by latent fingerprint examiners, inferences can be drawn as to the reliability of conclusions based on varying levels of available detail.

ACKNOWLEDGEMENTS

There are many people without whom I could not have completed this thesis and I would like to provide a show of thanks.

My committee chairman, Dr. Keith Morris, for providing the initial research idea and a great deal of guidance during the implementation.

My committee members, Dr. Afzel Noore and Mr. Jon Stimac for their valuable input in the writing.

Mr. Ming Hsieh and Cogent Systems, Inc. for the generous furnishing of the CAFIS operating system used in this research and support and maintenance thereof.

Lastly, to all the staff and other students for the WVU Forensic & Investigative Science graduate program, all of whose own experiences and insights have helped me grow immeasurably in the time I've spent here.

-Eric Widman

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Since their adoption by law enforcement agencies around the world in the early 1900's, fingerprints have long been considered a powerful tool of individualization. Early research demonstrated that a person's fingerprints would not change in appearance over time, and through years of comparisons it was accepted that no two people shared the exact same fingerprint. These principles of uniqueness and persistence have long supported the strength of fingerprint identification when made to determine identity of the contributor of an unknown source impression.

As fingerprints grew and gained popularity, so too were standards of conduct devised to assist in the practice. Perhaps the first codified standard for conclusions in the field was devised by French forensic scientist Edmond Locard in 1912[1]. In considering fingerprint comparison and identification, Locard created a tripartite rule which based possible conclusions in the number of corresponding Galton details, or minutiae, found between the unknown and known prints. Locard held if two fingerprints were found with twelve or more corresponding details were identifications "beyond reproach"[1]. Two fingerprints with a number of matching details above eight but below twelve required a more thorough approach. For these borderline identifications, Locard required that additional information be used to support the conclusion; for example, the quality of the prints in question, the pattern type, the visibility of pores or the shapes and edges of the ridges themselves[2]. Finally, if the amount of detail was limited to lower than those amounts, then there would be no certainty in an identification; however, a presumption of source could be made depending on the number and clarity of details[1]. In this way, Locard allowed the possibility of weighting one's conclusion. Locard's choice of threshold values for his tripartite rule were derived from early research from Galton and Balthazard, and in turn his work supported many later fingerprint researchers, such as Wilder and Wentworth, and Cummins and Midlo[2].

Locard's tripartite rule may have been pivotal in the basis for quantitative standards for fingerprint comparisons. For many years, it was required of fingerprint examiners to consider the exact number of matching details in support of a conclusion[3]. Often referred to as point standards, these held that fingerprints could only be declared matches when a certain number of details were found in correspondence, as a quality control measure. As an example, in the past the United Kingdom held identifications that would be used in court to require 16 corresponding points between the two prints[2]. This began to change in 1973, when the International Association for Identification passed a resolution stating that no valid basis existed for requiring some minimum number of corresponding details between fingerprints to effect a conclusion[4]. The position was further developed by a resolution presented at the International Symposium on Fingerprint Detection and Identification held in Ne'urim, Israel in 1995, which concluded:

> "No scientific basis exists for requiring that a pre-determined minimum number of friction ridge features must be present in two impressions in order to establish a positive identification."[5]

While there are a number of countries that still consider a point standard for identification, many allow for variability based on the examiner's judgment to subvert the required number of detail in borderline cases, and ultimately the purely quantitative approach to conclusions has largely fallen out of favor[4]. Instead, the quantity-quality approach to identifications has been adopted by fingerprint practitioners.

First formalized by David Ashbaugh in his fingerprint text *Quantitative-Qualitative Friction Ridge Analysis*, the quantity-quality model has formed the center of the ACE-V comparison methodology and provides a more flexible approach to identifying fingerprints. Rather than considering matches based on an arbitrary value of required points, this approach dictates that the examiner consider all aspects of the prints in question when formulating a conclusion[6]. This approach to fingerprint identification is ultimately reliant on the experience of the examiner in question; the more experienced examiner has a better understanding of the likelihood of finding similar details among a general population. This runs counter to the hard threshold set by the past use of minutiae; in the quantity-quality approach, the examiner is the one to determine sufficiency. Given a variable boundary that represents this determination of sufficiency, the argument is that the more experienced examiner can make decisions based on less available information[7]. This variability among examiners, though, has led to some criticism of the fingerprint discipline in the legal and scientific community, as others question what exactly is necessary to constitute an identification[8]. This divide between forensic and legal practitioners has further developed since the change in admissibility standards set by the Supreme Court of the United States in the 1994 ruling in *Daubert v. Merrill Dow Pharmaceuticals*.

The *Daubert* ruling came about in part to revise the previous federal admissibility standard of requiring general acceptance. In the ruling, the Court held that expert testimony to scientific techniques needed to meet a higher level of excellence, and in order to be admissible in court, a proffered technique or methods should meet several criteria: the technique should be peer-reviewed, with a known error rate, accepted standards, and accepted as valid by the relevant scientific community[9]. The decision incorporated changes made with the development of the Federal Rules of Evidence for expert testimony, and the new guidelines have since been incorporated into many state court judgments as well. And, in the forensic sciences, the change in ruling has led to challenges among many historically-accepted forensic disciplines[10]. The first legal challenge arguing the status of fingerprint evidence is held as *U.S. v.* Mitchell in 1999.

In this case, fingerprint examination was argued as being inadmissible under the *Daubert* standards in part because of the lack of a known error rate and the lack of standards in the field due to the reliance on the individual examiner's judgment[11]. Though fingerprints were deemed admissible in *Mitchell*, there have been many admissibility challenges since that case. The arguments levied against fingerprints have largely remained the same; critics contend that fingerprint identification has an error rate which is underplayed by examiners, and that standards across the field are nonexistent[8].

The question of reliability, though, has been addressed in part through past studies on the chance that two fingerprints from different sources would be found to be alike. One way in which this chance was sought was through basic statistical modeling; researchers tested, and attempted to determine the statistical likelihood of encountering two fingerprints that were the same. Models based on empirical observations allowed for the determination of how many details would be required before the chance of finding a similar configuration would be negligible. These initial models all assumed a base probability for encountering detail and assumed that, as minutiae were random, were statistically independent and carried the same weight. Statistical models proposed by Sir Francis Galton, Edward Henry and many others all developed work in this regard[12]. Some later models began to acknowledge that different minutiae arrangements could occur with varying frequency and adjusted the calculations to accommodate for this. A good example is the grid method proposed by Osterburg[13], in which the fingerprint is divided up into 1 mm^2 cells and the number of events counted. By determining how many cells contain an event of a specific type, the probability could be calculated in addition to that of finding however many cells containing uninterrupted ridge flow. As time passed, these models were acknowledged by the community only in that research on the subject had been done, but none of these saw noted use in practice.

Most of these early statistical models were developed to determine the chance of two fingerprints being found to be identical in their entirety; thus, the issue being addressed was the observed uniqueness of fingerprint ridge formations. However, in the years since the *Daubert* decision, statistical models have been developed that focus on approximating casework conditions. These models are designed to account for smaller numbers of details, as well as acknowledging the spatial relationships and direction between them, allowing for source intervariability (distortion) and uncertainty of detail location. For example, Pankanti, Prabhakar and Jain[14] showed a model that considers the number of minutiae found in correspondence between two prints and the total number of minutiae found in both impressions. The likelihood ratio model presented by Neumann[15] provides another example of finding the chance of association based on the spatial configuration of the selected details. There have been other models developed as well, and they show good promise in evaluating fingerprint matches based on what they are designed to consider. Still, since statistical models rely largely on minutiae, or level 2 details, none have seen a great deal of use in the fingerprint community as they are currently not able to approximate qualitative considerations of fingerprint examiners.

Given that a quantitative model for supporting fingerprint reliability has not been accepted to date, there has been a variable response from the fingerprint community in terms of clarifying the error rate of the discipline. In the early days of the admissibility challenges, there was the contention that the error figure be thought of as two distinct components: a methodological error rate and an examiner error rate[11]. According to this theory, if the principles of uniqueness and persistence were accepted as truth, the only possible source of error would be that of the examiner incorrectly applying the comparison methodology. Thus, for a time the methodology of fingerprint comparison was held to have a “zero error rate.”[8] This position has largely been abandoned due to increasing scrutiny to the claim, and publicized errors such as the Brandon Mayfield case[16]. Currently, there has not been an accepted answer to the issue of error rates and more research has been called for in terms of determining whether, among fingerprint examiners, one can be defined.

Based on the history and evolution of fingerprinting comparison up to now, it seems there has been some confusion over the meaning of error and its role in a method. In a scientific process, the degree of error is given by the variability between an observed value and its true state. No method can be without all error, as there additionally exists some limit of detection to which results can no longer be determined. The same can hold true for fingerprint examination; if certain latent prints are classified as no value due to insufficiency, this implies the limit of detection of the method applied, and therefore some error as a result. The goal is to attempt to quantify this error. In this case, one can consider error solely through a quantitative aspect. While a purely numerical approach to setting standards would neglect potentially useful information, evaluating qualitative considerations is slightly more problematic as there is not a universal agreement as to fingerprint quality classification[17]. If the case for quantitative thresholds is accepted at face value for the purposes of experimentation, there should exist some critical value where the presence or absence of one additional detail would make or break the identification, as it were. If this critical level were found, it would allow for a degree of support behind the determination of sufficiency as made by the fingerprint examiner. While examiner observation may influence the results based on the factors discussed previously, an automated system works well for limiting the observed results to conclusions based on second-level detail, which is easily quantifiable. Through analysis of a large number of results and looking for patterns within those, any trends in response can demonstrate whether such a value exists. This analysis can be done through application of signal detection theory.

Signal detection is commonly used in diagnostics and clinical medicine to measure the ability of some test to distinguish between measured states based on a given input. The original application was based on the ability to discern an incoming signal from background noise. At the most basic level, binary response systems always have four possible output indices.

		Ground truth	
		True	False
Response	True	True positive	False positive
	False	False negative	True negative

Figure 1. General confusion matrix format

In cases of false response, two possibilities are a given: the threshold for defining a response is set too low and a noise response is classified as a valid signal, resulting in a false positive. Alternatively, the threshold is set too high and a true signal is misinterpreted as noise, a false negative. Based on the values of yes-no responses given and by knowing the ground truth state of each entered test, it becomes possible to calculate these values to determine the strength of the applied method.

Another useful tool of signal detection theory is that of the receiver operating characteristic (ROC) curve. By plotting the true positive rate (sensitivity of the method) against the false positive rate (1 – specificity of the method), the ROC curve is found for a given evaluator[18]. As it is displaying the chance of true versus false results graphed across each response variable, the ROC curve acts as a normalized measurement of how that likelihood changes over input levels. Not only this, but the area contained under the ROC curve acts as a direct evaluation of how well the method correctly rates positive results[19]. The greater the area under the curve (AUC value), the better the method is able to discriminate between states, and the lower values show classifiers with less worth. If the area under the curve approaches

0.5, this signifies that the method is actually unable to separate accurate responses from false ones[20]. AUC values below 0.5 actually mean that the system is misclassifying results; the results would have to be reversed to increase the accuracy[18].

The application of receiver operating characteristic curves works well in considering binary response (yes-no) tasks, and fingerprint identification thus becomes an ideal candidate for this type of evaluation. In this case, searching latent prints through an automated matching system will provide data on response for multiple classes. This data in turn allows for a measurement of the ability of the system to determine the correct result given specific stimulus, which can then be extrapolated to considerations of fingerprint matching in its entirety. Notably, this method does not provide a statistical metric of the likelihood of finding similar fingerprints arrangements. Instead, this method provides a measure of how reliable the identification of a fingerprint is based on the availability of input information. Through this reasoned analysis of the signal response, the research will serve to highlight decision thresholds in latent fingerprint identification and examine their usefulness in discussions of reliability.

METHODS

Sample Collection

Fingerprints were collected from anonymous donors. For collection of latent impressions, the donor was instructed to touch the surface two to four times in a small area, with a specific finger. Depending on the type of surface, the latent impressions were processed with regular black fingerprint powder and brush, or magnetic powder. Upon developing the test latent prints, the multiple impressions were collected using tape and fingerprint lift cards, and then finger number was noted to keep track of the source finger. This was repeated two to three times for each using separate fingers, so that a large initial set could be collected. The donor's known prints were also collected on standard-size tenprint cards using ink. Upon collecting both latents and a tenprint card from each donor, the set was assigned a letter notation to note that the latents belonged to that tenprint card; along with the source finger, this was the only notation to keep the latents connected to their source tenprints.

Fingerprint Entry and Automated Searches

All latent lifts and tenprint cards were imaged using an Epson 4990 Perfection scanner. Scans of the tenprints cards were made at 500 pixels per inch resolution while latent scans were made at both 500 and 1000 pixels per inch, both standard qualities for digital imaging of fingerprints[21]. While higher resolution images are much more important when evaluating at the third level, this was not a factor in this study; the 1000 ppi images were recorded simply as a precautionary measure in case high resolution images were needed for some reason, and in practice the 500 ppi scans were appropriate for use in the automated systems. Additionally, the use of a scanner ensured that all prints would be truly sized 1:1 without the need for any subsequent image adjustment. After scanning the latent cards onto the computer, the lifted impressions on each card were reviewed to determine whether they were of value for comparison, and their overall suitability for inclusion in the test set. The latent prints for use in the test set were selected based on the degree of distortion present as well as the amount of detail. In order to obtain a wide range of results, latent prints which contained a larger amount of detail were preferable. Ultimately, no more than two latent impressions from each subject were selected, cropped down and inserted into the respective AFIS unknown database.

Two databases of record prints were used in the completion of this research, each associated with two different vendors of automated identification systems. A test set of approximately 1000 records was used with the automated system AFIX Tracker (version 5.0.0.77). The other automated system provided by Cogent (CAFIS version 5.1) was linked to a database containing approximately 1.8 million tenprint records. The entire test set of tenprints and latents was added to both systems.

After entry of the test set latents and known tenprints was completed, the minutiae details on each had to be encoded for system searching. Tenprint card records were entered into the respective system and the prints auto-encoded by the system; if necessary, human review was implemented only to clear away any spurious details seen by the system. However, based on the design of the experiment, the latent searches were encoded entirely by hand. It was decided early on in the research design that preparing an overly detailed plan for searching would not be necessary; rather, a simple guideline would allow for greater flexibility in dealing with variation between the latent prints. To start the search process, each test latent was marked with three minutiae. The starting minutiae were selected based on their location and proximity to one another; groups were sought that lay outside of major pattern areas such as a core or delta of a print. Additionally, as stated in the introduction, only minutiae that had a clear location and direction were used in the experiment, in order to limit uncertainty in the minutiae placement and increase the likelihood that minutiae marked on the rolled print would be found in the latent. When available in the system, adjusting the minutiae for a perceived level of quality was not performed.

After each completed search, an additional minutia was marked on the latent based on its proximity to those details already marked. Moving from the starting cluster to the closest visible detail, the process generally moved around the pattern before filling in the core, available delta and lower part of the latent if present. If the extreme tip of the finger was captured as part of the latent, this area was generally encoded last, as a rolled tenprint does not usually capture this portion. In this way, each latent obtained a candidate list result through many levels of encoded minutiae. For those latents that had more than 20 minutiae to mark for searching, the minutiae were added pairwise after going beyond 20; this was to prevent those prints with large amounts of minutiae from dominating the results, as well as to move the searching along expediently. In this way, minutiae were added to the latent until no more could be found that met the necessary criteria for their use.

In the total result set, all searches were performed without any other identifying information added to the record. Automated identification systems allow the user to input restrictions on the search parameter based on other knowledge, such as the possible type or possible source fingers. By limiting the work done by the server, results can be obtained with a shorter waiting time while decreasing the amount of non-match candidates returned. For the purposes of this research, though, the initial test data set was designed to be open set; that is, searched against the entire available population of the database.

Searches with Modified Parameters

Based on a preliminary testing of the method, it was hypothesized that the order of adding minutiae in the search process may change the results to an unknown degree. In order to test this theory, it was decided that resubmitting a small group of the test set under different search conditions, then comparing the new results against the initial results would provide an indicator of what additional factors may influence the system matching.

In total, ten prints were selected and resubmitted searches to test the effect of modifying the starting cluster of minutiae on the search results. Five of the prints were selected at random and resubmitted to the AFIX database with different seed clusters than those used in the initial research testing. While starting with a different cluster, the rule of working to the closest, clearly visible minutiae was still followed when marking these prints. The other five prints were selected based on the inclusion of a triradius in the impression. These five were searched in the Cogent database with marked minutiae starting in the delta and moving out towards the core of the print.

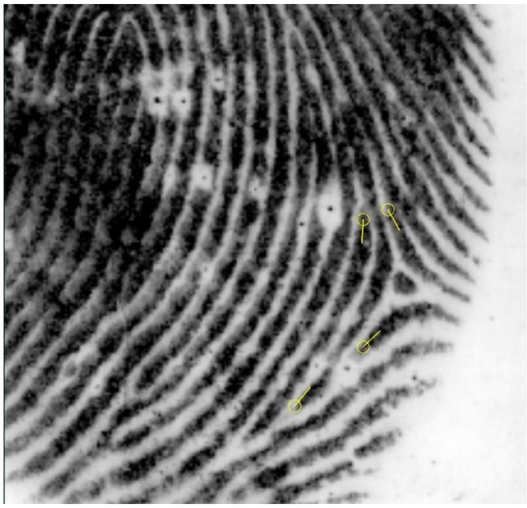


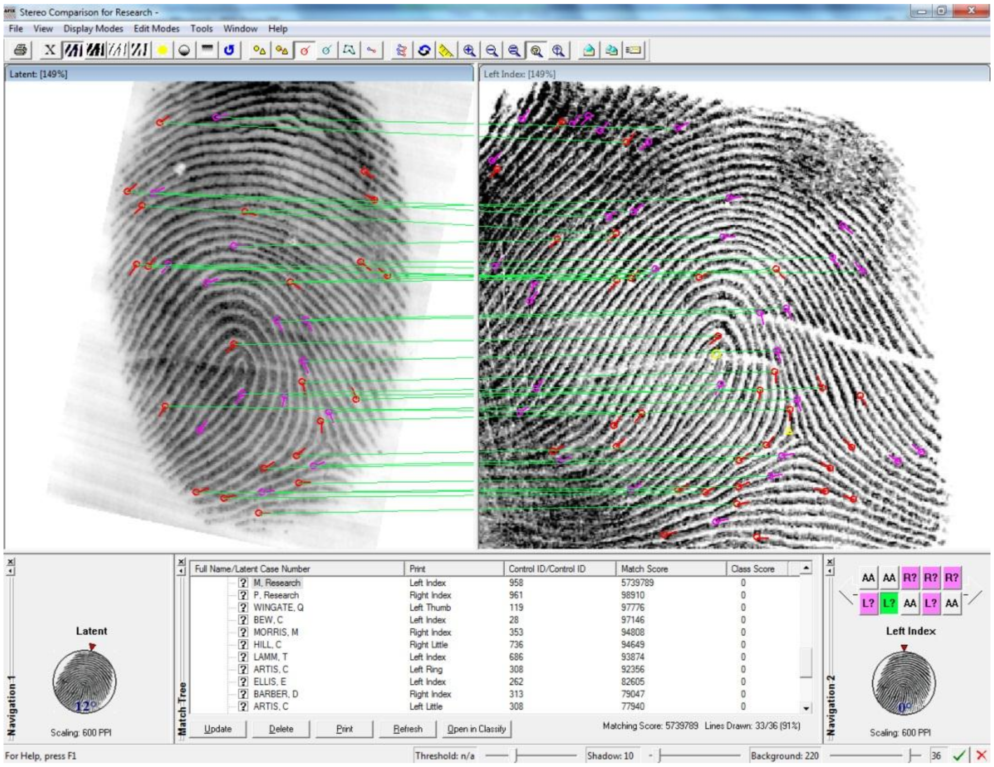
Figure 2. Example of a print marked starting in a delta area.

The results of these ten resubmitted prints were contrasted against those results obtained from searching the same prints in the main research set.

As a separate test, five randomly selected latent prints from the test set were resubmitted to the Cogent database based on modifying the search parameters instead of the marked minutiae. These five prints were searched using system restrictions that examiners might use to increase database penetration and improve search results. The fingerprint pattern was used to limit the search results, and the degree of available rotation for the entered minutiae was reduced to acknowledge proper orientation. Potential donor fingers were also limited based on inference from the observed pattern type; for example, a left-slant loop is much more common on left hand fingers, therefore searching a latent print of this pattern may allow the examiner to limit the search to fingers on the left hand. In this way, the database penetration is increased, filtering out obvious non-matches without expending the time to search them. In this case, at each level of minutiae each print was searched twice: once with and once without these limiting parameters. As the marked minutiae for each level remained the same, the results will allow for a measurement of how these limitations affect the results returned.

Data Analysis and Receiver Operating Characteristic Curves

After each completed search against the respective database, the results were recorded as follows: first, where applicable, only the top 20 candidates were recorded. This was a practical consideration to control the size of the data mining. For each candidate, the match score and the number of minutiae matched by the system were recorded, as well as whether the candidate was a match (the print was correctly mated to its source in the database) or non-match (all other results).



Full Name/Latent Case Number	Print	Control ID/Control ID	Match Score	Class Score
M. Research	Left Index	958	5739789	0
P. Research	Right Index	961	98910	0
WINGATE, Q	Left Thumb	119	97776	0
BEW, C	Left Index	28	97146	0
MORRIS, M	Right Index	353	94808	0
HILL, C	Right Little	736	94649	0
LAMM, T	Left Index	686	93874	0
ARTIS, C	Left Ring	308	92356	0
ELLIS, E	Left Index	262	82605	0
BARBER, D	Right Index	313	79047	0
ARTIS, C	Left Little	308	77940	0

Matching Score: 5739789 Lines Drawn: 33/36 (91%)

Scaling: 600 PPI

Scaling: 600 PPI

Threshold: n/a Shadow: 10 Background: 220 36

Figure 3. Results shown by system searching AFIX Tracker. The match score and number of matched minutiae are given by the system.

Other classifiers of the search results were extrapolated from the recorded data to see whether any could offer additional discriminating ability to the system. The difference between match scores was analyzed: dubbed Δscore, the value was found by taking the difference between the match score and the score from the next highest candidate. The reasoning is as follows: if matches exhibit higher system scores than non-match counterparts, the difference in score between a match and the non-match below it should be noticeably greater than those between non-matches. Another possible classifier was taken as the result of the match score divided by the number of system matched minutiae. A third potential classifier was calculated by first finding the fraction of minutiae matched by the system out of those entered, and multiplying this by the score. If matched prints exhibit higher percentage of minutiae matched to minutiae entered, the resulting value will be greater than those non-matches which exhibit a poor match-to-entered ratio.

For each of the analyzed classifiers, the observed values for matches and non-matches recorded at each level. By considering the number of these occurrences at an *input* level and that level *previous* to it, it can be determined how many results would be properly or improperly classified at the dictated level.

$$TP_{input} = Total\ Matches - \sum_{input} Matches$$

$$TN_{input} = Nonmatches_{input} + TN_{previous}$$

$$FP_{input} = Total\ Nonmatches - TN_{input}$$

$$FN_{input} = Matches_{input} + FN_{previous}$$

To classify the overall distinguishing power of each classifier, a receiver operating characteristic curve was constructed through calculating the true positive and false positive rate.

$$True\ positve\ rate = \frac{True\ positives}{Total\ positive\ returns\ (TP + FN)}$$

$$False\ positive\ rate = \frac{False\ positives}{Total\ negative\ returns\ (TN + FP)}$$

The area under the ROC curve can be found through a rectangular approximation. It should be noted that this approximation gives a slight underestimate for the actual area[22]. After calculating the AUC value and the corresponding standard error, significance testing was performed by finding the two-tailed *p* for comparisons between specific curves.

RESULTS

Results for AFIX Tracker Searching

AFIX Tracker searches were the initial evaluation of the research method, and were the first completed. 1158 total searches were completed, with 930 returned matching prints to 21704 non-match results. The following figures show the total matches made by the system based on score and the number of matched minutiae.

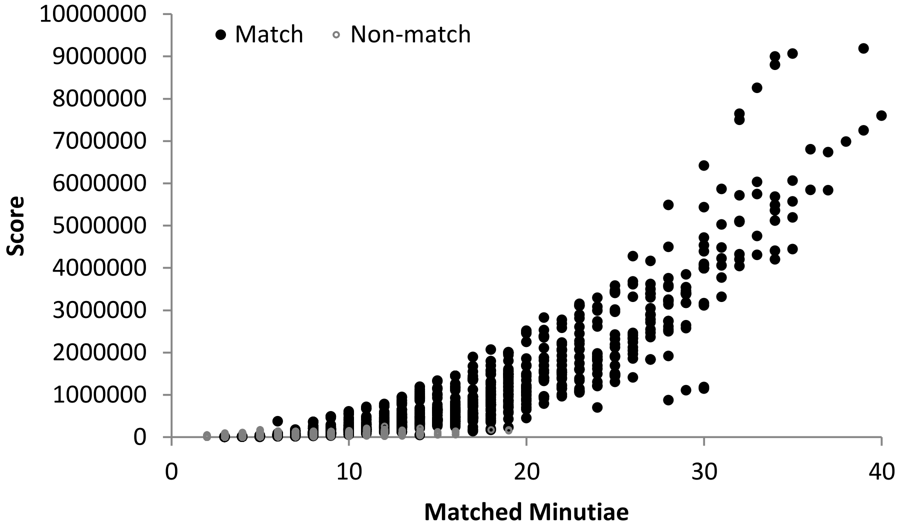


Figure 4. Total results of AFIX Tracker searches.

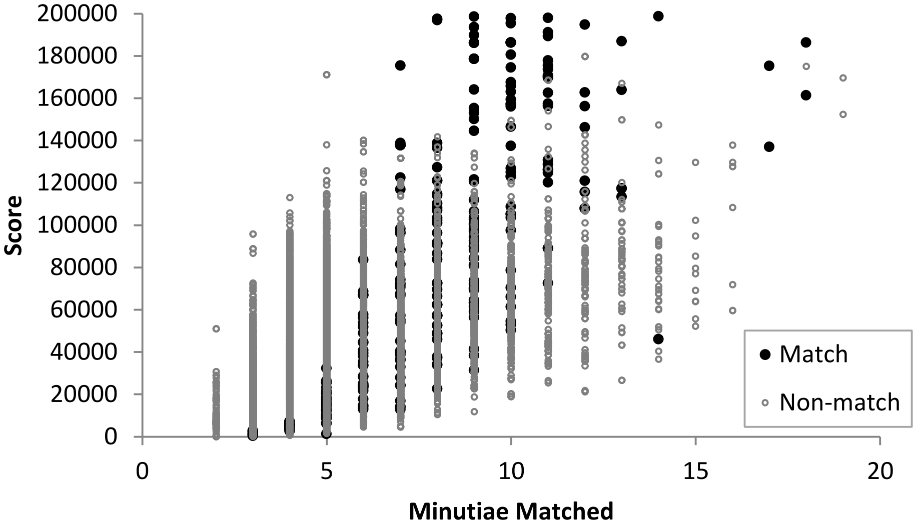


Figure 5. Attention to the non-match results from Figure 4.

Based on the observed number of matches and non-matches made by the system, values can be derived for the occurrence of true and false results, when considering responses at a given input value.

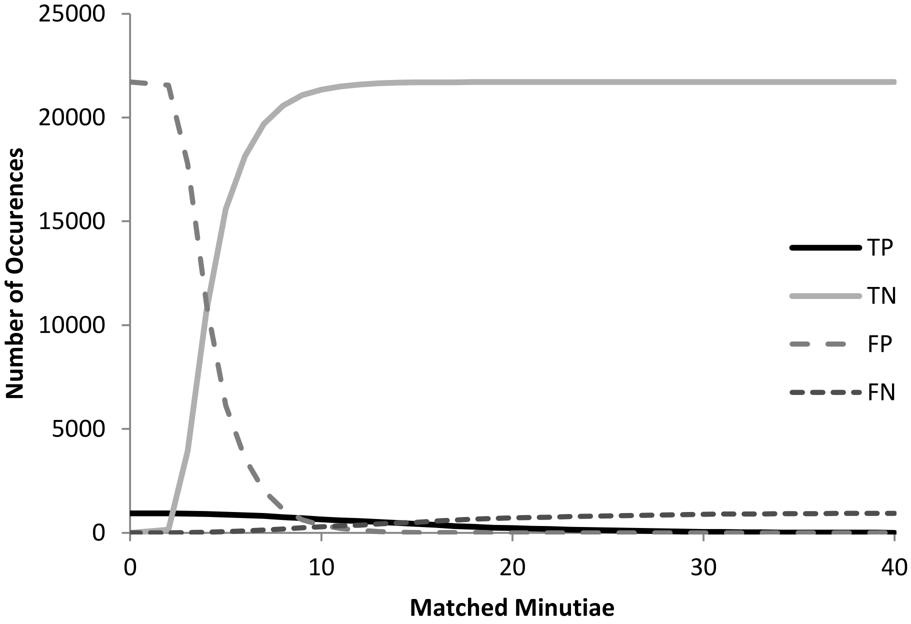


Figure 6. The evaluation of true and false responses over the number of system matched minutiae on AFIX Tracker.

This figure demonstrates the overall matches recorded on the system by calculating the rate of response at each instance level of input minutiae matched. Through observation of the points where responses begin and cease to occur, conclusions can be drawn as to the best location for placement of thresholds.

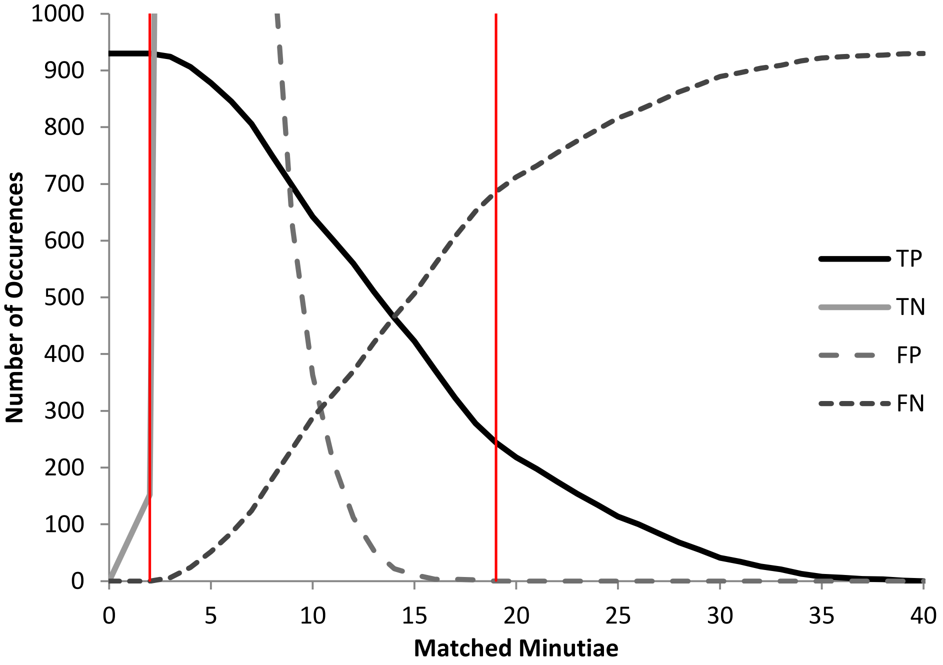


Figure 7. The placement of threshold cut-off based on the occurrence of false responses for AFIX Tracker.

In this figure, it can be seen that false negative responses begin at as low as three matched minutiae, and false positives cease at 19 matched. The point where a false positive or false negative shared the same number of occurrences was found at about ten matched minutia. Cut-off threshold values for all of the observed classifiers can be found in Appendix B.

The individual prints do show some degree of variability in the generated ROC curves. This is demonstrated by Figure 8.

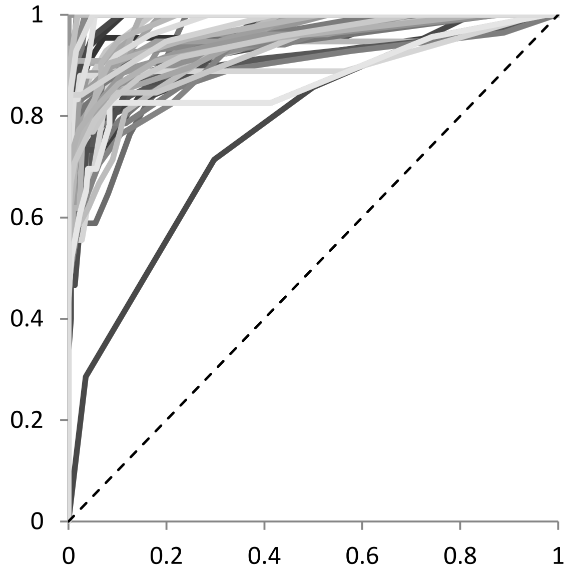


Figure 8. Individual ROC curves based on minutiae discrimination for AFIX Tracker.

This fluctuation is observed in all of the classifiers' results. Combining the results into the aggregate will compensate for this.

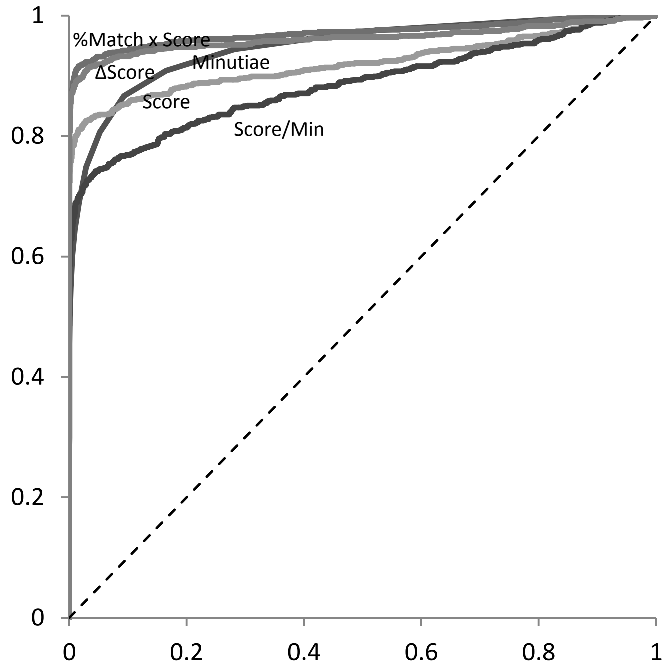


Figure 9. ROC curves for each of the evaluated classifiers. AFIX Tracker.

Results for CAFIS Searching

For the CAFIS searches, 1161 searches were completed, resulting in 1404 matching prints returned and 19787 returned non-matches. CAFIS results differed from the AFIX results in several aspects. Unlike AFIX, CAFIS did not give any results at all when searching clusters of only three or four minutiae; results only began to return at clusters of five entered minutiae or more. This meant that out of those 1161 input searches, two searches not giving returns for each of the 50 test prints gave 1061 candidate lists for use. Additionally, CAFIS reporting allowed for the return of plain and rolled impressions from the same subject if the match was high enough; AFIX only took the better match of the two, which gives CAFIS a higher number of matches at large amounts of input minutiae. Perhaps most notably, the score reporting for CAFIS is very different than that of AFIX: match scores are only four digit values. As seen below, this keeps matches and non-matches within a smaller range.

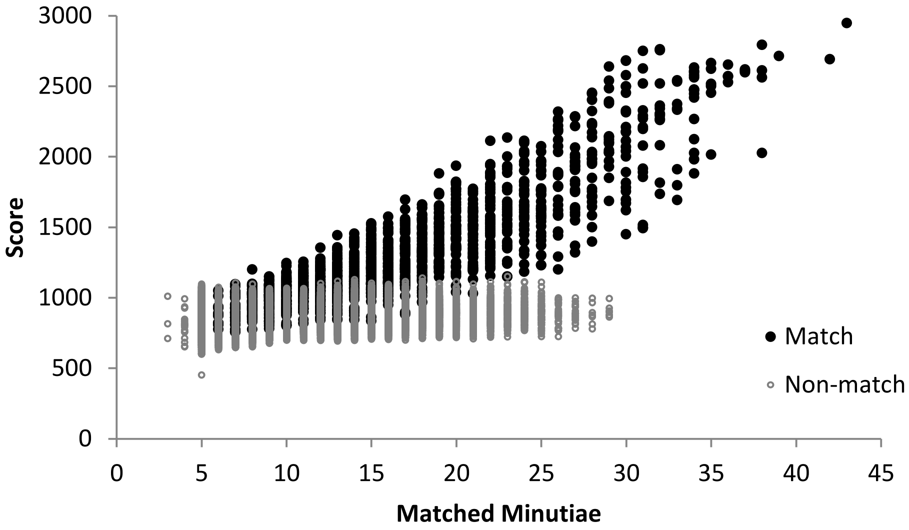


Figure 10. Total observed search results for CAFIS.

As with the results from searching AFIX Tracker, the number of matches and non-matches are used to extrapolate the occurrence of valid and errant response.

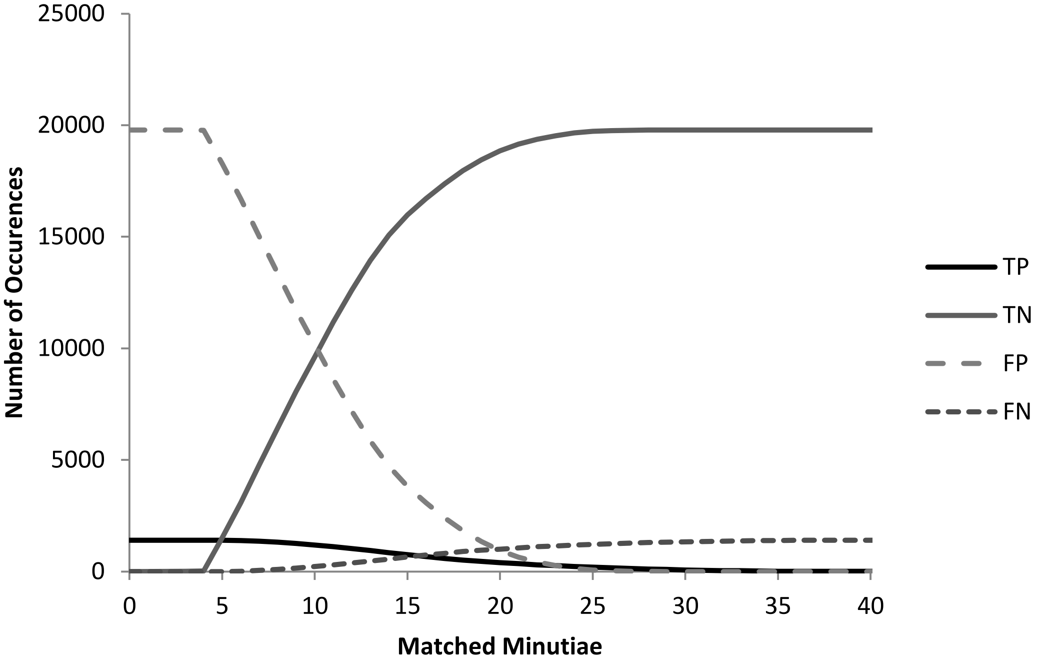


Figure 11. Total occurrence of valid and false responses for Cogent's AFIS evaluated based on system-matched minutiae.

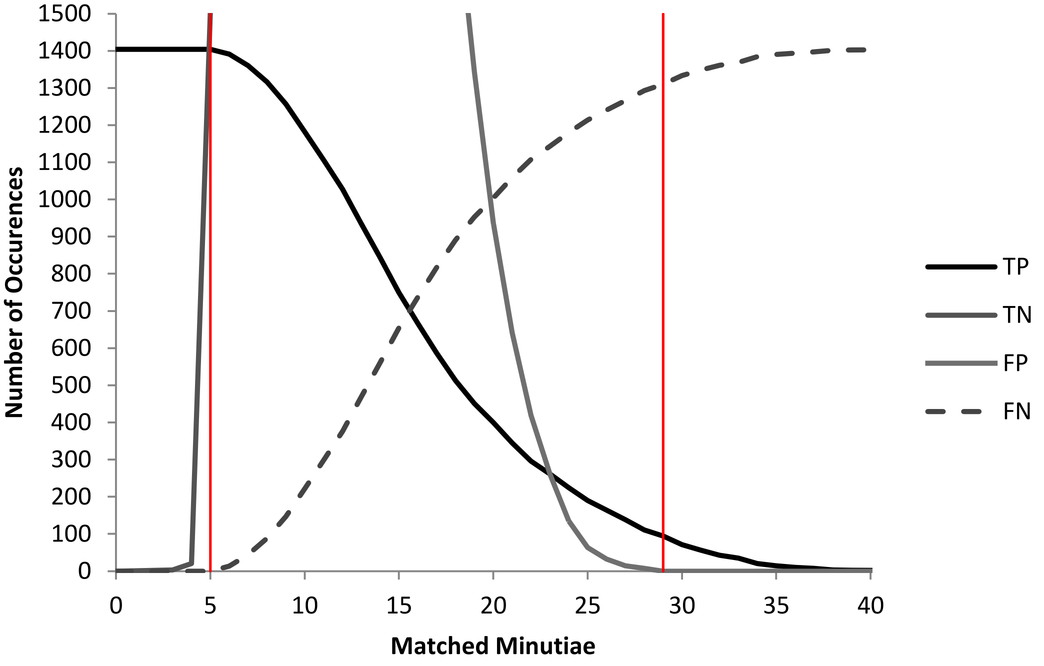


Figure 12. The placement of threshold cut-off based on the occurrence of false responses for CAFIS.

In searching CAFIS, the cut-off values for false negatives and false positives occurred at 5 and 29 matched minutiae, respectively. The number of false positives and false negatives was equal at 20 matched minutiae. Other threshold values are given in Appendix B.

Once again, by observing the ROC curves calculated based on the individual test prints, some variation can be seen between the curves themselves.

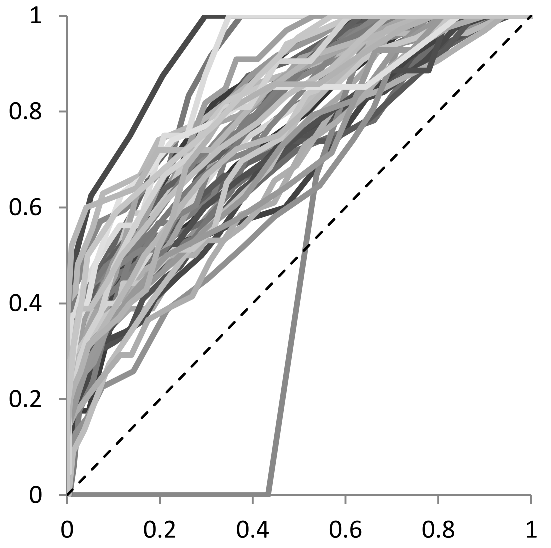


Figure 13. Individual ROC curves based on minutiae discrimination for Cogent.

Again, the variability seen is based on minutiae. The results are varying similarly across each classifier. The results taken in aggregate show differences between the method used as well, just as in AFIX Tracker, as seen in Figure 14.

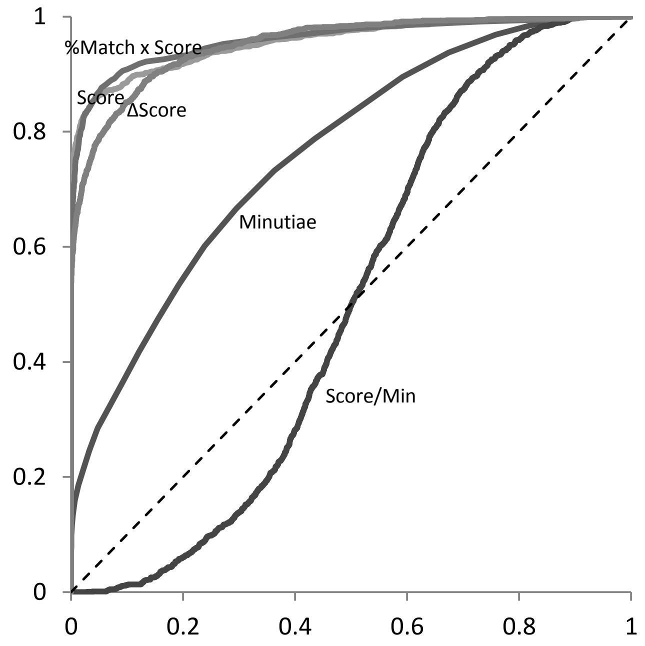


Figure 14. ROC curves for each of the evaluated classifiers. Cogent Systems.

System Contrasting

Comparing results between the systems themselves provides insight into how well each of the classifiers works. The ROC curves also allow for the comparison of the classifiers between both of the systems, to determine whether one or the other works significantly better. This is demonstrated in the following five ROC curves, where the AFIX curve is represented by the darker curve and the CAFIS curve is shown as the lighter curve. The *p*-value from significance testing is also shown for each appropriate graph.

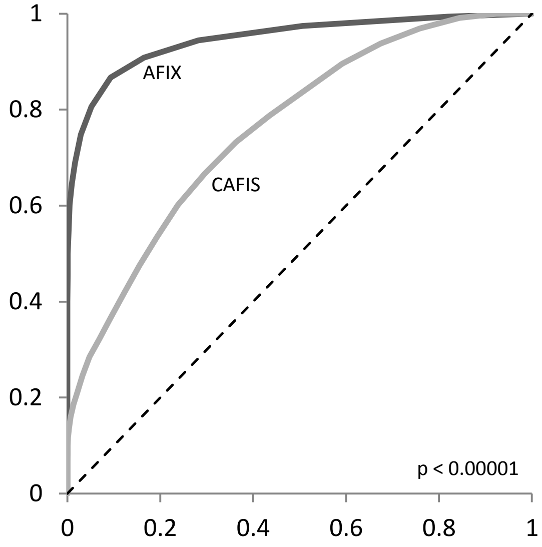


Figure 15. AFIX versus CAFIS when evaluated based on minutiae matched.

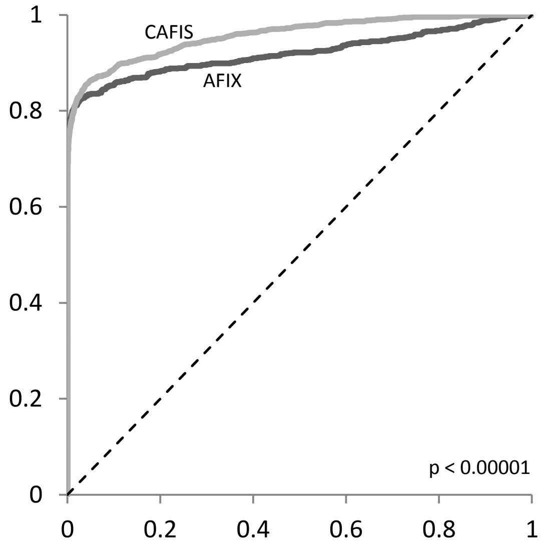


Figure 16. AFIX versus CAFIS when evaluated based on match score.

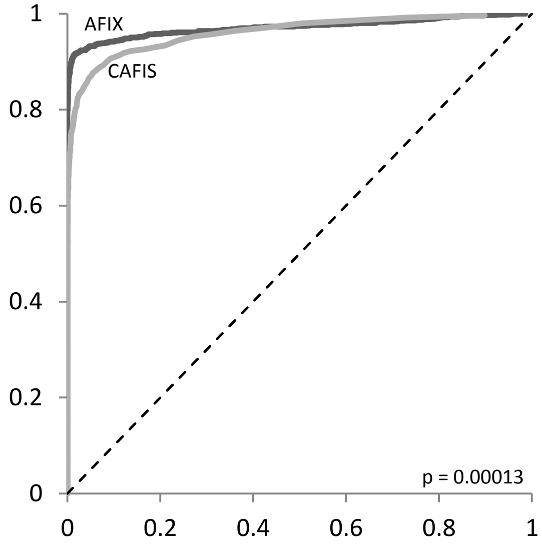


Figure 17. AFIX versus CAFIS when evaluated based on the change in match score.

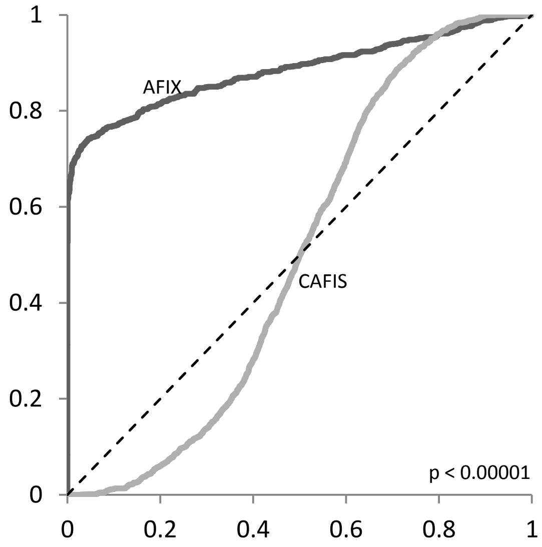


Figure 18. AFIX versus CAFIS when evaluated by the match score divided by number of matched minutiae.

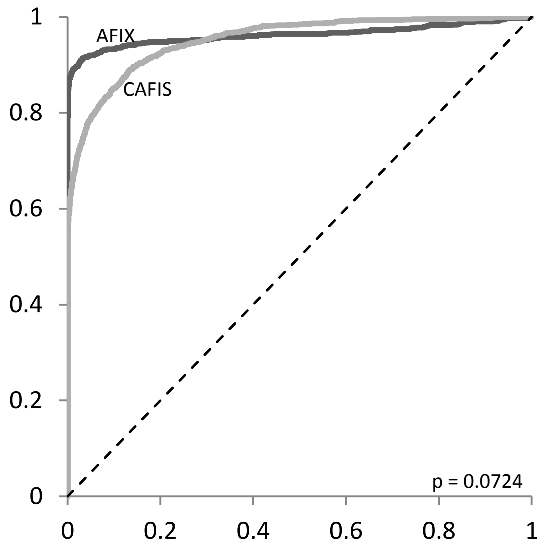


Figure 19. AFIX versus CAFIS when evaluated by the percentage of matched minutiae multiplied by match score.

Resubmitted Searches

The subset of fingerprints that were resubmitted to the database did show some variability as well, though none to a significant degree. The curves generated for these resubmitted results compared against the results from the original searches can be seen in Appendix C.

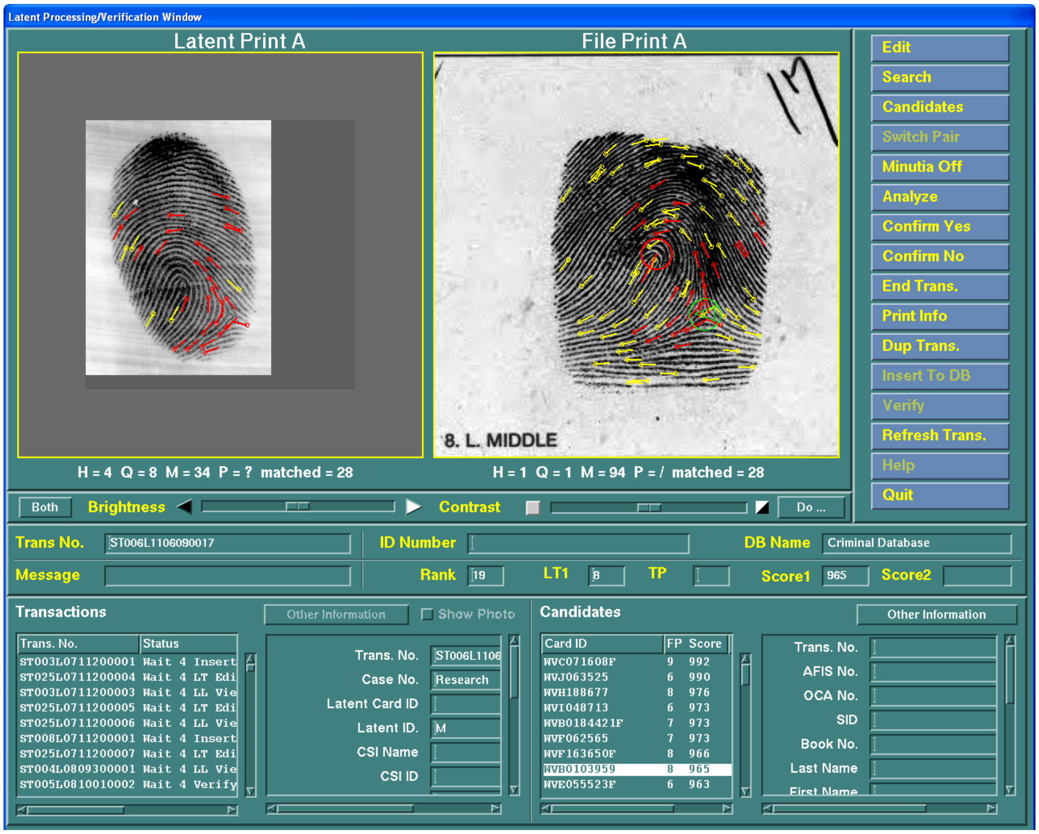
DISCUSSION

Based on observation of the search results, some considerations can be given on placement of decision thresholds for values appearing in the data. The ideal result would be a value where all negative results would fall below it and all positive results occurring at or above, resulting in perfect discrimination. Evaluating based on a practical basis, though, means that one must expect a varying degree of overlap between the accept and reject state, which may change depending on the parameter being studied. The various effects that searching has on the discriminations of the results can be explored in detail to determine whether or not one value can be found that both gives acceptably reliable conclusions and is applicable for implementing in casework.

From looking at the ROC curves derived from each individual latent print, some fluctuation of the curves is instantly evident. While most of the curves shown are very close to each other and show excellent discrimination, that slight degree of variability does exist between them. This signifies that the fingerprint itself has some effect on the ability of the system to properly match it; however, this is only an inference as to the observed change in the signal detection results. Based on general observation of the prints, it seems that those with more tightly clustered minutiae density are found first by the automated system. Reasonably, the system is able to properly match these prints since the denser minutiae allow for proper orientation to be determined more quickly. The effect of the variability of results between prints means that results taken in aggregate are more likely to provide an estimate of the system discriminability that is applicable to general cases.

The few outlier curves generated by the system also point towards the effect of the test print on the system results. Certain latent prints showed different results in regard to how soon the source tenprint began appearing in the search results; in some cases, a print matched well on one system did not match well on the other. The prints that required more input data to find the true match than other may either result from a poor choice of starting minutiae, or may simply be the result of the arrangement of the print itself. This distinction will be discussed at length later. In a specific case, one of the prints was not matched by either system. Based on further observation of the print in question, it seems that low latent quality from pressure distortion led to an inability to place minutiae in any kind of localized fashion, which impacted the results negatively. In the end, all that can be concluded is that quality considerations do affect the results as well, though it is harder to conclude exactly how from these results alone.

Considering the variation in ROC curves generated for minutiae based on the two AFIS vendors, a significant difference is found. Notably, the Cogent test set was much less able to properly discriminate as evidenced by the curve comparison in Figure 15. This does not mean, however, that the vendor itself is significantly worse than the other. Rather, the much greater size of the database appears to be the main factor in the shift in these results. As the Cogent database held 1.8 million records to AFIX's 1000, a greater number of potential corresponding configurations exist within the significantly larger data set.



Latent Processing/Verification Window

Latent Print A

H = 4 Q = 8 M = 34 P = ? matched = 28

File Print A

H = 1 Q = 1 M = 94 P = / matched = 28

Trans No. ST006L1106090017
ID Number
DB Name Criminal Database
Message
Rank 19
LT1 8
TP
Score1 965
Score2

Transactions

Trans. No.	Status
ST003L0711200001	Wait 4 Insert
ST025L0711200004	Wait 4 LT Edi
ST003L0711200003	Wait 4 LL Vie
ST025L0711200005	Wait 4 LT Edi
ST025L0711200006	Wait 4 LL Vie
ST008L0711200001	Wait 4 Insert
ST025L0711200007	Wait 4 LT Edi
ST004L0809300001	Wait 4 LL Vie
ST005L0810010002	Wait 4 Verify

Trans. No. ST006L1106
Case No. Research
Latent Card ID
Latent ID. M
CSI Name
CSI ID

Candidates

Card ID	FP	Score
WVC071608F	9	992
WVJ063525	6	990
WVH188677	8	976
WVI048713	6	973
WVB0184421F	7	973
WVF062565	7	973
WVF163650F	8	966
WVB0103959	8	965
WVE055523F	6	963

Trans. No.
AFIS No.
OCA No.
SID
Book No.
Last Name
First Name

Figure 20. Screenshot of a highly-rated non-match from the Cogent database.

This manifested in both the increased threshold for false positives and the greater result overlap and thus lowered AUC value determined for this database.

It bears asking whether implementation of thresholds that would be derived from this research would be feasible in casework. Based on the evaluation of a large database as seen in the Cogent experiment, the size of the database has a great effect on the number of highly-rated false positives. Again, this emphasizes an issue that does not see much mention in fingerprint literature anymore: the more comparisons that are made, the higher the likelihood of a false association[12]. If one wanted to eliminate the entirety of these false positive results when evaluating based on minutiae points alone, it would require that no identifications be made without 29 or more of those details in correspondence. When dealing with latent prints, however, this amount of detail is exceedingly rare; this high threshold level would not allow for the majority of casework prints to be evaluated properly. The exclusion of results can be seen in the search data itself; out of approximately 1400 true positive results from the CAFIS searching, 1310 would become false negatives if a 29-point standard was implemented. While it can be argued that every effort must be made to exclude the potential for false positive results, a limitation of this magnitude would all but eliminate the ability to reach match conclusions. Looking at things from another standpoint, if the examiner is making a conclusion based on a smaller number of tenprint comparisons, the threshold derived from the smaller AFIX Tracker database may be applicable. Using a threshold of 19 minutiae necessary for a purely second-level match conclusion is still somewhat high, but much more acceptable. This decision may be compounded by the other factors that examiners are trained to consider in second-level detail evaluations, as will be discussed later. On the other hand, a latent print with a very small cluster of minutiae, in this case three, will often be deemed no value for comparison. These amounts made up the majority of non-matches seen in the data set, so concluding such at this level appears justified.

Considering that the obtained threshold values are somewhat high for practical use in forensic casework, there remains another way to discuss the applicability of the results. Looking at the thresholds from the standpoint of the human fingerprint examiner, there are additional considerations that can be used, such as pattern considerations and third-level detail such as poroscopy and edgeology. Combining these with the minutiae configurations will liken the threshold values back to the tripartite rule that Locard discussed[1]. Using the AFIX Tracker data results, two fingerprints with 19 or more corresponding details could be matched based on the minutiae alone, with absolute confidence in the truth of the result. Below three minutiae matched, nothing can be said about the results (no value). Between these two values is the range that requires additional information that would support the decision. These values are supported by the data. In the case of the CAFIS data, the cut-off values would shift to 29 and five, respectively. As fingerprint examiners are trained to consider these additional means of support when making comparisons, these values may not be so different than what would be found from decision-making conducted in actual practice.

While the minutiae classifier is applicable to both AFIS and human examiners, the same is not directly true about those derived from match scores. As minutiae are the main link between AFIS search results and examiner observations, the threshold values they show are of special interest and have thus been elaborated on more extensively above. The system-generated scores do not share the same weight. Fingerprints examiners often do not give the system match score significant weight, as its meaning is lost in the conclusion; the examiner makes the call and wishes to exclude any undue influence on the decision-making from the machine results. Regardless of their usefulness, the discriminating power of the scores themselves gave good results when used in this experiment. Score ROC curves from both systems gave good values for the area under the curve, so the score values themselves do not appear to be entirely meaningless.

The observed change in match scores, or Δscore, also provides insight into the changes between true matches and non-matches made by the system. This classifier also shows difference between the two systems compared, with CAFIS providing better classification when Δscore is applied. While the scores presented by CAFIS are much narrower in range, so too are the variability in score. Often, single-digit changes were observed from non-match to non-match, making the generated match scores much more discriminating even for only a change in 100 or more. Comparatively, the AFIX Tracker results showed a much greater spread due to the larger score values; thus this vendor was also able to discriminate very well when this classifier was applied, but via a different rating metric. In all, this parameter provides some interesting results, and it is something that has not been detailed very often in literature. Some further investigation may be warranted on whether this classifier can be useful in allowing examiners more options when considering score values.

The use of the Score/Minutiae classifier became extremely problematic when applied in this case. The reasoning behind its initial use was that it appeared relevant in the trial experiment. However, this initial evaluation was conducted on AFIX Tracker, where match scores are allowed to reach the millions. Due to this, the impact of the denominator was negligible; dividing a match score of about 2800000 by 30 minutiae matched, just for example, makes little difference when even the closest search-yielded non-matches generally give a match score in the tens of thousands. In the case of CAFIS, though, the match scores are limited to four places, and are thus much closer together relatively. Using Cogent, a good match might have a generated score of 2000 or so, but when divided by the larger number of matched minutiae, it falls below lower scores that have a small number of minutiae. The matches actually fall below the non-matches due to this and the ability of this classifier to discriminate results suffers for it. As can be seen in Figures 14 and 18, the Score/Min classifier yielded the worst result for CAFIS classifiers and showed the worst discriminability out of all tests with an AUC value of approximately 0.5016. Such a low value signifies that the system is literally unable to distinguish between matches and non-matches when the metric is applied. Since the value seems to be entirely dependent on the size of the match score allowed by the system vendor, it seems safe to conclude that this classifier does not offer any advantage whatsoever in interpreting the search results.

Out of all the classifiers discussed, the parameter evaluating through multiplying the percentage of matched minutiae by the match score is the one that appears to work the best. In fact, this classifier is the only one that resulted in AUC values greater than 0.95 when evaluated on both systems as seen in Figure 19. It would be reasonable to assume that the number of minutiae matched out of those entered would be a factor in the calculation of the match score, though without knowledge of the specific system algorithms there is no way to say for sure. If this is assumed to be the case, then combining the factors is acting as a weighted response based on the expected high levels of corresponding information found in the true matches. This serves to further distinguish those results from other configurations in the database.

From significance testing of the aggregate results, some further support of the trends discussed can be gathered. In the AFIX results, the match score and score/min classifiers were significantly different from any other, as seen in Table 7. The minutiae, change in match score, and percent match multiplied by the match score all were not significantly different among themselves. In the CAFIS results, the difference between all classifiers was found to be statistically significant, with the exception of the difference between the match score classifier and the percent match multiplied by the match score. These greater differences also appear as a result of the larger known set and variation in reporting as described earlier. The differences in system matching also appear in the *p* values for testing across the two systems. From Figures 15 through 19 it can be seen that the only classifier not different between the two systems was the percent match multiplied by match score, further suggesting that this classifier works the best out of those used.

Looking at the results up to now, it can be seen that both the classifier chosen and the fingerprint itself will affect the ability of the system to return the correct result. However, this discussion can be extended to consider the variability between sections contained within the test fingerprint itself. Based on resubmitting a subset of the latent fingerprints marked with different minutiae, it appears that the order in which minutiae are added does have some effect on the search results. As can be seen in the tabulated results in Appendix B, the number of results as well as the area under the resulting ROC curves is slightly different in each case, each increasing or decreasing variably. This is also a byproduct of the number of fingerprints available to search against, as the ability of the system to match a cluster of minutiae depends on how many similar configurations are available in the database. Changing the input cluster will change this amount. If certain configurations of minutiae are accepted to be rarer than others[2], then accordingly some configurations must be more common as well. Additionally, the density of minutia in specific areas can have an effect, as discussed previously. Significance testing, though, did not show a statistical difference in the curves generated by these additional searches as seen in Appendix C. Based on the initial data set, it would appear that the effect is limited.

Referring back to the discussion on variability between latent samples, the above considerations of cluster variability introduce a certain complication to the results. If a poor result is observed when signal detection is applied to individual fingerprints, it becomes unclear whether the poor result is based on the fingerprint itself or based on a poor selection of minutiae clusters for searching. In fact, it may be a combination of both factors; again, though, attempting to account both possibilities at once may be beyond the scope necessary for human consideration. If an automated system could be designed to do this, conclusions could be drawn as to the potential weight of each aspect; this dichotomy is another area that may need future research.

Considering this variability in selecting different starting points only has effect on the results of the experiment; in casework examples, the standard procedure is to mark all the details that have confidence. Considering all possible variations in order for this experiment may be overcomplicating the data as well, since no significant difference was noted between the resubmitted prints and the originals. Taking an example print with 24 minutiae, if no order is considered when marking the detail, it results in 2024 $({}^{24}_{3}C)$ possible ways to mark the latent up for clusters of three minutiae. The possibilities are additive in ways to proceed in the markup from that point. In light of this, following a standard procedure for markup, as was done in this research, may be more valid than attempting to account for each possible configuration of minutiae. However, the issue that this introduces is that, for small number of detail entered, the results may not be applicable to searches from separate areas. A more detailed study on entering minutiae clusters from multiple areas of latent prints may provide better insight as to how the available area of said print will affect the results.

The resubmitted searches against the delta areas of the print also provide some insight into the variability of search results. When considering the delta area of a fingerprint, it is an area where ridge flows converge, and thus minutiae is very frequent in the triradius. Because of this fact, searching an AFIS in that area shows more similarities, as again, events appear more often.

Table 1. Excerpted results from the resubmitted latent prints marked in the delta area.

Minutiae Matched	Matches	Non-matches
4	0	1
5	0	150
6	4	122
7	6	140
8	4	152
9	7	178
10	10	160
11	9	175
12	12	154
13	12	123
14	11	133
15	8	98
16	9	89

Minutiae Matched	Matches	Non-matches
4	0	0
5	0	110
6	0	114
7	0	105
8	2	124
9	4	107
10	6	142
11	9	139
12	12	160
13	8	209
14	13	199
15	9	148
16	10	106

Table 1 demonstrates one issue that complicates the signal detection approach to identification. On the right, an excerpt of the results is shown from the five fingerprints that were resubmitted to the database starting from the delta area, while the corresponding original search results are shown on the left. From the data, it can be seen that those prints searched outside the delta showed matches earlier than those searched in the delta. To the fingerprint examiner, this is the better consideration of the two. The set shows that the correct results would be attainable in the results with less available detail, as may often be the case when working with casework latents. On the other hand, the fingerprints seeded in the delta require a larger number of details for the truth results to start appearing, which would mean that an identification might be missed if only the delta area was available. However, when considering these results via signal detection, the resubmitted prints show better discrimination than the original results for both minutiae matched and match score, as seen in Figures 26 and 27. This is because there is a lesser degree of overlap between the matches and non-matches, which in turn means better discrimination. The results from the other classifiers were slightly better in the original searches than those resubmitted, which suggests that the differences observed between minutiae and match score classifiers may be weighted out when used in one of the combinatorial applications. Of note, only the score classifier showed a significant difference from the *p* value, as seen in Figure 27.

This may be a result of the score reporting, as other classifiers which involved the match score did not show a significant difference.

While variation in the results seems a natural consequence of the order of marked minutiae, there are some additional limitations based on the searches themselves to consider. With the Cogent system, a subset of the database was searching by using certain limiting parameters that go along with the system. Being able to exclude searches based on the pattern type of the fingerprint, the potential source finger and degree of available rotation, it was found that the system could match latents correctly beginning one to two instances of input minutiae less than without the limitations using those same entered minutiae. The justification for this is the same as elaborated on for those fingerprints resubmitted for the delta-based searches. By limiting the search results with these parameters, the true matches are found sooner than without them, but the measured discrimination of the system lowers slightly as a result. This may seem counterintuitive; for an automated system, one might expect the potential for false results to decrease if obvious non-matches are eliminated from the search. The reverse may be the actual case; by limiting the search to those prints which would be more like the questioned print the likelihood of corresponding details increase, though this is only speculation based on observations of the results.

To this point, the various details concerning the applicability of these results have been weighed. Most of the considerations have been whether or not the evaluations hold up when considered for automated fingerprint matchers only. However, signal detection considerations should remain equally applicable to traditional fingerprint comparison as well. If an experiment was designed to control the amount of available input detail, then that detail would be similar to the multiple AFIS searches at varying input levels performed here. The results of the examiner decisions could then be directly compared to those from this research, to demonstrate the improvement gained from additional comparative considerations. Some variation can be expected based on the subjective judgment of the examiner as well. If the claim that experience determines the determination of sufficiency holds up, then one would expect the threshold values found for each latent print to fluctuate slightly depending on the training and experience of the comparing examiner. Additionally, the way in which examiners provide support for their conclusions might provide some insight as to how much weight the different details are given in the determination of a result. The less the determined thresholds vary, the more conclusions would show agreement across the field.

Again, the primary issue that complicates applying this method to human examiners is that they will look beyond the presence or absence of a minutiae as an AFIS might. Even if directly asked, examiners may not be able to purely limit themselves to looking at prints the way the automated systems did in this experiment. Examiners not only consider second-level detail as a position and direction, but its influence on the surroundings as well, the specific way in which the events occurs and the flow of the ridges around it. Furthermore, while an AFIS will simply ignore spurious of absent minutiae between a questioned and known print, to the examiner these issues easily allow for exclusion. It is these factors that form a part of evaluating the “totality of information available in the print” as dictated by quantity-quality[23]. As discussed previously, the inclusion of data beyond the simple presence of a detail should lessen the threshold dictated by the experimental set. The degree to which the threshold fluctuates, though, is variable and may depend on a number of factors that cannot be represented so easily by a forced yes-no decision. Allowing the fingerprint examiner to make determinations of sufficiency (value-no value) as well as inconclusive determinations, then analyzing the results with a weighted-decision multiclass ROC may be of benefit in interpreting how the examiner’s conclusions change based on the available information[18].

FUTURE RESEARCH

Based on the results seen in this research, there is some future work that can be done using this method to evaluate the results of latent matching. Submitting the latent prints to a group of fingerprint examiners will allow for a better measure of how well this method applies to actual practice. As has been stated previously, there are a number of methods by which human comparison could be likened back to the AFIS research done here. Providing latent prints of certain quantity of information, latent prints can be compared based on that level of quantity as was done here; additionally, a multiple class evaluation can allow for a weighted consideration of conclusions, with added attention to the considerations of the examiner.

Furthermore, some additional work can be done with the automated searching itself. In this case, testing was performed with two databases, one of very small size (1000) and one of a large size (1.8 million). Not only could data from databases sized in between those two be useful, but further investigation may be warranted into databases that are sized larger than that in this experiment. Additionally, it may be of interest to evaluate the results based on searching a number of test prints against the same known records on different vendor systems, instead of using separate databases on separate systems as was done here. One additional area of investigation could be whether a combinatorial approach to classification would be valid. In the discussion of the classifier using the percentage of matched minutiae and the match score, it was theorized that multiplying one by the other was weighting the response towards the true matches. From this, combining some of the other ROC results, perhaps using a decision matrix, may show improved results over applying each classifier separately. As it was shown that there was difference between the classifiers and some worked better than others with more or less information, combining classifiers in such a way may provide for a more reliable analysis overall.

Ultimately, signal detection shows good promise in classifying fingerprint comparison responses, in both automated comparison systems and quantitative considerations of latent fingerprint examiners. It demonstrates that fingerprint identification will show reliable results, so long as appropriate considerations are taken to the focus by which they are evaluated. The more support one gains for an identification, the better the conclusions that can be reached. Some future work remains to be done in determine whether the conclusions drawn from this research will hold up when applied to human examiners, however, the potential for a decision-making based evaluation is possible with reasoned application of this method.

Appendix A – Description of test set

Table 2. Raw data on latent test set.

Code	Source Finger	Pattern	Minutiae
A	Left middle	Central pocket loop whorl	26
B	Left index	Plain arch	16
C	Left middle	Left slant loop	32
D	Right index	Left slant loop	40
E-2	Right index	Plain whorl	38
E-3	Right middle	Right slant loop	30
F	Right middle	Right slant loop	28
G	Right thumb	Right slant loop	18
H	Left index	Left slant loop	38
I	Left middle	Plain arch	30
J	Right little	Right slant loop	24
K	Left index	Plain arch	24
L	Right ring	Plain whorl	34
M	Left index	Left slant loop	36
N	Left index	Left slant loop	24
O-2	Right index	Plain whorl	24
O-7	Left index	Plain whorl	26
P	Right ring	Right slant loop	40
Q	Right index	Left slant loop	26
R	Right index	Left slant loop/tented arch	32
S	Right little	Right slant loop	32
T	Left thumb	Left slant loop	46
U	Right index	Right slant loop	40
V	Right thumb	Double loop whorl	34
W	Left ring	Central pocket loop whorl	22
X	Left index	Double loop whorl	40
Y	Right middle	Right slant loop	24
Z	Left ring	Plain whorl	45
AA	Left index	Left slant loop	24
BB	Left middle	Left slant loop	30
CC	Right ring	Plain whorl	14
DD-3	Right middle	Right slant loop	23
DD-4	Right ring	Right slant loop	36
EE	Left thumb	Double loop whorl	46
FF	Left middle	Left slant loop	20
GG	Left ring	Plain whorl	7*
HH	Right index	Plain whorl	26
II	Left thumb	Left slant loop	45
JJ	Right middle	Left slant loop/tented arch	30
KK	Right index	Left slant loop	38
LL	Left thumb	Plain arch	45
MM	Left ring	Central pocket loop whorl	45
NN	Left index	Plain whorl	35
OO	Left index	Left slant loop	22
PP	Left index	Left slant loop	31
QQ	Right index	Plain whorl	28
RR	Right thumb	Right slant loop	22
SS	Right middle	Central pocket loop whorl	30
TT	Left thumb	Plain whorl	34
UU	Left ring	Plain whorl	36
VV	Left index	Left slant loop	24

*Excluded from test set – insufficient detail

Appendix B – Summary of system searches and thresholds

Table 3. Summary results for AFIX Tracker searches.

	Minimum	Maximum	False Neg. Cutoff	False Pos. Cutoff	AUC	SE
Minutiae	2	40	2	19	0.9469	0.0051
Score	64	9184032	318	254184	0.9203	0.0062
Δscore	0	9051564	5	219870	0.9575	0.0046
Score/Min	32	264381	106	34227.2	0.8848	0.0072
%Match*Score	42.66667	8811663.2	318	254184	0.9621	0.0044

Table 4. Summary results for Cogent Systems AFIS searches.

	Minimum	Maximum	False Neg. Cutoff	False Pos. Cutoff	AUC	SE
Minutiae	3	43	5	29	0.7614	0.0076
Score	451	2946	752	1150	0.9571	0.0038
Δscore	0	1768	0	199	0.8592	0.0064
Score/Min	27.78571	337	46.05555556	337	0.5016	0.0079
%Match*Score	223.8947	2759	416.25	1617.076923	0.9516	0.0040

Table 5. Results of the original searches and resubmissions of varying the starting minutiae cluster.

		Minimum	Maximum	False Neg. Cutoff	False Pos. Cutoff	AUC	SE
Original searches	Minutiae	2	32	2	16	0.9398	0.017
	Score	96	7495232	699	135516	0.9116	0.021
	Δscore	0	7393729	6	59284	0.9350	0.018
	Score/Min	48	234226	233	25967.5	0.8777	0.023
	%Match*Score	64	7495232	528	76815.6667	0.9319	0.018
Resubmissions	Minutiae	2	31	3	17	0.9711	0.012
	Score	636	4672668	4144	136150	0.9091	0.020
	Δscore	0	4565748	3	49622	0.9512	0.015
	Score/Min	212	166881	1036	22516.2	0.8390	0.026
	%Match*Score	564.75	4088584.5	3315.2	74417.7778	0.9645	0.013

Table 6. Results of the original searches and resubmissions of clustering minutiae in the delta area.

		Minimum	Maximum	False Neg. Cutoff	False Pos. Cutoff	AUC	SE
Original searches	Minutiae	4	38	5	29	0.7123	0.022
	Score	658	2662	765	1110	0.9270	0.013
	Δscore	0	1544	0	103	0.8554	0.017
	Score/Min	30.2069	247.75	48.05	247.75	0.5679	0.023
	%Match*Score	369.6786	2451.8421	569.318	1087	0.9529	0.011
Resubmissions	Minutiae	5	38	7	29	0.7490	0.023
	Score	693	2769	930	1155	0.9809	0.0074
	Δscore	0	1578	0	116	0.8389	0.019
	Score/Min	30.03571	215.2	52.7	215.2	0.5425	0.024
	%Match*Score	381.3333	2856.5	650.088	1140	0.9489	0.012

Table 7. Results of the original searches and resubmissions using parameter-limited searches.

		Minimum	Maximum	False Neg. Cutoff	False Pos. Cutoff	AUC	SE
Original searches	Minutiae	3	42	5	28	0.7319	0.022
	Score	661	2752	752	1072	0.9494	0.011
	Δscore	0	1622	2	153	0.8759	0.017
	Score/Min	30.64	271.66667	56.8333	271.6667	0.5502	0.023
	%Match*Score	286.6111	2509.7333	585.375	1064	0.9456	0.012
Resubmissions	Minutiae	1	42	4	28	0.7121	0.022
	Score	663	2760	730	1105	0.9312	0.013
	Δscore	0	1623	0	118	0.8657	0.017
	Score/Min	32.4	848	55.4	848	0.5594	0.023
	%Match*Score	65.23077	3297.25	588.923	1049	0.9526	0.011

Table 8. Statistical testing, two-tailed p-values from aggregate AFIX results.

	Minutiae	Score	ΔScore	Score/Min
Minutiae	-	-	-	-
Score	0.000912	-	-	-
ΔScore	0.124	<0.00001	-	-
Score/Min	<0.00001	0.000181	<0.00001	-
%Match*Score	0.0234	<0.00001	0.468	<0.00001

Table 9. Statistical testing, two-tailed p-values from aggregate CAFIS results.

	Minutiae	Score	ΔScore	Score/Min
Minutiae	-	-	-	-
Score	<0.00001	-	-	-
ΔScore	<0.00001	<0.00001	-	-
Score/Min	<0.00001	<0.00001	<0.00001	-
%Match*Score	<0.00001	0.322	<0.00001	<0.00001

APPENDIX C – Resubmits

Resubmitted Searches – Alternate Cluster

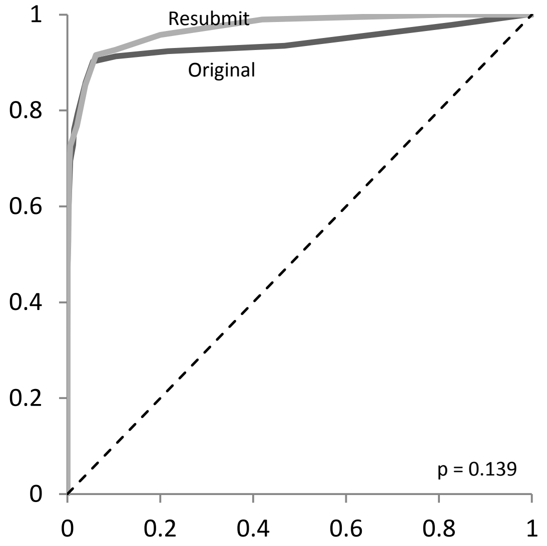


Figure 21. ROC curves for cluster resubmissions when evaluated based on minutiae matched.

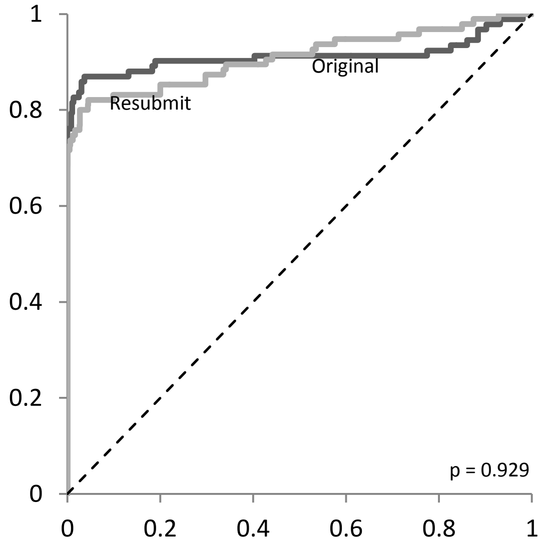


Figure 22. ROC curves for cluster resubmissions when evaluated based on match score.

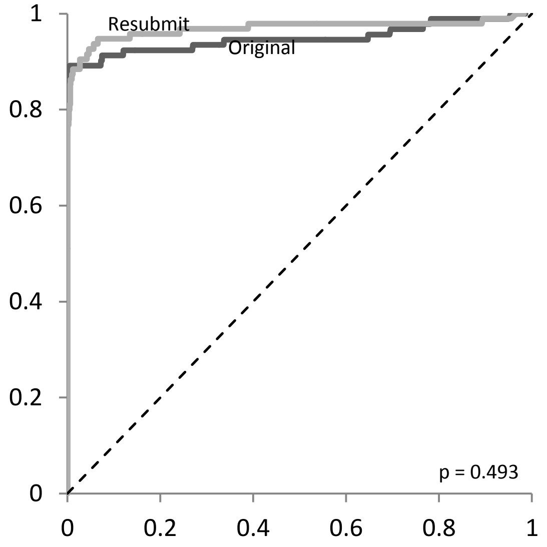


Figure 23. ROC curves for cluster resubmissions when evaluated based on the change in match score.

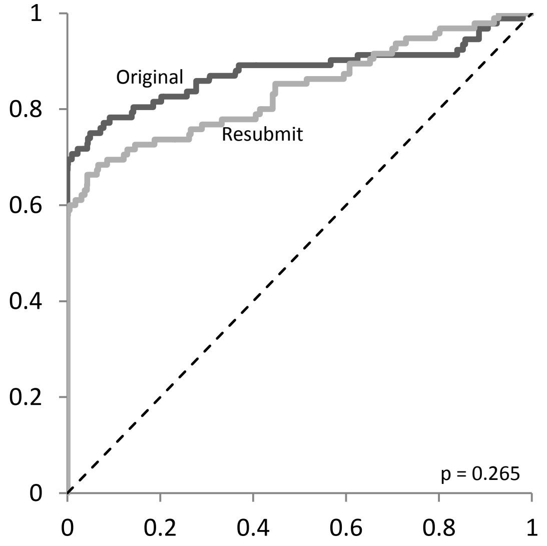


Figure 24. ROC curves for cluster resubmissions when evaluated based on match score divided by number of matched minutiae.

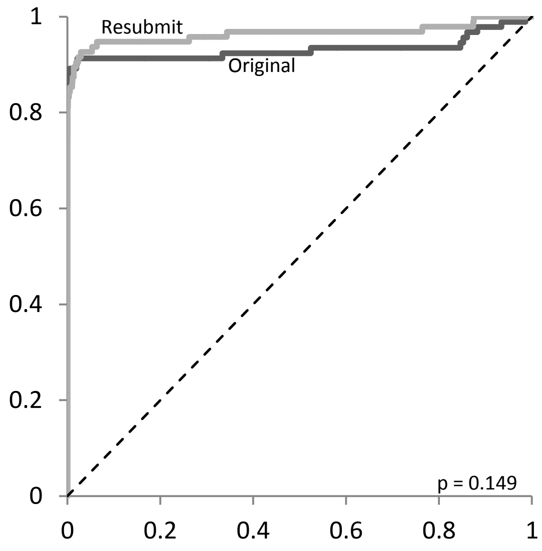


Figure 25. ROC curves for cluster resubmissions when evaluated based on the percentage of matched minutiae multiplied by match score.

Resubmitted Searches – Delta Searching

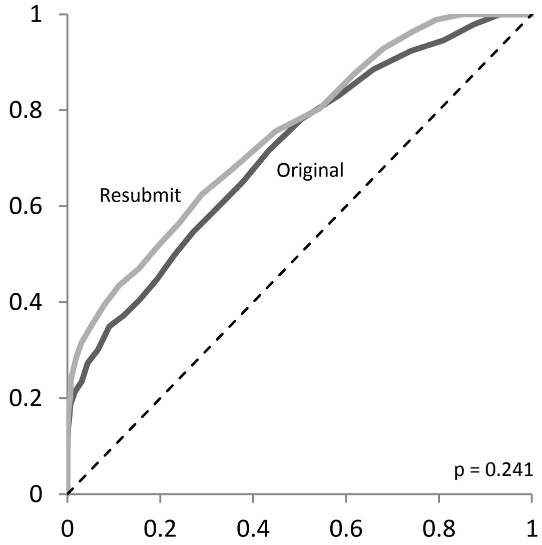


Figure 26. ROC curves for delta resubmissions when evaluated based on minutiae matched.

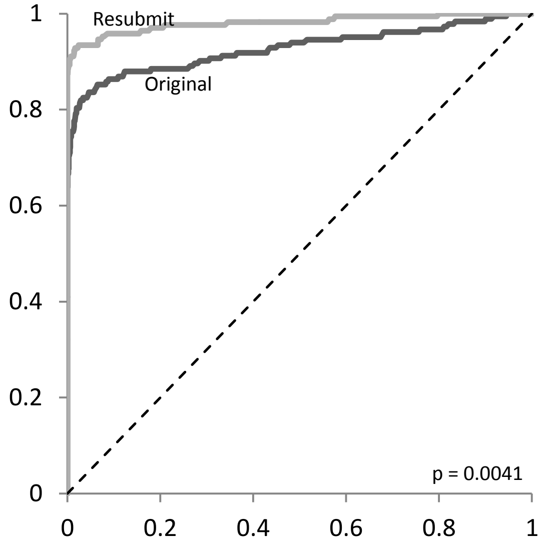


Figure 27. ROC curves for delta resubmissions when evaluated based on match score.

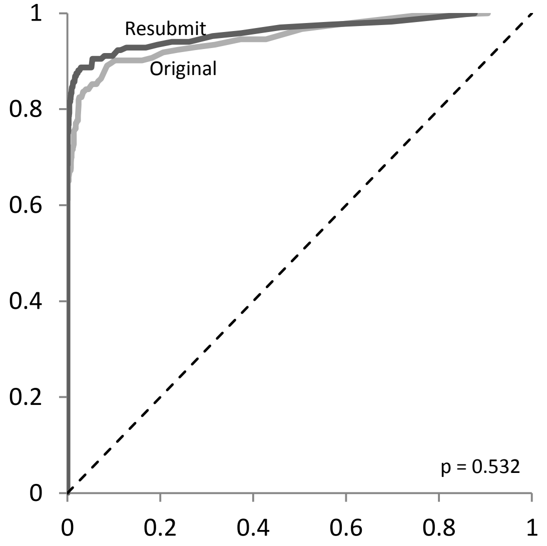


Figure 28. ROC curves for delta resubmissions when evaluated based on the change in match score.

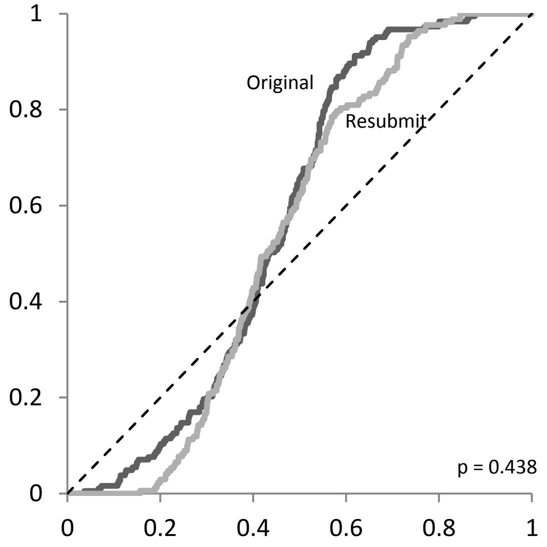


Figure 29. ROC curves for delta resubmissions when evaluated based on match score divided by number of matched minutiae.

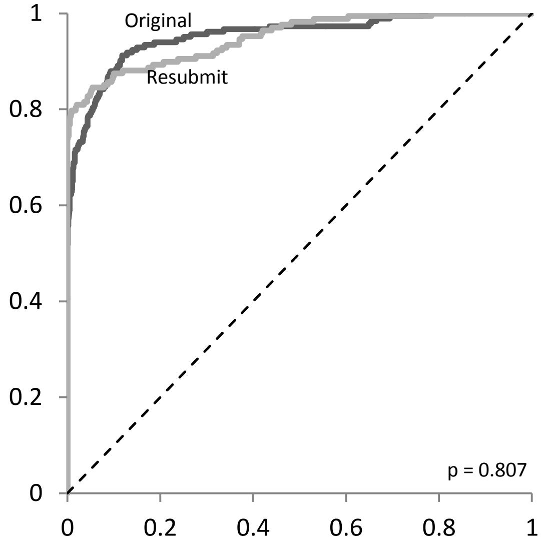


Figure 30. ROC curves for delta resubmissions when evaluated based on the percentage of matched minutiae multiplied by match score.

Resubmitted Searches – Parameter Limitations

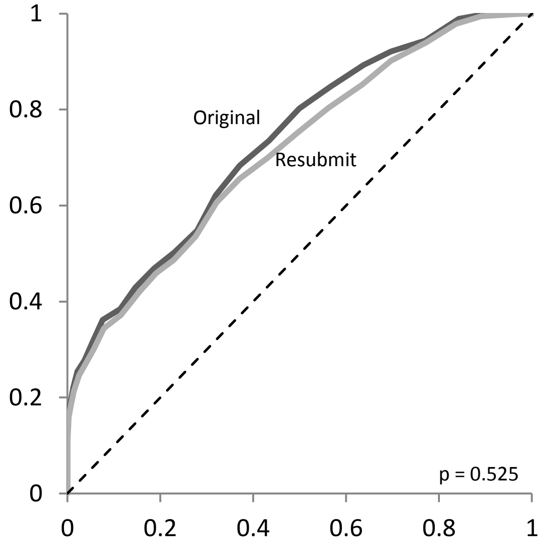


Figure 31. ROC curves for parameter-limited searches when evaluated based on minutiae matched.

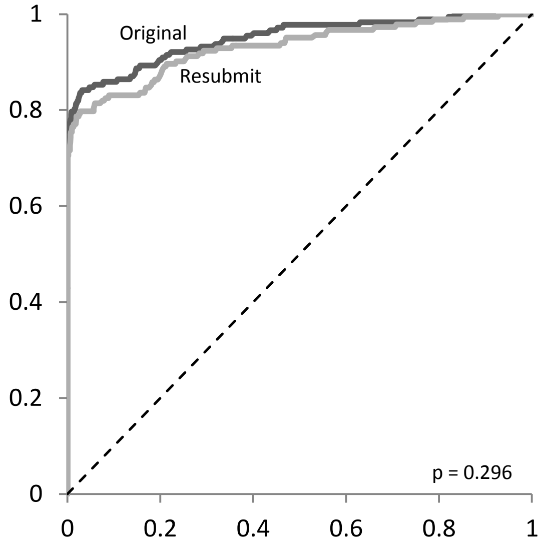


Figure 32. ROC curves for parameter-limited searches when evaluated based on match score.

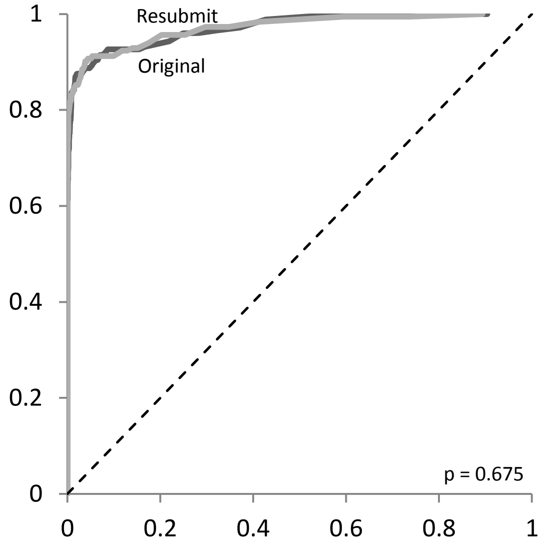


Figure 33. ROC curves for parameter-limited searches when evaluated based on the change in match score.

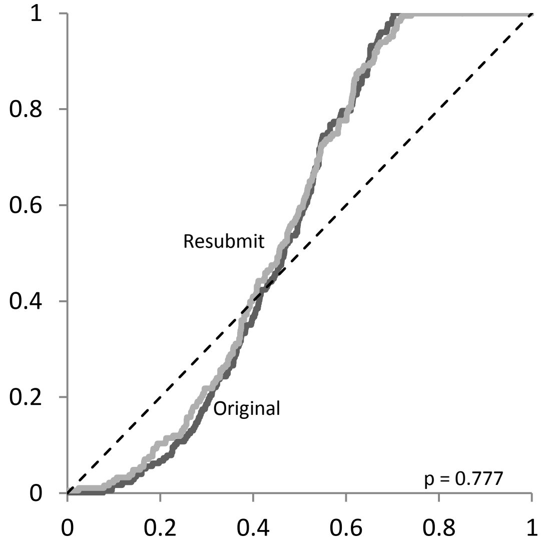


Figure 34. ROC curves for parameter-limited searches when evaluated based on match score divided by number of matched minutiae.

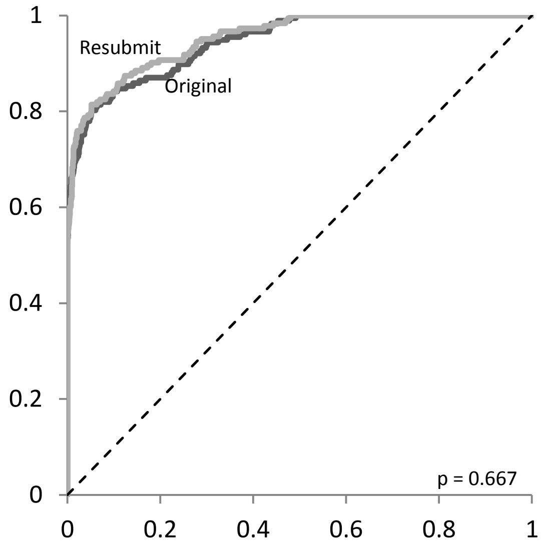


Figure 35. ROC curves for parameter-limited searches when evaluated based on the percentage of matched minutiae multiplied by match score.

REFERENCES

1. Champod, C., *Edmond Locard – Numerical Standards & 'Probable' Identifications.* Journal of Forensic Identification, 1995. **45**(2): p. 136-63.
2. Champod, C., et al., *Fingerprints and Other Ridge Skin Impressions*. 2004, Boca Raton: CRC Press.
3. *An analysis of standards in fingerprint identification.* FBI Law Enforcement Bulletin, 1972. **June**.
4. *Report of the Standardization Committee of the International Association for Identification* Fingerprint and Identification Magazine, 1973. **55**(4): p. 11-16.
5. *Ne'urim declaration*. in *International Symposium on Fingerprint Detection and Identification*. 1995. Ne'urim, Israel.
6. Ashbaugh, D., *Quantitative-Qualitative Friction Ridge Analysis: An Introduction to Basic and Advanced Ridgeology*. 1999, Boca Raton: CRC Press.
7. Vanderkolk, J., *Levels of quality and quantity in detail.* Journal of Forensic Identification, 2001. **51**(5): p. 461-468.
8. Cole, S., *More than zero: accounting for error in latent fingerprint identification.* Journal of Criminal Law & Criminology, 2005. **95**(3): p. 985-1078.
9. *Daubert v. Merrell Dow Pharmaceuticals*. 1993, Supreme Court of the United States.
10. Saks, M.J. and J.J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science.* Science, 2005. **309**(5736): p. 892-895.
11. *U.S. v Byron Mitchell*. 1999, United States District Court for the Eastern District of Pennsylvania.
12. Stoney, D.A. and J.I. Thornton, *A Critical Analysis of Quantitative Fingerprint Individuality Models.* Journal of Forensic Sciences, 1986. **31**(4): p. 1187-1216.
13. Osterburg, J.W., et al., *Development of a Mathematical Formula for the Calculation of Fingerprint Probabilities Based on Individual Characteristics.* Journal of the American Statistical Association, 1977. **72**(360): p. 772-778.
14. Pankanti, S., S. Prabhakar, and A.K. Jain, *On the individuality of fingerprints.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. **24**(8): p. 1010-1025.
15. Neumann, C., et al., *Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae.* J Forensic Sci, 2007. **52**(1): p. 54-64.
16. General, O.o.I., *A Review of the FBI's Handling of the Brandon Mayfield Case (Unclassified Executive Summary)*, D.o. Justice, Editor. 2006.
17. Maltoni, D., et al., *Handbook of Fingerprint Recognition*. Second ed. 2009, London: Springer.
18. McNicol, D., *A Primer of Signal Detection Theory*. 1972, London: Allen & Unwin.
19. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**: p. 29-36.
20. Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861-874.
21. ANSI/NIST, *Image resolution requirements*, in *American National Standard for Information Systems— Data Format for the Interchange of Fingerprint Facial, & Other Biometric Information – Part 1*. . 2007. p. 10-12.
22. Hanley, J.A. and B.J. McNeil, *A method of comparing the areas under receiver operating characteristic curves derived from the same cases.* Radiology, 1983. **148**: p. 839-843.
23. Hawthorne, M., *Fingerprints: Analysis and Understanding*. 2009, Boca Raton: CRC Press.