Year : 2020

UNDERSTANDING, EXPANDING, AND PREDICTING THE SUITABILITY DECISION IN FRICTION RIDGE ANALYSIS

Eldridge Heidi

UNIL | Université de Lausanne

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

ECOLE DES SCIENCES CRIMINELLES

UNDERSTANDING, EXPANDING, AND PREDICTING THE SUITABILITY DECISION IN FRICTION RIDGE ANALYSIS

THESE DE DOCTORAT

présentée à la Faculté de Droit, des Sciences Criminelles et d'Administration Publique de l'Université de Lausanne

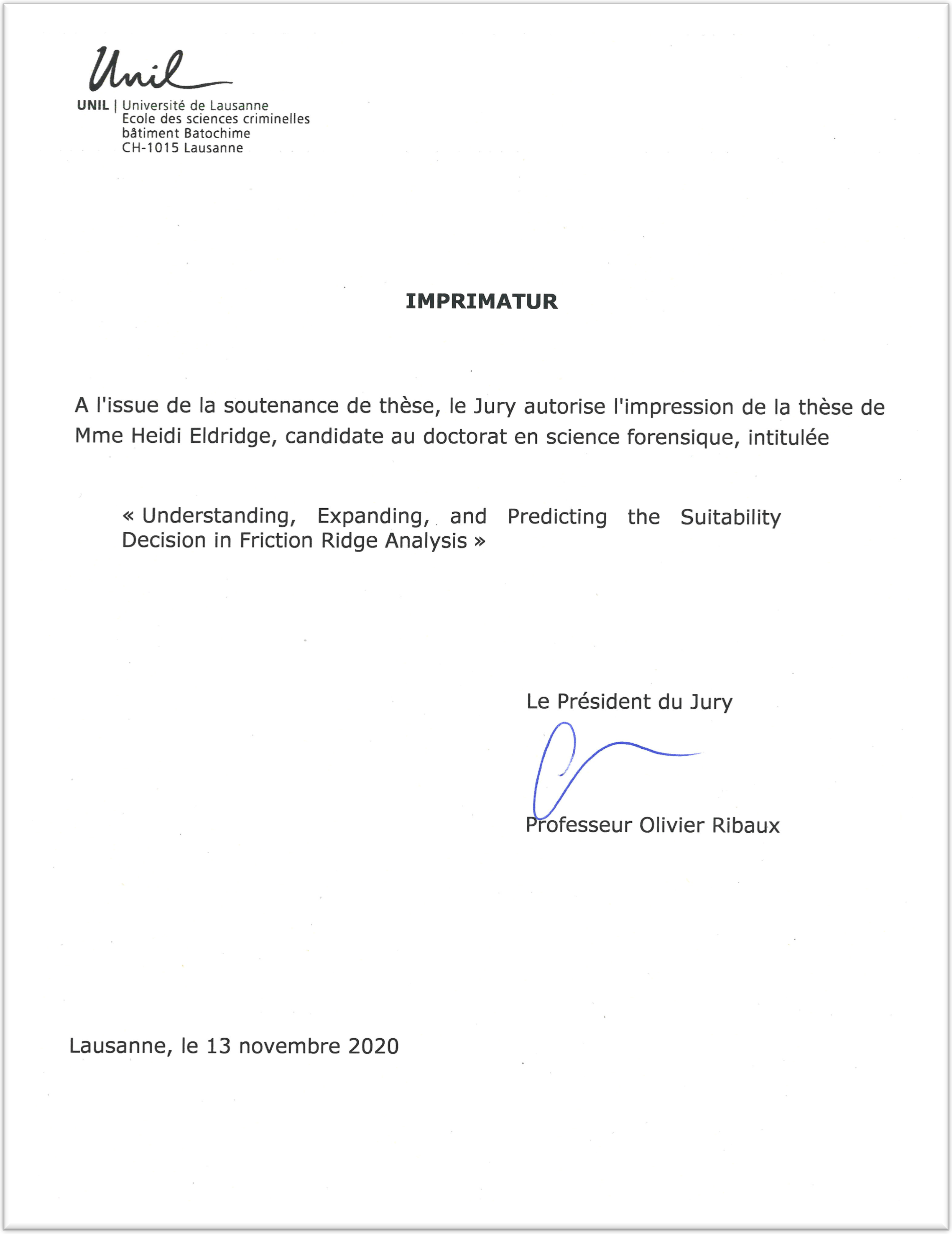
pour l'obtention du grade de Docteure en science forensique

par

HEIDI ELDRIDGE

Directeur de thèse : Prof. Christophe Champod

Lausanne (2020)



UNIL | Université de Lausanne
Ecole des sciences criminelles
bâtiment Batochime
CH-1015 Lausanne

IMPRIMATUR

A l'issue de la soutenance de thèse, le Jury autorise l'impression de la thèse de Mme Heidi Eldridge, candidate au doctorat en science forensique, intitulée

« Understanding, Expanding, and Predicting the Suitability Decision in Friction Ridge Analysis »

Le Président du Jury

Professeur Olivier Ribaux

Lausanne, le 13 novembre 2020

Acknowledgements

This dissertation has been eight years in the making. Completing a dissertation while working full-time and trying to be a wife and mother is no easy task, and I would not have gotten through it without the proverbial village supporting me.

First and foremost, I must extend my heartfelt thanks to Professor Christophe Champod, my thesis advisor, mentor, and friend. Christophe has been an ideal mentor, allowing me space to develop my ideas, challenging me to defend my decisions, requiring me to learn R, patiently explaining the machine learning parts to me time and again, collaborating with me on other interesting research projects outside of this work, and always making me feel welcome on my trips to Lausanne.

Next, this project would not have happened without the assistance of Marco De Donno. I cannot begin to describe the work he has put into customizing the PiAnoS interface, pulling data when I needed it, troubleshooting as problems came up, and supporting the development of the journal manuscripts produced in association with this research. I would like to thank my committee: Professors Christophe Champod and Alex Biedermann and Drs. Glenn Langenburg and Nicole Egli Anthonioz for their valuable input. In particular I would like to thank Glenn for his encouragement way back in 2008, when contacting Christophe about starting a PhD project was just a tiny idea germinating in the back of my mind.

I have held two different jobs during the completion of this project, and both have been instrumental in my success. I would like to thank Alice White (then Maceo) for helping me get access to the images I needed, as well as just being a good friend. I would also like to thank RTI International for providing internal funding that allowed me both labor time and travel funds to get to Lausanne when I needed to work with Christophe in person. I am fully cognizant of what a privilege it was to have this level of support. I'd also specifically like to thank Donia Slack for her unwavering moral support, and Vicki McCall and Ashley Cochran who served as my confidential liaisons.

The friendship and support of many people helped on this journey. I would like to especially acknowledge Joëlle Vuille, Heather Conner, Alicia Wilcox, Henry Swofford, Penny Dechant, and Carey Hall, who were always there when I needed to bounce an idea, get away from it all, or just have a good chat. I'd also like to thank those who made me feel at home in Lausanne, in particular Tacha Hicks Champod; my fellow students Lydie Samie-Foucard, Ilaria DeMarch, and Emmanuelle Erne; Bill Thompson, who I saw more in Lausanne than I do in the US; and Everett Peachey and Suliana Manley, two dear friends I've had for decades who by pure coincidence happen to live in Geneva and Lausanne respectively. I'd also like to thank Tom Busey and Bobbie Spellman, who, in addition to being fantastic collaborators on another project, have been great sounding boards, cheerleaders, and friends. Last, but certainly not least, I need to thank all the study participants who gave so generously of their time to make this project possible, and my family: my husband Mike and daughter Tyche, who put up with my stress and frequent prolonged absences while working on this and who believe that I am the smartest woman in forensic science and make me feel like a rock star.

Abstract

This PhD thesis is an examination of the concept of “suitability” in friction ridge analysis. The decision of whether or not a friction ridge impression is forensically useful, and what specifically it is useful for, can have far-reaching consequences to the criminal justice system since it is the gate through which marks must pass to proceed in the examination at all. This thesis unpacks suitability into component parts, questioning what different decisions examiners may face in determining what a particular mark may be useful for and to what degree it may be useful, as well as investigating what features they use to support those decisions. It proposes four distinct scales of suitability that have applications for policy, practice, quality assurance, research, testing, training, and testimony.

The work is divided into two main parts: first, a white box study to better understand the information that is most considered by examiners when making decisions; and second, development and validation of a predictive suitability model that relies on both key observations from a human expert and automated measures from existing quality tools.

This thesis introduces the benefits of considering suitability in an expanded and more nuanced way. It also demonstrates the performance that can be achieved by a hybrid examiner-algorithm model that leverages the strengths of both to provide consensus-based guidance.

Résumé

Cette thèse de doctorat étudie le concept de "suffisance" associé à l'analyse des traces papillaires. La décision de savoir si une trace est utile ou non d'un point de vue forensique, et à quoi elle sert spécifiquement, peut avoir des conséquences considérables pour le système de justice pénale, puisqu'il s'agit de la porte d'entrée par laquelle toutes les traces doivent passer pour pouvoir éventuellement continuer vers la phase de comparaison. Cette thèse décompose le concept de suffisance en plusieurs composants, en s'interrogeant sur les différentes décisions auxquelles les examinateurs sont confrontés lorsqu'ils déterminent la qualité d'une trace particulière et son degré d'utilité, et en étudiant les caractéristiques qu'ils utilisent pour étayer ces décisions. Ce travail propose quatre échelles distinctes de suffisance qui ont des applications pour la politique systémique, la pratique, l'assurance qualité, la recherche, les tests de compétence, la formation ou le témoignage de l'expert.

Le travail est divisé en deux parties principales : premièrement, une étude de type "boîte blanche" pour identifier quelles sont les informations les plus prises en compte par les examinateurs lorsqu'ils prennent leurs décisions ; et deuxièmement, le développement et la validation d'un modèle prédictif de la suffisance qui repose à la fois sur les observations clés d'un expert humain et sur les mesures automatisées des outils de qualité existants.

Cette thèse présente les avantages de considérer le concept de suffisance d'une manière élargie et plus nuancée. Elle démontre également les performances qui peuvent être atteintes par un modèle hybride examinateur-algorithme qui exploite les forces des deux pour fournir des évaluations basées sur le consensus.

List of Acronyms

AAAS – American Association for the Advancement of Science

ACE-V – Analysis, Comparison, Evaluation, and Verification

AFIS – Automated Fingerprint Identification System

AQ – AFIS Quality

AUC – Area Under Curve

CTS – Collaborative Testing Services

ESLR – Estimated Score-Based Likelihood Ratio

FBI – Federal Bureau of Investigation

FRS – Friction Ridge Subcommittee (of OSAC)

GT – Ground Truth

ID – Identification

L3D – Level 3 Detail

LFIQ – Latent Fingerprint Image Quality

LO – Lights Out

LOOCV – Leave One Out Cross Validation

LPE – Latent Print Examiner(s)

LVMPD – Las Vegas Metropolitan Police Department

ML, MLA – Machine Learning, Machine Learning Algorithm

NAQ – Not AFIS Quality

NCFS – National Commission on Forensic Science

NIST – National Institute of Standards and Technology

NFIQ – NIST Fingerprint Image Quality

NRC – National Research Council

NV – No Value

OSAC – Organization of Scientific Area Committees (see Appendix A – Terminology)

PCAST – President's Council of Advisors on Science and Technology

PI – Principal Investigator

PiAnoS – Picture Annotation System

ppi – pixels per inch

QA – Quality Assurance

RF – Random Forest

RFE – Recursive Feature Elimination

ROC – Receiver Operating Characteristic

RTI – Research Triangle Institute

SWGFAST – Scientific Working Group on Friction Ridge Analysis Study and Technology

ULW – Universal Latent Workstation

UQM – User Quality Metrics

VB – Value for Both (see Appendix A – Terminology)

VID – Value for Identification (see Appendix A – Terminology)

VIDO – Value for Identification Only (see Appendix A – Terminology)

VEO – Value for Exclusion Only (see Appendix A – Terminology)

Table of Contents

1 Introduction

Friction ridge comparison is done using a process that is often articulated in four phases: Analysis, Comparison, Evaluation, and Verification (ACE-V). During each phase of this process, the human examiner makes decisions in which subjectivity is unavoidably introduced. This subjectivity has led to variability, which has resulted in multiple examinations of the same friction ridge images yielding different conclusions (Ulery et al. 2011, 2012; Swofford et al. 2013; Neumann et al. 2013; Ulery et al. 2016), including several high-profile disagreements (United States Department of Justice and Office of the Inspector General - Oversight and Review Division 2006; Campbell 2011). These disagreements have manifested both as between-examiner variability and as within-examiner variability upon viewing the same evidence at different times.

This potential for variability can reflect negatively on friction ridge comparison science in a variety of ways. First, consistency (repeatability and reproducibility) is a component of the reliability of a method (Langenburg 2009). Without demonstrated reliability of results and conclusions, friction ridge examinations may not meet the threshold for admissibility in a court of law (The National Judicial College & Justice Speakers Institute 2019). Second, variability leads to professional disagreements. If two experts cannot agree on their conclusions after reviewing the same data using the same methodology, both the methodology itself and the expertise of the experts are called into question. Third, without consistency in methods and terminology, experts are not speaking the same language, and thus cannot communicate effectively, or conduct and report meaningful research. Without consistency of methods and terminology, there will always be some level of confusion regarding *which* version of the method was being used, or *what* level of difficulty was tested, or *how* the conclusion was reached, to name only a few.

(A brief note given the importance of terminology: throughout this work, terms will be used that may be unfamiliar to the lay reader, or that we are using in very specific ways that may differ slightly from the way some examiners may understand them. Whenever a term that needs defining is used for the first time, it will be *underlined and in italics*. This will alert the reader to terms that are defined or described in Appendix A.)

The observed variability caused by subjectivity in every step of the friction ridge comparison process has been a source of concern to observers and has been commented upon in several notable publications (Edwards 2009; President's Council of Advisors on Science & Technology 2016; AAAS 2017). This work aims to reduce variability in friction ridge comparisons by making a deep exploration into the *suitability*—or "*value*"—decision that comes at the end of the initial Analysis phase after the quality of a *mark* has been assessed to determine whether to continue the ACE-V process. There are a number of decision points within the Analysis phase, including the selection and *weight* assignment of minutiae and other features; the interpretation of distortion and noise; and the ultimate determination of suitability, each of which is an opportunity for variability to enter the process. This research seeks to reduce this variability to the extent possible by first understanding what information is most important to latent print examiners (LPE) in reaching a suitability decision, then building a predictive hybrid examiner-algorithm model that combines that highly diagnostic information with automated measures to produce standardized guidance on suitability. The overall approach to the research is described in Section 1.4.

There are additional factors that add to the challenge of understanding how examiners reach suitability decisions, such as the lack of clear thresholds and the fact that very little is known about how the information that supports the suitability decision is weighted. These factors, which increase the complexity of the undertaking, are described in Sections 1.1 and 1.2.

Finally, a central question to this thesis is, “What is suitability?” We will argue that there is much more to this question than a simple determination of value versus no-value. This argument is broached in Section 1.3, then explored in greater detail in Chapter 2.

The outcome of this work will be a greater understanding of the nature of the suitability decision and how examiners reach that decision, the introduction of new scales and conclusions to give LPEs a more nuanced way to think about suitability, and the development of a software tool capable of providing consensus suitability guidance across several dimensions following the input of only a few key observations. Although some commentators feel that all of these questions could be resolved by the simple application of an automated, probabilistic assessment tool, this research takes the stance that there may still be value to the intervention of the human examiner and that the field is not yet ready for a fully probabilistic approach. Thus, we approach the field through the lens of taking the next logical step away from binary conclusions toward conclusion scales with more gradients, while evaluating the contribution of the human examiner to the decision-making process.

1.1 Thresholds remain undefined

When considering suitability, there is one obvious threshold—value or no-value, usually taken to mean whether or not a mark will be taken forward to the next step in the comparison process. But where does that threshold lie? Each examiner has a personal threshold for where value is achieved, and many also differ on their definition of value. Value can variably refer to a mark that contains sufficient information to be identified to a source, to a mark that contains sufficient information only to definitively exclude a subject, or simply a mark that contains sufficient information to provide some useful information to inform an investigation such as providing an investigative lead.

And once value has been established, there can be the issue of *complexity*—the of-value mark represents a wide range in quality of marks from the barely-of-value mark to the mark that is so clear and complete, it may exceed the corresponding *print* in quality. Within this range, there exists a hidden threshold separating marks that are complex from those that are not. With no clear definition of what constitutes a “complex” mark, this decision is essentially subjective and thus expected to vary widely from examiner to examiner.

Many reports addressing bias, error rate, and variability in decisions (Dror et al. 2005; Langenburg 2012; Ashbaugh 1999), have noted the greatest fluctuation on "complex," "difficult," or "borderline" marks, but these terms are not defined. Most often in studies, the difficulty of the test mark is determined by a consensus of some number of experts, but no objective measure of how those experts came to their determination is given, nor is discourse typically had on the treatment of outliers. Likewise, one of the most prevalent criticisms of commercial proficiency tests, such as that offered by Collaborative Testing Services (CTS), is that the difficulty of the marks given on the test is not representative of casework, or is not challenging enough (Max et al. 2019; Kelley et al. 2020; Koertner and Swofford 2018). Although Koertner and Swofford do measure the clarity of images to reach this conclusion, they also note that other factors in addition to clarity likely contribute to subjective assessments of difficulty and complexity, and they do not set thresholds or definitions for either term, simply noting that as clarity measurements decreased, difficulty assessments increased. Thus, terms are being used in the literature that have no real agreed-upon meaning. What is a "difficult mark"? What is "representative of normal casework"? Without objective thresholds that define and standardize these terms, meaningful communication on the issues surrounding them seems a hopeless task.

Marks of varying qualities should not be treated equally. Complex marks should be more carefully analyzed and more thoroughly documented and reviewed (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2013; Dror 2009). Policy should require increased quality assurance (QA) measures or prohibit examiners from making identifications as mark complexity increases, whereas time-savings will result from loosening regulations on high-quality marks that carry a low chance of error (Ulery et al. 2013; Kellman et al. 2014; Ashbaugh 1999). However, without a standardized method to define these thresholds and ascertain complex marks, any such policies would be arbitrary.

In essence, without some means of standardizing thresholds, whether or not a mark will be declared of value and compared or declared no-value and discarded, or whether a mark will be recognized as complex and treated with extra care or designated as high quality and given superficial review is largely a matter of chance. One examiner will reach one conclusion regarding the suitability of the mark, another examiner may reach a different conclusion entirely. This jeopardizes our criminal justice system, as decisions that could affect the outcome of a case are left to little more than the luck of the draw—which examiner will be assigned the case? Will they be having a "good eye day" when they are?

One goal of this research will be, not to define thresholds directly, but to create a tool that can put a given mark on one side or the other of these thresholds in such a way that a consensus of experts would likely agree with the decision.

1.2 Information supporting decisions is largely unknown

It is not understood what specific factors go into determinations of value, complexity, and assignment of weight. Ashbaugh suggests factors that should be considered during Analysis (substrate, matrix, red flags, etc) (Ashbaugh 1999), but their interpretation and weighting is left up to the individual examiner.

Although the Analysis phase has been described as an intelligence-gathering process where the quality of the mark is assessed, and the areas to consider during Analysis have been well laid-out (e.g. Levels One, Two, and Three detail, distortion factors, anatomical source, orientation, size of the impression, etc) (Ashbaugh 1999), nobody has broken down the steps, or phases, of Analysis itself.

This work proposes that there are three distinct steps, or tasks, that occur while gathering *information* about the mark during Analysis:

- **The Observation Task – What do you see?**

  The first task of Analysis is to catalog the observations that are made of the mark. This may be done sub-consciously, or explicitly; mentally, or with extensive documentation. Regardless of how it is done or how it is documented, the first step is always the same—look at the mark and determine what information is visible.

- **The Assigning Weight Task – Is it useful?**

  During this task, the features that were noted are assessed for their value to the mark, and ultimately, to a comparison. The assignment of weight to each piece of observed data is determined by answering (again, whether sub-consciously or explicitly) two questions:

  - How distinctive is it?

    This concept is often expressed by use of the terms “selectivity” or “discriminability”. However, selectivity is a statistical concept that is not often clear or linguistically accessible to jurors, along with being frequently misused by latent print examiners themselves and discriminability is a term that is commonly misunderstood as “discrimination” in the sense of racial discrimination and negatively interpreted by laypersons. Hence, clearer terminology is desirable.

    The author prefers the use of the term “distinctive”. This is a word that is in common usage in the English language and will be more familiar to the average juror. Its definitions include “markedly individual,” “notable,” and “serving to differentiate.” Thus, it neatly encapsulates the real crux of the concept of selectivity—when a feature is being selected, the examiner is considering its rarity (how markedly individual it is); whether it is notable (if there is something quirky or unusual about the feature that makes it stand out and grab one's notice—that feeling that the examiner would be certain to recognize it if they saw it again); and, overall, its capacity to differentiate (whether it can help to discriminate *this particular mark* from other marks that may have a similar appearance).

- How confident am I?

  Because every touch differs slightly in pressure, deposition matrix, substrate, etc, no two impressions can ever be exactly alike. Thus, every time an LPE looks at a mark, there is some amount of distortion present and the examiner is doing some level of interpretation of the differences between two impressions and assigning *tolerances* for how much dissimilarity is acceptable. They are also assessing the *reproducibility* of the features, considering whether what they see is likely to appear in other impressions of the same area.

  Although they may not be aware of it, when the examiner is interpreting distortion and assigning tolerances, what they are really doing is assessing the *risk* of error associated with each feature determination. How certain are they that something is there? How certain are they of the identity of the thing they see (e.g. is it a ridge ending, or a bifurcation? Is that pattern a loop, or could it be a whorl?). What is the chance the examiner has misinterpreted what they think they see? What may be the consequences of such a misinterpretation (e.g., could it lead to a false identification or a false exclusion? Would it simply make the search or comparison more difficult?). The murkier the image, the more difficult it is to answer these questions, and thus, the higher the chance of making an incorrect determination.

  On top of this, there are cultural, policy-driven aspects to be considered. Some agencies (for example, the Dutch experts discussed by Langenburg (2012)) put a premium on only utilizing minutiae that are clear and unambiguous. Others may have additional documentation requirements for marks that have low clarity or are deemed "complex." For an examiner operating in such an agency culture, using a minutia in a smudged or low-contrast area would carry considerably higher risk of *negative outcomes*, ranging from being required to complete additional documentation to being subject to discipline or re-training if the agency determined their decisions were frequently not supported by a consensus. Thus, they must decide for each feature that they *think* they see, how high is the risk associated with using this feature? Is the risk worth the potential benefit (i.e. having the use of that feature during the

Comparison phase, or having sufficient features to even proceed to the Comparison phase)?

This concept of risk has a flipside with which most examiners are more familiar—that of *confidence*. Risk and confidence have an inverse relationship—the lower the risk, the higher the confidence. Thus, when an examiner assigns high confidence to a feature, they have *de facto* determined that the risk of an error (or the risk of an unacceptably negative outcome, should there be an error) is low. Conversely, if the examiner assigns low confidence to a feature, they are indicating that there is a high chance the feature is not what they think it is, or is not present at all, and that there is a higher risk of a negative outcome associated with its use.

In the assigning weight task, the two questions work in concert to settle on an overall, appropriately balanced, opinion of the value, or usefulness, of the observed data. Does this mark contain sufficient weight, to be worth carrying forward into a comparison? This brings us to the third task of Analysis.

- **The Decision Task – What decision do you make?**

  In this final step, the actual Analysis decision is made. The examiner considers the observed data, the cumulative weight of those data, along with any associated concerns about the *reliability* of the data, and makes a predictive determination about the anticipated usefulness of the mark without having ever seen any relevant print(s). Depending upon the particular agency policies, the examiner may consider whether the mark can be used to identify; whether it is of sufficient quality to exclude only; or whether it can merely be used to inform an investigation, even in the absence of the ability to render a definitive conclusion. This consideration results in the determination of value or no-value.

Two types of information that may affect the Assigning Weight task are rarity and clusters. Examiners tend to give additional weight to what they perceive as rare features. Simply put, every examiner knows that a plain ridge ending by itself is more common than a compound feature, such as a spur or an enclosure. Most examiners would even agree that some compound features, such as enclosures, are more common than others, such as trifurcations. Thus, when a rare compound feature is seen during comparison, it is given relatively more weight than a common feature would be given.

Likewise, examiners may factor the presence of highly *distinctive clusters* or *target groups* into their suitability decisions. Groupings of highly distinctive features carry more weight than groupings of common features much in the same way that a cluster of two or more simple minutiae in extremely close proximity may form a more distinctive compound feature[1]. Thus, the value of a target group is closely linked with the concept of rarity.

Whether the target group is noted during Analysis for the searching benefit it may provide, or for the value in distinctiveness offered by its components, it seems clear that the presence of one or more target groups is a criterion that contributes to the overall usefulness of the mark, as assessed during Analysis. A mark that has one or more easily-recognizable target groups should be easier to locate and identify, and likewise may be identified with greater confidence, than a mark that has a spattering of solo minutiae spread throughout the mark without any neighbors to anchor them.

The research approach of this project is designed to take these factors along with other analysis factors into account. A custom web interface (PiAnoS – Picture Annotation System, version 4.2.2-h0.2) was created to capture the information examiners considered while reaching their suitability decisions, including minutiae, clarity, distortion, target groups, feature clusters, rarity, incipient ridges, scars, pores, and confidence in features. These tools and the reasoning behind their inclusion are described in Sections 3.1 and 3.2.1.

1.3 Suitability is multi-faceted

Suitability is often thought of as a binary proposition—either a mark is of value, or it is not. However, in reality, the question of suitability—or what uses a mark may be good for and the level of its goodness for that use—is much more nuanced. Even confining ourselves to the concept of value, suitability is considered by many laboratories to be more than a binary decision. Some laboratories allow for designations of no value, value for identification (*VID*), and value for exclusion only (*VEO*), a trinary choice; others label these distinctions as value for identification versus no-value (Approach 1), whereas still others prefer value for comparison versus no-value (Approach 2) (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2013).

Once a mark has been declared as suitable for *something*, there are still additional things about the usefulness of the mark that the examiner might wish to know. For example, is it complex or not? Is it suitable for AFIS entry or not? The answers to these questions will affect the way LPEs proceed through their workflow and may invoke different policies and requirements. Thus, they are all part of the overall decision of whether a mark is "suitable" and what it is suitable *for*.

The multiple facets of the concept of suitability will be explored in much greater detail in Section 2.2. Because this research seeks to develop a tool capable of measuring the consensus perception of usefulness of a mark along several different dimensions, or scales, this work will refer to the resulting tool throughout as a *utility tool*. The term "*utility*" throughout this work

[1] For example, a spur is more distinctive than a bifurcation and a ridge ending standing alone (Gupta 1968; Osterburg et al. 1977).

will be used as synonymous with “usefulness” and is not meant to invoke or imply utility functions as related to decision theory (Biedermann et al. 2008).

1.4 Objectives of the Thesis

The main objective of this work is to measure and potentially reduce variability between examiners in the suitability decision. We seek to meet this objective through a multi-step process, which is outlined in this section. One obvious solution to mitigate the challenge of variability is to embrace a fully automated, objective process. Although this would indeed reduce variability, it would not necessarily improve overall outcomes (one could imagine a model that predicted the exact same, wrong response every time). Additionally, there are advantages in observation and interpretation skills that the human examiner brings to the process that would be lost in a fully-automated system.

Ways to mitigate inter-rater human variability

- Introduce standardized tools;
- Incorporate *objective data*;
- Introduce standards and guidelines for policy and training;
- Focus attention on a limited number of highly diagnostic features; and
- Increase transparency to allow review.

The box above and to the right lists a number of potential strategies that could be employed to mitigate the impact of the variability that is introduced by a human examiner. This research capitalizes on several of these suggestions. We introduce a standardized tool that focuses the attention of the examiner on highly diagnostic information and away from less fruitful areas. This information is then combined with objective data to offer guidance on the suitability decision along four scales. The information that was noted by the examiner will be recorded and transparent and the tool can be useful to agencies in setting policies and as a training tool.

1.4.1 Demonstrate the current level of variability in decision-making

The first step in the research was to demonstrate the current level of inter-rater variability in decision-making. We began by undertaking a white-box study in which LPEs were requested to annotate only the information they used to reach their suitability decision, to answer a few questions about the quality of the image, and to record their suitability decisions along four scales (described in Section 2.2). We then evaluated the trends in agreement on suitability decisions.

1.4.2 Characterize the information examiners use to make decisions

Next, we used descriptive analyses and machine learning to understand what information examiners were relying upon to reach conclusions. Here, we tested how well each participant’s annotations predicted their *own* suitability decisions along the four scales described in Section 2.2.

1.4.3 Identify the key predictive variables

From this point forward, we worked on designing a model that could to a high degree predict the consensus suitability decision along each scale. Using well-known and common machine learning techniques, we identified the key predictive variables under both idealized (unlimited resources of examiner time and computational resources) and operational (limiting examiner input to minimize time requirements and avoiding expensive software development or heavy computational load) conditions.

1.4.4 Evaluate the benefit of human examiner input

A main question of this thesis was whether the human examiner adds anything to the suitability decision process, or whether a fully automated process would yield more consistent results against a decision made by a consensus of examiners. We tested this question by creating models using examiner-only input, automated-only input, and combinations of examiner and automated input to find which yielded the best performance.

1.4.5 Develop and optimize a model to predict consensus decisions

Once we had finished exploratory modeling, we optimized the model to reach the highest predictive ability using the fewest required inputs and tested the completed model on a new set of participants and new images that were not used in the initial study and model building.

1.5 Structure of the Thesis

Taking an exploded view of suitability…………………………………………………………………..Chapter 2
This chapter explores the idea that suitability is not a single, binary decision, but a multi-faceted one that we will explore along four scales: value, complexity, AFIS quality, and difficulty.

Understanding how examiners evaluate suitability………………………………………………..Chapter 3
In this chapter, we present the manuscript of a published journal article (Eldridge et al. 2020) that describes the white box study used to examine the information examiners use to reach suitability decisions. This chapter provides additional depth and examples that are not included in the article for publication.

Predicting consensus suitability decisions…………………………………………………………………Chapter 4
In this chapter, we present the manuscript of a published journal article (Eldridge et al. 2021) that describes the process of developing, optimizing, and testing the predictive model. This chapter provides additional in-depth discussion of the development and decision-making processes that were part of this research but were not included in the published article for publication.

Limitations, Recommendations, and Future Research……………………………………………….Chapter 5
This chapter describes limitations of the present study and makes recommendations for policy, practice, and future research.

Conclusion…….Chapter 6
This chapter is a standalone summary of the entire dissertation, providing a high-level overview of the purpose, methods, and results of the research and presenting our vision for the future of the field.

2 Taking an exploded view of suitability

An exploded view generally refers to a drawing in which the component parts of an item are shown separately, and their relative position is maintained. For example, Figure 1 is an exploded view sketch of a Swiss Army Knife. In forensic science crime scene work, exploded views are often used to show both the floor and the walls of a scene, giving a simple way of understanding the relationships between items of evidence on the walls and those on the floor or furniture (Figure 2).

In this chapter, we will take an exploded view of the notion of suitability. Rather than considering it as a single thing, we will examine its component parts and consider how they may fit together in the decision-making process that occurs during analysis of a mark. First, we will review the literature that has already considered the question of suitability. We will then describe in detail the four scales of suitability proposed by this work. Finally, we will touch on the implications of these scales for crime laboratory workflow and policy and for the criminal justice system.

Figure 1 Exploded view of a Swiss Army Knife (copyright Matteo Garbi, http://www.studentshow.com/gallery/24369919/swiss-army-knife, used with permission under CC BY-NC 3.0.)

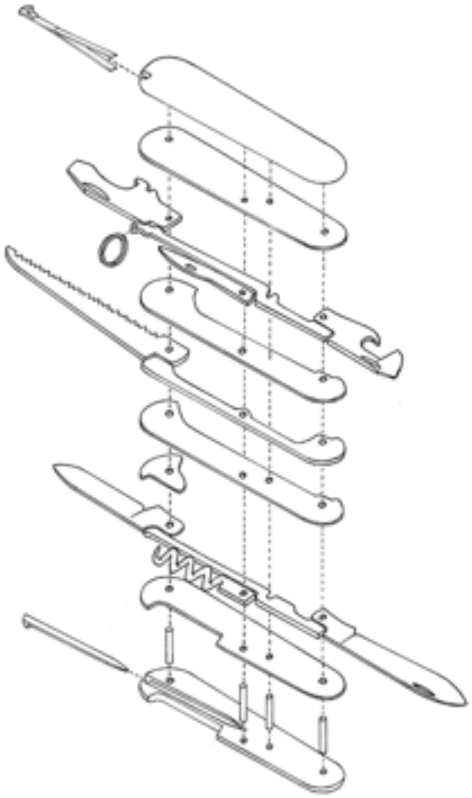
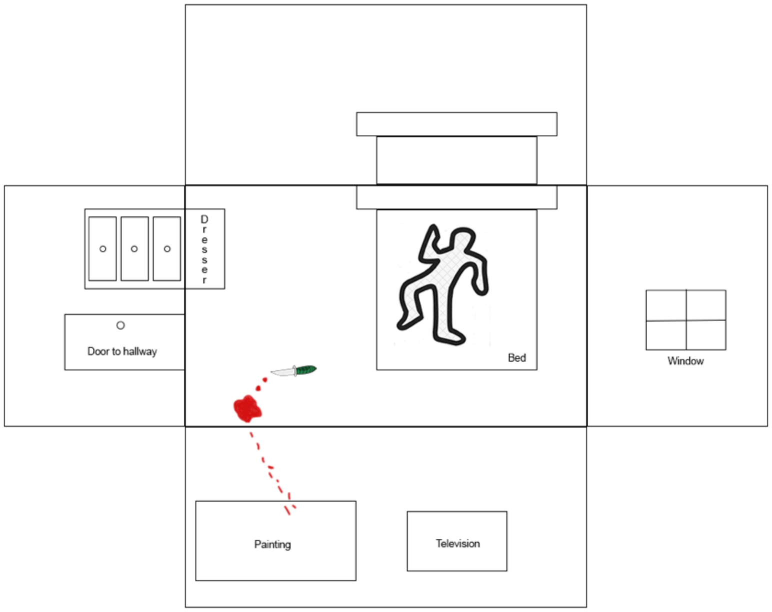


Figure 2 An exploded view of a hypothetical crime scene.



2.1 Existing Quality Metrics

This research does not represent the first attempt to understand the nature of quality assessments or to develop a metric of quality; however, it may be the most holistic and focused on the needs of *manual comparisons* in an operational laboratory. Most previous research in contrast has been focused on the quality needed for an AFIS search and in fact some have focused only on the quality of prints and are not adequate for assessing marks. However, several groups have been working on various ways to measure quality and automate the suitability decision. Here we briefly summarize their previous work.

In 2011, Hicklin et al. used an examiner survey approach to create an automated metric of mark quality (Hicklin et al. 2011). However, they explicitly stated that to them, quality was equated with clarity ("…determined in terms of the confidence that the presence, absence, and details of features can be precisely detected") only and no other factors affecting the quality of a mark (such as specificity or quantity of features) were considered. Furthermore, they explicitly instructed their participants to ignore any agency notions of *utility*, saying:

> "The participants were instructed to base their assessments on their fundamental understanding of friction ridge impressions with no operational goals or legal consequences, not to invoke any agency practices or policies for the analysis of a latent print, and not to consider whether they would testify in court to their assessments" (p. 393).

These results were incorporated into their follow-up research in 2013, in which they presented a clarity map tool that utilized a color-coding scheme for local clarity assessments (Hicklin et al. 2013). This scheme was incorporated into the Extended Feature Set (EFS) that is part of the

Universal Latent Workstation (ULW) and will be tested as a potential predictor variable in early phases of the model development work in the present research.

This research, as with their previous 2011 work, was at pains to distinguish between quality and clarity, and emphasized that this software focused on only the *clarity* of the mark, not its overall quality (a label that they define to include the quantity and distinctiveness of features). Local clarity scores were aggregated into an overall clarity score for the mark that was thresholded into four bins (no value; VEO; very difficult or difficult; and easy or very easy). These results were compared to subjective results from an examiner survey with good correlation. However, this study did not take rarity of features or minutiae counts into consideration. It did suggest that the quality maps could be used for designating complex comparisons and for determining areas of a mark that may not be given weight due to deficiencies in the corresponding areas of the print. Langenburg (2012) suggested this tool could be used for quantifying the number of pixels in each quality category versus the total area of the image to calculate an overall quality score for the image.

The research group of Yoon et al. published two studies in 2012 and 2013 that sought to develop a lights-out approach to measuring quality for use in an AFIS environment (Yoon et al. 2012; Yoon et al. 2013). Their approach focused on average ridge clarity and minutiae count as the main factors driving quality. Rarity of features was not considered in this approach and the focus of the research was on looking at the probability of getting a hit in the top-100 candidate list. The algorithm generated by this research (LFIQ) generated overall quality assessment scores, which will be tested as possible variables for inclusion in the model being tested in this research.

Research in 2012 by Murch et al. also focused on AFIS applications and sought to develop software for automated feature extraction for use in a lights-out AFIS system (Murch et al. 2012). This study considered “rare features,” which they defined not as minutiae types that were less frequent in the population, but as rare configurations of multiple minutiae that were designated by creating triangles between minutiae and measuring the ridge counts between minutiae pairs to come up with statistically unusual compound features.

Additionally, in 2013, a group at Pennsylvania State University published on a quality grading system that utilized three readily-available software programs to produce a measurement of the quality of a mark (Pulsifer et al. 2013). This process utilized clarity, identifiable minutiae, and percentage of area with identifiable minutiae to calculate the grade of the mark. As with the others, this metric focused on the *quality* of the mark (i.e. its clarity and minutiae count), but did not address its overall *usefulness* (i.e. what can reasonably be *done* with the mark), nor did it provide quality thresholds for different decisions.

Another model published by Neuman et al. in 2016 predicted whether a mark should be entered into AFIS based upon number of minutiae marked by a human examiner, specificity of the spatial relationships between the minutiae (represented by the proportion of exemplars in a reference database that shared at least the same number of minutiae in common with the mark as the true source did), and several measures of local clarity (Neumann et al. 2016).

Additional approaches to automated quality metrics for AFIS have included evaluating variations in ridge direction (Tabassi et al. 2004; Yen and Guzman 2007) or frequency magnitude (Fierrez-Aguilar et al. 2005; Nill 2007) to distinguish ridge flow from background and reach estimates of image clarity. Later work has approached the problem using a receptive field approach that relies on a self-organizing map and random forest algorithms (Tabassi et al. 2013; Danov et al. 2014; Wang et al. 2014). The work of Tabassi et al. has resulted in metrics known as NFIQ and NFIQ2, which are widely used, but utilized prints rather than marks in their development with biometric applications in mind and thus have limited applicability to manual comparisons.

Finally, Chugh et al. (2018) developed a model capable of assigning a quantitative value score to a latent mark, rather than a print, which could then be used to predict AFIS rank. This model was trained based upon crowdsourced expert examiner opinions of the quantity and quality of information present in the mark holistically. However, it did not take specific features or rarity into account, nor did it consider the value of a mark for a manual comparison.

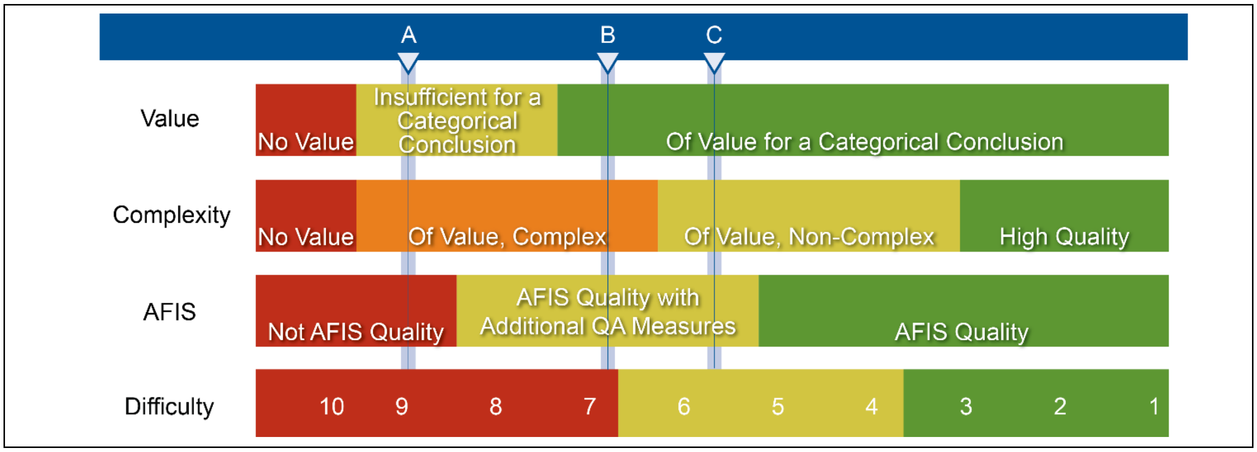
Although these approaches all represent steps in the right direction, most focused on the mark's clarity or quantity without considering the distinctiveness of the information contained therein, which we posit is an important component of both the suitability decision and the weight given to the ultimate *sufficiency* decision. Although some of the models did incorporate a measure of rarity, or specificity, of the features, those that did did not conceive of distinctiveness in the same way that examiners think about it or take the importance that human examiners put on it into account. Additionally, most of these models put a heavy emphasis on AFIS applications, if not focusing there exclusively, without consideration for how suitability thresholds differ for manual comparisons and several used prints exclusively in their development, reducing their usefulness with marks. Furthermore, none of these models offer guidance on the overall utility of the mark for different applications, and most do not provide thresholds for more than binary decisions.

Our work seeks to break the notion of the utility of a mark into four distinct scales which cover applicability for both manual and AFIS comparisons, and to provide guidance in sorting marks across multiple categories on each of these four scales, providing flexibility for different needs of operational laboratories. Additionally, we will incorporate the notion of distinctiveness into our assessments and utilize real-world marks in the development of our model.

2.2 Four Scales of Suitability

Although the suitability decision is often approached as a single concept—value or no value—this work proposes four aspects of suitability that are useful in forensic science work, training, testing, and research. These aspects have been formalized as four separate *scales*: Value, Complexity, AFIS Quality, and Difficulty. The conceptualization of these four scales and their relationship to one another is novel and represents a paradigm shift in how examiners and supervisors can think about the "value" of a mark. Because each of these four scales measures something different about the usefulness, or utility, of a mark, a single mark's value along each scale may vary. The theoretically overlapping nature of the four scales is illustrated in Figure 3. This research creates a tool that is capable of sorting marks into the overlapping categories along each scale.

Figure 3 A hypothetical depiction of the possible relationships between the four scales of utility, showing how a single mark can rank differently on each scale. Mark (A) is simultaneously insufficient for a categorical conclusion; of value, complex; and not AFIS quality and encompasses three different quality rating colors (where green is the highest quality and red, the lowest). (B) and (C) are both of value but only AFIS quality with additional QA measures; however, (B) is complex, and (C) is not. The exact relative relationships of these scales are unknown; this figure shows one way the thresholds could be placed to illustrate the concept. Note that this image is from an early conceptualization of the project and in the final model, the Difficulty scale comprised 3 categories (as represented by the 3 colors below) but was not resolved into 10 Difficulty scores.

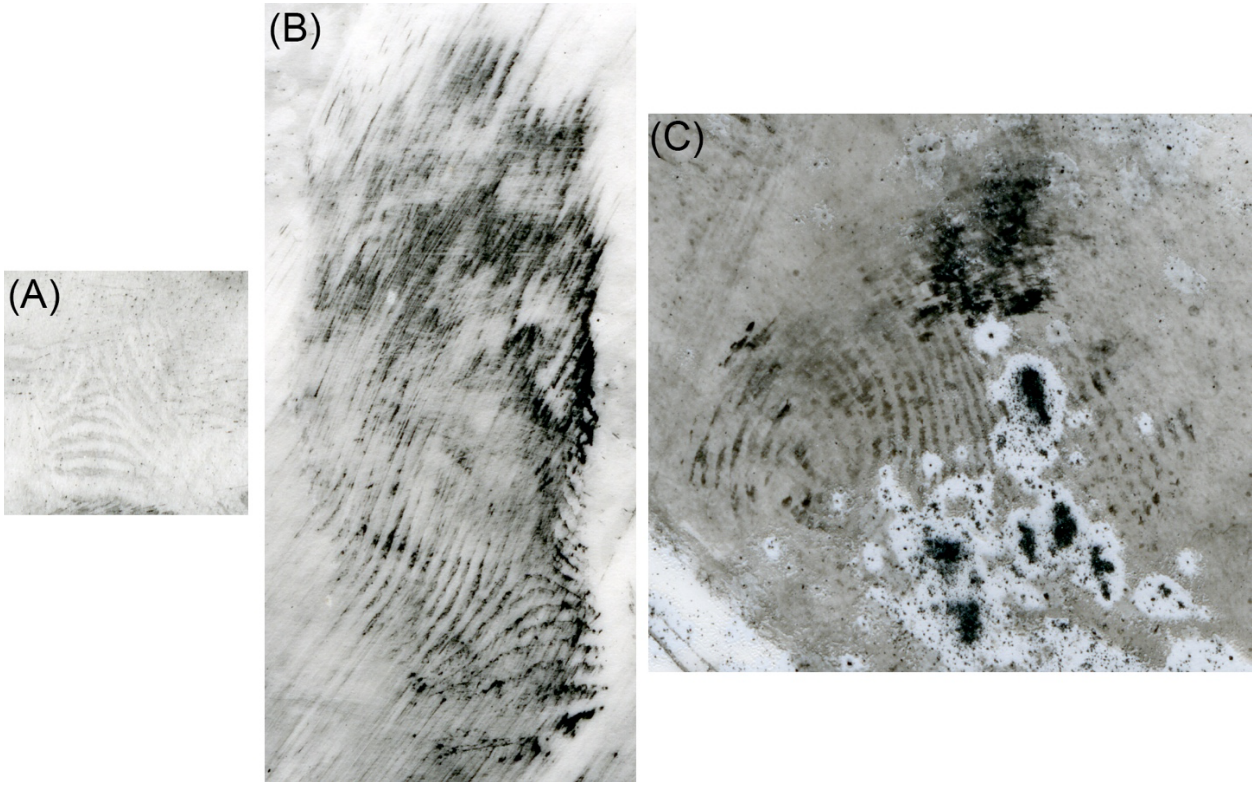


2.2.1 The value scale

The value scale considers whether or not a mark should be used in a comparison, and if so, how strong a conclusion it has the potential to reach. Although some laboratories currently consider a "value for comparison" decision, which is separate from a "value for identification decision," many laboratories that have separated the VID decision from the VEO decision have done so in a hierarchical manner. That is, it is assumed that if a mark is VEO (value for exclusion only), it is inferior in quality to a mark that is VID (value for identification and presumed to *also* be suitable for exclusion).

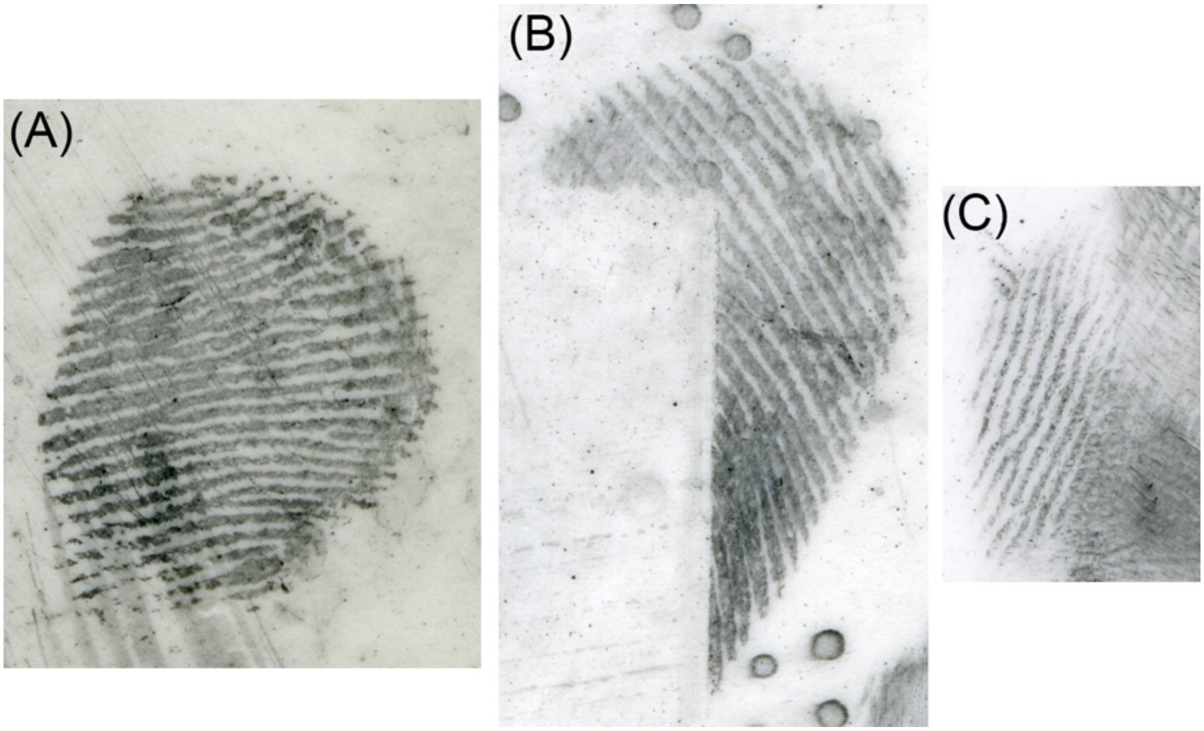
This brings up a series of questions, the first of which is: are there marks that can be excluded that can't (or shouldn't) be identified? Figure 4 presents 3 marks that may fall below many examiners' thresholds to identify, yet they could be excluded from most clear prints. Figure 4(A) and (B) presents marks with very few clear minutiae, yet these are located right next to an *anchor*, so they could be easily excluded. Figure 4(C) shows a mark that appears to be a whorl pattern. Although it is highly distorted and so could potentially be a loop, all arches could easily be excluded. This brings up an additional question: are marks that are of value to exclude suitable for excluding *all* non-donor prints, or are some suitable only for excluding *some* non-donor prints? If they can exclude *any* non-donors, then shouldn't they be considered "suitable" for exculpatory purposes?

Figure 4 Three marks that could be easily excluded from at least some prints but may be more difficult to identify. Marks (A) and (B) are focused around an anchor. Mark (C) could easily be excluded from some prints, but not necessarily from all.



The converse question is more interesting: are there marks that can be identified that can't be excluded? We propose to represent this using the label "Value for identification only" (VIDO), which is not in common use in laboratories today. However, let's examine the concept. Figure 5 presents 3 marks that are suitable to identify but may pose a practical challenge for exclusion. Figure 5(A) and (B) presents marks that likely come from the tip of a finger and it is unclear how far they are from the core. The examiner might request better standards, but they might never feel comfortable excluding, since they might never be certain that *all* the necessary ridge detail had been clearly recorded far enough up and out, even if major case prints were requested. Figure 5(C) presents a mark with very few orientation and location clues. This mark could be identified if it was found using a brute force search. But again, if every part of available ridge detail had been searched and this mark wasn't found, could the examiner be certain and confident that the prints they received truly recorded *all* the subject's friction ridge skin? In this case, the mark could be easily identified, but if it wasn't, Inconclusive or the Support for Different Sources conclusion proposed by the Organization of Scientific Area Committees Friction Ridge Subcommittee (OSAC-FRS) might be better conclusion options than Exclusion.

Figure 5 Three marks that could be easily identified if located but might pose practical challenges for exclusion. Marks (A) and (B) appear to be from the tips of fingers. Mark (C) has few orientation or location clues. In all 3 cases, it might be difficult to be certain sufficient exemplars had been received.

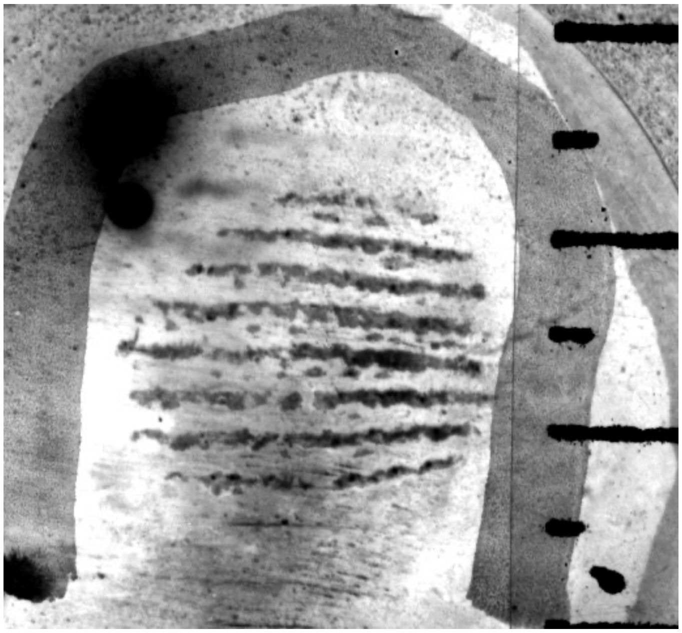


What do the examples in Figure 4 and Figure 5 have in common? All the marks in Figure 4 that could be excluded but not identified (VEO) lacked quantity or quality. All the marks in Figure 5 that could be identified but should maybe not be excluded (VIDO) lacked anchors. However, although all the example VEO marks had low quantity or quality, it is not necessarily true that all marks with low quantity or quality are VEO. In other words, particularly when anchors are missing, there are situations in which a low quantity or quality mark also can be identified but not easily excluded.

This is important to the assumption that VEO marks are somehow inferior in quality to VIDO marks because there are times when you would actually need *more* information (i.e. a higher quality or quantity mark) to be able to exclude than what is required to identify. For example, in the well-known “zero-point identification” reported in the Journal of Forensic Identification (Reneau 2003), a mark was presented that appeared to be from the tip of a finger (Figure 6). This mark was not, in fact, a tip, but came from the side of a loop in the middle. If it had been excluded because it was not present in any fingertip exemplars, an error would have been made. Yet, this mark was identified by the author of the article. Whether or not that identification was sufficiently supported is not generally agreed upon within the latent print community.

Another well-known example of this phenomenon is a case shared by David Ashbaugh (1999), in which an identification was made on a telephone wire, based upon poroscopy. Naturally, this mark could not be excluded with so little reproducible information available and no anchors.

Figure 6 A "zero-point identification" made using exclusively level 3 detail. Reprinted with permission from JFI (Reneau 2003).



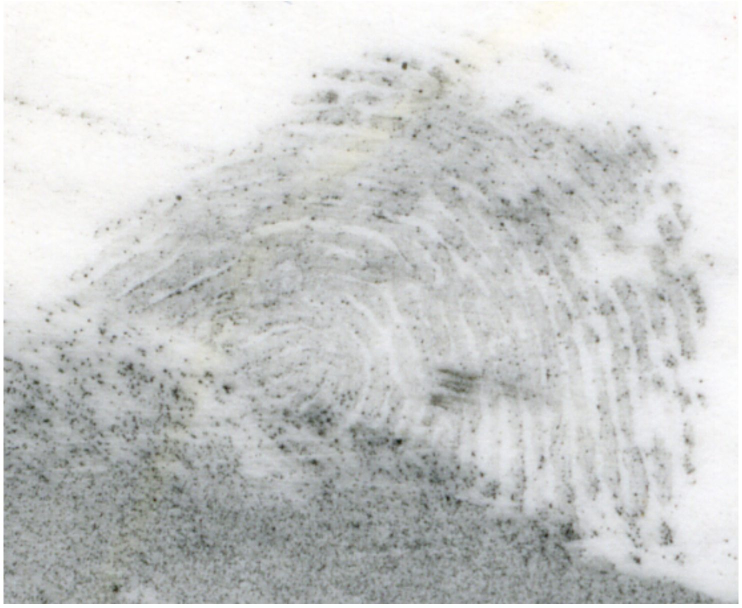
Both of these two examples involve identifications made exclusively using *Level 3 Detail*, a risky practice. The very reason this practice is risky is because Level 3 Detail is reliable, but not always reproducible. That is, *if* it is present in both impressions, you may rely upon it, but it does not always record. Thus, we find ourselves in a situation in which if an examiner finds the detail in both impressions, they may feel comfortable reaching an identification, but "absence of evidence is not evidence of absence," thus one cannot exclude on the same information.

Just because a pore, or an incipient ridge, did not record, it does not necessarily follow that it is not present in the source skin. In these cases, a VIDO mark could actually be considered inferior to a VEO mark because there is not enough information here to successfully exclude.

Finally, Figure 7 presents a mark that *does* include an anchor, orientation information, and even a core and possible *pattern type*. However, this is another example that could be fairly easily identified to a clear print but might be difficult to exclude (thus VIDO). This is due to reliability issues with the observed data and falls under the same philosophy as the above. If one marks out ambiguous minutiae during analysis that are *weakly believed* to be present, then they are found in the print, this is seen as a confirmation of their existence and an Identification is effected.

However, if the ambiguous minutiae are *not* found, the examiner is not confident enough that they truly existed to reach an exclusion decision. Once again, the mark is suitable for ID (Identification), but *more* information, or *higher quality* information is needed for an exclusion.

Figure 7 A mark that could be identified but may be difficult to exclude (VIDO) due to reliability issues with the information in the mark.



The salient point when considering VEO versus VID (in the traditional sense, which really implies Value for both identification *and* exclusion) versus VIDO is not that one value designation requires *more* information, but rather that they require *different* information.

Some laboratories, such as Arizona Department of Public Safety, already take this into account, and have set separate criteria for a mark to be suitable for identification or suitable for exclusion, and a mark may be suitable for one, the other, both, or neither [personal communication].

Table 1 illustrates the different criteria that are *logically* needed, at a minimum, to reach an Identification decision versus an Exclusion decision, without reference to any particular agency policy.

It can be easily seen from this table that the necessary attributes to reach an Identification are nearly opposite to those necessary to reach an Exclusion. For an Identification, one must reach a certain accumulation of data in agreement. That data can be accumulated due to the *quantity* of information in agreement, or due to the *distinctiveness*, or rarity, of that information. The exact opposite is true of exclusions. It doesn't matter at all how rare a feature is if it is not present in both impressions. And it doesn't matter *how much* information is in disagreement if the impressions are both clear and the location of the data is unambiguous.

Similarly, location information is irrelevant to reach an Identification. It makes the search easier, but an Identification can be made using a brute force search. For an Exclusion, on the other hand, location information is critical unless every bit of potential ridge detail has been clearly and completely recorded, which is operationally uncommon.

Finally, to reach an Identification decision, the data in agreement must be at least at Level 2 or 3—one cannot make an Identification based solely on Level 1 detail. However, to reach an Exclusion decision, the data in disagreement should be at Level 1 or 2—one should not exclude based solely on Level 3 detail.

Table 1 The criteria necessary to effect an identification versus those necessary to effect an exclusion. The type of information needed for each is fundamentally different and, in some cases, opposite.

Suitable to Identify	
What is needed	*What is NOT needed*
Reliable data in agreement (which must include at least Level 2 or Level 3)	Location information • Anatomical source • Orientation • Anchors • Target Groups
Data in agreement must be: • High quantity; or • High distinctiveness; or • Both high quantity and high distinctiveness	Completely recorded exemplars • Only the relevant area is needed
Suitable to Exclude	
What is needed	*What is NOT needed*
Reliable data in disagreement (which must include Level 1 or Level 2)	High quantity
In addition: • Completely recorded exemplars; or • Clear location information and the relevant exemplars	High distinctiveness

It is because different information is needed for Identifications versus Exclusions and because sometimes a VEO mark is "better" than a VIDO mark that we have elected in this research to create 5 categories on the Value scale so we can examine the properties of marks that are considered to be VEO, VIDO, or *VB*.

In addition to the question of whether there is a difference in what is needed for VEO, VID, or VIDO, there is also the question of whether all marks that are VID are really equally suitable, or looked at another way—are all identifications equal? Due in part to a philosophy that was codified in 1979, when the International Association for Identification (IAI) passed Resolution VII[2] (Moenssens 1979; Davis 1979), expressly forbidding its members to testify to probabilistic

[2] This resolution was hotly debated at the time and was eventually passed despite the arguments of objectors such as Moenssens and Davis, who argued that the Resolution represented a step backward for the legitimacy of friction ridge examination as a science, noting particularly that no information on similarity at all could be offered in court without an absolute conclusion under this Resolution (Moenssens) and that marks have exculpatory as well as incriminatory value and that information less than "proof-value" could still be probative (Davis). Interestingly, these arguments are still being made today and are at the backbone of the pressure that resulted in the IAI rescinding this Resolution in 2010.

conclusions (Champod 1995), all identifications were, in the words of the Resolution, “positive” and to the exclusion of all others.

During this time, it was widely held within the discipline that the only valid reason for an Inconclusive decision was that better standards were needed. It was thought that if one had appropriate exemplars, the competent examiner should be able to conclusively determine whether the mark did, or did not, originate from a particular individual. If the problem was with the quality of the mark, then that mark was clearly not suitable and should not have been designated as such in the first place.

This philosophy created a culture in which all identifications were treated as equal. A mark of very high clarity that was identified was given the same label (Identification) as a mark of marginal quality. Once the “Identification” threshold was crossed, all Identifications carried the same amount of weight, were reported the same way, and were presented in court the same way. This practice created two phenomena: First, examiners were forced to claim the same amount of confidence in the Identification to the poor-quality mark as they had in the high-quality mark—a situation that was very uncomfortable for many and that seemed to defy logic and common sense. Second, examiners were often compelled to identify marks that they probably should not have been identifying. Poor quality marks where some detail could be found in common with the corresponding print were being identified because there was institutionalized pressure to reach a definitive conclusion and the mark could not be excluded.

Now, after the release of the NAS Report, with the advent of probabilistic models, and with the IAI rescinding their prohibition on probabilistic conclusions, there is a growing trend toward creating more bins of conclusions, or shades of grey[3] (Champod 1995; Neumann 2012).

Not all identifications are created equal. Some carry more information than others, and therefore, more confidence. Similarly, there is sometimes not enough information present that an Identification decision is warranted, yet the fact that some information was found in common is still probative. Some of these differences in strength of comparison conclusion can be anticipated during the analysis phase. If a mark on its own clearly does not contain sufficient information to support a full Identification decision, yet may be sufficient to provide some probative or investigative value to the case, it could be so-labeled prior to ever viewing any prints, thus limiting the allowable conclusions from such a comparison.

Some agencies and standard-setting bodies are beginning to embrace this distinction in comparison conclusions, creating sub-categories of Inconclusive. For instance, the Friction Ridge Subcommittee (FRS) of the OSAC has proposed conclusion language on a 5-point scale, which includes the category “Support for Same Source” between Inconclusive and Source Identification (OSAC friction ridge skin subcommittee 2019; Carter et al. 2020). The Las Vegas

[3] This was argued by Champod well before the development of the current statistical models. He characterized the interpretation of friction ridge evidence as “an increasing scale [that] runs from exclusion to identification”. More recently, Neumann has also argued against the dichotomous model of conclusion reporting.

Metropolitan Police Department (LVMPD) has an allowable conclusion of "Cannot Exclude," which indicates that detail was found in agreement between the mark and the print, but not enough to rise to the threshold of an Identification [personal communication]. This conclusion is in keeping with the third prong of Locard's 1914 recommendation (Champod 1995) that:

(1) if more than 12 concurring points are present and the fingerprint is sharp, the certainty of identity is beyond debate

(2) if 8 to 12 concurring points are involved, then the case is borderline and the certainty of identity will depend on:

   (a) the sharpness of the fingerprint
   (b) the rarity of its type
   (c) the presence of the center of the figure and the triangle in the exploitable part of the print
   (d) the presence of pores
   (e) the perfect and obvious identity regarding the width of the papillary ridges and valleys, the direction of the lines, and the angular value of the bifurcations

(3) if a limited number of characteristic points are present, the fingerprints cannot provide certainty for an identification, but only a presumption proportional to the number of points available and their clarity.

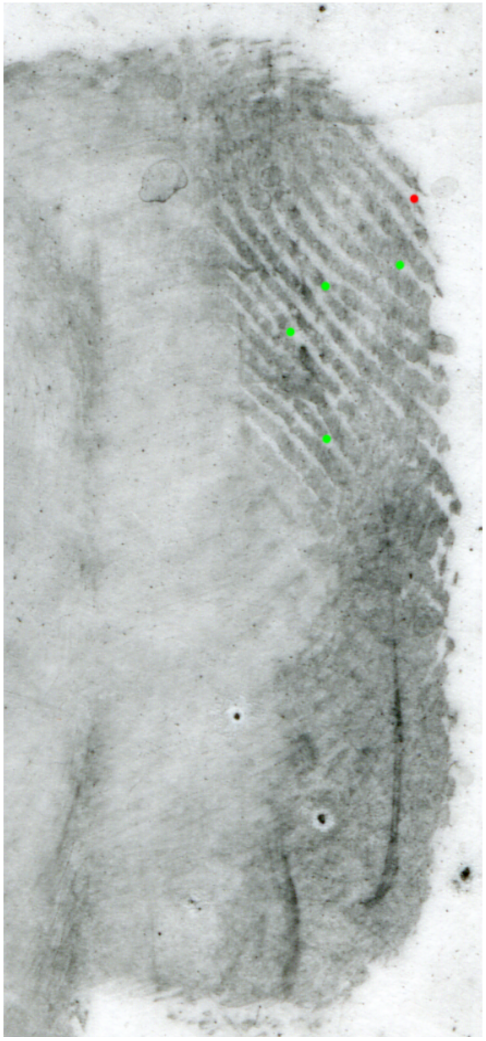
Although conclusions such as "Cannot Exclude" or "Support for Same Source" do not incorporate a numerical proportion[4] (Lennard 2013), they do at least convey some presumptive information to the jury. This information is typically lost entirely under the current leading practice, where the results of such comparisons would either be reported as Inconclusive, and would likely never make it into a courtroom for further explanation; or would be reported as an Identification, which would be overstating the strength of the evidence.

Thus a question of interest is: can it be determined during analysis whether a mark will be suitable for identification, or only for a less definitive, but still probative, conclusion, such as "Support for Same Source"? The answer to this question may depend on the quantity of the information available in the mark. For example, Figure 8 presents a mark with 4 clearly discernable, reliable, but not particularly distinctive, minutiae (marked in green) and 1 more clearly discernible but less reliable, and still not particularly distinctive, minutia (marked in red). These may not be sufficient to reach an identification decision, but if all 4 (or 5) minutiae were found in agreement with a print, it could provide an investigative lead to a detective, or some probative information that a person of interest *could* have left the mark, although due to the

[4] Lennard espouses a compromise of sorts, in which the expert's opinion is presented along with statistical information to back it up; but he also presents the research of Martire, which suggests that jurors are poor Bayesians and may struggle to understand evidence presented in a statistical framework. Nevertheless, he conclusively states that the Inconclusive category is "overly broad and uninformative".

low quantity and specificity of the information, it is quite reasonable to assume that a number of other people could also have left the mark.

Figure 8 A mark that may only ever rise to the threshold of "support for same source" or "cannot exclude" regardless of the print it is compared to, due to the low quantity and distinctiveness of the information it contains. This mark could be classified as "investigative value."



In addition to quantity of information, a mark may be insufficient for identification (but sufficient to provide probative information) due to the reliability of the information in the mark. In some cases, this may be unknown until after the print has been seen. For example, Figure 9 shows a mark that contains a great deal of information—but most of it is in the form of incipient dots, detail that is notoriously unreliable in its recording. It is clear that the dots are present in the mark, but are they also present in the print? If not, the few minutiae visible might lead to a "support for same source" conclusion, but if they *are* present in both impressions, an identification decision may be warranted.

Because the value decision is so nuanced, in this research, we have investigated a scale that includes five possible decisions on the value scale: No value; Some probative or investigative value, but insufficient for an identification or exclusion; Value for exclusion only (VEO); Value for identification only (VIDO); or Value for both identification and exclusion (VID).

These five options were explained to study participants in the PiAnoS User Manual they were provided and a table summarizing the five options was also provided and is reproduced here as Table 2.

If adopted, this expanded scale will provide additional guidance and nuance to the determination of value, both limiting conclusions when insufficient information is available during analysis and expanding the number of marks that will be considered "of value" for *something* by separating the notions of VEO and VIDO. This research will test both whether the proposed conclusions can be successfully modeled, and whether the new conclusions would be embraced by the latent print examiner community.

Figure 9 A mark that may be VID or value for investigative value only depending on the print to which it is compared. Here, the limitation of the data is not the quantity of the information, but its reliability (in this case, its likelihood to robustly reproduce).

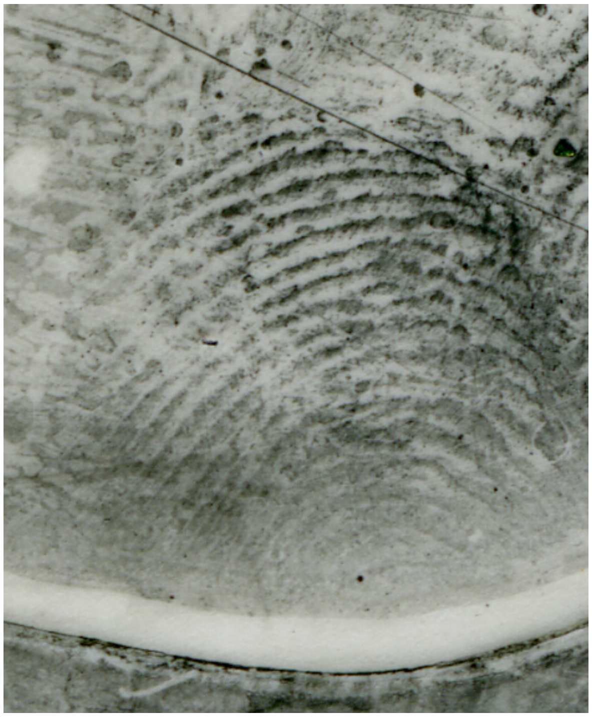


Table 2 The five possible decisions of the value scale along with their brief definitions and some clarifying text, reproduced from the User Manual provided to study participants.

Decision	Definition	Clarification
No value	The mark does not contain sufficient information to proceed with a comparison	This decision indicates that the mark does not contain enough information to be searched or compared at all, even to a single 10-print card; or if it was compared, it is not expected that the comparison would yield any useful information to inform an investigation or provide support for one source proposition over the other.
Some probative or investigative value, but insufficient for an identification or exclusion	The mark contains sufficient information to proceed with a comparison, but insufficient information to reach either an identification or exclusion decision	This decision indicates that the mark contains sufficient minutiae and orientation/location information that it could be effectively searched; however, the information is limited enough that if an association were made, it would not be strong enough to rule out the possibility that someone else could share the same features.

		Similarly, if the features in the mark were not found in an exemplar, or if apparent differences were found, they would not be sufficient to rule out the possibility that the mark *could have* been made by that source. This may occur, for instance, in marks with sufficient distortion that the features or anchors are somewhat unreliable. While there is not enough information to reach the highest level of reported conclusions, these marks still may be useful in providing investigative leads, or may provide support for one source proposition over the other.
Of value for a *categorical conclusion**	The mark contains sufficient information to potentially reach **either** an identification or exclusion decision, or both	These decisions indicate that the mark ***is*** or ***may be*** identifiable or excludable. One conclusion is not considered to require better, or more, information than the other. If the mark is in this category, you will specify whether it is of value for an Exclusion only, an Identification only, or if it could be used for both an Identification and an Exclusion (assuming legible and completely recorded exemplars for each circumstance).
• Value for Exclusion only • Value for Identification only • Value for **both** Identification and Exclusion		

* It is recognized that some laboratories *never* reach a categorical conclusion; that is, they report only on the strength of the association without taking the final step of declaring an identification (Swofford 2015). For those laboratories, the term Identification, as used here, will represent their strongest level of association.

It is recognized, of course, that none of these decisions exist within a vacuum. Although there are general ideas about what "should" be enough to reach a definitive conclusion, or to use AFIS without additional QA measures, or even to keep a mark for comparison at all, in reality these decisions are colored by the agency culture in which the examiner is operating. So, although it is true that "not all identifications are created equal" in the sense that some have higher quantity and quality of information, it may also be said that "not all agencies are created equal". What may be an acceptable practice in one agency culture may be considered shockingly reckless by another.

Put back into an Analysis perspective, these cultural differences may contribute heavily to observed differences in value decisions. A large agency that deals predominantly with violent crimes may have a significantly different threshold for value decisions than a small local agency that sees a high volume of property crimes, which may again differ from a culture such as that of the Dutch experts reported by Langenburg (2012), who place a premium on consistency of minutiae selection, but consequently allow many marks to go uncompared that would likely be considered by other agencies.

Even within the same agency, there will be some differences in perspective that will allow two examiners (or the same examiner at two different times) to reach different value decisions on the same mark (and thus, the same available information). For example, an examiner who has recently had a false positive identification discovered in his work may become for a time much more conservative, and let many marks go that he would have consistently called “of value” prior to the error being discovered.

These decision criteria can be properly explored using a decision theory framework, as described by Biedermann et al. (2008). This research does not directly exploit the decision theory framework, although we do note how it can be incorporated into the final model to best reflect the priorities of an agency. Instead, we focus on exploring the expansion of the Value scale into a 5-category scale and learning what information examiners use to support their Value scale decisions. Our hope is that the introduction of this expanded scale will lead to its adoption and to more nuanced thought in the field about both the limitations of marks that can be ascertained during analysis and the inclusion of marks for comparison that might be suitable for only exclusion or only identification.

2.2.2 The complexity scale

Although we spent a good deal of time and energy exploring the nuances of the value scale, a determination of value is not the only dimension of suitability with which we concern ourselves. A mark that has been declared to be suitable for comparison or database search may still be situated anywhere along a continuum of quality for “of value” marks. A mark’s position along this continuum is often referred to as its “complexity.” Typically, a mark is referred to as complex, or non-complex, yet there is no set threshold to distinguish between the two.

Once again, we are faced with a situation in which multiple viewings of the same mark (by the same examiner or different examiners) may result in different assessments of the mark. This challenge is not unique to friction ridge examinations. Other pattern-comparison disciplines, such as document examination, are similarly working on automated or standardized methods to assess sample complexity (Found and Rogers 1996; Found et al. 1998; Stern et al. 2018) although in the case of signature complexity, higher complexity is judged to allow for *more* certain conclusions, in contrast to how complexity affects friction ridge conclusions. Nonetheless, and despite a small sample size (n = 5 examiners) document examiners were found to vary considerably in their subjective judgements of signature complexity (Stern et al. 2018).

The literature has been clear and consistent in recommending that more complex marks should be afforded more time for consideration, be analyzed with more care, be documented more thoroughly, and be reviewed more closely (Ashbaugh 1999; Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2012; Forensic Science Regulator 2015). It may be reasonable to require additional quality assurance (QA) measures or that examiners not use conclusions stronger than support for same source or support for different sources in the case of increasingly complex marks.

At the other end of the spectrum, there are marks of exceptional quality that should not require such close scrutiny. These marks might justifiably require less QA oversight and less documentation than the average mark, resulting in time- and cost-savings to the agency. It may be reasonable, then, to set a threshold of high quality, beyond which abbreviated documentation procedures are followed. Furthermore, the level of scrutiny applied to these high-quality marks in court may be reasonably expected to be different than with poor quality marks that score lower on the quality metric scale.

However, since the determination of complexity varies from examiner to examiner, it is difficult to construct and consistently apply policies to govern the examination of a complex mark.

The complexity scale considers the chance that two examiners will disagree about the suitability or interpretation of features and distortion in the mark or the sufficiency of conclusions reached after comparison of the mark. The complexity scale aims to predict the marks that are prone to causing disagreements, which will in turn influence the quality assurance (QA) measures that will be appropriate for the mark.

The complexity scale is broken into three main categories (rather than simply complex or non-complex). These categories represent marks that are complex; marks that are non-complex, but also not high-quality (this should be the bulk of the routine, non-complex casework marks); and marks that are of exceptionally high quality. There is a fourth category for marks that are not of value at all, which corresponds to the “no value” category on the value scale.

The four options of the complexity scale were explained to study participants in the PiAnoS User Manual they were provided and a table summarizing the four options was also provided and is reproduced here as Table 3.

Table 3 The four possible decisions of the complexity scale along with their brief definitions and some clarifying text, reproduced from the User Manual provided to study participants. Note that the text for “No value” is exactly the same as that provided for the value scale.

Decision	Definition	Clarification
No value	The mark does not contain sufficient information to proceed with a comparison	This decision indicates that the mark does not contain enough information to be searched or compared at all; or if it was compared, it is not expected that the comparison would yield any useful information to inform an investigation or provide support for one source proposition over the other.
Of value, complex	The mark contains sufficient information to proceed with a comparison, but due to limitations of quality, a high	This decision indicates that the mark contains sufficient information to search and compare, but the high degree of interpretation required

	degree of interpretation is required	means that there is a high chance of variability in judgments between examiners. Disagreements may occur regarding suitability, sufficiency, or existence and type of features and distortion in this mark. The mark may be at or near the value/no value threshold. This mark should be subject to additional documentation[5] and quality requirements. It is critical to thoroughly demonstrate the basis for any conclusions rendered on this mark.
Of value, non-complex; requiring documentation	The mark contains sufficient information to proceed with a comparison, and although some interpretation may be required, no significant disagreements over suitability, sufficiency, or interpretation are anticipated	This decision indicates that the mark contains sufficient information to search and compare, and the quality and quantity of information available is about average. This mark may require some interpretation, and there may be limited disagreements between examiners on suitability, sufficiency, or interpretation. This mark represents the average, run-of-the-mill mark encountered in casework. This mark should be subject to standard documentation and review requirements.
Of value, non-complex; self-evident	The mark contains sufficient information to proceed with a comparison and requires minimal or no interpretation	This decision indicates that the mark contains sufficient information to search and compare, and that the overwhelming quality and quantity of information available makes the basis for any conclusions virtually self-evident. There should be no expectation of disagreement between examiners regarding this mark. This mark can be subject to reduced documentation (e.g., only name the mark and indicate orientation and anatomical source) and review requirements. If this was a casework mark, you would feel that the documentation you have done to this point due to the requirements of the research was a waste of time (thank you for doing it anyway – we are going to use it!)

[5] We recognize that different agencies have different standards of documentation. For purposes of this study, "additional documentation requirements" presumes that standard documentation is fairly minimal. If an agency already does full documentation of every mark, they would not need to do MORE for complex marks, although they might do LESS for very high-quality marks under this philosophy.

2.2.3 The AFIS quality scale

In addition to being complex or non-complex, an of-value mark may also be suitable or unsuitable for search in an AFIS. In some agencies, these value decision and the AFIS decision are congruent—that is, it is considered that if a mark is suitable to be compared, it is also suitable for AFIS search. However, in the majority of agencies, these are two distinct decisions with different thresholds. This makes sense, since whether a mark is AFIS quality may vary according to the size of the reference database, the quality of the particular AFIS vendor's matching algorithms, or the workload of a particular agency. Thus, individual determinations of AFIS quality could rely on very different criteria than those used to determine value.

Interestingly, the AFIS decision is one place where most agencies do embrace a numerical threshold (those that will acknowledge a numerical minutiae threshold for the value decision are in the minority in the United States). This threshold is frequently set at 8 minutiae, although it does vary by agency and is often set in accordance with the technical recommendations of the agency's AFIS vendor. There is also often a requirement that orientation and core location be known for AFIS entry of distal phalanges. Some AFIS systems have required a putative pattern type to be entered as well, but this is becoming less common as AFIS algorithms and internet bandwidths improve, reducing the need to increase penetration of the database.

Indeed, some agencies, such as LVMPD, have *two* AFIS thresholds—one for regular casework, and one for the rapid AFIS screening program (which they call Administrative AFIS) [personal communication]. In this program, a higher threshold of 12 minutiae is set. This allows the agency to "cherry-pick" only the very clear marks with an abundance of minutiae from a case and get those run through AFIS very quickly in order to generate investigative leads without the case waiting in the backlog line to even be glanced at. If no AFIS hit is made with these best marks, the case is returned to the regular queue and when it is fully worked at a later time, additional marks that did not meet the Administrative AFIS criteria may also be searched.

Similar to the way all Identifications are currently treated as equal, all AFIS quality marks tend to be treated as equal. Once a mark has been declared to be AFIS quality, it is searched thorough AFIS, and the resulting candidate list is screened, the same way, regardless of the quality of the mark. Furthermore, any Identifications effected as the result of an AFIS search are treated as equal in strength and certainty to those that came from comparing a known subject. Dror et al. (2012) have demonstrated that a risk of erroneous conclusions due to bias created by list position may exist with the use of AFIS (although, see (Kukucka et al. 2020) for an argument that evidence lineups, such as occur with an AFIS may reduce the effects of bias when compared to evidence showups, such as occur when a single known candidate is compared); whereas Dror and Mnookin (2010) have outlined reasons to exercise caution when making identifications using AFIS; and numerous sources (see, e.g. (Neumann 2012; Langenburg 2012; United States Department of Justice and Office of the Inspector General - Oversight and Review Division 2006; Lennard 2013)) have pointed out the statistical dangers of making an Identification decision based upon selecting a candidate out of the large pool provided by an

AFIS. Some qualities of a mark, such as low minutiae count, low specificity minutiae groupings, or high interpretation areas, can increase this risk.

In light of these concerns, it may be prudent to establish a quality threshold below which Identifications made using AFIS must be subjected to additional QA or documentation requirements. Additional QA measures could include ideas such as additional or *blind verifications*; using poor quality marks to generate investigative leads, but not to identify; or requiring that additional minutiae *not* entered in AFIS be found in agreement between mark and print before an identification may be declared.

The AFIS scale considers whether a mark should be entered into an AFIS system, and if so, whether additional QA measures are warranted (expanding the traditional decision scale from two options to three). The QA measures to be implemented will be determined by agency policy, but this scale seeks to measure which marks should be subject to *some* additional QA measure, whatever it may be.

The three options of the AFIS quality scale were explained to study participants in the PiAnoS User Manual they were provided and a table summarizing the three options was also provided and is reproduced here as Table 4.

Table 4 The three possible decisions of the AFIS scale along with their brief definitions and some clarifying text, reproduced from the User Manual provided to study participants.

Decision	Definition	Clarification
Not AFIS quality	The mark is not suitable for entry into an AFIS system	The mark may or may not be of value for comparison, but this decision indicates that it either does not meet your agency's threshold for AFIS entry, or if AFIS criteria are left to analyst discretion, you do not feel that the mark contains sufficient information for an effective AFIS search in your primary search system.
AFIS quality, with additional QA measures	The mark is suitable for entry into an AFIS system, but due to the presence of risk factors, additional QA measures should be used	This decision indicates that, while it meets your minimum agency or personal criteria for AFIS entry, this mark contains risk factors (e.g. low minutiae count, low specificity groupings, or high interpretation) for a coincidental match or contains unreliable information. You would enter this mark into AFIS, but feel it would be prudent to apply additional QA measures as a precaution.

AFIS quality	The mark is unconditionally suitable for AFIS entry	This decision indicates that the mark meets or exceeds your agency or personal criteria for AFIS entry and does not contain risk factors or unreliable information. This mark is suitable for a standard AFIS entry procedure without need for additional caution beyond accounting for the size of the database in your decision-making process.

2.2.4 The difficulty scale

The final suitability scale we will consider is the difficulty scale. The difficulty scale considers the mark for training, testing, testimony, and research purposes and predicts how difficult it would be to compare. Section 2.3 discusses in more depth the applications of the difficulty scale.

This scale can be used to assign difficulty levels to marks such that examiners can be trained to specific levels, tested at those levels, and ultimately testify at specific levels (i.e. by stating the difficulty level of a case image in relation to their tested proficiency level. This could provide jurors with a framework by which to judge how much weight to give the evidence). The difficulty scale will be resolved into three bins: low, medium, or high difficulty.

Low difficulty marks are those that should require very little effort to compare (because this research is focused on the analysis decision, the difficulty of the *comparison* is not considered, and a high-quality exemplar is assumed for all of these categories). A low difficulty mark has high quality and quantity and may also have highly distinctive features.

Medium difficulty marks are the average marks that make up the bulk of the impressions that are *compared* in casework (not the marks that are *seen*, which are largely no value). They are not exceptionally clear, nor are they exceptionally distorted. They may have some clear areas and some distorted areas. Their comparison may be expected to require some effort, but not to be so difficult that the examiner wishes they did not have to compare them at all.

High difficulty marks are the worst marks encountered in casework. At the extreme end, they are not of value at all. When they are of value, the examiner may wish they weren't. These are marks that are retained for comparison because they meet an agency's minimum suitability criteria, or if the agency doesn't have a stated threshold, because the examiner feels they *could* be compared and would feel guilty if they didn't try, even though they anticipate that it will be a very difficult task. They may have low quantity, low clarity, or both. They may be severely distorted. They may lack a target group. They may lack an anchor. They may lack cues for orientation or anatomical source.

2.3 Potential implications for policy and practice

The expansion of the suitability decision into four scales and development of a utility tool that can place marks along each of those scales has multiple potential applications and benefits for laboratories, researchers, and the criminal justice system. We will address these benefits in detail in Section 5.2. Here, we present only a brief list of the areas where the tool can potentially provide practical benefits to reduce variability, provide guidance, support quality assurance, and improve communication and efficiency. Many of these benefits may not be readily measurable and have not been quantified in this work. Nonetheless, they can provide tangible improvements and transparency to forensic operations and research.

The five main areas this work will address that the creation of a utility tool will benefit are: Research, Proficiency Testing and Training, Testimony, Quality Assurance (QA), and Providing a Consensus.

2.4 Chapter 2 summary

In this chapter, we have explored the idea of deconstructing the suitability decision and considering it along four separate dimensions, or scales, which are: Value, Complexity, AFIS, and Difficulty. In addition to introducing new scales, this work will introduce new conclusion options on several of the scales, which will add additional nuance to the suitability decisions they represent.

Previous research into developing a quality tool has focused on a single suitability determination and has been strongly focused upon AFIS applications. It has also predominantly utilized prints (not marks) and dealt with rarity superficially, or not at all. The current research will consider among its four scales both suitability for AFIS entry and for manual comparison. It will also incorporate a measure of rarity directly into the model through use of ESLR assessments. This research will also materially differ from previous efforts by assessing whether examiners alone, lights-out methods alone, or a hybrid of the two produce the best results.

During the white box study portion of the research, examiners will be asked to annotate the information they use to reach their suitability decisions, as well as to render those decisions on each of the four scales for each impression they view. These data will be used to identify the information that is the most diagnostic in reaching decisions for each of the four scales. This will assist us in better understanding how the four scales differ in terms of the utility they evaluate and the underlying information that supports suitability decisions for each use.

In the second part of the study, the same data will be used to predict consensus *ground truth* suitability decisions for each of the four scales and a model will be developed and optimized to maximize accuracy in these predictions while minimizing necessary user input and computational load. This optimized model will be externally validated with new participants and images to test for generalizability.

The completed model will be a valuable tool that will reduce variability between examiners and provide guidance in five key areas: Research, Proficiency Testing and Training, Testimony, Quality Assurance (QA), and Providing a Consensus.

3 Understanding how examiners evaluate suitability

The first step to developing a model that can predict examiners' suitability decisions is to understand the information they rely upon when reaching those decisions. In the first paper concerning this research, "Examining and expanding the friction ridge value decision," we present the results of a white box study conducted to explore just that. Section 3.1 presents the manuscript of that paper, which has been published in *Forensic Science International* (Eldridge et al. 2020), here with figure and table captions prefixed with "3.1 –" to integrate with this dissertation. Note that citations have been changed to author-date format to be consistent with this dissertation and footnotes have been re-named to simple asterisks. Figures may appear in different locations in the manuscript due to journal formatting.

For length and to appeal broadly to the friction ridge examiner community, the main findings of the white box study were incorporated into the article, but some additional discussion of why particular research design choices were made and in-depth review of some additional specific image examples, were omitted. Section 3.2 of this chapter dives a little deeper into the design, observations, and data of the white box study as well as taking a deeper look at the general underlying philosophy of suitability.

3.1 Examining and expanding the friction ridge value decision

Examining and expanding the friction ridge value decision

Heidi Eldridge, MSc[a, b], Marco DeDonno, MSc[b], Julien Furrer, PhD[b], Christophe Champod, PhD[b]
[a]RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC, 27709 USA

[b]School of Criminal Justice, Faculty of Law Criminal Justice and Public Administration. University of Lausanne, 1015 Dorigny, Switzerland

Abstract

The first step of a friction ridge examination involves determining the suitability—or value—of an impression. Often, this is interpreted as whether the impression is suitable for comparison. However, examiners tend to be variable in their suitability determinations, and suitability itself can be a multi-faceted decision, comprising suitability for comparison, suitability for exclusion, suitability for identification, suitability for AFIS entry, complexity, and others. We undertook a white box study to explore the different facets of suitability determinations and to measure the specific categories of information upon which examiners most heavily rely when reaching these decisions. Although minutiae count was the best indicator of a value determination, clarity and distortion were better predictors of complexity determinations. Examiners were found to be highly variable in their determinations, as well as in their annotations of what information they relied upon. Some unanimous decisions were reached for only high-quality impressions; there was never unanimity on "no value" determinations. Examiners tended to use high-confidence minutiae markers, even when there was connective ambiguity or low clarity. Several new suitability categorizations were introduced and had good usage from study participants, indicating that they might have some value for inclusion in routine casework.

Keywords friction ridge, latent prints, suitability, value, variability, standardization

Introduction

It has been well-established through both structured research (Ulery et al. 2011; Hicklin et al. 2011; Ulery et al. 2012; Ulery et al. 2013; Ulery et al. 2014, 2016, 2017, 2015; Pacheco et al. 2014; Neumann et al. 2013) and anecdotal experience, such as the well-known Mayfield and McKie cases, that friction ridge examiners are variable in their decision-making. At every decision-point in the comparison process, from minutiae selection (Ulery et al. 2013; Ulery et al. 2014, 2016, 2015; Neumann et al. 2013; Swofford et al. 2013; Langenburg 2012), to suitability[*] determination (Ulery et al. 2011; Hicklin et al. 2011; Ulery et al. 2012; Ulery et al. 2013; Ulery et

[*] Although the terms "suitability" and "sufficiency" are often used interchangeably in the friction ridge community, this paper recognizes a distinction between the two that will be maintained throughout. "Suitability" refers to the decision that is reached at the end of the Analysis phase—is the unknown mark suitable for some particular purpose, most often comparison. "Sufficiency" refers to the decision that is reached at the end of the Evaluation phase—is there sufficient information present in two impressions to support a particular source conclusion.

al. 2014, 2015; Pacheco et al. 2014; Neumann et al. 2013; Langenburg 2012), to interpretation of distortion (Neumann et al. 2013; Maceo 2009), to comparison conclusion (Ulery et al. 2011, 2012; Ulery et al. 2014, 2017, 2015; Pacheco et al. 2014; Neumann et al. 2013), examiners have displayed variability with their peers and even with themselves, when presented more than once with the same evidence.

Although black box studies (focused on the decision outputs without eliciting the reasons or features used to make them) are invaluable for illuminating this variability, they are unable to explain *why* it exists or suggest ways to reduce it. White box studies, on the other hand, allow researchers to understand the reasoning that goes into the decision-making process by asking participants to record the observations they made as they were going through the decision-making process.

Here, too, there are challenges. Although participants can be asked what they considered during their decision-making, there is evidence in the cognitive psychology literature (Nisbett and Wilson 1977) to suggest that people tend not to be aware enough of their higher-order processes to report on them accurately. In other words, when problem-solving or decision-making, people are likely to give very high confidence answers when asked “how did you reach this decision?” or “why did you think this was of value?” but those answers are also likely to be poorly reflective of the actual stimuli that were used to reach their decisions. Not only are humans generally poor at knowing what went into their decision-making, but they are similarly poor at recognizing that they are poor at it.

Despite these challenges, white box studies can help to illuminate some small part of the decision-making rationale, and the data thus obtained can be tested for their predictive value. If examiners are asked both what information they considered in reaching their decision and what that decision was, we can then test to see how well the information that was claimed to be diagnostic actually predicted the decisions that were made.

This paper reports on the first part of a research study that will attempt to link elicited information by experts and their analysis decisions. First, we asked participants to record the observations they relied upon in making a series of suitability decisions, then we looked for both consistency between examiners on those decisions, and relationships between the observations made and the decisions reached.

Minutiae count in suitability determinations

Within the literature, there is a strong claim that minutiae count is the driving factor behind suitability determinations. In fact, multiple studies (Ulery et al. 2013; Ulery et al. 2014; Langenburg 2012) have observed a so-called “operational tipping point” of 7 or 8 minutiae at which, on average, examiners feel comfortable finding value at the analysis phase, or declaring an identification at the evaluation phase.

However, although minutiae count is undoubtedly an important variable, it doesn't tell the whole story. If minutiae count were all that mattered in determining suitability or sufficiency, we would expect to see a sharp threshold at the "magical" number. Everything below 7 minutiae would be declared of no value, then suddenly, everything above 7 minutiae would be declared of value. But that's not what the data show. There is always a gradual accumulation of "of value" calls as the number of minutiae observed increases.

Similarly, there are far too many outliers in the data to support a minutiae-count-only model of suitability. For instance, in one of the studies from the FBI/Noblis series (Ulery et al. 2013), marks in which participants annotated as many as 12 minutiae were declared of no value, whereas marks with as few as 0 minutiae annotated were accepted as of value for Identification and marks with up to 27 minutiae annotated were declared as of value for exclusion only. In another study, Neumann et al. (Neumann et al. 2012) demonstrated that highly discriminating configurations of 3 minutiae could provide a strength of evidence equal to or exceeding configurations of 12 very common minutiae.

Although it is not the intent of this paper to dispute the importance of minutiae count in reaching a suitability determination, we posit that there are other factors that must also influence this decision and help to tip the balance, particularly in ambiguous cases. But which factors are they, and how much influence do they have? Does everyone rely on the same factors consistently enough that they can be used to predict people's suitability decisions?

The four scales of suitability

When latent print examiners speak about suitability, or value, they tend to confine themselves to a single binary decision—an unknown mark is either of value, or no value. Some get slightly more granular and draw a trinary distinction between no value, of value for identification (VID), and of value for exclusion only (VEO). This last example is often referred to as the Approach I/Approach II distinction, as described by the Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2013). This paper however introduces the idea that multiple distinct suitability determinations can be made for a single unknown mark, and thus presents four scales of suitability: Value, Complexity, AFIS quality, and Difficulty. We will describe each one in turn below. In our study, we will explore experts' decisions in the context of these four scales while recognizing that they are correlated.

We acknowledge that in introducing new scales, and new decision options on familiar scales, we introduce some risk that variability between examiners will be increased due to a lack of familiarity with the new options. However, we feel that this risk is outweighed by the potential improvements to operational quality that the new options may support. By encouraging examiners to make deliberate and nuanced decisions about the suitability of a mark, we hope that both variability and the risk of error will be reduced. Additionally, there is previous evidence in the literature (Ulery et al. 2013; Neumann et al. 2013) that examiners do not fully comprehend or consistently apply the current suitability scales. Thus, the introduction of new scales, which have been fully described in the instructions for the study, should not materially increase the risk of variability caused by lack of familiarity.

The two new scales that are being introduced are the Complexity scale and the Difficulty scale. The familiar Value scale and AFIS scale have had additional conclusion options added to their traditional ranges.

Value

The value scale considers whether the unknown mark should be used in a comparison, and if so, how strong a conclusion it has the potential to reach. The five categories that are provided on the value scale are:

- No value
- Some probative or investigative value, but insufficient for an identification or exclusion
- Value for exclusion only
- Value for identification only
- Value for both identification and exclusion

The categories in this scale indicate two things that differ from more traditional value scales. First, there are marks that simply do not contain enough information to reach a categorical source conclusion such as identification or exclusion, but which can still be compared and may yield some information that could aid an investigation or may provide some (albeit weaker) probative value in favor of one proposition[*] or the other. These marks are represented by the second category listed above and are the ones that may end up with comparison conclusions such as “cannot exclude” or “insufficient detail in agreement to identify” in some agencies.

Second, a mark that can be used for an identification is not inherently *better* than one that can be used for an exclusion. Because identifications and exclusions require different kinds of information, it is both possible to have a mark that can be excluded but not identified, *or* to have a mark that can be identified but not excluded. For example, a mark showing a fully blurred arch pattern will allow an exclusion (of all other general pattern but arches), but the quality of the ridge pattern may not be sufficient for an identification. Conversely, a small mark of the tip of a finger with excellent legibility and numerous minutiae will be retained for identification purposes but may not be excluded by an expert fearing that the corresponding area of the print may simply not be recorded on the available exemplars. Thus, these two situations have been separated from those marks that are suitable to *both* identify and exclude.

[*] The propositions referred to are commonly called the prosecutor’s proposition (i.e. the defendant left the mark) and the defense proposition (i.e. some unknown person left the mark). The observed evidence must lend support to one or the other or be entirely inconclusive.

Complexity

The complexity scale considers the chance that two examiners will disagree about the suitability, sufficiency, or interpretation of features and distortion in the mark. The complexity scale aims to predict the marks that are prone to causing disagreements, which will in turn influence the quality assurance (QA) measures that will be appropriate for the mark. As marks become more complex, agency policy may dictate that they be subject to additional documentation requirements, additional verifications, blind verifications, or other quality measures. On the other hand, marks of very high quality ("of value, non-complex; self-evident") may enjoy policies requiring reduced documentation and review. The four categories that are provided on the complexity scale are:

- No value
- Of value, complex
- Of value, non-complex; requiring documentation
- Of value, non-complex; self-evident

The category "of value, non-complex; requiring documentation" is intended to represent the bulk of marks encountered in casework—those that are neither complex, nor exceptionally clear. The notion that was presented to study participants is that these marks should require some minimum, standard level of documentation to support conclusions, but not the enhanced documentation that would be required of a complex mark. Thus, there are two levels of non-complex marks; those that require "standard" documentation, and those that warrant reduced documentation because they are self-evident.

AFIS Quality

The AFIS scale considers whether a mark should be entered into an AFIS, and if so, whether additional QA measures are warranted. In large databases, the chances of a coincidental match can be much greater than in comparisons to a single or a few known subjects. Some attributes of a mark, such as low minutiae count, low specificity minutiae groupings, or areas requiring a high degree of interpretation, can increase this risk. Thus, examiners should adjust their decision thresholds for making an identification to an exemplar located through an AFIS search and should likewise consider the threshold for AFIS entry separately from that which is used for a manual 1:1 comparison (Dror and Mnookin 2010). Although some agencies and AFIS vendors impose a minimum minutiae count threshold for AFIS entry, there is value to considering overall what characteristics of an impression make it more, or less, appropriate for AFIS entry and which increase the risk of a coincidental match. Additional QA measures to mitigate this risk can include measures such as: additional or blind verifications; using poor marks to generate investigative leads, but not to identify; or requiring that additional minutiae not entered in AFIS be found in agreement between mark and print before an identification may be declared. The additional category in this scale (AFIS quality, with additional QA measures) is intended to identify those marks that have sufficient information to enter into an AFIS but should nonetheless be viewed with caution and subjected to additional QA safeguards due to their perceived increased risk of a coincidental match. The three categories that are provided on the AFIS scale are:

- Not AFIS quality
- AFIS quality, with additional QA measures
- AFIS quality

Difficulty

The difficulty scale considers the difficulty level of the mark for research, training, testing, and testimony purposes more than for casework/comparison applications. If marks could be consistently categorized by difficulty, these categorizations could be used to design proficiency tests at known, stratified difficulty levels and to design progressively more challenging training curricula. They could also be used in research to ensure that there is consistency across research projects such that the results obtained by different researchers could be compared (e.g., if two studies claimed to have tested examiners using "difficult" images, there would be agreement in the community about what that means and that the images were, in fact, difficult). Finally, this information could be used in testimony to inform the fact-finder of the difficulty of the images in the case as well as the relative proficiency level of the examiner (e.g., "the mark in this case was of medium difficulty, but I have successfully completed proficiency tests at a high difficulty level"). The 3 categories that are provided on the difficulty scale are:

- Low
- Medium
- High

Difficulty refers to the anticipated difficulty in comparing the mark to appropriate exemplars. For example, a "low" difficulty mark should require little or no interpretation and should present a straightforward comparison with a clear exemplar. On the other hand, a "high" difficulty mark will likely have clarity or distortion issues, small area, few reliable minutiae, few distinctive minutiae, or other factors that make it more challenging to search, compare, and reach supportable decisions about. Difficulty does *not* refer to the difficulty of reaching a suitability decision (for instance, a very poor mark might be rated as "high" difficulty because it would be nearly impossible to compare, but the suitability decision is easy to make—the mark is no value).

Although the Complexity and Difficulty scales may frequently align, their intended applications are quite different (quality assurance guidance versus standardization of research, training, and testing levels) and there will be cases where the same mark may not fall into the most analogous categories on both scales (for instance, a "low" difficulty mark may fall into either the "non-complex, documentation required" or "non-complex, self-evident" complexity categories depending on whether its low difficulty rating is due to the extremely high clarity that tends to move marks into the "non-complex, self-evident" category on the Complexity scale.

Methods

Selection of marks

Marks were obtained from casework at a large, metropolitan police laboratory. The marks came from cases that were past the statute of limitations and would not be used in any criminal proceedings. All were on lift cards and were used with permission of the police agency. The primary author selected the lift cards for inclusion in a research database to represent a wide range of quality and quantity and to include a large number of impressions with interesting interpretative issues or minutiae configurations. 1,633 impressions were selected and scanned as .tiff documents at 1,000 ppi using an Epson Perfection 2200 flatbed scanner. Palm impressions were removed, and the pool was selected down to 1,259 impressions.

Marks used in the study were selected from that pool of 1,259 impressions. These were selected pseudo-randomly to produce a study pool of 100 images. The constraints put upon the draw were that the quality proportions should be 60:30:10 of low-, medium-, and high-quality, according to scores from running all impressions through ULW software (Hicklin et al. 2013). The pool of 100 was manually verified to ensure it included a wide range of characteristics of interest. Once the 100 images had been selected, pseudo-random draws of 30 were drawn to create each user set according to the same quality constraints as the overall pool.

The complete set of marks that were selected for use in the study can be viewed at: https://doi.org/10.5281/zenodo.3716428.

Participant recruitment

Participants for the study were recruited via multiple methods. Emails were sent out to professional email lists as well as directly to professional contacts of the primary author. Additionally, the primary author announced the study and invited participation during several presentations given at professional educational meetings.

Any latent print examiner signed off for independent casework was eligible to participate, either from the USA or abroad. All study materials were reviewed by RTI International's Institutional Review Board, and informed consent was completed by all participants prior to being granted study access. Participants' confidentiality was guaranteed to the extent allowable by law and no compensation was provided. Thus, participants were voluntary and self-selected. A demographic and agency policy survey was also completed online prior to beginning the study (Figure 3.1 - 1).

Data were collected during an approximately 4-month period in 2017. A total of 186 users enrolled in the study, of whom 116 completed at least one trial. 105 participants completed all

30 trials assigned to them. At the completion of the data collection period, all completed trials (n = 3,241) were exported into R (R Development Core Team 2017) for data analysis. Each of the 100 study images was analyzed by between 26 and 41 examiners.

Data collection using PiAnoS

Data collection was done using PiAnoS, a web-based user interface platform developed by the University of Lausanne that allows users to view a mark on the screen, annotate it, and answer questions prompted by the system. A custom modified version of PiAnoS (release 4.2.2 b5eb0fcc) was created with tools and questions tailored to the needs of the research. The user manual for the study version of PiAnoS included instructions for completion of the study and descriptions of all tools and can be found at: https://doi.org/10.5281/zenodo.3716427.

Participants were asked to annotate the information they considered when reaching a suitability decision, and *only* that information. They were not asked to annotate everything they could see because this was not a vision test, but an attempt to understand *what* information was important to examiners when considering decisions about suitability. For example, if a mark was very clear and contained 50+ minutiae, but the examiner had determined that the mark was suitable for identification after marking only 15 minutiae, they were asked to stop annotating minutiae at that point. Likewise, if they noted third level details such as pores, but did not consider those in making their decision, they were instructed not to annotate the pores specifically.

In order to fully annotate the features that could be considered, several new tools were incorporated into this version of PiAnoS. Two new tools allowed participants to group minutiae together: the Target Group tool (Figure 3.1-2) and the Combined Groups tool (Figure 3.1 - 3). These tools were intended to capture the extra weight that an examiner might put on a cluster of minutiae that stood out to them as being unusual, or distinctive. In both cases, minutiae that had already been annotated could be grouped together. Use of the Target Group tool indicated a cluster of minutiae that the examiner would use during their initial search for the mark against a print, or collection of prints. Use of the Combined Groups tool indicated a cluster of minutiae that the examiner felt was unusual, distinctive, or stood out in some way. Using this tool was a way of communicating to the researchers that the examiner would put extra weight on this cluster, if they were to locate it in a print.

Figure 3.1 - 1 Participant responses to demographic and agency policy questions.

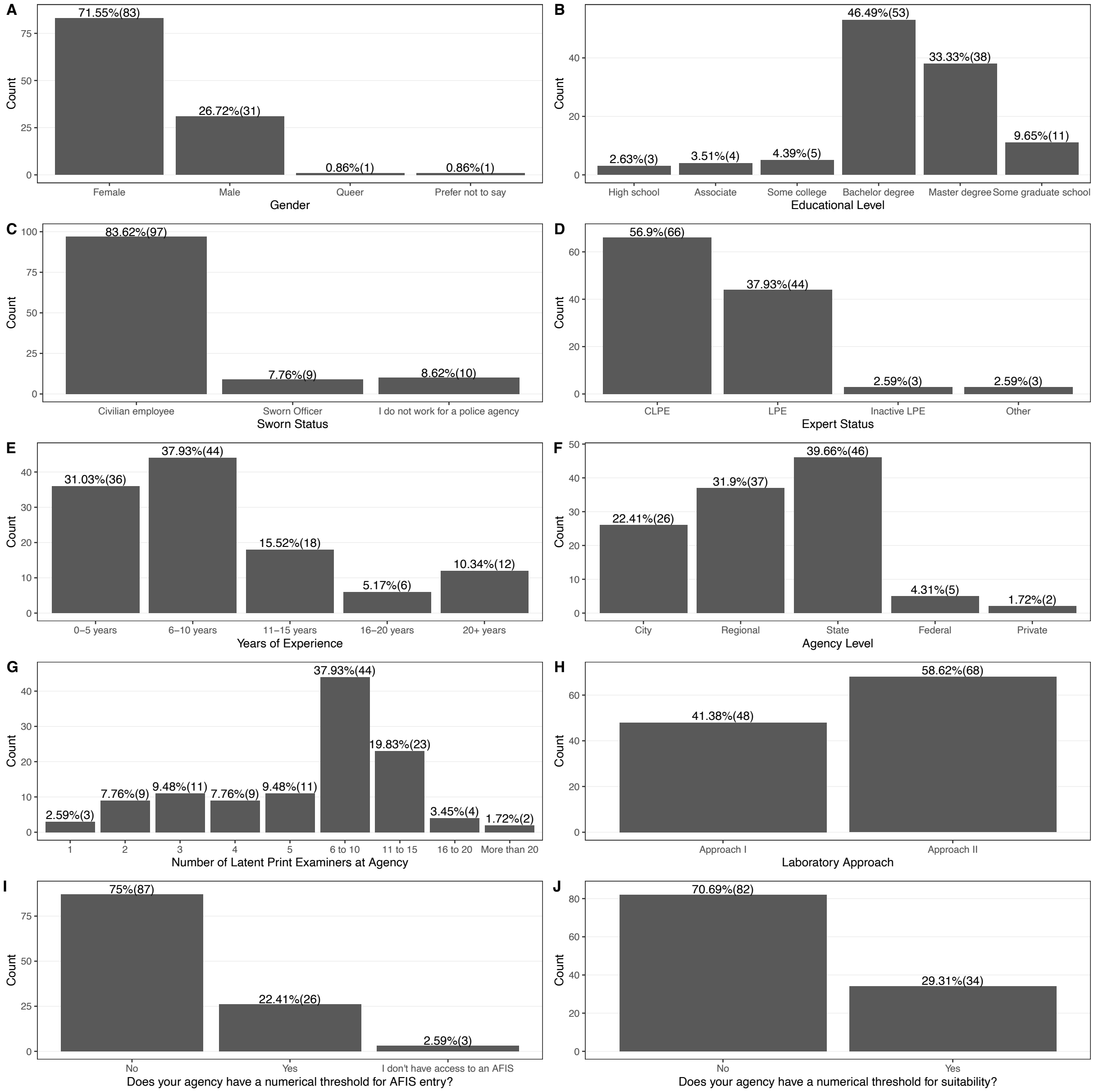
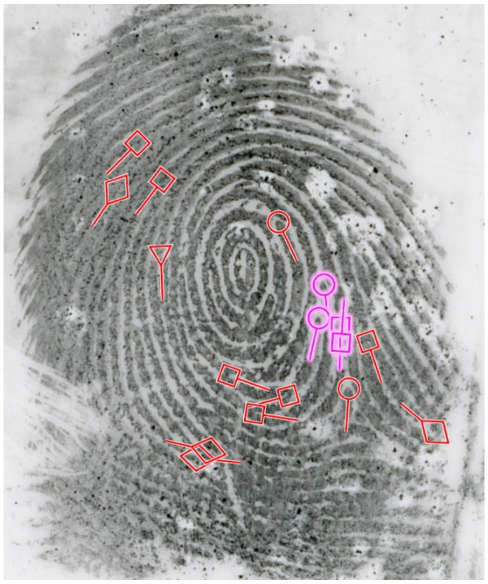


Figure 3.1 - 2 The Target Group tool in PiAnoS. This tool was used to indicate a cluster of minutiae that would form the participants' initial search image. This tool could only be used once per image. The glowing pink minutiae in the image are those that have been selected as a target group in this example.



Thus, it was critical that all examiners viewing the same image (leaving aside the constrains of screens and their calibration) were seeing the same stimuli. If one examiner did more enhancement than another,
they might clarify additional features, which might then influence their suitability decisions.

The Incipient Ridges tool and Pores tool (Figure 3.1 - 4) allowed participants to separately annotate incipient ridges or pores, if they felt they would assign extra weight to incipient ridges in making their suitability determinations.

No tools for image enhancement, such as contrast adjustments or the ability to invert light and dark ridges, were included. Although many examiners use these tools regularly in their casework and many expressed some frustration at their exclusion, these tools would introduce unwanted additional variables into the research. The research was not designed to test peoples' ability at using enhancement tools, but to test what information they saw that contributed to their suitability decisions.

Figure 3.1 - 3 The Combined Groups tool in PiAnoS. This tool was used to indicate a cluster of minutiae that was unusual in the participants' opinion, and would be given additional weight if found in a candidate exemplar. This tool could be used as many times as the participant wished to form different combined groups. The glowing blue minutiae in the image are those that have been combined into a group in this example.

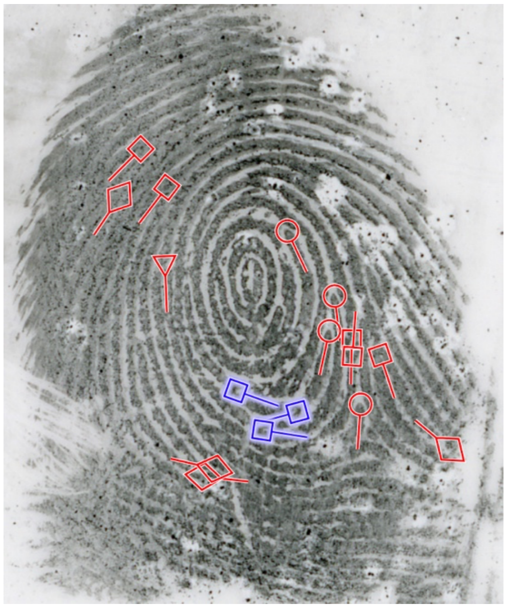
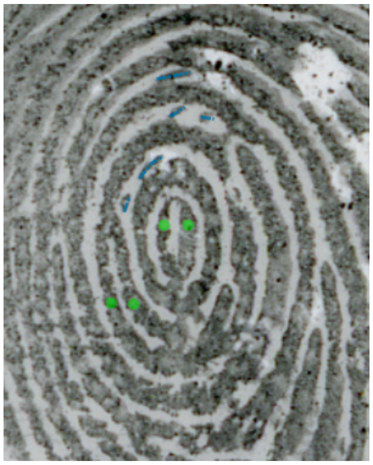


Figure 3.1 - 4 The Incipient Ridges tool and Pores tool in PiAnoS. The Incipient Ridges tool produced thin, blue lines and could be used to annotate incipient ridges or small dots that the participant used in forming a suitability decision. The Pores tool produced small, green dots and could be used to annotate pores that the participant used in forming a suitability decision.

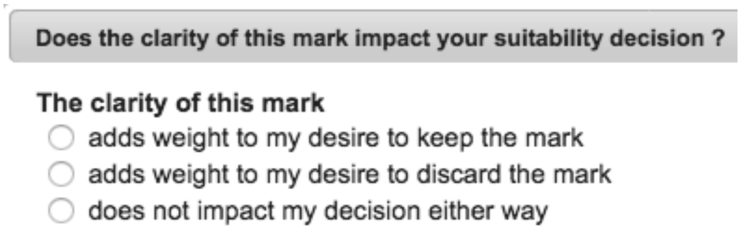


In addition to annotating features, participants were asked a series of questions about their observations regarding the clarity, distortion, pattern type, and level 3 detail of each mark.

Examiners tend not to have a consistent language or training to describe levels of clarity, distortion factors, or level 3 detail. In fact, previous research (Neumann et al. 2013; Anthonioz et al. 2008) has shown that examiners are not consistent in their interpretations of the types of distortion they see, or in what they label as level 3 details. In order to avoid this problem, and because we were less interested in what examiners *called* the things they saw than we were in how much those observations *impacted* their decisions, we refrained from asking them to catalog distortion or level 3 details they saw and instead confined ourselves to asking about how their observations affected their analysis of the mark.

We did not ask if the clarity of each impression was high or low, but instead asked whether it made the examiner more inclined to keep, or to discard, the mark (Figure 3.1 - 5). We were careful to clarify with the examiners that their *desire* to keep the mark didn't necessarily have to align with their *decision* to keep the mark—every examiner has had to keep marks in their career that they wish they didn't!

Figure 3.1 - 5 The clarity question asked during the analysis of each impression to gauge how the clarity of the impression impacted the participants' suitability decisions.



Does the clarity of this mark impact your suitability decision ?

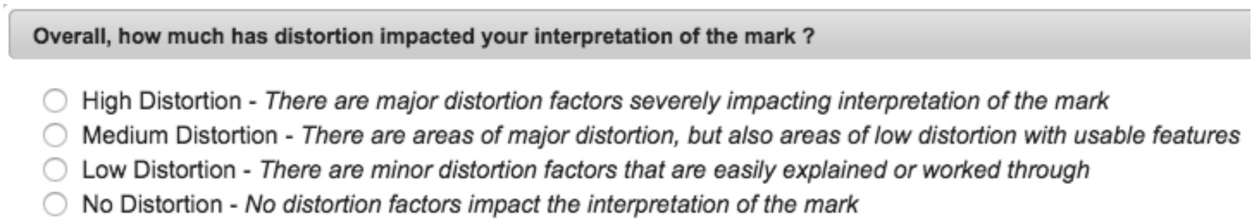
The clarity of this mark

○ adds weight to my desire to keep the mark
○ adds weight to my desire to discard the mark
○ does not impact my decision either way

The only question asked about distortion was whether the distortion observed (if any) was low, medium, or high, and we defined these in terms of how much of the mark required extra interpretation due to the distortion (Figure 3.1 - 6).

For level 3 details, we only asked whether any level 3 details were seen and given weight in the suitability decision (Figure 3.1 - 7). Again, we did not ask what kind of level 3 details they were, but only whether they mattered in the decision. If level 3 details were noted, but did *not* factor into the suitability decision, there was a response option to cover this.

Figure 3.1 - 6 The distortion question asked during the analysis of each impression to gauge how the distortion in the impression impacted the participants' suitability decisions.

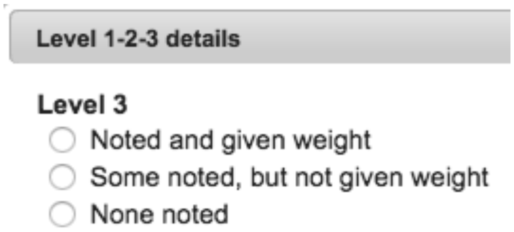


Overall, how much has distortion impacted your interpretation of the mark ?

○ High Distortion - *There are major distortion factors severely impacting interpretation of the mark*
○ Medium Distortion - *There are areas of major distortion, but also areas of low distortion with usable features*
○ Low Distortion - *There are minor distortion factors that are easily explained or worked through*
○ No Distortion - *No distortion factors impact the interpretation of the mark*

Finally, participants were asked to indicate the pattern type, or types, that they believed the mark could be, and to rank them in order from most to least likely, if more than one type was judged possible (Figure 3.1 - 8). They were also asked if there was anything distinctive about the shape of the pattern that would make it stand out from other marks of the same pattern type.

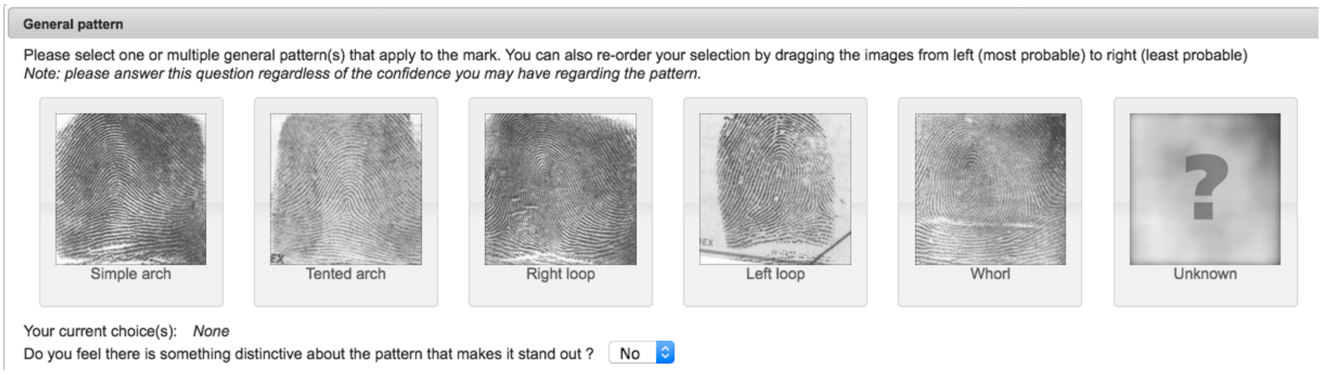
Figure 3.1 - 7 The Level 3 Details question asked during the analysis of each impression to gauge whether the participant relied on Level 3 Details in reaching their suitability decisions.



Level 1-2-3 details

Level 3
○ Noted and given weight
○ Some noted, but not given weight
○ None noted

Figure 3.1 - 8 The pattern type selection questions. Participants were asked to select any pattern types they thought the impression could be, to rank them if more than one was considered possible, and to note whether anything about the shape of the pattern was distinctive.



General pattern

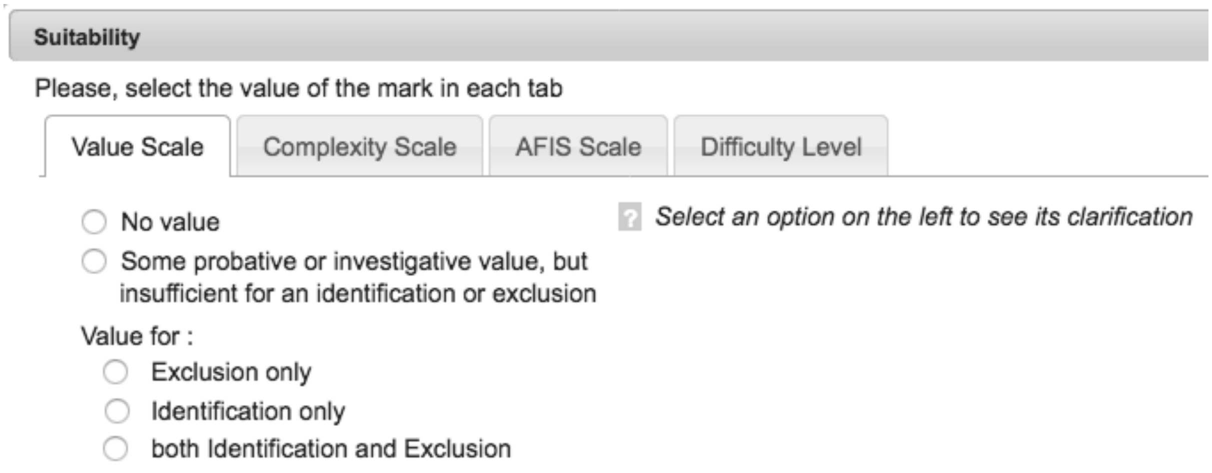
Please select one or multiple general pattern(s) that apply to the mark. You can also re-order your selection by dragging the images from left (most probable) to right (least probable)
Note: please answer this question regardless of the confidence you may have regarding the pattern.

Your current choice(s): *None*
Do you feel there is something distinctive about the pattern that makes it stand out ? No

Once these questions had been answered, participants were required to make a suitability determination along each of the four scales described above (Figure 3.1 - 9).

Prior to beginning the study, each participant completed 2 practice trials in order to try out the new tools and become familiar with the new questions and scales. Feedback was provided anonymously to each user prior to granting them access to the full study. This allowed a chance to clear up any confusion about the tools and terminology prior to collecting the study data.

Figure 3.1 - 9 The suitability decisions question. Each of the four tabs shown was selected in turn, and a suitability decision was entered for each of the four scales. Here, the expanded options for the value scale are shown. As soon as a radio button was selected, help text appeared on the right of the screen to remind the participant of what that selection meant.



Please, select the value of the mark in each tab

○ No value

○ Some probative or investigative value, but insufficient for an identification or exclusion

Value for :

○ Exclusion only

○ Identification only

○ both Identification and Exclusion

Select an option on the left to see its clarification

Statistical analysis

Statistical analysis of the results was carried out in R (version 3.5.3) (R Core Team 2018) coupled with RStudio Version 1.2.5033 (RStudio Team 2015) using the following packages: *tidyverse* (Wickham et al. 2019) for data wrangling and graphics, *caret* for machine learning (ML) and computing associated error statistics (Kuhn 2020, https://CRAN.R-project.org/package=caret) and *vip* for variable importance (Greenwell et al. 2019, https://CRAN.R-project.org/package=vip). Note that *caret* will call for the required packages for each ML model used.

Machine learning was used in the same way as in a previous white box study on fingerprints carried out by Neumann et al. (Neumann et al. 2013). In the present study, we tested a set of ML classifiers that could act as rational proxies for human decision-making. The most accurate ML model will allow us to find a set of reasonable predictors for the decisions made by the study participants based on their annotations. We used all survey variables from the participants and all variables from the annotations they made as input variables to the classifiers. The predicted output is the conclusion reached by the examiner for each scale. We selected the best performing classifier based on its accuracy. We then computed the importance of the predictors against the individual decisions made by the participants. That ranking allowed us to identify predictor variables that have the most impact on the decisions reached.

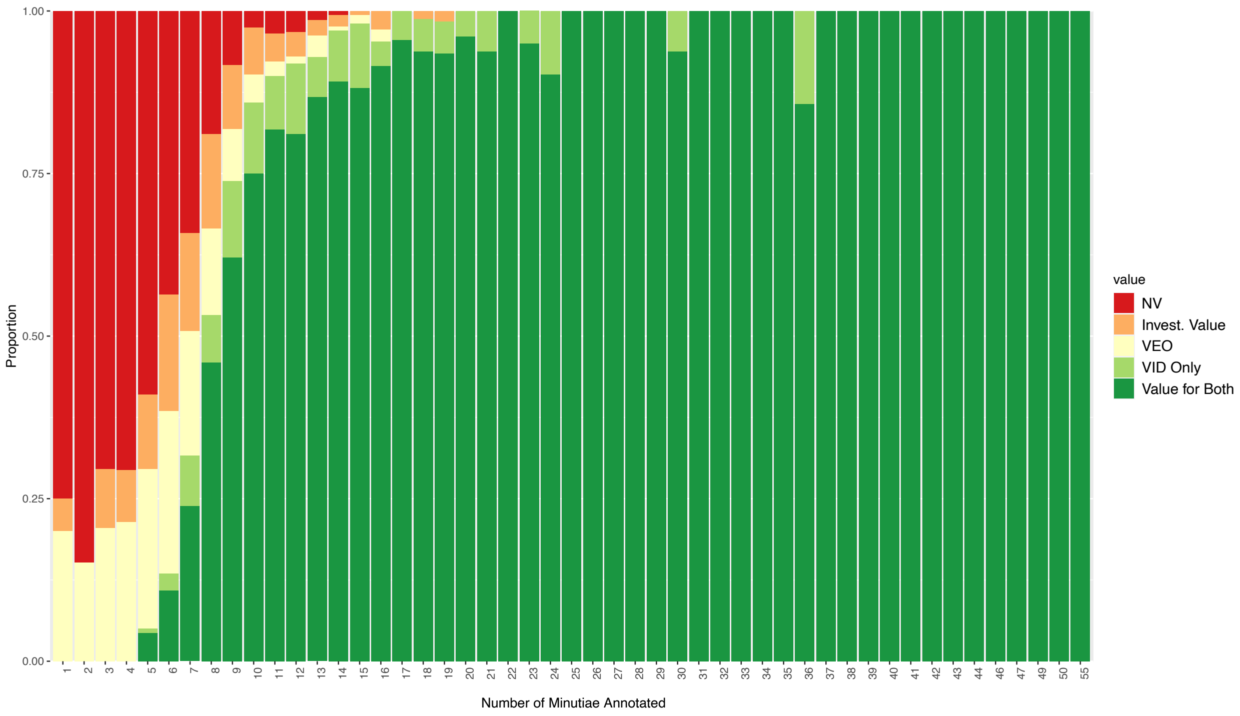
The retained classifiers ranged from simple tree model (CART), to K-Nearest Neighbors (KNN), Random Forest (RF), Neural Networks (NN), boosted trees (C5.0), and Support Vector Machines (SVM). The entire dataset was used for the machine learning. To avoid overfitting, cross-validation using 10 folds and repeated 5 times was systematically used for each classifier.

Results and Discussion

Consistency of variable selection on four scales

We explored the data to identify trends in which variables best captured the information used to reach conclusions along each scale. Beginning with an assumption that minutiae count was the most important factor for predicting all suitability decisions, we ordered the data by minutiae count and plotted the conclusions that were rendered along each scale. Figure 3.1 - 10 and Figure 3.1 - 12 show that minutiae count was a good predictor of the value decision but did not perform as well on the complexity decision. This is intuitive, because when considering complexity, we would expect examiners to be thinking about more than just minutiae count, such as the distortion or clarity of the image. Minutiae count was a similarly poor predictor of decisions on the AFIS and Difficulty scales where, likewise, one would expect other factors to more heavily influence decisions.

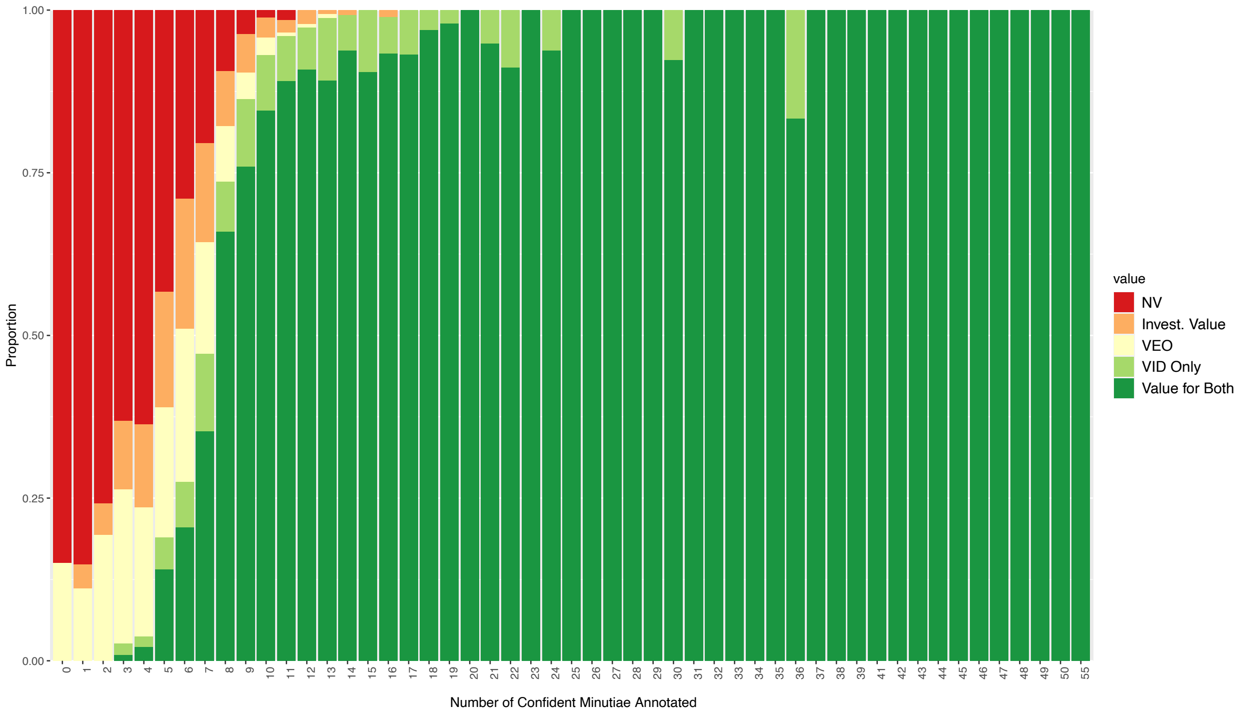
Figure 3.1 - 10 Suitability decisions rendered along the value scale. The data are ordered according to the number of minutiae annotated. Note that because different participants annotated different numbers of minutiae on the same image, these data do not reflect agreement on any particular image, but rather, what value decision individual examiners made on any image when they annotated the number of minutiae noted on the x-axis.



Interestingly, in contrast to earlier findings (Ulery et al. 2013; Ulery et al. 2014; Langenburg 2012) these data do not initially support the 7-to-8 minutiae operational tipping point for value. Figure 3.1 - 10 shows that around 7 or 8 minutiae only about 25 to 50 percent of value decisions are either “Value for Both” or “VID Only”. It is not until the 9 minutiae mark that these two categories combined are exclusively over 50%, and not until 11 minutiae or higher that they are exclusively over 75%. Additionally, since the instructions for this study explicitly asked participants to only annotate the minutiae they used to make their decisions, not all minutiae they saw, these data may more closely reflect decision thresholds.

However, Figure 3.1 - 11 presents the same data, but excluding minutiae marked as “uncertain” by the examiners. We can see that when using only confident minutiae, the “tipping point” does fall back to an approximately 8-minutiae threshold. It is unclear how these results align with previously-reported data, since the other studies did not draw a distinction between certain and uncertain minutiae in their analyses and thus their thresholds may also move upon closer analysis.

Figure 3.1 - 11 Suitability decisions rendered along the value scale excluding minutiae marked using uncertain minutiae marker types. The data are ordered according to the number of confident minutiae annotated.



We then considered combinations of variables that may explain complexity decisions better than minutiae count did. Figure 3.1 - 13 shows the effect of perception of distortion on complexity decisions. When there is high distortion, even high minutiae count can rarely overcome it to reach a decision of non-complex. However, with low or no distortion, very few images are considered complex unless minutiae count is very low. This same effect was more weakly seen with the perception of clarity (Figure 3.1 - 14), suggesting that distortion plays a bigger role than clarity in overcoming minutiae count for complexity decisions.

Figure 3.1 - 12 Suitability decisions rendered along the complexity scale. The data are ordered according to the average number of minutiae annotated.

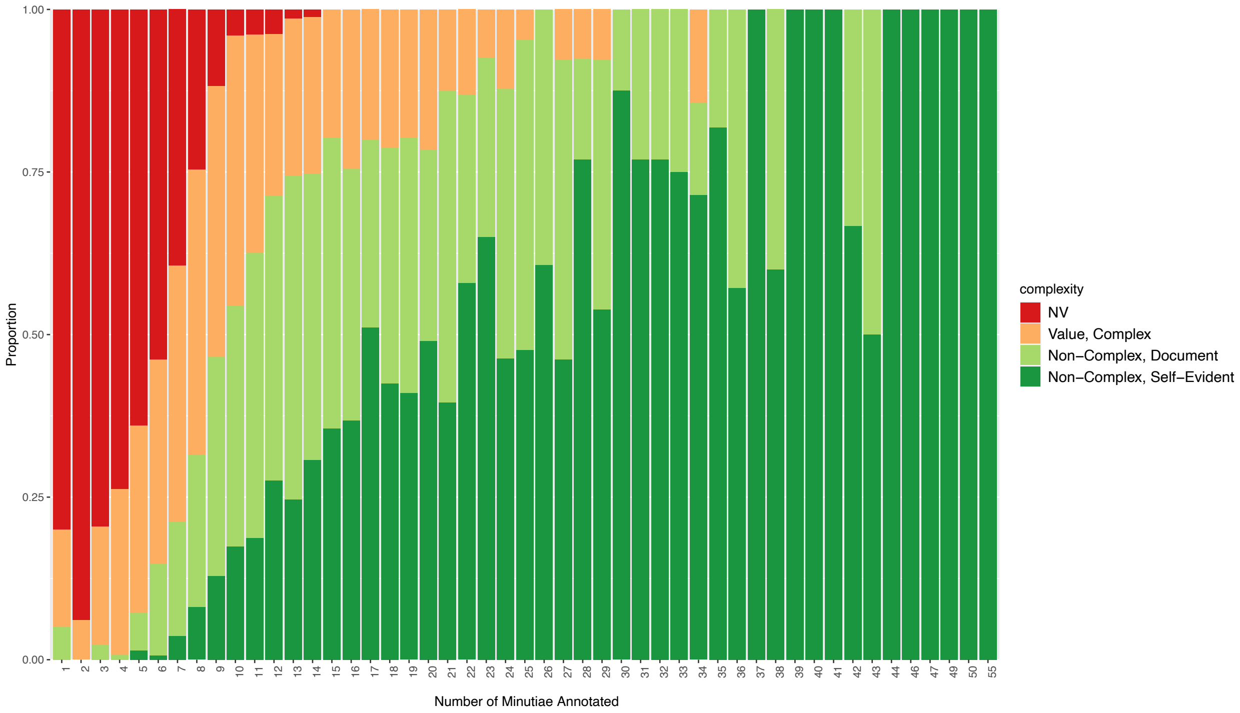


Figure 3.1 - 13 Suitability decisions rendered along the complexity scale. The data have been separated according to the participants' responses to the distortion question to see how the distortion perceived in the image may have affected complexity decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.

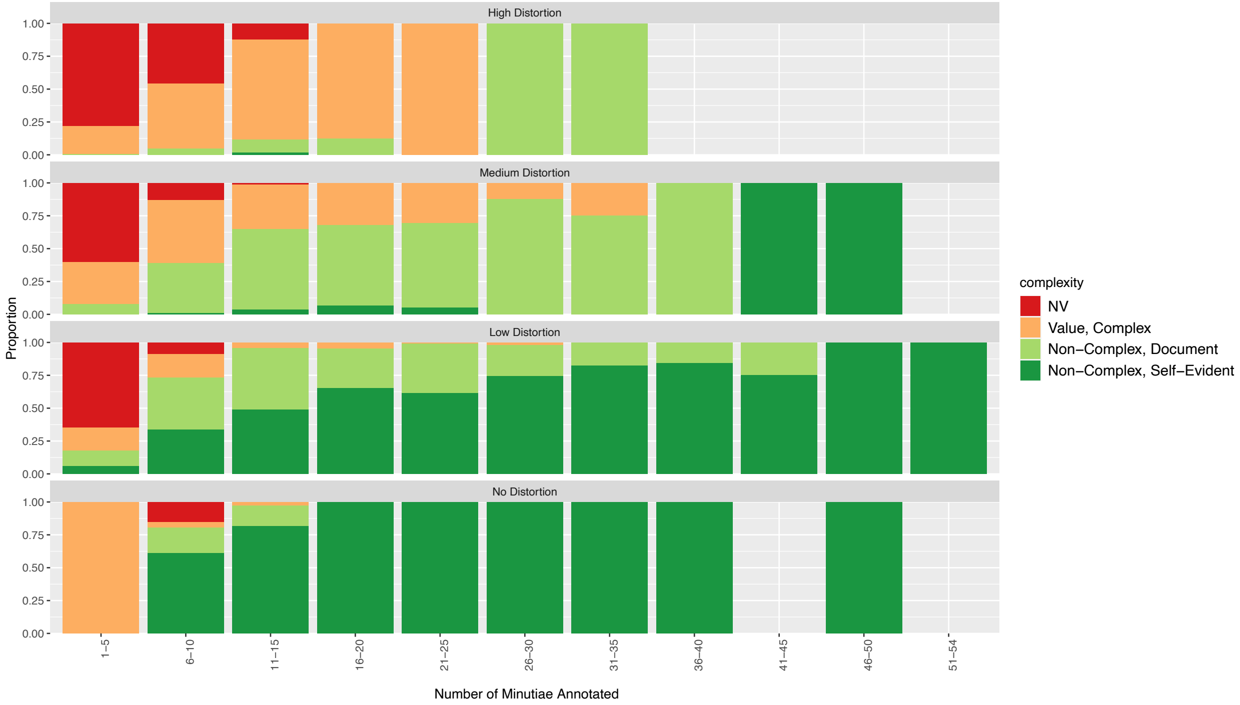
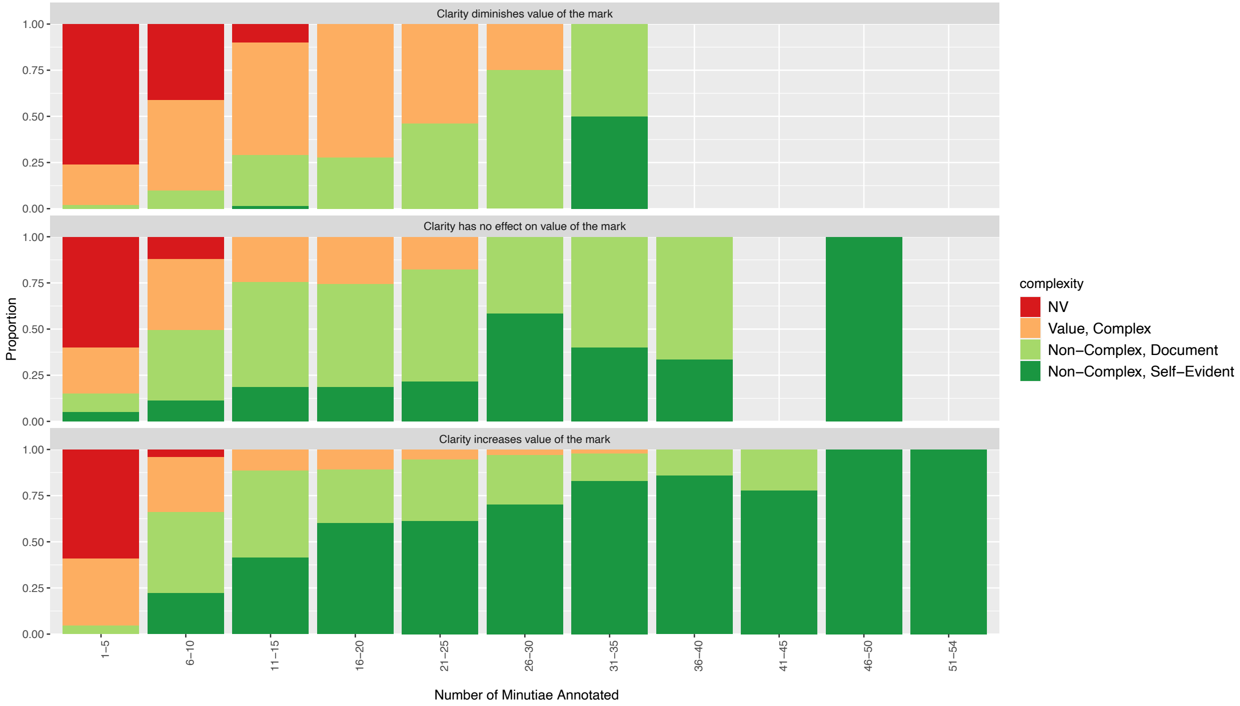


Figure 3.1 - 14 Suitability decisions rendered along the complexity scale. The data have been separated according to the participants' responses to the clarity question to see how the perceived clarity of the image may have affected complexity decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.



We then considered combinations of variables that may explain complexity decisions better than minutiae count did. Figure 3.1 - 13 shows the effect of perception of distortion on complexity decisions. When there is high distortion, even high minutiae count can rarely overcome it to reach a decision of non-complex. However, with low or no distortion, very few images are considered complex unless minutiae count is very low. This same effect was more weakly seen with the perception of clarity (Figure 3.1 - 14), suggesting that distortion plays a bigger role than clarity in overcoming minutiae count for complexity decisions.

The AFIS and Difficulty scales showed a very similar pattern of distortion and clarity's effects on overcoming minutiae count. Figure 3.1 - 15 and Figure 3.1 - 16 illustrate the effects of distortion and clarity on the AFIS scale, and Figure 3.1 - 17 and Figure 3.1 - 18 illustrate their effect on the Difficulty scale. It can be seen in Figure 3.1 - 15 and Figure 3.1 - 16 that on the AFIS scale, distortion once again seems to have a greater impact on decisions than clarity does, although both distortion and clarity's effects are somewhat more pronounced on the AFIS scale than they were on the Complexity scale. Figure 3.1 - 17 and Figure 3.1 - 18 show that the same pattern (of distortion having a greater impact than clarity) still holds, but that on the Difficulty scale distortion appeared to almost completely drive the decision.

Figure 3.1 - 15 Suitability decisions rendered along the AFIS scale. The data have been separated according to the participants' responses to the distortion question to see how the distortion perceived in the image may have affected AFIS decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.

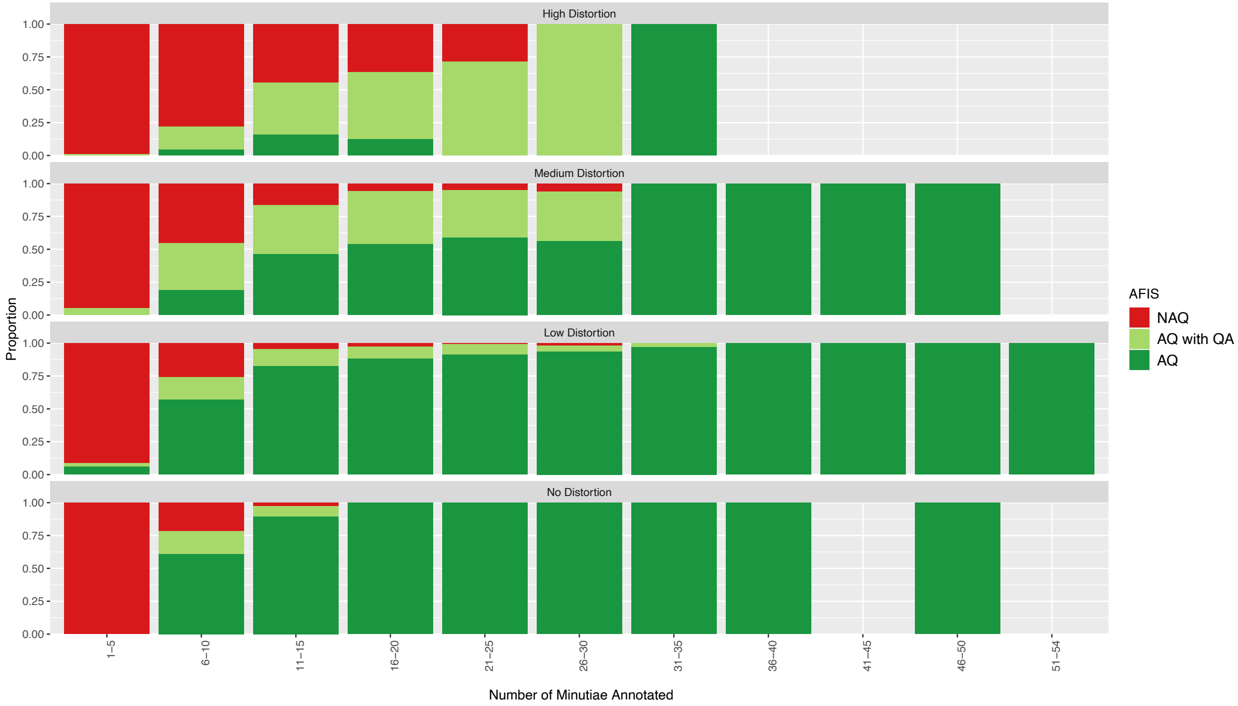


Figure 3.1 - 16 Suitability decisions rendered along the AFIS scale. The data have been separated according to the participants' responses to the clarity question to see how the perceived clarity of the image may have affected AFIS decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.

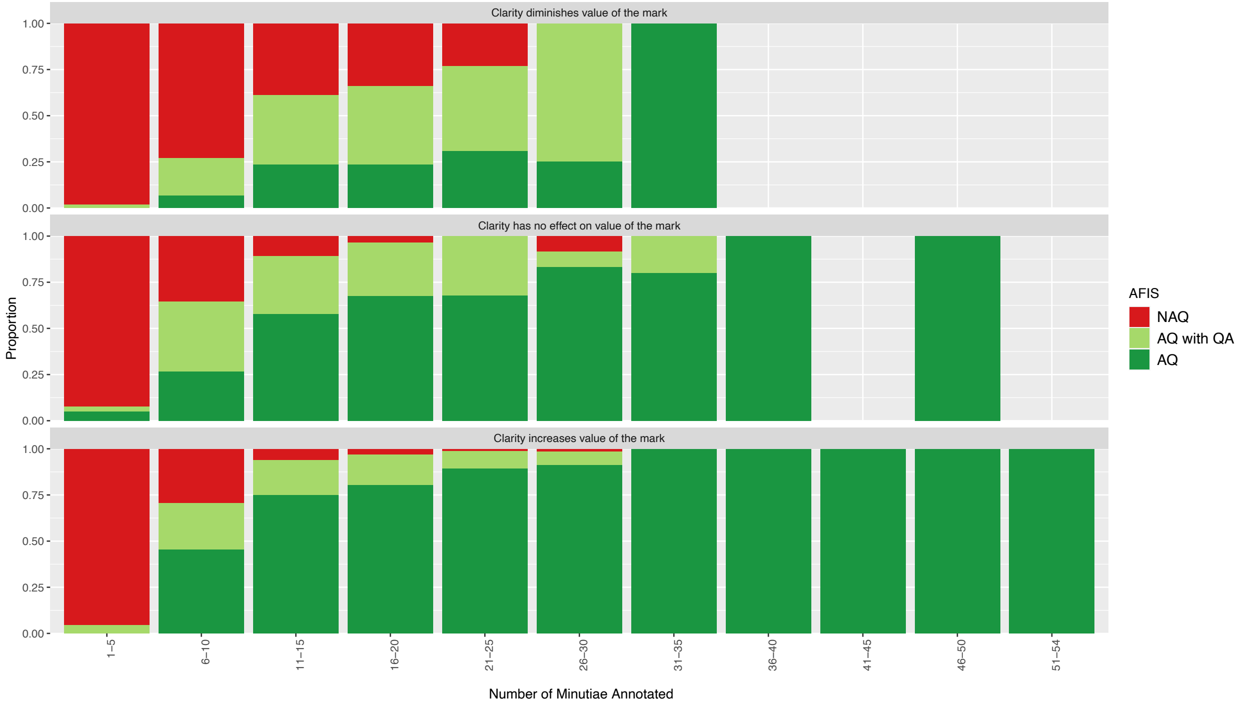


Figure 3.1 - 17 Suitability decisions rendered along the difficulty scale. The data have been separated according to the participants' responses to the distortion question to see how the distortion perceived in the image may have affected difficulty decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.

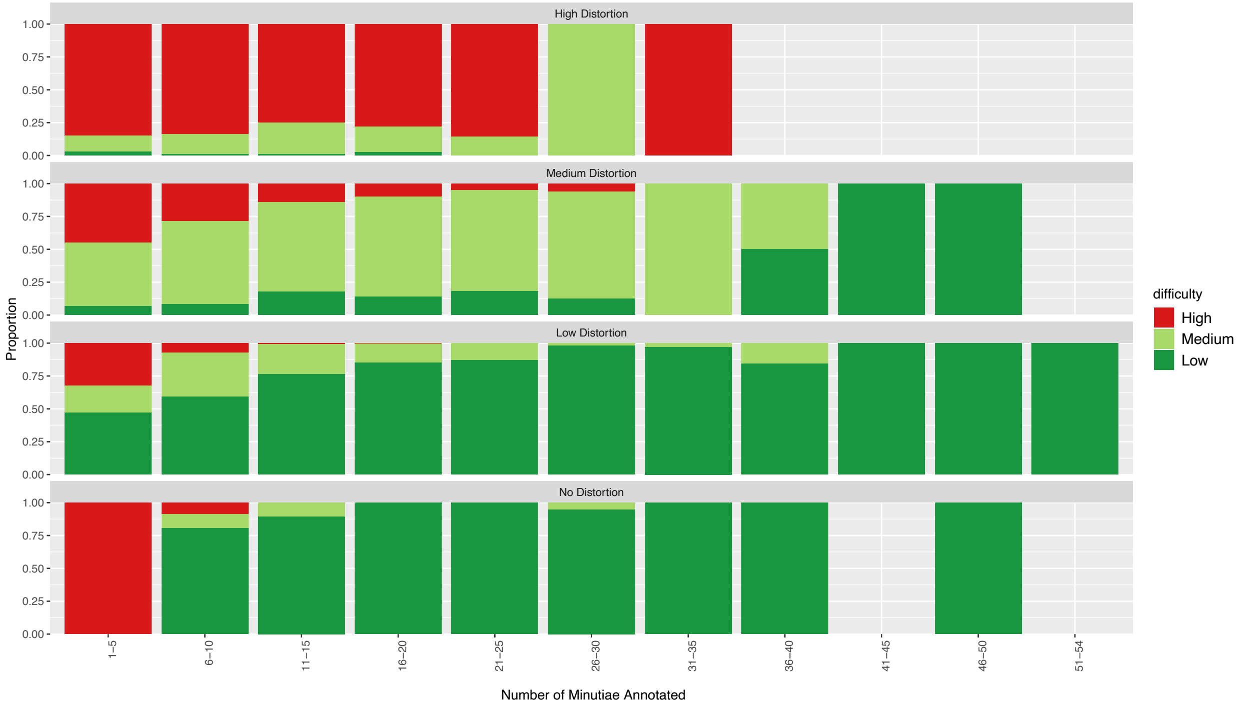


Figure 3.1 - 18 Suitability decisions rendered along the difficulty scale. The data have been separated according to the participants' responses to the clarity question to see how the perceived clarity of the image may have affected difficulty decisions. The data are ordered according to the number of minutiae annotated and binned in groups of 5 minutiae.

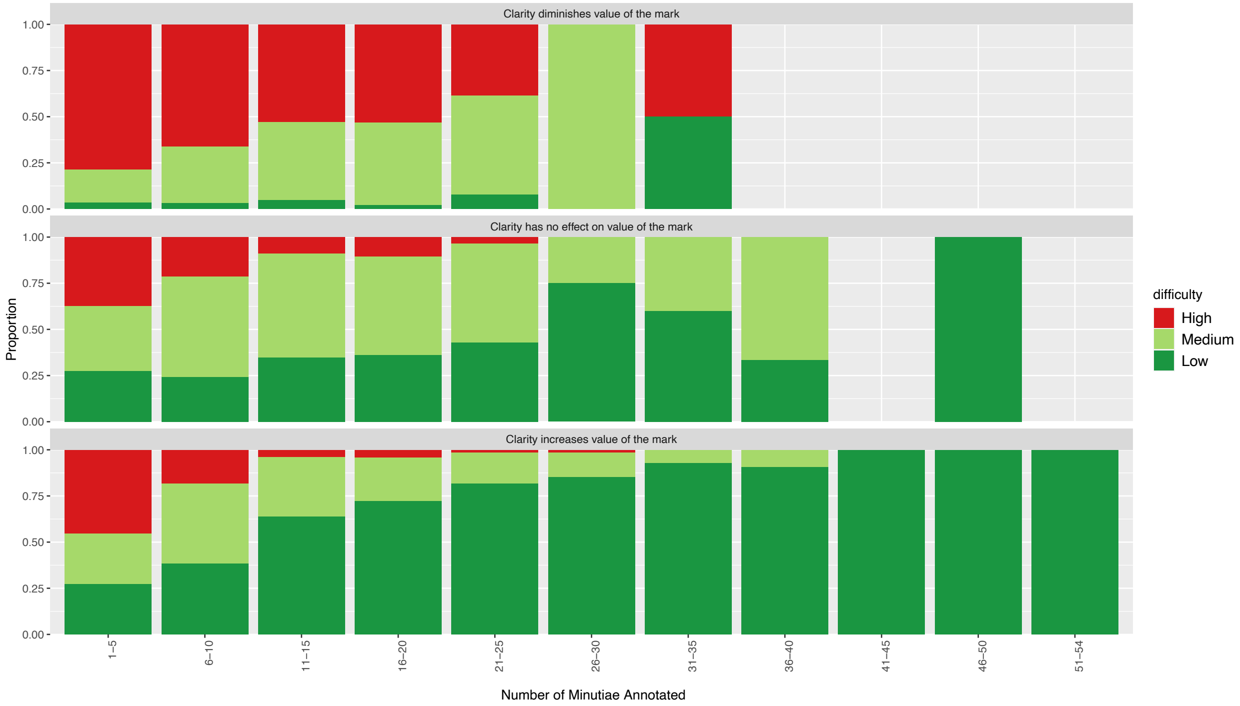
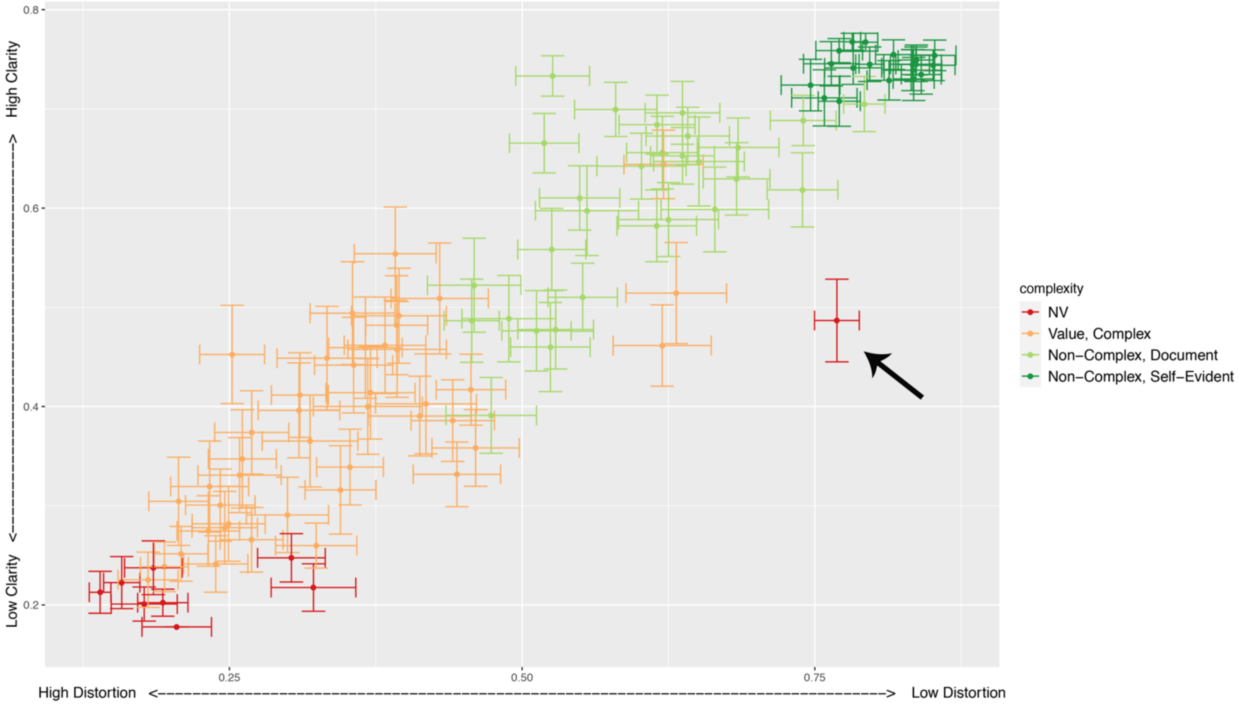


Figure 3.1 - 19 and Figure 3.1 - 20 show how well the variables Distortion and Clarity together predict conclusions on the complexity and value scales[*]. Note that each data point represents the average response of all users who viewed a single image. The error bars represent the standard error of the mean observed on each exercise. Figure 3.1 - 19 illustrates that, with one exception (indicated by the arrow), distortion and clarity behave as expected in relation to complexity decisions; high distortion, low clarity images tend to be called no value whereas low distortion, high clarity images tend to be called non-complex and self-evident and the transition between the two extreme categories is smooth. The outlier (Figure 3.1 - 21), is an image in which there is low distortion and medium clarity, but very few minutiae and no anchors, which may explain why the consensus response for this image was No Value.

Figure 3.1 - 19 Suitability decisions rendered along the complexity scale for all 100 study images. Each dot represents one image and is placed at the average value for clarity and distortion responses among the participants who viewed that image. The arrow indicates a single outlier of low distortion and average clarity that was nonetheless determined by consensus to be no value.



In contrast, Figure 3.1 - 20 illustrates that distortion and clarity don't predict the value decision nearly as well. Although the same outlier exists for the same reasons, note how much messier the color progression of the conclusion responses is, compared to Figure 3.1 - 19. Value determinations are not following a tidy progression based upon distortion and clarity. Thus, it would seem that value was, in fact, better predicted by minutiae count (Figure 3.1 - 10), whereas complexity was better predicted by distortion and clarity.

[*] The variables Distortion and Clarity were recorded as ordered factors. For data analysis, they were converted to numerical values on a continuous scale between 0 and 1 using ridits (Bross 1958) to empirically determine the weighted space between each value on the scale.

Figure 3.1 - 20 Suitability decisions rendered along the value scale for all 100 study images. Each dot represents one image and is placed at the average value for clarity and distortion responses among the participants who viewed that image.

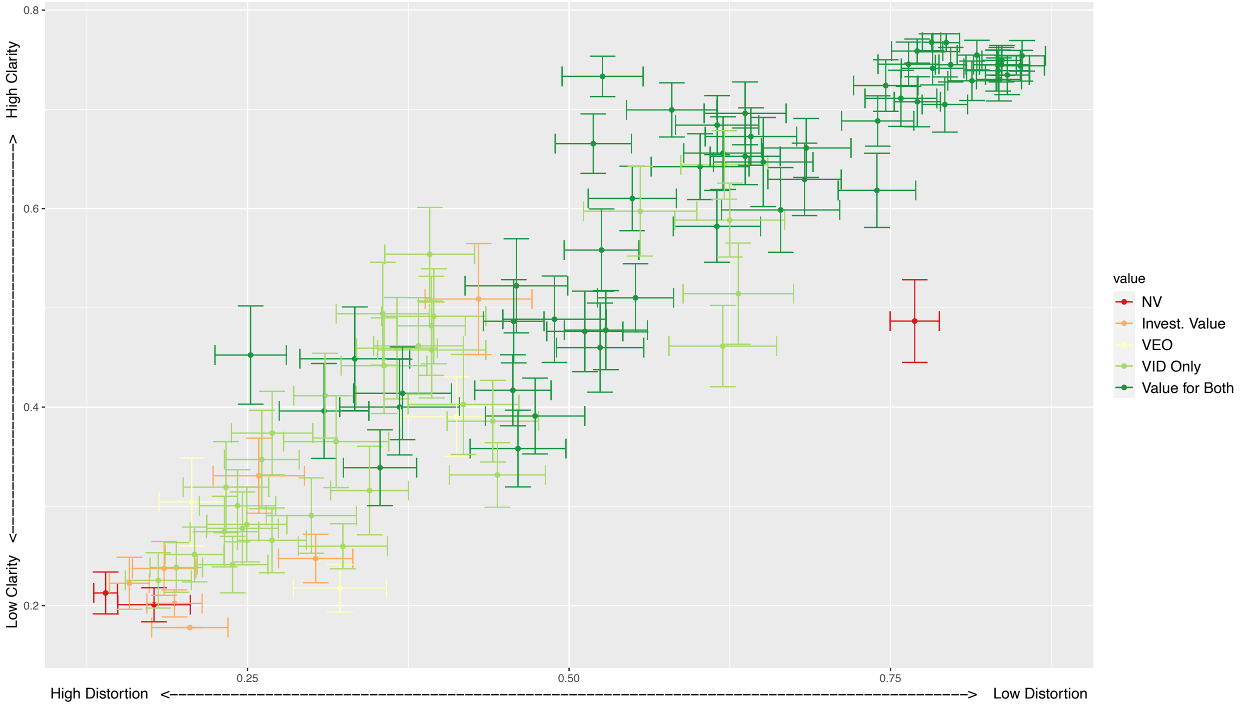


Figure 3.1 - 21 Image 32 in the study. This is the image indicated by the arrow in Figure 3.1 - 19. While this image has low distortion and average clarity, there were few minutiae present and no core or anchors. Some of the participants' comments indicate their reasoning in selecting NV for this mark: "not enough 2nd level detail for an ID, not enough 1st or 2nd level detail for exclusion", "Latent contains insufficient minutiae and wouldn't be effectively searched due to ambiguity of location on fingertip. There is the potential that another individual may share same features due to lack of rarity in ridge events. No reliable target group. No value decision more heavily based on the lack of ridge events, not clarity."

However, there is more to unpack from Figure 3.1 - 19 and Figure 3.1 - 20. Although the overall trends indicated in Figure 3.1 - 19 fall in line nicely with what would be expected, the variability represented in these data is quite high. The vertical and horizontal bars around each data point represent the standard errors of each mean (or an estimate of how far some respondents stray from the population mean for each image) taking into account the number of participants who viewed each image. The placement of each dot represents the average response from all users who viewed the image.

This means that although consensus observations may be reliable predictors of consensus decision outcomes, for each image there will be examiners who disagree with that consensus, in some cases quite strongly. In order to capture the consensus response, the opinions of individual examiners at the extremes will be lost in models based upon average examiner observations. Because of this, any models based upon these data will be unable to please all the examiners all the time; there will always be cases where the examiner disagrees with a consensus-based model in the same way there would always be cases where an individual examiner disagrees with a consensus of other examiners. It is up to the practitioner community to decide whether they are willing to trade sensitivity (or the ability of some examiners to make correct decisions on more marginal data) for reliability (or security in knowing that decisions are defensible and a consensus of experts is likely to accept them).

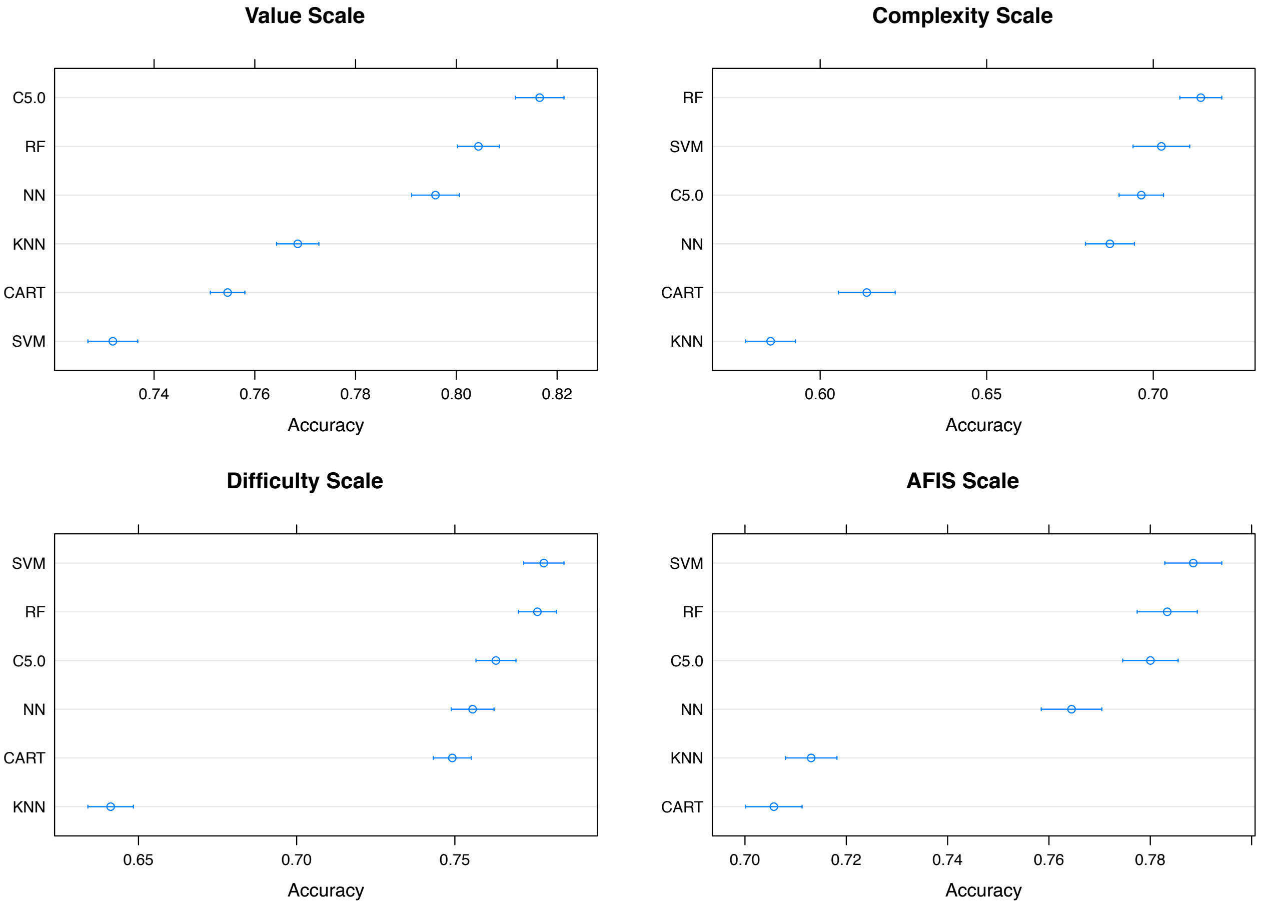
Because there is no ground truth available for suitability decisions, we have chosen in this research to consider the majority vote to be the “ground truth” along each of the four scales for each mark. It should be understood that we could just as easily have chosen to always take the lowest opinion, or the highest, but this would skew the data toward or away from risk levels that the majority would not share.

To identify the variables for each scale upon which examiners most relied to reach their suitability decisions and to test the impact (if any) of demographic and agency policy factors, we used machine learning algorithms to find the most important variables that predicted a participant’s *own* suitability decisions using their own annotations. In other words, this analysis did not consider the consensus ground truth suitability determination for each mark, but only which variables most impacted a participant’s own decisions. In all, 36 variables were tested.

Figure 3.1 - 22 shows the accuracy of the ML models tested for the four scales. Overall, random forest is the best performing across the scales. Random Forest offers a robust model-based assessment of the ranked importance of variables. Variable importance is shown for each scale in Figure 3.1 - 23 for the 10 highest impact variables.

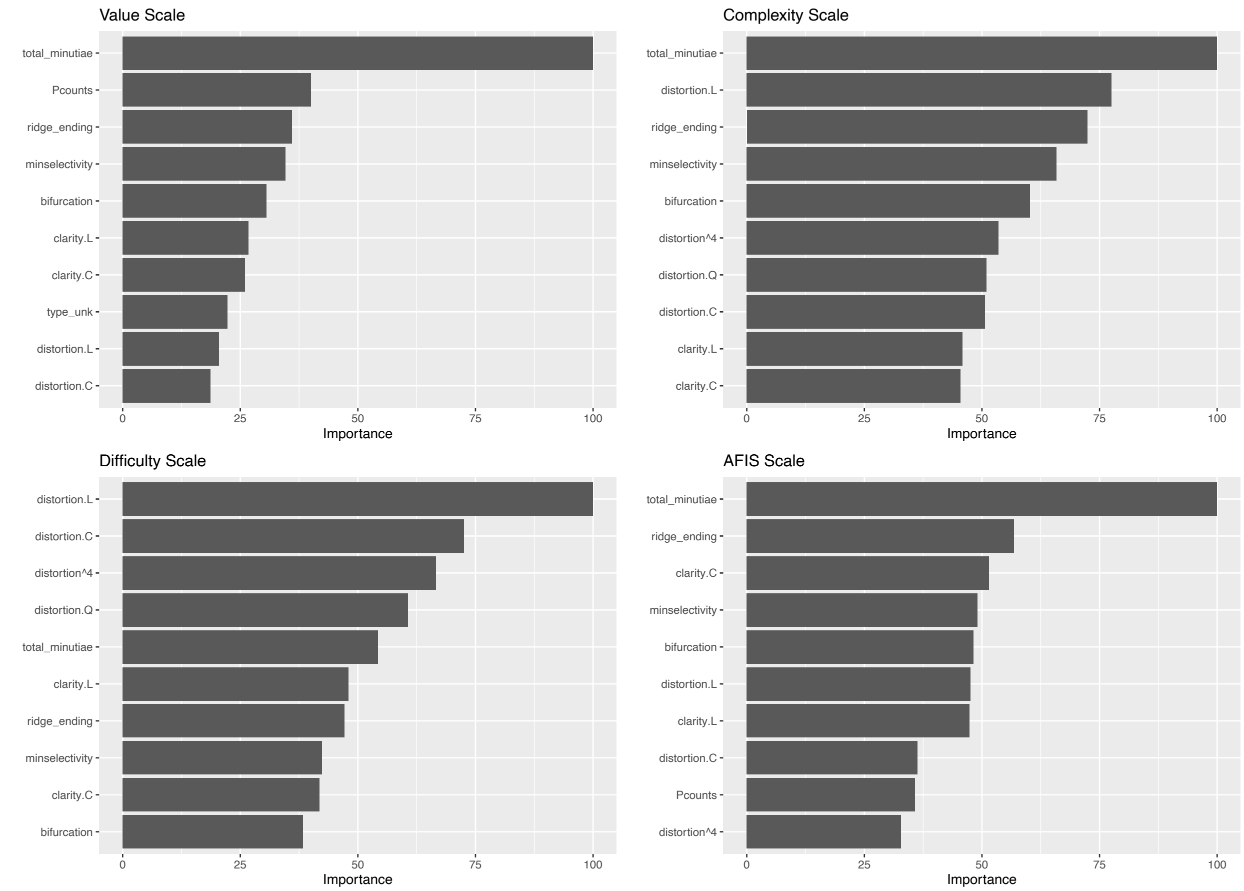
Overall, and across all four scales, the variables that consistently performed the best in predicting suitability decisions were: total number of minutiae marked, number of confident minutiae marked (as opposed to uncertain minutiae marked), the clarity of the image, the level of distortion in the image, whether the pattern type could be determined with confidence, and the selectivity of the minutiae*. Weight of level 3 features, annotation of incipient ridges, pores, or target groups, and demographic factors including gender, years of experience, certification status, educational level, and agency approach type had no influence on suitability decisions for any of the four scales.

Figure 3.1 - 22 Comparison of the accuracy of the ML models tested for each of the four scales. The intervals around each dot give the 95% confidence interval based on the 5 repeats carried out.



* Selectivity of the minutiae is represented in the variables of importance as "minselectivity," which is a measure of the perceived selectivity across the entire impression. It takes into account the ratio of confident total minutiae, the number of highly distinctive combined minutiae groups noted, and whether a target group was noted. These last two variables, although important parts of the minselectivity variable, did not prove to be highly diagnostic on their own.

Figure 3.1 - 23 Comparison of the relative importance of the variables[*] for each of the four scales. The 10 highest impact variables are shown ordered by importance. The model chosen for all four scales is Random Forest (RF).



Consistency of suitability decisions

As was previously noted, examiners can differ in their suitability determinations. However, it has often been posited that although examiners may vary on ambiguous images, they are going to be unanimous on suitability decisions at the extreme ends of the spectrum (i.e. on very clear, or very degraded images). However, our data do not provide evidence for this. Although there are some images where there is unanimity in the "Value for Both" determination, there are many images where the responses are nearly equal across all 5 choices, and further, there is not a single case in which there is unanimity for a no value ("NV") decision (some examples are shown in Figure 3.1 - 24). This suggests that although examiners may agree on what a high-quality mark looks like, they do not agree on what a low-quality image looks like.

[*] The odd appearance of some of the variables (e.g., "distortion.L," "distortion.C", etc) is an artifact of the machine learning software. These are dummy variables with their root (e.g., "distortion") being the true root variable. For variables with multiple possible responses (distortion has 4), each of the possible responses represents a sub-category that can be present, or not. When several of these dummy variables appear in the Variables of Importance figure (such as distortion does in this figure), it indicates that the root variable (distortion) is globally important.

Figure 3.1 - 24 The value decisions from all participants who viewed the image for four select images from the study. These graphs represent some of the range of variability in value decisions. Panel (A) shows the results for Image 40 – an image for which the "Value for Both" decision was unanimous (see Figure 3.1 - 25). Panel (B) shows the results for Image 77 – an image for which there was a strong consensus that the image was "NV," yet there were also many participants who did not agree (see Figure 3.1 - 26). Panels (C) and (D) show two examples of the results for images where the votes for the five possible value determinations were nearly evenly split. The images corresponding to panels (C) and (D) are presented as Figure 3.1 - 27 and Figure 3.1 - 28.

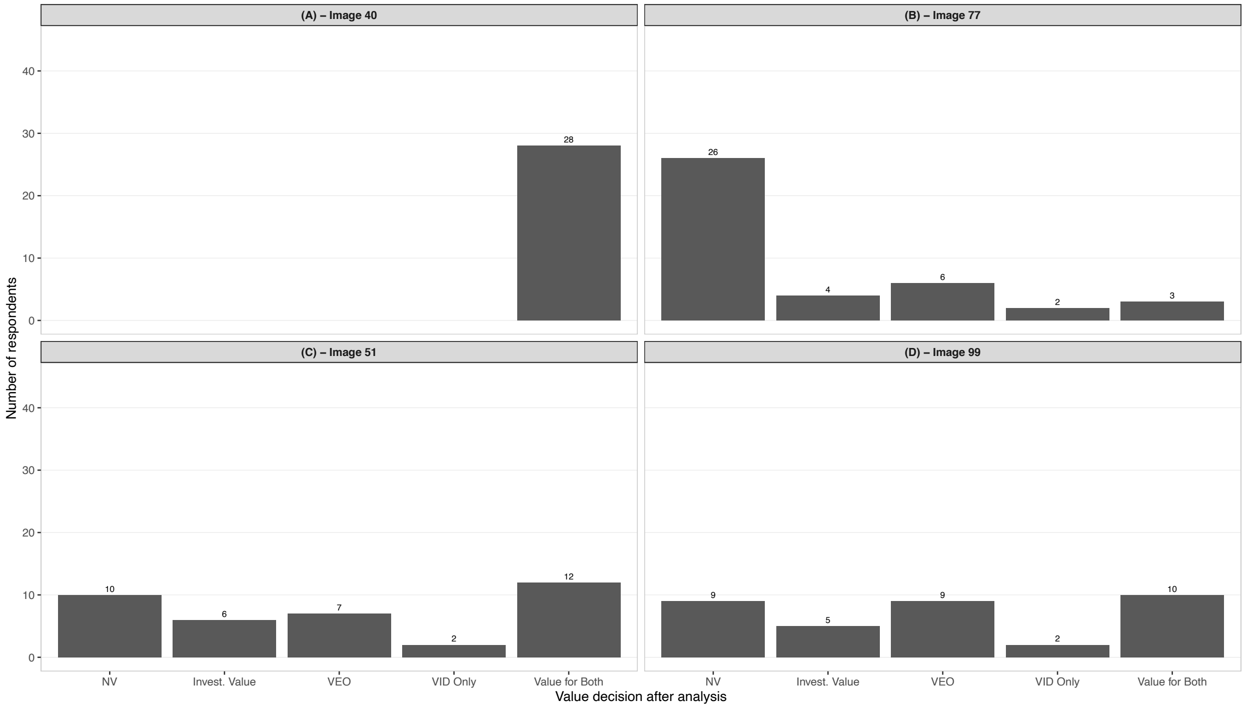
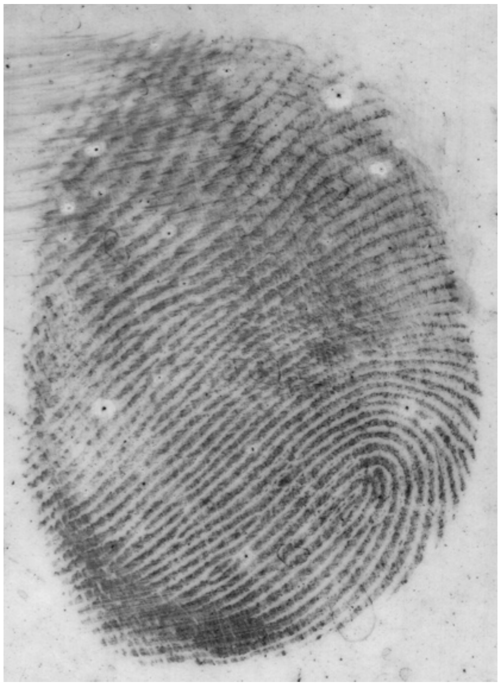


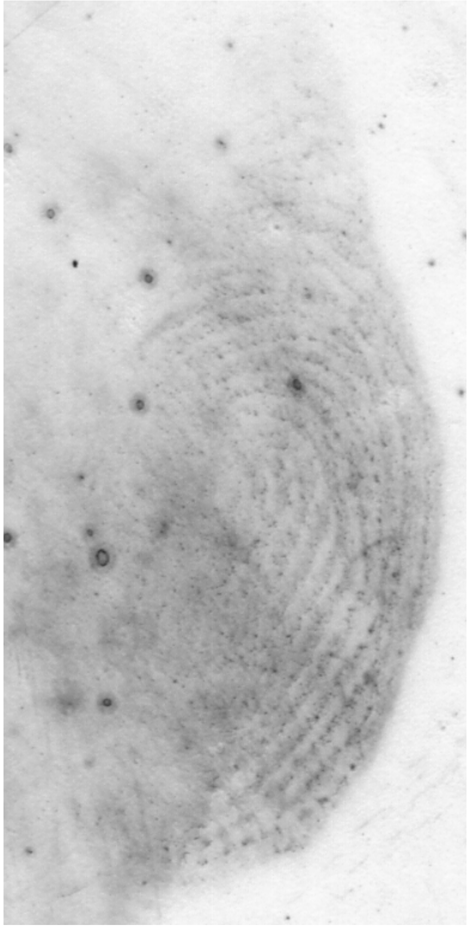
Figure 3.1 - 25 This is Image 40 from the study. It is one of 3 images in the study for which there was 100% agreement from all participants who viewed it at the top end of both the value and complexity scales. That is, every participant who viewed this image selected "Value for Both" on the value scale and "Non-Complex, Self-Evident" on the complexity scale.



At the high end of the value scale, there were 17 images (out of 100) in which the "Value for Both" decision was unanimous and 13 in which only one person disagreed. However, in 7 out of 8 situations where only one participant chose VID (all others chose Value for Both), that participant was User_144. These data support that, with the exception of one participant who seemed to favor the use of the VID category, participants are fairly consistent at determining when a mark is of very high value (for example, Figure 3.1 - 25).

However, this does not mean that if many people agree that an impression is of Value for Both, then nearly all people will agree. In 18 of the 100 images in the study, 20 or more participants agreed upon Value for Both, yet more than 3 participants disagreed (Figure 3.1 - 30 shows one of these cases). In fact, in 13 of those 18 cases (72%), at least one person thought the same image was no value.

Figure 3.1 - 26 This is Image 77 from the study. It is one of 7 images in the study for which there was a strong consensus (over 20 votes) for "NV" on the value scale, yet in every case at least 5 people disagreed with that assessment. The votes for this particular image were: NV (26); Investigative Value (4); VEO (6); VID only (2); Value for Both (3).



Consensus seemed to be even more difficult to reach at the low end of the value scale. As previously noted, there was not a single image for which there was a unanimous "NV" decision. Even when there was strong agreement on NV (defined as at least 20 votes for "NV"), there were always between 5 and 15 participants who disagreed with the NV designation and in every case, there was at least one vote for "Value for Both" (one example is given in Figure 3.1 - 26).

An interesting observed phenomenon was some participants' inconsistency in rendering a NV decision on the same image. In the PiAnoS interface, participants were asked to first annotate the image, then answer questions about clarity, distortion, etc. and finally to render their suitability decisions along the four scales. The four scales question was a single question with four tabs (see Figure 3.1 - 9)—one tab for each scale. Thus, once the participant was finished with their annotations and got to the suitability questions, they would select their conclusion for the value scale, click the next tab to enter their conclusion for the complexity scale, and so on. For most participants, it should have been a matter of seconds between when they made their selections on the Value tab and the Complexity tab.

Yet, it was observed that in 41/100 images, at least one person changed their mind about whether the image was NV or not between the Value scale and the Complexity scale (e.g., selected "NV" on the Value scale, then selected "Of value, complex" on the Complexity scale moments later). This was the same person, viewing the same image, rendering two different decisions moments apart. It is true that the Value scale and the Complexity scale were *different* so an argument could be made that people set their personal NV thresholds differently on the two scales. However, the definition of NV that was provided to participants was *exactly the same* for both scales[*] and help text was provided at each decision point as a reminder. We could argue that, if something is not of value, it's not of value independent of whether one is considering its complexity or only its value. Our participants didn't respond following that expected logic. Unfortunately, the collected data do not illuminate the reason for this observed discrepancy.

Figure 3.1 - 27 This is Image 99 from the study. It is another example of an image that had nearly even votes across the value categories. The votes for this particular image were: NV (9); Investigative Value (5); VEO (9); VID only (2); Value for Both (10).

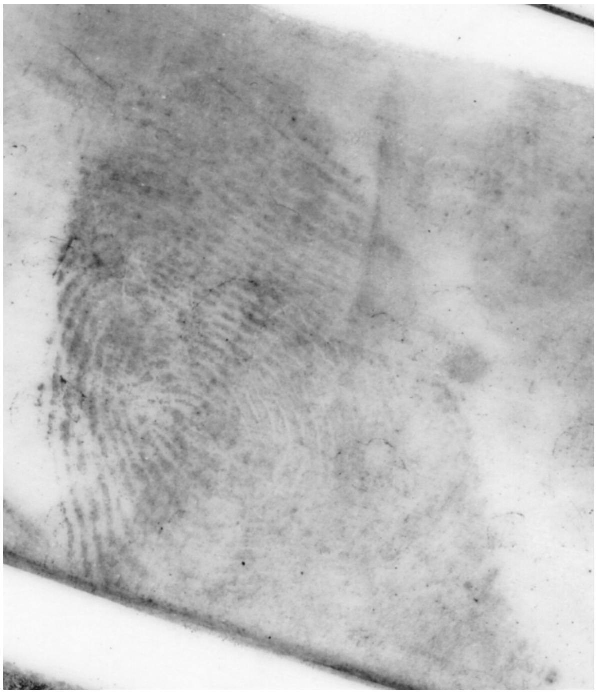
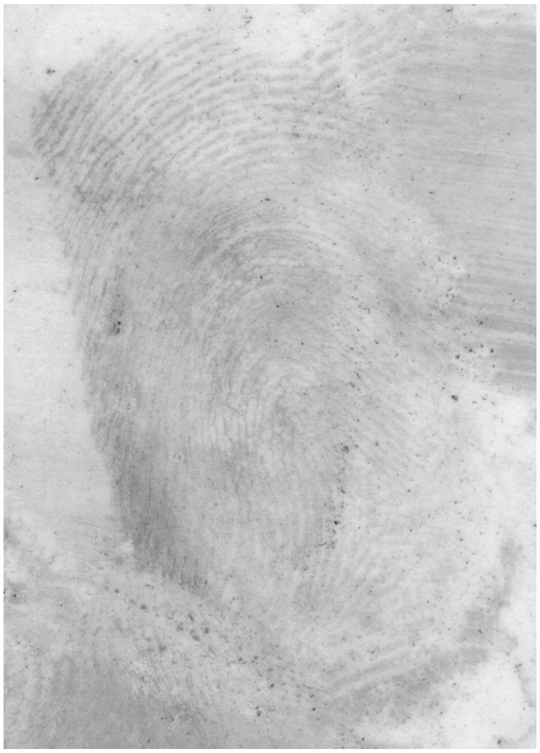


Figure 3.1 - 28 This is Image 51 from the study. It is an example of an image that had nearly even votes across the value categories. The votes for this particular image were: NV (10); Investigative Value (4); VEO (7); VID only (2); Value for Both (12).



Interestingly, among the 41 images where someone changed their NV determination between the Value and Complexity scales, most of the time between 1 and 4 people changed their minds. The highest affected image recorded 7 individuals changing their minds (an example is

[*] The definition that was provided for No value on both the Value and Complexity scales was "The mark does not contain sufficient information to proceed with a comparison."

Figure 3.1 - 29). Overall, 47 individual participants changed their NV determination between the Value and Complexity scales on at least one image; most did this on 1, 2, or 3 images throughout the study, but one user each changed their mind on 4, 5, and 6 images.

Figure 3.1 - 29 This is Image 100 from the study. It is the image for which the highest number of people (7) changed their mind regarding the NV decision between the value scale and the complexity scale.

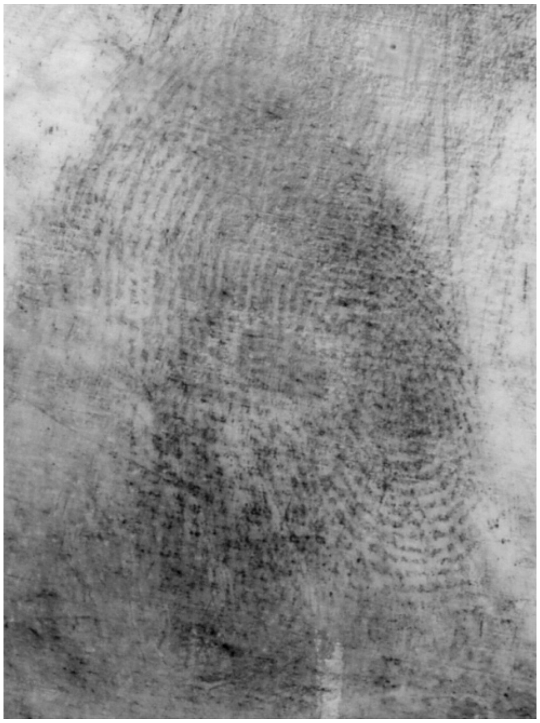
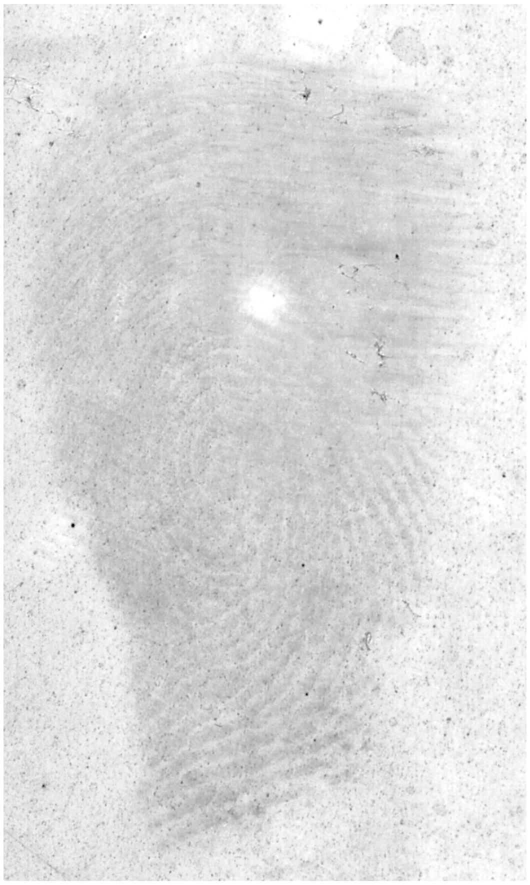


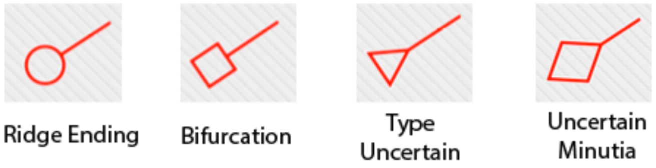
Figure 3.1 - 30 This is Image 87 from the study. It is one of the 18 images in the study for which there was a strong consensus (over 20 votes) for "Value for Both" on the value scale, yet still a high number of people who disagreed with that assessment. The votes for this image were: NV (4); VEO (4); Value for Both (23).



Certainty of minutiae selection

Tools were provided for participants to not only annotate the minutiae they took into consideration in making their suitability determinations, but also to indicate the type of minutia and their certainty about the minutia type. The 4 minutiae marker types (Figure 3.1 - 31) were: ridge ending, bifurcation, “type uncertain” and “uncertain minutia” (used when the participant was not confident that a particular minutia was actually present).

Figure 3.1 - 31 These are the four minutiae marker types that were available to participants to annotate minutiae. The first two, bifurcation and ridge ending, are “certain” marker types where the participant is indicating that they are sure the minutia is of the type selected. The second two, type uncertain and uncertain minutiae, are “uncertain” marker types where the participant is unsure of the type of the minutia, or uncertain whether the minutia is even present, respectively.



Overall, the 2 uncertain minutia markers (triangle and diamond) were seldom used. The mean number of uncertain minutiae annotated per image ranged from 1 to 5, whereas the mean number of certain minutiae annotated per image ranged from 1 to 25. The mean total number of minutiae annotated per image ranged from 4 to 26. The percentage of uncertain minutiae annotated per image ranged from 5 percent to 69 percent.

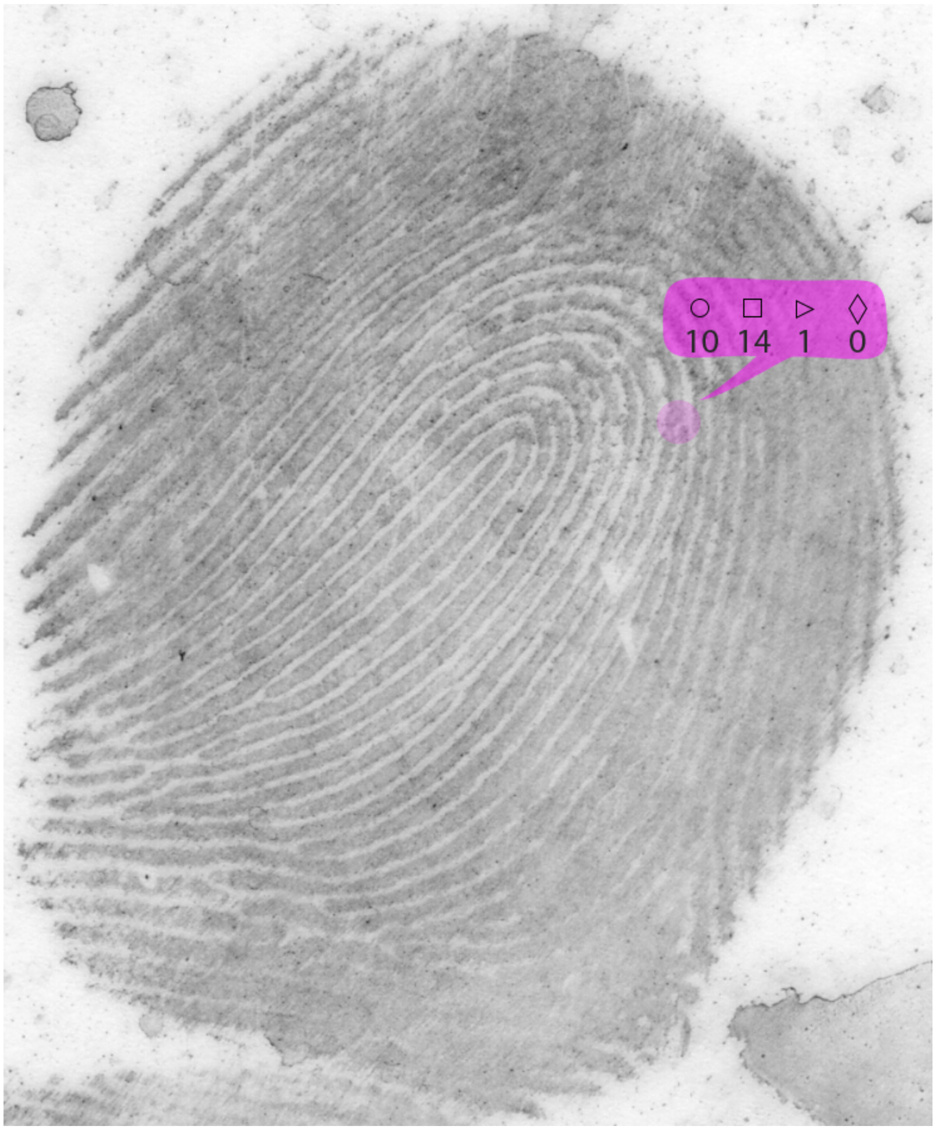
Despite the low usage of uncertain minutia marker types, there were many minutiae annotated with certain minutia markers that nonetheless lacked consensus on minutia type. In other words, many minutiae were annotated confidently as ridge endings by some participants whereas the same minutiae were annotated confidently as bifurcations by other participants. In many cases, these “votes” were very uneven, with the vast majority of participants choosing 1 minutia type whereas only a few chose the other. However, for many annotated minutiae, the votes were nearly even.

If approximately half of the experts viewing a particular minutia state that it is clearly a ridge ending, and approximately half of the experts viewing the same minutia state that it is clearly a bifurcation, there is *de facto* connective ambiguity (as defined by Stoney and Thornton (Stoney and Thornton 1986)) to the minutia. It does not necessarily follow that these ambiguities only occur in low quality marks, or in degraded local areas within otherwise clear marks. In many cases, there was a nearly even split of votes even in clear areas of high-quality images.

Figure 3.1 - 32 illustrates a minutia that was annotated in almost equal numbers as a certain ridge ending or a certain bifurcation. Only one participant marked it as an uncertain type. This image is a clear impression and was tied in the rankings as the image with the lowest percentage of uncertain minutiae annotated, at 5%. The consensus response on clarity for this image was “adds weight to my desire to keep the mark” and the consensus response on

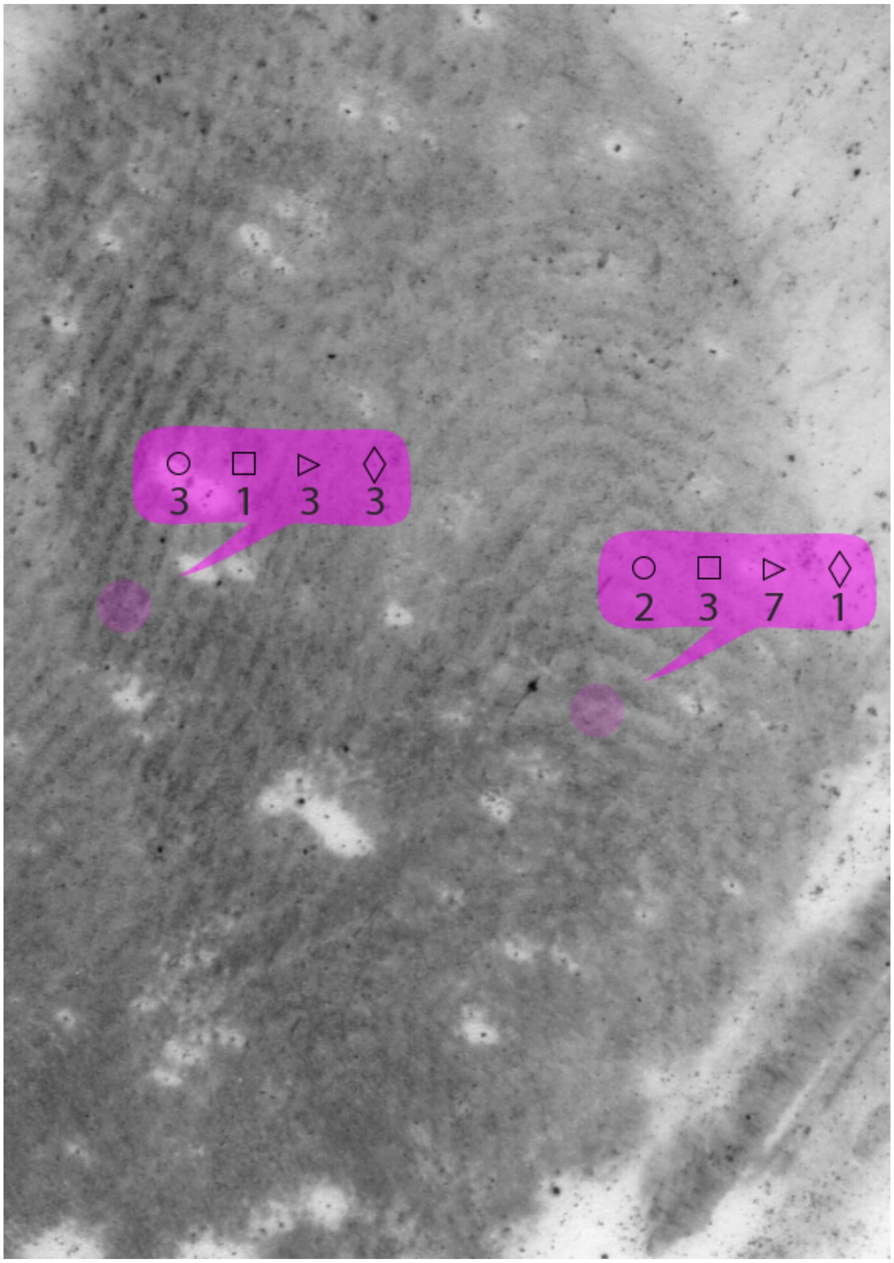
distortion was "Low Distortion." Although the overall clarity of the mark is high and even the local clarity of the noted minutia is high, there is nonetheless connective ambiguity around that minutia. It seems from multiple examples such as this one that when the area of the mark is clear, participants tend to express a high level of confidence in minutia type, even if there is connective ambiguity and a lack of consensus.

Figure 3.1 - 32 This is Image 61 from the study. The highlighted minutia is an example of a minutia in a clear local area that nonetheless suffers from connective ambiguity, as evidenced by the nearly even vote for ridge endings versus bifurcations. The text bubble shows the number of each minutia marker type that was used to annotate this minutia.



At the other end of the spectrum, Figure 3.1 - 33 illustrates the impression in the study that had the highest percentage of uncertain minutiae annotated, at 69%. The consensus response on clarity for this image was "adds weight to my desire to discard the mark" and the consensus response on distortion was "High Distortion." In this figure, we draw the reader's attention to two different minutiae that both represent the full range of possible minutia marker types. Participants who viewed this image appropriately used far more uncertain minutia types than they did on clearer images. Yet, with an image as highly degraded as this one, it could be argued that none of the minutiae should be marked with certainty as to its type.

Figure 3.1 - 33 This is Image 57 from the study. The two highlighted minutiae are examples of minutiae for which all four minutia marker types were used for a very degraded image. The text bubbles show the number of each minutia marker type that were used to annotate each minutia.



The lack of consensus on many minutiae annotations that were made with certainty indicates two things. First, there are no clear definitions to guide examiners on when a minutia is a ridge ending versus when it is a bifurcation. Second, it is highly probable that the uncertain minutia markers are being under-utilized. There is a debate needed regarding whether the minutia type actually *matters* for the comparison process to be completed reliably. We do not enter into this debate here. We do note that depending on the education and training of examiners, a different emphasis may be put on minutiae type. For instance, AFIS operators are often instructed to disregard the type of minutiae as it is not relevant in the encoding used by the systems. However, if practitioners cannot come to a consensus on what types of minutiae they are observing, we recommend that until clear definitions exist and are followed, distinct minutia types should not be recorded.

Use of new scale categorizations

Three new suitability scales (Complexity, AFIS, and Difficulty) were introduced in this research and participants were asked to make judgements on each. In addition, 3 new determinations[*] for use on the Value scale and one new determination[*] on the AFIS scale were introduced. Although there were specific reasons each of these new determinations were included in the research, it was uncertain whether they would be embraced and used by the participants in the study, or whether they would be summarily ignored.

<u>Investigative or Probative Value only</u>

The suitability determination "Some probative or investigative value, but insufficient for an identification or exclusion" was selected by at least one participant in 59 out of 100 images in the study. In 50 of the 59 images in which it was selected, more than one participant selected this option. The highest usage for this option was 9 participants selecting it for the same image. 74 unique participants selected this option at least once, and 49 of those used it more than once. The highest usage by a single participant was 10 times.

<u>Value for Identification only</u>

The suitability determination "Value for Identification only" was selected by at least one participant in 69 out of 100 images in the study. In 41 of the 69 images in which it was selected, more than one participant selected this option. The highest usage for this option was 13 participants selecting it for the same image. 48 unique participants selected this option at least once, and 31 of those used it more than once. The highest usage by a single participant was 18 times.

[*] These were: (1) Some probative or investigative value, but insufficient for an identification or exclusion; (2) Value for Identification only: and (3) Value for both Identification and Exclusion.

[*] AFIS quality with additional QA measures

Value for both Identification and Exclusion

The term Value for Both was technically new, but since it had been separated from "Value for Exclusion only" and "Value for Identification only," it was essentially equivalent to the "value for comparison" or "value for identification" terms currently in use in most laboratories. That is to say, it represents the standard, or common, usage of the term "of value." Because of this, this determination was very commonly used and specific numbers on its usage are not included.

AFIS Quality with QA

The AFIS scale determination "AFIS quality with additional QA measures" was selected by at least one participant in 85 out of 100 images in the study. In 79 of the 85 images in which it was selected, more than one participant selected this option. The highest usage for this option was 20 participants selecting it for the same image. 106 unique participants selected this option at least once, and 99 of those used it more than once. The highest usage by a single participant was 16 times.

Although these categories were new and unfamiliar to the participants, it seems that all were used more than sporadically. Shifts in thinking can take time as can new ways of expressing conclusions, but it does appear that there is value to thinking about suitability determinations as more than just binary, to designating marks upfront that should not be used to support categorical conclusions, and to considering whether it would be prudent to institute additional QA measures to guard against error for marginal AFIS impressions. It is encouraging that participants were willing to try out these new categories and we recommend examiners begin to incorporate these categories into their everyday casework as a way to think deliberately about risk and identify those marks that should be treated more cautiously. Further research in this area should investigate whether the use of these categories leads to a decrease in error rates in comparison decisions.

Conclusion

Research was undertaken to try to understand the variables that latent print examiners most consider when making suitability decisions and the extent of the variability observed between examiners when it comes to these decisions. The notion of "suitability" was also expanded, and a total of four scales of suitability (value, complexity, AFIS, and difficulty) were explored, along with several new conclusions that could be reached along each scale.

Examiners were found to be variable in the features they relied upon, their perceptions of amount of clarity and distortion, and their ultimate decisions regarding suitability. Consensus observations were fairly good predictors of consensus decisions; however, the variation in the data suggested that individual examiners would not agree with the consensus opinion in many cases. Although variability will unavoidably be present to some extent in human endeavors that rely upon subjective assessment of visual cues, the discipline should nonetheless make efforts to reduce this variability to the extent possible because differences in opinion over the suitability of a mark for comparison or AFIS could have real-world consequences to the criminal justice system. In addition, a reduction in variability around the complexity decision would allow for the logical application of common-sense QA procedures to reduce the chances of error.

Minutiae count was the strongest driver of value decisions, whereas clarity and distortion together better explained decisions on the other scales. Overall, the variables that were most consistently relied upon to reach suitability decisions along all four scales were: total number of minutiae marked, number of confident minutiae marked (as opposed to uncertain minutiae marked), the clarity of the image, the level of distortion in the image, whether the pattern type could be determined with confidence, and the selectivity of the minutiae.

Although examiners tended to agree on which images were very high quality, there was no consensus on no value images. Also, if an image was not extremely high quality, it was likely that many examiners would assign it to the highest value category, whereas many other examiners would disagree. Many images had nearly equal votes across all value categories.

Examiners tended to express strong confidence in minutia type, even when there was connective ambiguity. More degraded images resulted in a higher use of uncertain minutia markers, but certain minutia marker types were still used.

The new suitability categories that were introduced along the four scales of suitability were chosen often by participants in this study. There appears to be value in expanding the notion of "suitability" of latent marks and considering different uses for which a mark may be useful as well as considering more granular conclusion options that may suggest additional quality assurance measures.

In light of our findings, we make the following recommendations:

(1) Laboratories should consider adopting the 3 scales of Value, Complexity, and AFIS as ways of thinking about the suitability decision and should develop criteria for assigning marks to the categories in each;
(2) Researchers and Laboratories should consider the use of a standardized Difficulty scale to assist in uniformity between research projects, training, testing, and testimony. The thresholds for this scale will ideally need to be agreed upon by stakeholders prior to implementation in order to be useful across agencies and research groups;
(3) Quality managers should implement QA policies that reflect the position of marks along the Complexity scale, requiring enhanced documentation for complex marks and allowing reduced documentation on those that are so high quality as to be self-evident;
(4) Laboratory management should encourage the adoption of an "AFIS quality with additional QA measures" category to flag those marks that are suitable for entry into AFIS but should be subject to additional QA measures due to risk factors. This category should be implemented with specific criteria to define its thresholds;

(5) Examiners should not use specific minutiae marker types unless and until specific criteria are developed to define each because the current practice lends a misleading veneer of certainty to what is often an arbitrary decision; and

(6) Examiners should document the features and observations they relied upon to reach their suitability decision(s) because it is known that these decisions can vary widely and without a way to substantiate *why* a particular decision was reached, it is opaque and arbitrary.

Many of these suggestions will be easiest to implement once criteria are established for the thresholds of different categories, ideally based upon modeling of the consensus of experts. A project to provide proof-of-concept for such a model has been undertaken by this research group and the results of that work are presented in Chapter 4.

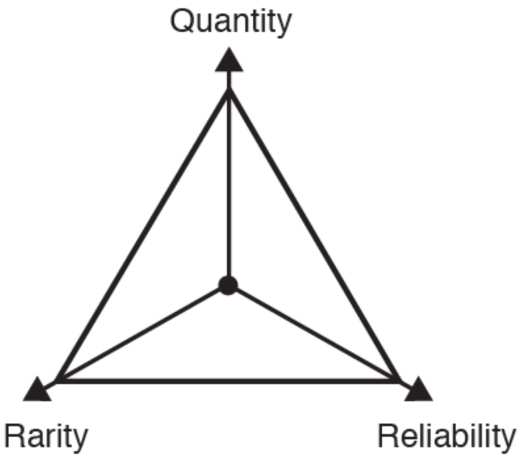
3.2 Additional insights and commentary on the white box study

This section first provides additional insight and explanation of the rationale behind the experimental design (i.e. why we asked the questions we asked), then goes on to examine additional images of interest that were not included in the article and provide the full conclusion results for each of the four scales (in the article, these were only summarized and a few examples given).

3.2.1 Evaluating suitability and the Sufficiency Triangle

In designing the white box experiment, choices had to be made regarding what information would be collected through the PiAnoS interface to capture the information examiners use in forming their suitability decisions. The guiding principles behind these choices were outlined above in Section 1.2 under "the assigning weight task." In order to assign weight to a feature, an examiner ought to be considering both how distinctive the feature is, and how confident they are in the feature's existence and type.

Figure 10 The sufficiency triangle. There must be a sufficient total area between quantity, rarity, and reliability in the features contained in an impression to support a suitability determination.

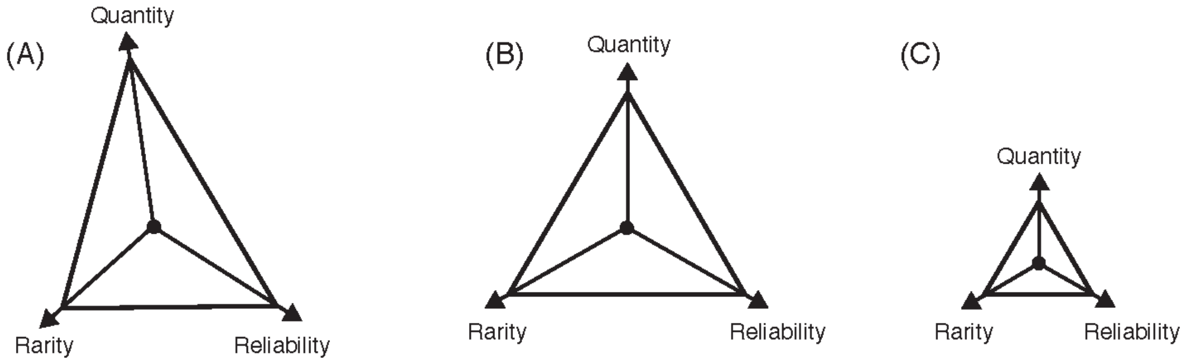


Although suitability is often thought of as predominantly related to quantity—i.e. is there *enough* information in the mark for it to be suitable—we propose that suitability is supported by three pieces of information: quantity, rarity, and reliability. We call this the "sufficiency triangle" (Figure 10) both because it deals with how much total weight is sufficient to reach a decision, and because it supports the concept of sufficiency generally (that is, both at the analysis (suitability) and evaluation (sufficiency) stages).
For a mark to be suitable, a specific number of minutiae is truly *not* sufficient justification, unless we are talking about numbers in excess of 30 or 40 minutiae. In addition to quantity, examiners should be considering both whether the features are distinctive (rarity), and whether they are confident in those features (reliability) as part of their weighting process. We already know that a few highly distinctive features can carry more evidential weight than several very common features. However, we also need to consider our confidence that those features are what we believe them to be. One cannot put heavy weight on a feature for being rare if it turns out the feature was actually not the rare feature they thought, but a common one, or even worse, if the feature is not actually present. Reliability is intrinsically linked to the clarity of the mark.

Thus, the sufficiency triangle can be thought of as having an ideal total area. If the triangle skews toward one of the three apexes, and away from one or more of the others, but the total area remains over the ideal threshold, the mark is still suitable. However, if any (or all) of the three apexes are so little represented that the total area of the triangle falls below the threshold, the mark is not suitable (Figure 11).

We tried to capture both the notions of distinctiveness and confidence for minutiae and pattern type. For minutiae, we approached this in two ways. First, we provided 4 different minutia marker types—two certain (ridge ending and bifurcation) and two uncertain (type unknown and uncertain minutia). These allowed us to capture the confidence level of the examiner of the type of minutiae and whether the examiner was even confident that the minutiae was actually there (uncertain minutia). Then, we provided the new grouping tool that allowed examiners to group together clusters of minutiae that they found to be particularly distinctive and would give extra weight. This addressed the distinctiveness question by allowing us to count the number of minutiae clusters examiners chose to group.

Figure 11 The sufficiency triangle. Triangles (A) and (B) have the same area. Although Triangle (B) has roughly equal quantity, rarity, and reliability whereas Triangle (A) has lower rarity and reliability, but high quantity, both meet the threshold for suitability. In contrast, Triangle (C) has roughly equal quantity, rarity, and reliability, but not enough of any of them for the mark to be suitable (hence its smaller total area).



For pattern type, we asked examiners to select all the pattern types they believed the mark could possibly be. The number of patterns chosen (Pcount) gave us an indication of their confidence in the pattern type. We also asked a separate Y/N question about the *shape* of the pattern and whether they found anything to be distinctive about it that would make it stand out from other patterns of the same classification. This allowed us to capture the perceived distinctiveness of the pattern.

For finer detail, like pores and incipient ridges, we combined the notions of distinctiveness and confidence into the single concept of weight. Because these fine details often do not reproduce reliably, because examiners have very different approaches on whether or not they rely on these types of detail at all, and because examiners are not consistent in their interpretations of whether incipient ridges and pores are Level 3 Detail or not, we took two approaches to capturing this information, and its importance to examiners.

First, we provided them with specific tools for marking both incipient ridges and pores, but we asked them *not* to annotate these features *unless* they factored into their analysis decision. For example, if they were like me, I don't compare pores so they don't make much difference to me in the suitability decision, other than as a general indicator of clarity. I would not mark pores. But some examiners would compare pores and get excited when they see them—they would be welcome to annotate the pores. Thus, rather than measuring whether examiners *noticed* the pores and incipient ridges, we were instead measuring the degree to which they *cared* about them. This was captured by the total number of each feature they annotated on each mark.

Second, we asked a general question about whether Level 3 Detail was noted, and relied upon, in reaching the suitability decisions. We asked this question without defining Level 3 Detail and left it to the examiner to determine whether the features they observed qualified. This gave us another indication of the extent to which Level 3 Detail (however defined) was important to examiners in reaching their suitability determinations.

Although the sufficiency triangle is presented here as a conceptual description of the thought process examiners should be going through to assign weight to features in a mark, and to the mark overall, it could be developed by future research into a usable model. As noted above, we collected data on quantity, rarity, and reliability of minutiae and pattern type. However, for things like level 3 detail, we combined rarity and reliability into a single concept of weight (which makes these data impossible to map onto the triangle). If numerical measures of perceived rarity and reliability were captured separately, in addition to quantity, for all data types, sufficiency triangles could be built for each mark as a visual representation of suitability.

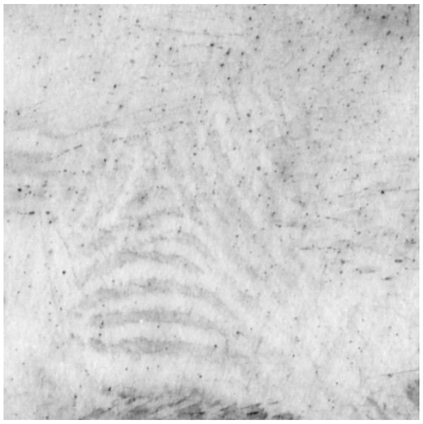
As it turned out (refer to Section 3.1), the only variable measuring distinctiveness and confidence that really contributed to predicting examiners' suitability decisions was the variable Minutiae Selectivity (MinSelect—more on this variable in Chapter 4). MinSelect was a variable that we created in order to synthesize the annotations we received into a measure of the weight given to the minutiae in a given mark overall. MinSelect took the proportion of certain (ridge ending or bifurcation) minutiae marker types to the total minutiae marked to estimate the examiners' confidence in their selected features overall. It then added a weighted factor for whether or not a target group was annotated and the number of combined groups annotated, to estimate the distinctiveness of the selected features according to the examiner.

From these results, it appears that minutiae (their quantity, distinctiveness, and reliability) were the main driving factor in decision-making. Although participants were provided with tools to annotate other kinds of information and were encouraged to do so, the predominant annotations we received were those relating to minutiae. This is not unexpected, considering that examiners tend to be accustomed to working in a minutiae-focused way, in which other features are considered, but the main focus is typically on minutiae.

3.2.2 Interesting case examples that highlight examiner variability

The article reproduced in Section 3.1 provides some examples of specific cases in the study that illuminate the lack of examiner consensus on particular value decisions and cases that illustrate people's inconsistent or incoherent choice of minutiae marker types. This section presents some additional examples that did not make it into the article but either illustrate additional specific areas of examiner inconsistency or have comments or annotations that prompt additional discussion.

Figure 12 img060, a mark presented in the white box study that generated commentary on the dangers of pattern force areas. This mark previously appeared in the thesis as Figure 4(A).



The first case involves img060 (Figure 12), a small delta area with few clearly discernible minutiae despite its relative clarity and few orientation/location cues. Thirty-seven participants viewed this mark, and their conclusions are summarized in Table 5.

Eleven of the 37 participants who viewed this mark left written comments and 5 of those specifically commented on the dangers of *pattern force areas*, and deltas in particular, as being at high risk of repetition of features. An additional 4 mentioned that the minutiae present in this image were not particularly distinctive (a more indirect reference to the possibility of repetition).

Table 5 Examiner suitability decisions for img060 across the four scales of suitability. Abbreviation Key: NV – no value; IV – Investigative Value; VEO – Value for exclusion only; VIDO – Value for ID only; VB – Value for both ID and exclusion; C – Complex; NC:DOC – Non-Complex, Document; NC:SE – Non-Complex, Self-Evident; NAQ – Not AFIS quality; AQ w/QA – AFIS quality with quality assurance measures; AQ – AFIS quality; H – High; M – Medium; L – Low.

Scale	**Value**					**Complexity**				**AFIS**			**Difficulty**		
Concl.	NV	IV	VEO	VIDO	VB	NV	C	NC:Doc	NC:SE	NAQ	AQ w/ QA	AQ	H	M	L
Votes	14	3	7	3	10	15	10	10	2	25	8	4	16	11	10

It is encouraging that so many of the participants took the trouble to note the dangers of pattern force areas and to apparently take them into account in their decision-making processes. Less encouraging is the fact that people who made these observations still reached different value decisions. Of the five participants who specifically noted the challenges of pattern force areas, 2 decided this mark was no value, 1 assigned it investigative value, and 2 determined it was of value for both identification and exclusion. Their comments, conclusions, and minutiae counts are summarized in Table 6.

Table 6 Conclusions, minutiae counts, and written comments from the 5 participants who noted the challenges that could be posed by pattern force areas upon viewing img060.

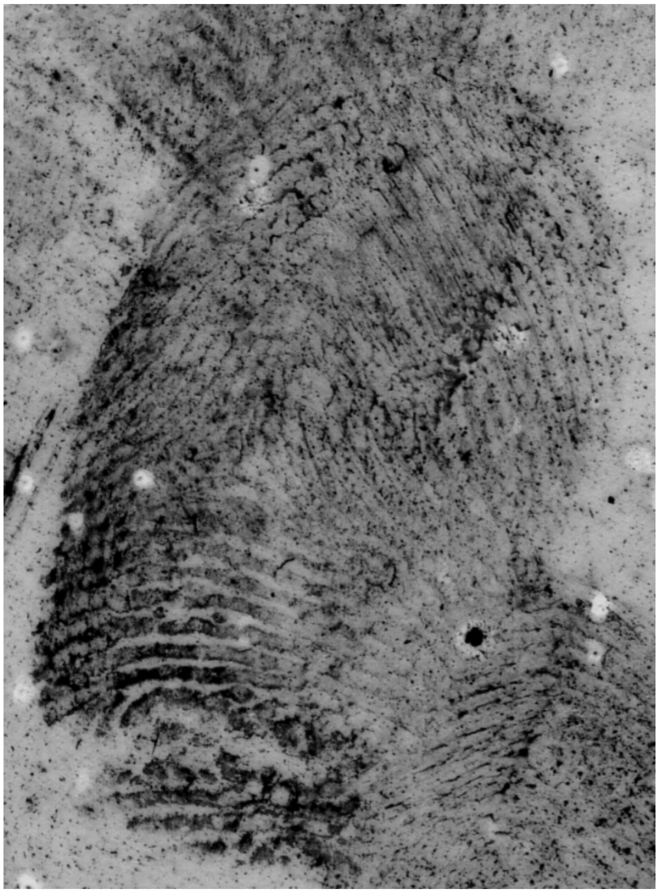
Username	Conclusion	Minutiae Count	Comment
User_024	NV	8	Impression is a tri-radius area. Whereas there are several minutiae, there is insufficient minutiae outside the delta area. Delta areas can have 'similar' formations.
User_148	NV	8	This mark is in a pattern force area. This makes the mark's features not unique.
User_034	Invest. Value	8	Forced areas, such as deltas, often have minutiae that could be valid, and could not. Caution when using only deltas.
User_111	Value for Both	15	unique ridge arrangements at center of delta. unable to anatomically orient. would consider sufficient for comparison – caution required if considering identification decision during evaluation due to pattern forced area and limited amount of information present
User_114	Value for Both	11	Complexity is due not to clarity, but to the limited amount of information present and the fact that the latent is in the delta area, where you are more likely to see similarities between knowns.

The fact that these 5 participants all had such similar comments yet chose 3 different conclusions (2 of which are diametrically opposed) highlights that there is still a need for clear criteria for the value decision. However, it is worth noting that the 2 participants who decided on Value for Both found more minutiae than the other 3, suggesting that total minutiae count is still a highly influential driver in this decision (although for User_114, 7 of the 11 annotated minutiae were uncertain).

The next case we will review involves img066 (Figure 13), a very messy mark with distortion, lateral pressure, tonal reversal, and one very small, very clear area with two dots. This image makes an interesting case study because it is such an awful mark generally, but the clear information it has is very valuable, if present in a compared print. To see how participants reacted to this mark,

Table 7 summarizes the conclusions of the 28 participants who viewed it.

Figure 13 img066, a mark presented in the white box study that provokes thought about the weight that can be assigned to highly distinctive features in a small area of clarity.



The minutiae count for this mark ranged from 1 to 27, whereas the minutiae count for those who gave a value decision of NV ranged from 1 to 8 and those who gave a value decision of Value for Both ranged from 6 to 27. Only 6 participants left written notes, and 3 of these mentioned the 2 dots at the bottom of the impression. The decisions of those 3 participants were Investigative Value (5 minutiae annotated), VEO (4), and Value for Both (12). Once again, although there are interesting things going on in this mark and participants are taking note of them, the ultimate decision appears to be driven largely by perceived number of minutiae suggesting that although distinctiveness is valued by examiners, it does not trump minutiae count in their minds.

Table 7 Examiner suitability decisions for img066 across the four scales of suitability.

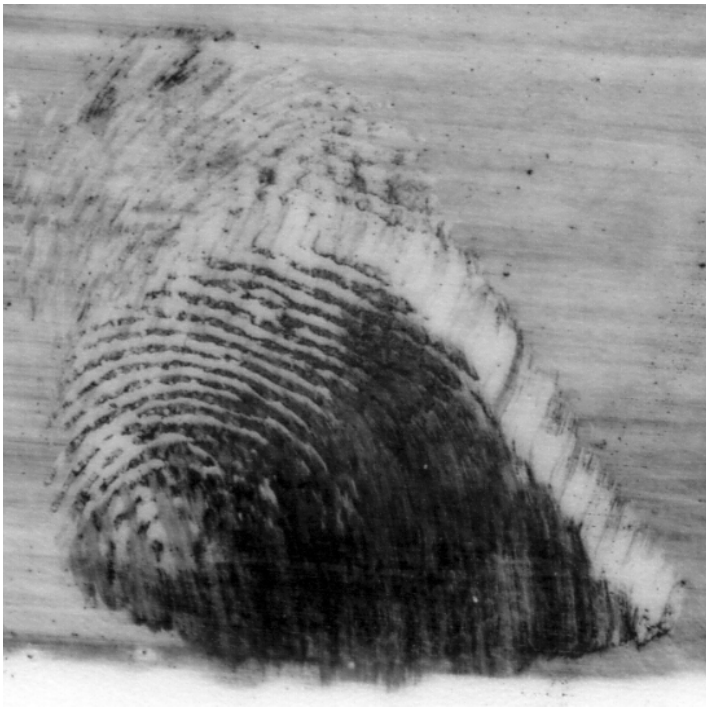
Scale	**Value**					**Complexity**				**AFIS**			**Difficulty**		
Concl.	NV	IV	VEO	VIDO	VB	NV	C	NC:Doc	NC:SE	NAQ	AQ w/ QA	AQ	H	M	L
Votes	12	4	4	2	6	13	13	2	0	21	6	1	21	7	0

Finally, we will examine img068 (Figure 14), a tip with few discernible minutiae, but a distinctive group near the top and a possible dot at about the 11 o'clock position. 28 participants viewed this mark and minutiae counts ranged from 5 to 13. However, only 3 participants annotated fewer than 8 minutiae, making this mark fairly stable in terms of minutiae count. Table 8 summarizes the conclusions of the 28 participants who viewed img068.

With minutiae counts fairly stable, decision-making on this mark appears to have been largely driven by clarity and distortion. Most people rated the distortion of this mark as "medium," but the 5 who rated it as "high" all reached value determinations in the bottom 2 categories. Furthermore, all 5 participants who reached a NV value decision chose "clarity increases my desire to discard the mark" as their clarity assessment. Although clarity assessments were mixed in the higher value categories, there was a marked increase in "clarity increases my desire to keep the mark" assessments.

Interestingly, despite the distinctive cluster at the top of the mark and the possible dot at 11 o'clock, these didn't seem to resonate strongly with participants, at least according to their annotations. Five participants annotated the dot (3 VIDO, 2 VB) and an additional 1 mentioned it in their written notes. Every participant marked all or part of the cluster in the tip, but there was no discernible relationship between value decision and whether the cluster was annotated using the grouping tool, so this cluster did not appear to be driving the value decision, other than insofar as it contributed to the overall minutiae count and was located in a clear portion of the image.

Figure 14 img068, a mark presented in the white box study that had a fairly stable minutiae count and owed more to clarity and distortion for decision-making. Like img066, it presents a small number of highly distinctive minutiae.



One thing that *was* mentioned by participants on this image was Level 3 Detail in the form of ridge shapes and pores. Eight participants marked pores in this image, with 5 of them marking more than 10 and 2 marking 30 or more. Six participants left written notes on this image and 3 of them mentioned pores specifically whereas 2 mentioned L3D or ridge shapes. Interestingly, one participant selected NV and made the comment, "Despite the number of minutiae [8 annotated by this participant] and the presence of L3D, I would not keep this mark. There are plenty of pores visible [30], but I don't find them helpful for identification. This is the type of mark that false positives are made of."

Table 8 Examiner suitability decisions for img068 across the four scales of suitability.

Scale	**Value**					**Complexity**				**AFIS**			**Difficulty**		
Concl.	NV	IV	VEO	VIDO	VB	NV	C	NC:Doc	NC:SE	NAQ	AQ w/ QA	AQ	H	M	L
Votes	5	4	2	7	10	5	11	12	0	17	4	7	9	15	4

Together, these additional examples illustrate how participants behave when confronted with particular types of limited-information marks. In cases with a wide range of perceived minutiae counts, participant decisions seemed to be largely driven by the minutiae count. When the minutiae count was more stable, clarity and distortion assessments seemed to have a greater influence. Some participants are aware of both dangers (pattern force areas) and high-value features (dots) and were thinking about fine details such as ridge shapes; however, these did not seem to drive the decision for most of the participants—at least to the extent that they faithfully recorded their thought processes and annotated the features upon which they relied.

It is clear that although minutiae count is a major driving factor in value determinations, examiners *are* noticing and considering a wide range of other information. It is likely that these other factors come into play when decisions are near a threshold, but we were unable to demonstrate this through the machine learning, which relied on average responses and thus lost the nuances of individual observations.

It is further clear that until and unless specific criteria are developed to (1) guide examiners in what information to weigh heavily during analysis and (2) specify an objective level of clarity at which a minutia "counts," we will continue to see wide variability in both what information is considered/weighted and in ultimate suitability determinations.

3.2.3 Summary of all responses along the four scales

The white box article reproduced in Section 3.1 presented examples (Figure 3.1 - 24) in the section titled "Consistency of suitability decisions" of the range of decisions examiners made for some of the marks in order to illustrate the levels of consensus at the high and low ends of the spectra.

In this section, we present the full conclusion data for each of the four scales, showing histograms of all the responses received for each of the 100 images in the study along each of the four scales. These are presented in Figure 15 through Figure 18 and a glance at their overall shape gives an overview of the level of agreement and disagreement between examiners. In these figures, it is easy to observe that there is rarely consensus on decisions, except at the high ends of the scales, and that there is frequently almost even distribution between opinions.

Figure 15 Responses received for all 100 images along the value scale.

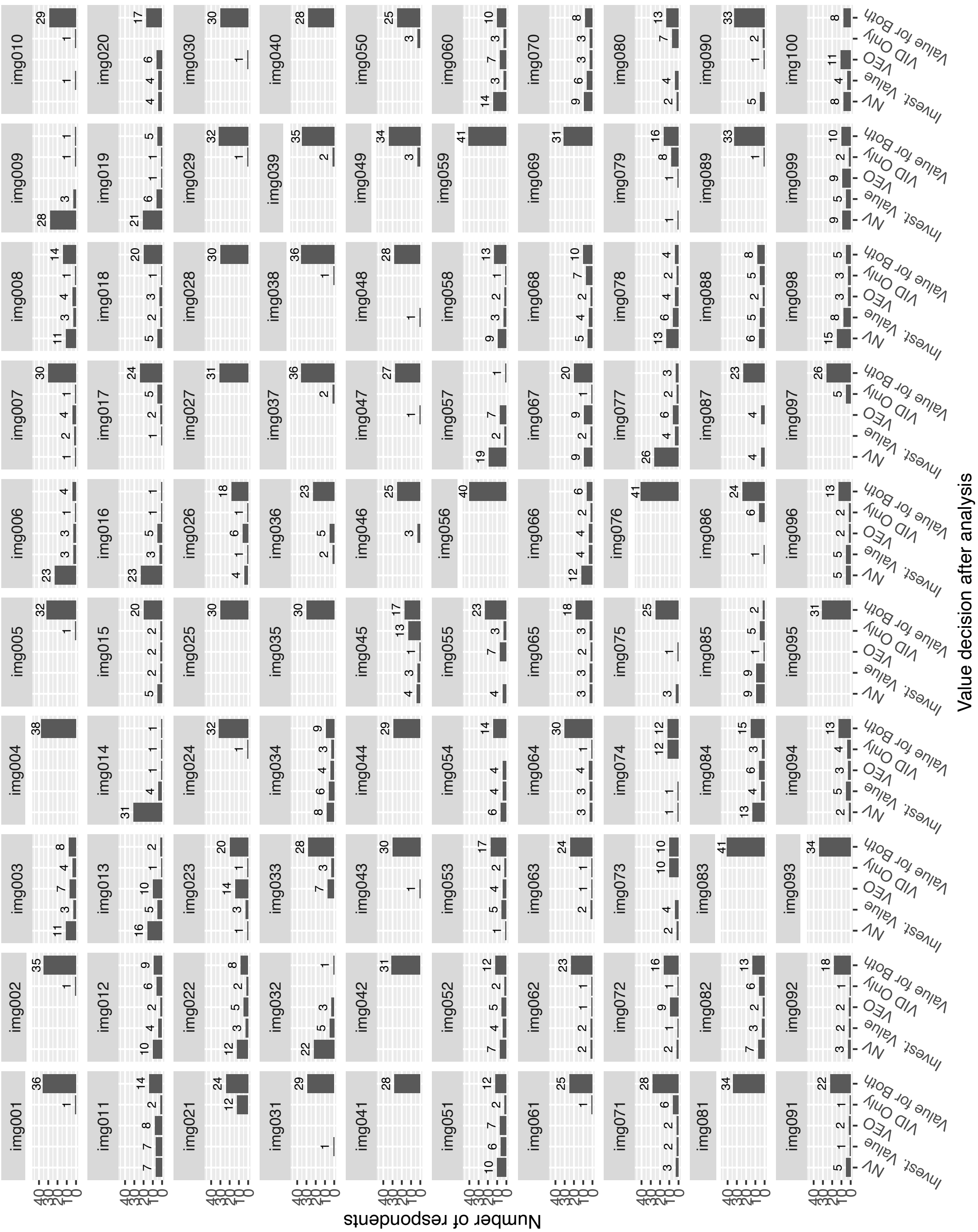


Figure 16 Responses received for all 100 images along the complexity scale.

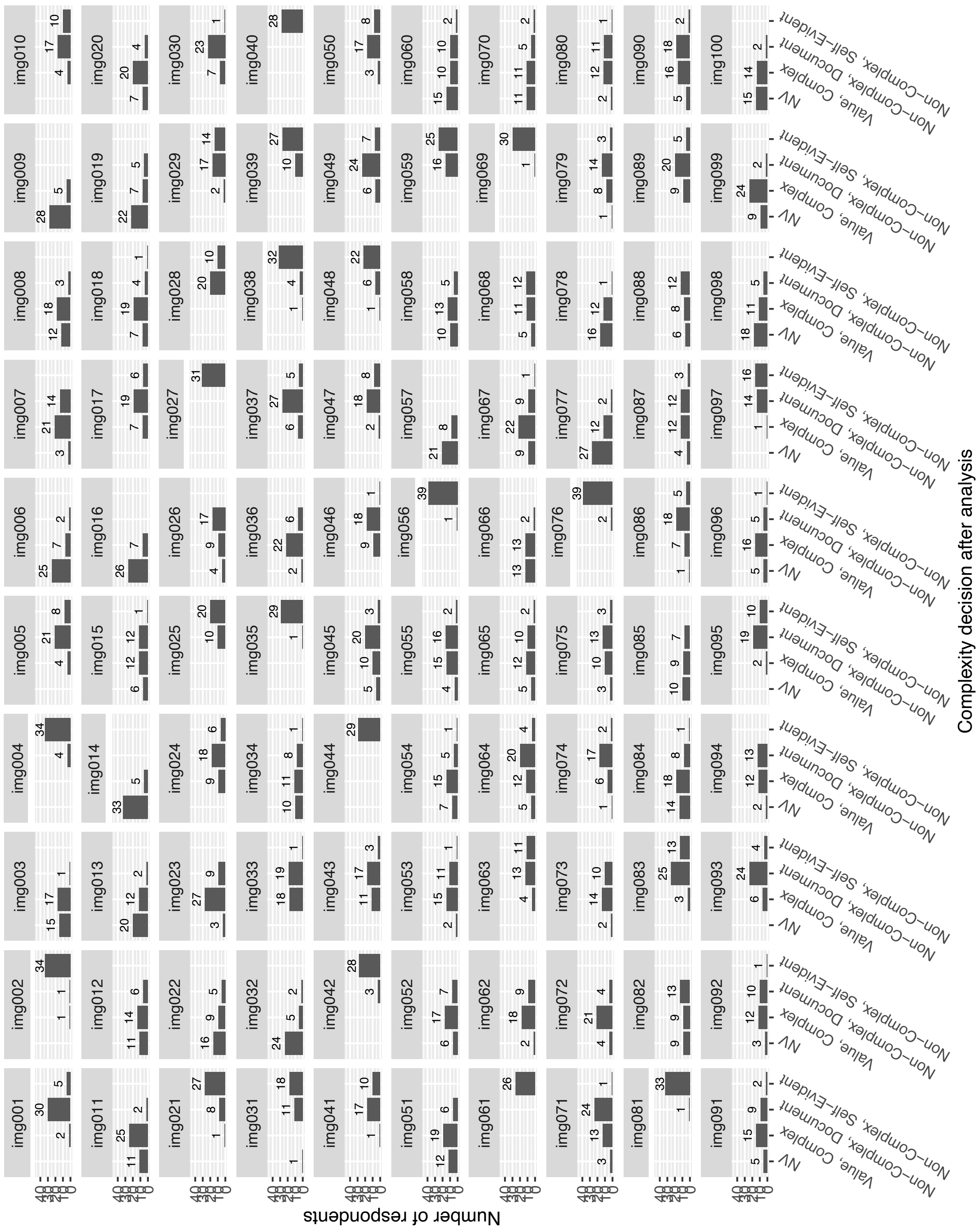


Figure 17 Responses received for all 100 images along the AFIS scale.

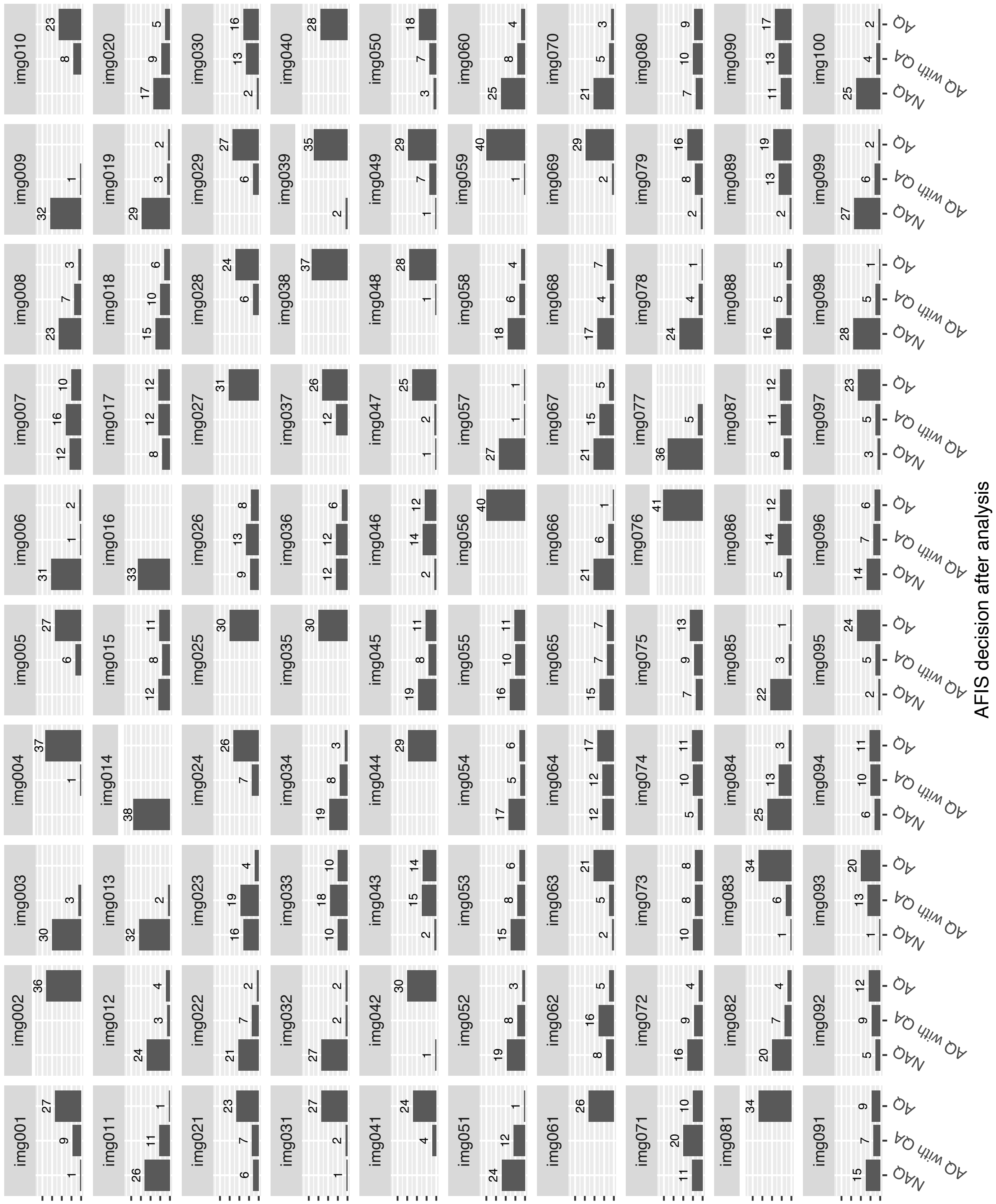
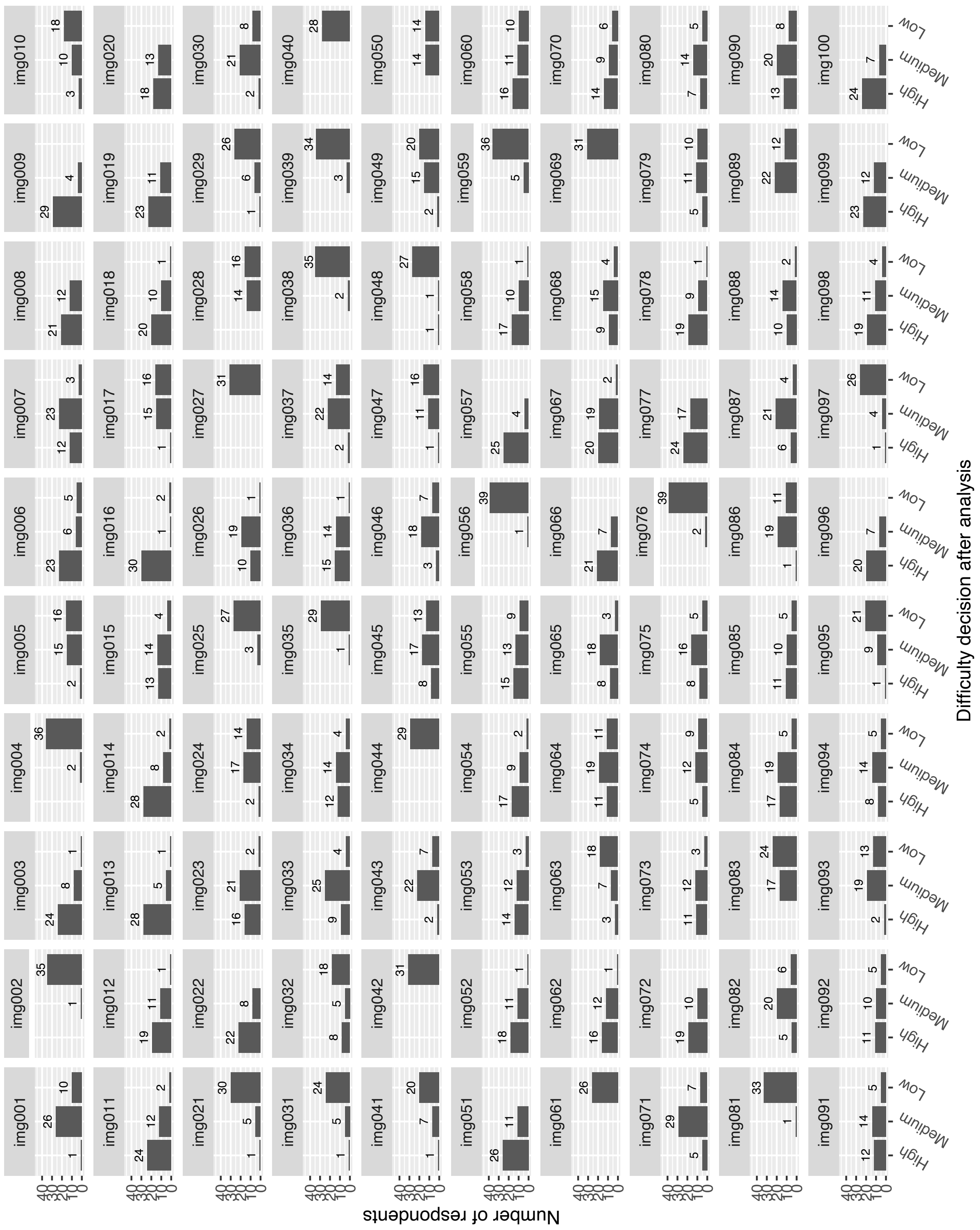


Figure 18 Responses received for all 100 images along the difficulty scale.



3.3 Chapter 3 summary

Chapter 3 presented a published white box article that explored a four-scale conceptualization of suitability and the information that examiners rely upon to support decisions along each of those four scales (Value, Complexity, AFIS, and Difficulty) and dove deeper into the philosophy of suitability and the rationale behind many of the choices made in the experimental design for the white box study.

Examiners are variable at every key decision point in the friction ridge comparison process, including the selection of features and the ultimate suitability decision. We sought to link examiner annotations of the information upon which they relied to their own analysis decisions on each of the four scales to determine the most predictive information. (These four scales were previously described in detail in Sections 2.2 And 3.1).

Data were collected from 116 friction ridge examiners who were each assigned sets of 30 mark images from a pool of 100, resulting in 3,241 completed trials. Participants were requested to annotate only the information they used to reach their conclusions and to answer questions about their perceptions of clarity, distortion, pattern type, and level 3 detail in the marks.

As found in previous literature, minutiae count was the best overall predictor of decisions on the Value scale, whereas distortion and clarity together were better predictors of decisions on the Complexity, AFIS, and Difficulty scales. The most diagnostic variables for predicting decisions were: total number of minutiae, total number of confident minutiae, clarity, distortion, number of possible pattern types, and minutiae selectivity, a measure that incorporated both the ratio of confident minutiae to total minutiae annotated and counts of target groups and distinctive minutiae clusters.

The concept of the Sufficiency Triangle was introduced. The Sufficiency Triangle is a visual way of representing that 3 factors must be considered to support a decision of suitability or sufficiency: quantity of information, rarity of that information, and reliability of that information. Although one or more of these factors may be lower or higher than the others for any given mark, the total perceived contribution of all 3 factors must result in a triangle with a large area in order for a mark to be considered suitable.

The experimental design of the white box study captured information on quantity, rarity, and reliability of minutiae and pattern type directly, and of other types of observable features (such as creases, scars, or level 3 detail) more indirectly. The strong performance of the minutiae selectivity variable as a predictor of decisions (a variable that took into account all three factors of quantity, rarity, and reliability) supports that all three are important. It also supports that minutiae (their quantity, rarity, and reliability together – not just their quantity alone) are the most influential driver of suitability decisions.

Image examples presented in the chapter illustrate that in addition to minutiae, participants did consider other types of information when reaching their suitability decisions, but that these factors were not as influential as minutiae, except in cases where the minutiae count was fairly stable between examiners, in which cases other factors seemed to have more influence.

A review of the overall levels of agreement between examiners on all four scales revealed that there was generally good agreement on high quality images, but not on low- or mid-quality images, and that there were many images where decisions were nearly evenly split across all possible categories, demonstrating that stringent decision criteria are needed to reduce variability between examiners.

Minutiae marker types were often used with certainty, even in areas of low clarity or when there was clear connective ambiguity. Additionally, different minutiae marker types were frequently used for the same minutiae by different examiners, both with high confidence, illustrating that the assignment of minutiae type (ridge ending versus bifurcation) is largely arbitrary and should not be presented as a decision that has been made with confidence until very clear criteria are in place for each type.

Two new suitability scales (Complexity and Difficulty) and several new conclusions along all four scales were introduced. Usage of these new conclusions was good across examiners and indicates that examiners found them useful and their adoption in operational casework should be considered.

4 Predicting consensus suitability decisions

The data collected in Chapter 3 were further analyzed to develop a predictive model that could categorize marks along each of the four suitability scales, based upon a few key observations from the examiner and some automated measures. Chapter 4 describes the process of developing, optimizing, and validating this model. In the second paper concerning this research, "Predicting suitability of finger marks using machine learning techniques and examiner annotations," we present the work that was done to create this model. Section 4.1 presents the manuscript of that paper, published in *Forensic Science International*, here with figure and table captions prefixed with "4.1 –" to integrate with this dissertation. Note that footnotes have been re-named to simple asterisks. Figures may appear in different locations in the manuscript due to journal formatting.

For length and to appeal broadly to the friction ridge examiner community, the process of model development, optimization, and validation was summarized in the article, but some additional discussion of why particular research design choices were made was omitted. Section 4.2 of this chapter goes a little further behind the scenes of this process.

4.1 Predicting suitability of finger marks using machine learning techniques and examiner annotations

Predicting suitability of finger marks using machine learning techniques and examiner annotations

Heidi Eldridge, MSc[a, b], Marco DeDonno, MSc[b], Christophe Champod, PhD[b]
[a]RTI International, 3040 E. Cornwallis Rd., Research Triangle Park, NC, 27709 USA

[b]School of Criminal Justice, Faculty of Law Criminal Justice and Public Administration. University of Lausanne, 1015 Dorigny, Switzerland

Abstract

Previous research has established the variability of examiners in reaching suitability determinations for friction ridge comparisons. Attempts to create predictive models to assist in this determination have been made, but have been largely confined to fully automated processes that focus on suitability for AFIS entry. This work develops, optimizes, and validates a hybrid predictive model that utilizes both examiner-observed variables and automated measures of quality and rarity to arrive at suitability classifications along four scales that have been proposed in our previous research: Value, Complexity, AFIS, and Difficulty. We show that a model based only on automatically extracted quality or selectivity measures does not perform as well as when used in conjunction with a limited set of user inputs. The model is then based on a limited set of input from the users while taking advantage of automatic measures with a view to limit the user encoding effort while maintaining accuracy. The developed model is able to make predictions at up to 83.13% accuracy when using full study data and maintains similar levels of accuracy in an external validation study. The model achieved accuracy at a similar level to that of examiners asked to make the same suitability determinations across all scales. The model can easily be introduced into an operational laboratory with very little additional operational burden to provide guidance on suitability, complexity, AFIS, and quality assurance decisions; to assist in designing testing and training exercises of progressive difficulty; to describe the difficulty of a mark in testimony; and to provide a consensus-based opinion in laboratories where a second opinion is desired but the laboratory lacks sufficient personnel to form a consensus panel.

Keywords

latent prints, friction ridge, value determination, identification

Introduction

Latent print examiner variability in the suitability* decision has been well-documented (Ulery et al. 2011, 2012; Ulery et al. 2013; Ulery et al. 2014, 2015, 2016; Langenburg 2012; Neumann et al. 2013; Pacheco et al. 2014; Eldridge et al. 2020).

Inter- and intra-examiner variability have been observed, both in the selection of features used to support the decision and in the ultimate determination regarding suitability. Practically speaking, these differences in opinion necessarily lead to an uneven application of justice-the same mark may be compared, or not, depending on which examiner evaluates the mark and even on which day the evaluation occurs.

The decision of whether or not to compare a particular mark could have a grave impact on a case, whether by failing to compare a mark that could have either implicated the accused or implicated a different party, potentially exculpating the accused; or by performing a comparison on an unreliable mark that could lead to a false identification or exclusion.

Beyond the decision of whether or not to compare a given mark, we have argued (Eldridge et al. 2020) that the notion of "suitability" is actually multi-faceted, as a mark can be more or less suitable for a variety of applications, and we have proposed four scales of suitability: Value, Complexity, AFIS, and Difficulty. Each of these scales provides information that may be useful to the forensic laboratory, researcher, or criminal justice system by providing guidance on when to compare a mark, enter it into AFIS, or apply enhanced or reduced Quality Assurance (QA) measures; or to use in the design of training materials, proficiency tests, and research samples of known difficulty; or to present information on difficulty level to a fact-finder in court. In that previous work, we also have described the variability of examiners in the decisions they reach along these four suitability scales and measured how well their annotated observations predict their own individual suitability decisions.

In this paper, we describe the second part of this study, in which we develop and validate a model using examiner observations in combination with automated measures of quality and rarity to predict the consensus suitability response along each of the four scales.

This model requires the examiner to enter only an image of the mark and three key observations (number of minutiae, and a global assessment of both clarity and distortion), then provides guidance on the suitability of the mark according to what a consensus of experts would likely support.

* Although the terms "suitability" and "sufficiency" are often used interchangeably in the friction ridge community, this paper recognizes a distinction between the two that will be maintained throughout. "Suitability" refers to the decision that is reached at the end of the Analysis phase-whether the unknown mark is suitable for some particular purpose, most often comparison. "Sufficiency" refers to the decision that is reached at the end of the Evaluation phase-whether there is sufficient information present in two impressions to support a particular source conclusion.

Study Description and Methods

The data used in this study were obtained in our previous work and the methods used to obtain them are fully described in our previous publication (Eldridge et al. 2020). However, we offer here a brief summary of the methods used to obtain the data. 100 study marks were selected to represent a range of quality from a pool of 1,259 casework images. 116 latent print examiners completed 3,241 trials in which they completed analysis of these 100 images following specific instructions we provided and resulting in suitability decisions along the four scales referenced above. Participants were requested to annotate only the information they considered in forming their suitability determinations, not all the information they could discern. Participants were provided with tools to annotate minutiae type and location, pores, and incipient ridges; indicate confidence in their minutiae markings; and select target groups and other highly distinctive minutiae groupings. They were further asked questions about possible pattern types, and asked to assess the level 3 detail, clarity, and distortion present in each mark. Finally, they completed a survey that collected demographic information and information on their policies and practices. All these responses, annotations, and suitability decisions made up the dataset used for the model development described in this article.

This study was completed in two parts: *Model Development and Optimization* and *External Validation*. In the first part, the best variables to predict consensus decisions and the best performing machine learning algorithm (MLA) were selected. The selected MLA was then optimized by first testing its performance on the consensus data set of participants who agreed with consensus ground truth to get an idea of the model's best possible performance then testing it again using the full data set of all participants to see how the model could be expected to perform in a more real-world situation where not all examiners are likely to agree with the consensus. In the second part of the study, the model was externally validated using a newly recruited set of examiners and a mix of new and old images to test for generalizability of the results.

Model Development and Optimization

Machine learning and subsequent statistical analysis were carried out in R version 3.6.3 RC (2020-02-21 r77847) (R Core Team 2018) coupled with RStudio Version 1.2.5033 (RStudio Team 2015) using the following packages: *tidyverse* (Wickham et al. 2019) for data wrangling, *caret* for machine learning, computing confusion matrices and associated error statistics (Kuhn 2020, https://CRAN.R-project.org/package=caret). *caret* was also used to investigate variable importance and to reduce their number by adopting recursive feature elimination (RFE) (Kuhn and Johnson 2013).

To develop the predictive model, we determined both which MLA would provide the best performance and also which predictors were most diagnostic in predicting the consensus ground truth decisions.* There are many MLAs available for these kinds of analyses and each

* Because there is no objectively true answer for the question of whether a mark is suitable for a particular use, we used a majority voting system to assign the "ground truth" expectation for each of the 100 study marks along each of the 4 scales.

has strengths and weaknesses, with some being more or less computationally expensive and some being more or less opaque in terms of interpretability. To select the best MLA for our needs, we evaluated the performance of a suite of commonly used MLAs.

We considered Classification and Regression Trees (CART), Random Forest (RF), K-nearest neighbors (KNN), Neural Networks (NN), Average Neural Networks (AvNN), Gradient Boosting Machine (GBM), C5.0, Support Vector Machine (SVM), Principal Components Analysis Nearest Neighbor (PCANN), multinomial regression (MULTINOM), and XGBoost Linear (XGB). The associated R packages giving access to these classifiers were all loaded when required through *caret* (Kuhn 2020, https://CRAN.R-project.org/package=caret).

All MLAs have been trained adopting a leave-one-out validation scheme.

When needed we will present the performance of the models using ROC curves taking advantage of the MultiROC package (Wei and Wang 2018) based on (Van Asch 2013). The MutliROC analysis allows construction of ROC curves for classification problems with more than 2 outcomes.

We trained the potential MLAs using only the data from users who agreed with the consensus ground truth decision for any given mark/scale combination. This was done because the model could not be expected to learn how to accurately predict an outcome based on annotations made by people who did not agree with that outcome. For each mark/scale combination (e.g., image 88 on the Complexity scale), the predictor responses of the concurring participants were averaged to create a single dataset of the consensus observations that could be used to predict the consensus decisions. These were referred to as "average user data." In our model development strategy, we wanted to train a model based on the predictors obtained from an ideal user who would make decisions along with the majority vote (our ground truth by proxy) and assessed each predictor in line with the average among all examiners considered.

At this stage we selected the MLA based on its accuracy. Then, we selected the best set of predictors that were most diagnostic to take into Step Two. This was done by recursive feature elimination (RFE), using model accuracy as the deciding metric.

In Step Two, we tested different combinations of variables using the average user data to test the best-case scenario of how that model could be expected to perform using data where every user agreed with the consensus. The combinations of variables are all based on sets of predictors selected after RFE, but included different scenarios such as only automatically-computed predictors (such as quality metrics); or only user-input predictors, from minimal input to full input of all predictors considered; or combinations of the two. Step Two allowed us to assess the potential loss in terms of accuracy when adopting a less computer intensive or less user intensive approach in the establishment of the predictors.

Finally, in Step Three, we tested the same combinations of variables with the selected MLA using all individual user data to represent a more messy, real-world scenario and used these results to select the variable set that would be used in the final model due to its optimized performance using variables that were operationally feasible.

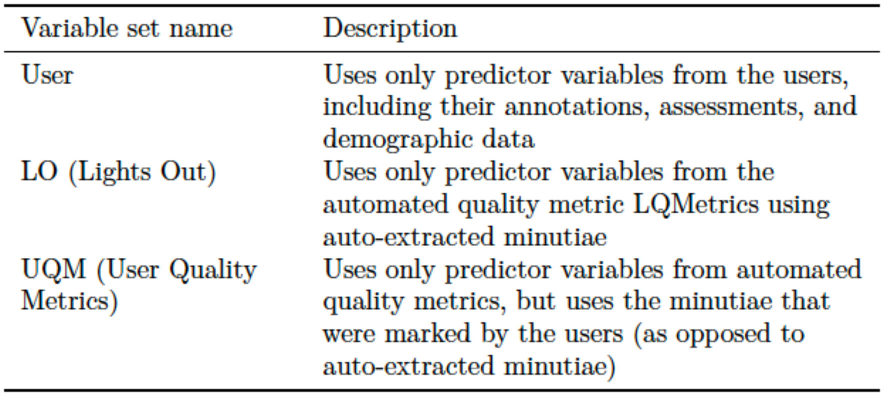
For this step, average user data was no longer used. The actual user variables from each user who agreed with the consensus ground truth was used along with the user input variables from all the users who did not agree with the consensus ground truth. It was expected that the model performance would be noticeably lower once noisier data were included compared to the ideal benchmark point set following Step Two.

Step One – Model Selection

The average user data for all user variables were run in all tested MLAs using 10-fold cross-validation with 50 repeats. This allowed us to directly compare the performance of multiple MLAs on the data while minimizing the risk of overfitting.

Each potential MLA was tested on each of the four scales using six different variable set combinations, resulting in 24 total assessments of the suite of potential MLAs. Each of these variable sets represents a different scenario for model use. One set considers all available predictors, whereas the others make some operational choices between variable sets. One question this research was designed to address was whether a human examiner brought any value to the analysis of marks for suitability, or whether it would be better to use a fully automated process. This step tested that question by evaluating variable sets that were fully human, fully automated, and some combinations of the two. The combinations of variable sets that were tested were: User, User/LO, User/UQM, UQM/LO, LO, and ALL (User/UQM/LO). LO is a term used to refer to a process that is completely automated, without any user input, a.k.a "lights out". Table 4.1 - 1 summarizes the variable sets that were tested.

Table 4.1 - 1 Description of the variable sets



Variable set name	Description
User	Uses only predictor variables from the users, including their annotations, assessments, and demographic data
LO (Lights Out)	Uses only predictor variables from the automated quality metric LQMetrics using auto-extracted minutiae
UQM (User Quality Metrics)	Uses only predictor variables from automated quality metrics, but uses the minutiae that were marked by the users (as opposed to auto-extracted minutiae)

After selecting an MLA, we used RFE analysis to determine which predictor variables achieved the highest possible model accuracy with the fewest number of variables by considering the variables of highest importance for each combination of scale and variable set then determining which were the most diagnostic overall. This was desirable because one goal of the model is to save analysis time for examiners by requiring them to enter only a few key observations in order to receive guidance, rather than having to do an exhaustive analysis of all information present.

User variables that were tested for their predictive value included number of minutiae annotated, their type, and the confidence associated with them; the number of target groups annotated and the minutiae that made them up; the number of distinctive minutiae clusters annotated; users' assessment of level 3 detail, clarity, and distortion; users' assessment of pattern type; and annotations of the presence of creases, scars, incipient ridges, or pores. In addition to user variables related to each impression, we also tested a number of user demographic variables to see whether they influenced examiners' suitability determinations.

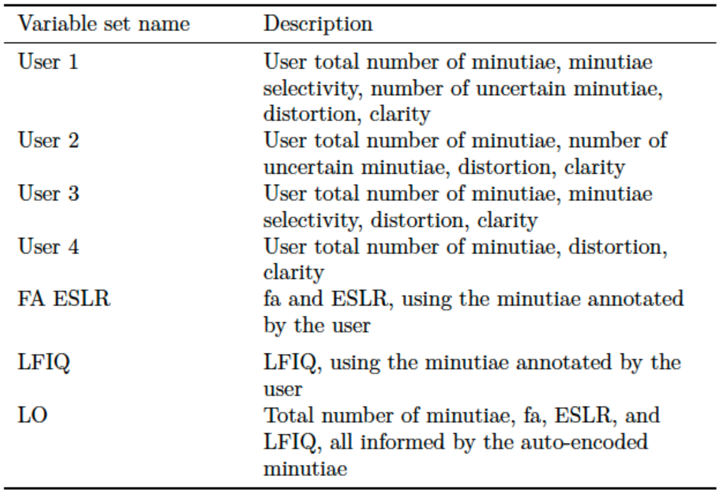
LO variables considered as predictors included automated measures of quality, clarity, and good ridge flow from LQMetrics (Hicklin et al. 2013); fa, which is a variable calculated based upon the AFIS matching score;* and ESLR, which is an Estimated Score-Based Likelihood Ratio based upon the mark alone that predicts how strong a match could be made *if* an appropriate same source exemplar was provided (Stoney et al. 2020).

UQM variables considered as predictors included the total number of minutiae marked by the examiner, the numbers of minutiae marked by the examiners in areas of different quality, and fa and ESLR based only upon the minutiae entered by the examiner.

Step Two – Ideal Performance Testing

Once the MLA (it turned out to be Random Forest) and predictor variables that provided the best overall solution were selected, we optimized the model using a multiple step process. In the first step (Step Two), we took the 12 selected predictor variables and combined them into variable sets, listed in Table 4.1 - 2. As in Step One, these represented variables from the users only, fully automated variables, and combinations of the two. These variable sets were then tested in 21 different combinations and the performance of each combination was compared to the others to measure any significant differences in their accuracy. A Random Forest (RF) cross-validation model was used with 2,000 trees, 100 repeats, and 10 folds.

Table 4.1 - 2 Description of the variable sets retained after RFE.



Variable set name	Description
User 1	User total number of minutiae, minutiae selectivity, number of uncertain minutiae, distortion, clarity
User 2	User total number of minutiae, number of uncertain minutiae, distortion, clarity
User 3	User total number of minutiae, minutiae selectivity, distortion, clarity
User 4	User total number of minutiae, distortion, clarity
FA ESLR	fa and ESLR, using the minutiae annotated by the user
LFIQ	LFIQ, using the minutiae annotated by the user
LO	Total number of minutiae, fa, ESLR, and LFIQ, all informed by the auto-encoded minutiae

* The fa variable is calculated as follows, with s being the AFIS score for the transaction:

$$fa = \frac{s - 1500}{500} - 1$$

The 12 predictive variables that were selected (represented in variable sets User 1, FA_ESLR, and LFIQ in Table 4.1 - 2) were chosen after Step One to give the best compromise across all four scales between accuracy of classification and number of variables. Our aim was to reduce the number of user variables to the extent possible to minimize the amount of time examiners needed to spend on annotations, particularly those that did not have a significant impact on the accuracy of the classifier.

Minutiae selectivity was measured by the variable "minselectivity". This was a variable that we created in order to synthesize the annotations we received in the white box study into a measure of the weight given to the minutiae in a given mark overall. It took the proportion of certain (ridge ending or bifurcation) minutiae marker types to the total minutiae marked to estimate the examiners' confidence in their selected features overall. It then added a weighted factor for whether or not a target group was annotated and the number of combined groups annotated, to estimate the distinctiveness of the selected features according to the examiner.

Each combination of variable sets was evaluated for its ability in terms of accuracy to predict the consensus ground truth using the chosen MLA and the average user data.

This gave an estimate of both which combination of limited variables provided the highest predictive accuracy and of how well the final model might be able to perform under idealized circumstances-that is, when users of the model agreed with the consensus ground truth. The results of this step represent the situation where we have superusers (those who always agree with the consensus determination), unlimited access to computing resources, and variables that have been selected while maintaining accuracy. We refer to this situation as Optimal. We also explore how the accuracy of the model would vary when less rich sets of predictors are used with a view to reducing the computing burden or the burden put on the examiners' annotations. We refer to this situation as Operational. The operational solution is then a compromise between computing resources (to obtain the quality measures), the required user input, and the accuracy obtained by the model.

Step Three – Model Optimization in realistic applications

Each of the 21 variable set combinations tested in Step Two was again tested along each of the four suitability scales in 3 different ways: how well each potential model predicted ground truth outcomes based on the inputs of each user per image, how well each predicted ground truth outcomes based on the inputs of each user per user, and how well each potential model performed compared to how well the examiners performed at choosing the consensus ground truth response. Accuracy of the models was measured using a leave-one-out cross-validation over the 100 images.

The performance of each potential model at predicting ground truth per user was selected as the basis for choosing a final model because this is the situation that most closely resembles the real world-if an examiner in a laboratory was looking to the model to assist them in making decisions on a case, what they care about is how well the model will work for them as an individual user, not how well it works on a particular case.

These per user data were examined to determine the best performing variable set combination under two circumstances: Optimal and Operational. The Optimal performance reflects how well the model could be expected to perform in the real world, where data is messy, but assuming unlimited resources of computational power and time. The Operational performance reflects how well the model could be expected to perform in the real world, but limiting the variables available to those that can be incorporated into the model without huge computational or development costs. The best Operational model is the one that was ultimately selected and taken forward to the final step for external validation.

External Validation

Once the model was optimized, we performed a final experiment to validate it externally. Although the model predicts the consensus suitability decisions for the 100 images used and the examiners involved in the study, and although we did use a cross-validation scheme to ensure we did not overfit the model to the data, we still wanted to verify that its performance was generalizable to new images and new examiners who were using the model directly, as opposed to analyzing data that had been selected down to only the model variables.

To do this, we created a new version of PiAnoS that had limited functionality to only allow the three most predictive user variables to be recorded. Examiners were recruited to participate, and 51 images were randomly selected for the study. To compare the model's performance with new examiners on new images against its performance with new examiners on images from the original study, we re-used 20 images from the original study and incorporated 31 new images. This design helped to ensure that any difference observed in model performance from the optimization phase was a true difference and not a consequence of using images that had not been used in the model's development.

The study participants were a self-selected convenience sample of latent print examiners who had previously participated in one or more of the Principal Investigator's (PI) studies, or who had previously expressed interest in doing so. They were solicited via an email list and responded to a confidential liaison to enroll, who assigned them an anonymous username. All participants completed informed consent, which had been reviewed and approved by RTI International's Institutional Review Board, prior to beginning participation.

Participants were alternately assigned to one of two groups as they enrolled-a ground truth defining control group (GT) and an experimental group (Exp). Because not all enrolled participants completed the trials, we ended up with different numbers of completions in each of the two groups. 39 examiners completed all trials in the GT group, and 43 in the Exp group.

The two groups had slightly different workflows. Each group was presented with all 51 images, one at a time. The GT group was asked to mark the minutiae they would use to reach a suitability decision, then answer two questions to describe their perception of the level of clarity and distortion present in the impression. They were then asked to provide a suitability decision on each of the four scales as described in our previous study (Eldridge et al. 2020). Once the four suitability decisions had been rendered, the model combined the entered data with the three most predictive automated assessments of quality and rarity (which had been pre-calculated for each image) to return the model's prediction of the consensus response for each of the four scales. Participants were then asked whether or not they agreed with the model's assessment, and if not, were required to enter a reason why they did not agree in a free text box. The Exp group followed the same workflow with the exception that they did not enter initial suitability determinations but moved straight from minutiae marking and answering the clarity and distortion questions to receiving the model's predictions and commenting on them.

The GT group was asked to provide their four suitability determinations because there are 31 new images presented in this experiment that were not used during the original study. Thus, those images did not yet have a ground truth by consensus designation established. We needed the opinions of these examiners to define the suitability ground truth for each of the four scales so it could be used during data analysis to evaluate the performance of the model. These decisions were recorded prior to the GT participants receiving the model's predictions so that the participants would not be influenced in their decisions by knowing what the model predicted.

Resampling simulations from the acquired data were used on each of the four scales to determine how many participants were needed in the GT group to ensure with a high likelihood that the consensus ground truth would be "accurate" in the sense that even in a worst-case scenario (i.e. the votes for the different possible conclusions were nearly even), the consensus decision for this group would be the "correct" decision. With 20 examiners in the GT group (which we nearly doubled), all four scales were expected to achieve over 90% accuracy.

Results and Discussion

Model Development and Optimization

Model development and optimization was accomplished in three broad steps. In the first, an MLA was selected and the best predictor variables were identified. In the second, the selected MLA was tested using only the user data from users who agreed with the consensus ground truth (average user data) under several combinations of variables to see which combinations had the best predictive accuracy under idealized conditions. Finally, the selected MLA was tested using the same variable combinations but user data from all the users to see how the model was likely to perform under more real-world conditions with noisy data and the final model, incorporating operationally affordable variables, was selected for external validation.

Step One – Model Selection

A suite of 11 MLAs was trained on each of the four suitability scales using six different combinations of variable sets (see Table 4.1 - 1). This resulted in 24 full sets of results, summarized in Figure 4.1 -1 to Figure 4.1 -4. Because each of the four scales was intended to measure different things, it was not expected that the same MLA would necessarily exhibit the best performance under each of the 24 combinations of predictors and scale. Nonetheless, as

Figure 4.1 -1 to Figure 4.1 -4 show, the performance of RF is the most consistently at or near the top across all four scales and nearly all variable set combinations. SVM, XGB, and GBM also performed well overall, but they were less consistently in the top 3 and are also more computationally expensive and opaque.

The LO variable set was always by far the worst performer, thus model selection decisions were not made based on LO results. We decided to select Random Forest (RF) as the model of choice for its achieved accuracy and known robustness. Even when RF is not the top performer for a particular scale/variable set combination, its performance is always very close to that of the top performer. In cases where RF was well below the top three performers (such as the UQM_LO variable set on the AFIS scale), the spread between top performer and RF was only the matter of a couple percentage points of accuracy. Thus, RF was selected to be the most flexible solution capable of producing good results across all four scales.

Another possible solution to selecting an MLA would have been to select the best performing MLA for each of the four scales individually, and had there been great differences in performance, this would have been done. However, it was decided that there was no great loss in accuracy and operational gain in simplicity in adopting RF as the MLA of choice for all four scales.

Figure 4.1 - 1 shows that, for the Value scale, many MLAs achieved 100% accuracy at predicting consensus value decisions. During previous machine learning work on these data (Eldridge et al. 2020), it became apparent that the MLAs could not well handle a five-option Value scale. This was because, although some examiners embraced the use of the three middle conclusion options, they were not used often enough for the consensus opinion to be one of these middle options in very many cases. The Value scale consensus ground truth results for the 100 images in the study are shown in Table 4.1 - 3. With so few examples of what a ground truth Investigative Value only, Value for Exclusion Only, or Value for ID Only impression should look like, the MLAs did not have enough information on which to accurately predict these outcomes. Thus, the decision was made to remove these four cases from the Value scale only and collapse the predictive options to Value or No Value. Once this was done, the prediction became a relatively easy task for the MLAs.

Figure 4.1 - 1 Value scale: Accuracy achieved by each MLA with indications of the 95% confidence interval.

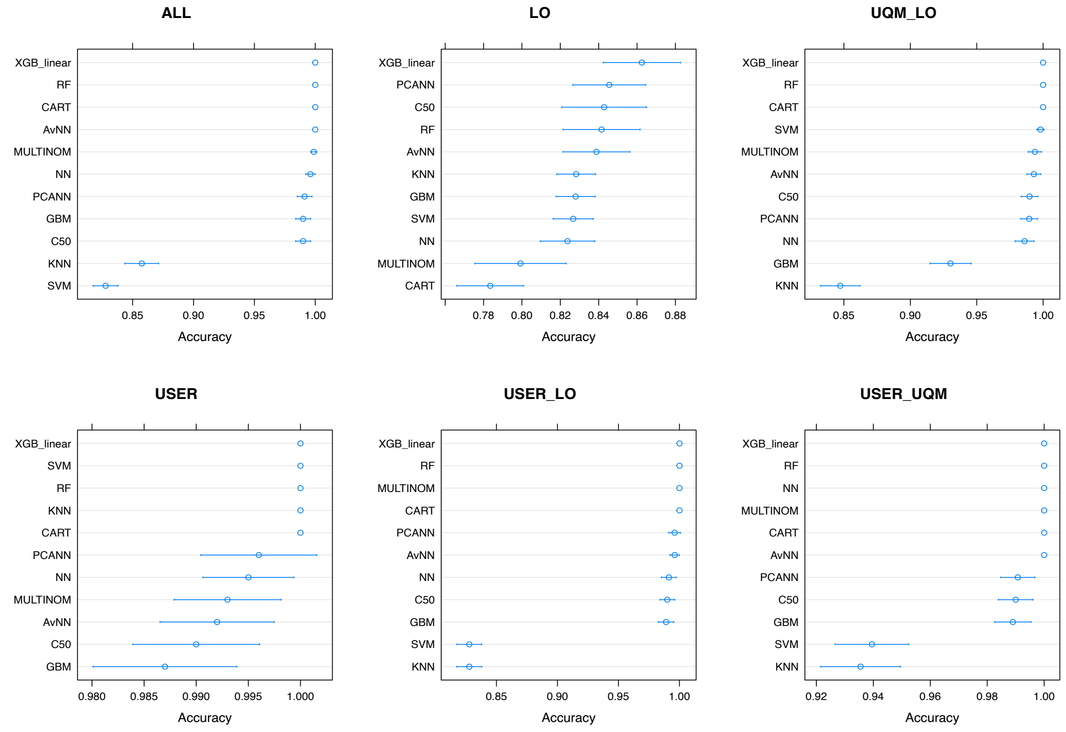


Figure 4.1 - 2 Difficulty scale: Accuracy achieved by each MLA with indications of the 95% confidence interval.

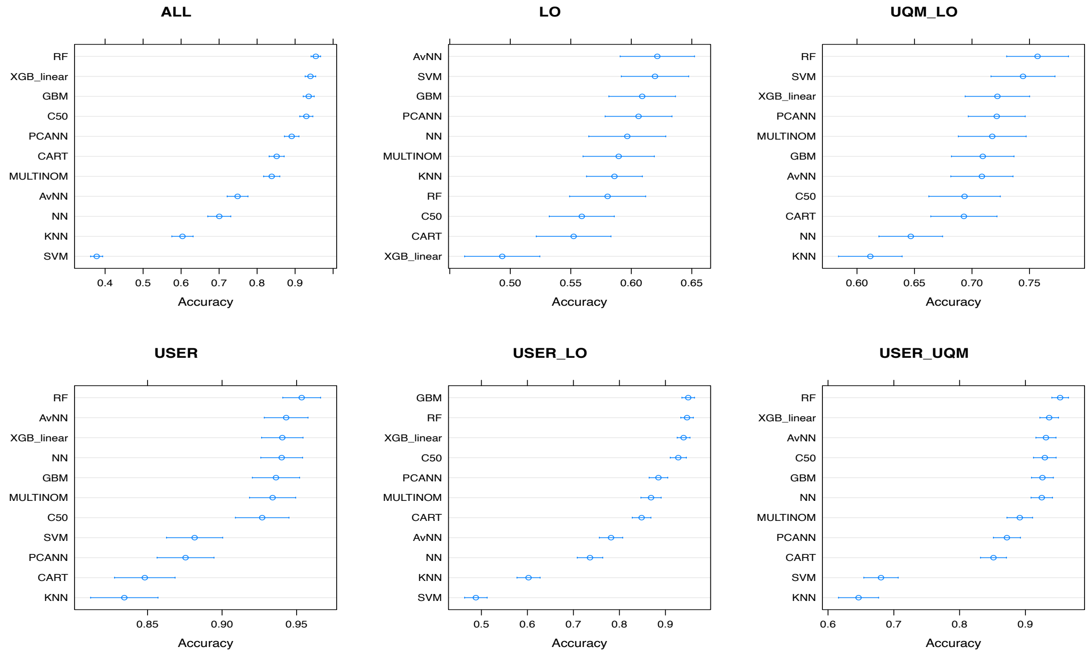


Figure 4.1 - 3 Complexity scale: Accuracy achieved by each MLA with indications of the 95% confidence interval.

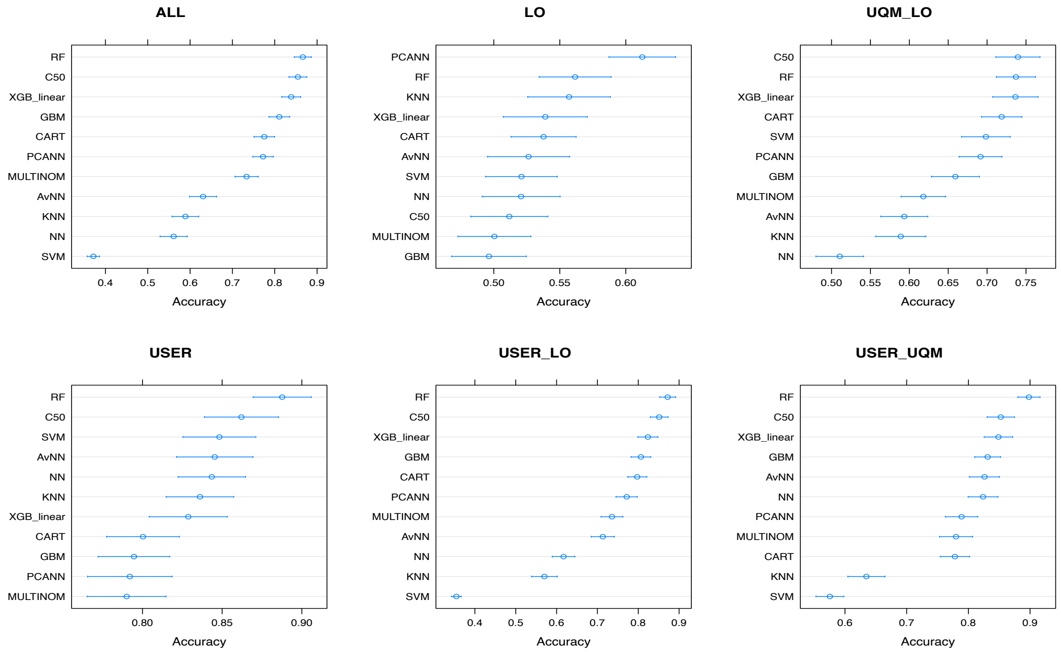
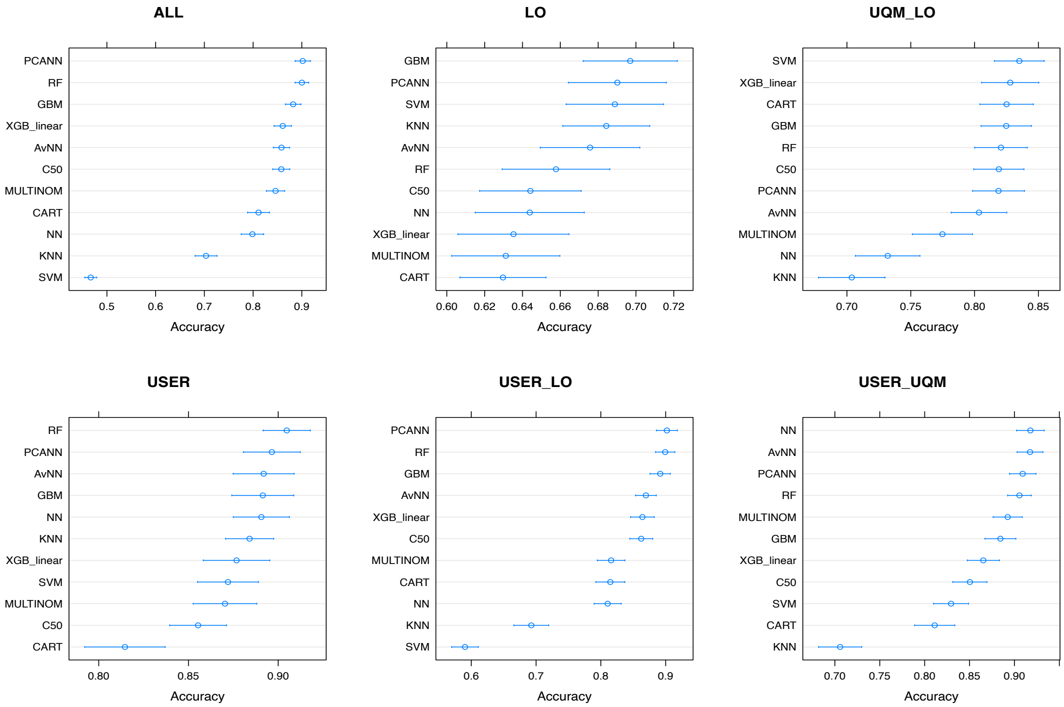


Figure 4.1 - 4 AFIS scale: Accuracy achieved by each MLA with indications of the 95% confidence interval.



Once RF had been chosen as the MLA, it was necessary to select the predictors that best contributed to the accuracy of the model. We used RFE analysis to select a limited set of predictors while maintaining accuracy. One output of the RFE process is a ranked list of the variables of importance. Figure 4.1 - 5 to Figure 4.1 - 8 present the top variables of importance for each variable set on each of the four scales with an indication of the achieved accuracy and the number of predictors (n_Preds) retained after RFE.

Table 4.1 - 3 Consensus value decisions on the Value scale for each of the 100 study images.

NV	Invest.Value	VEO	VID	Value for Both
17	1	1	2	79

Figure 4.1 - 5 Value scale: Variable importance from the most impacting (100%) to the least impacting for each set of variables.

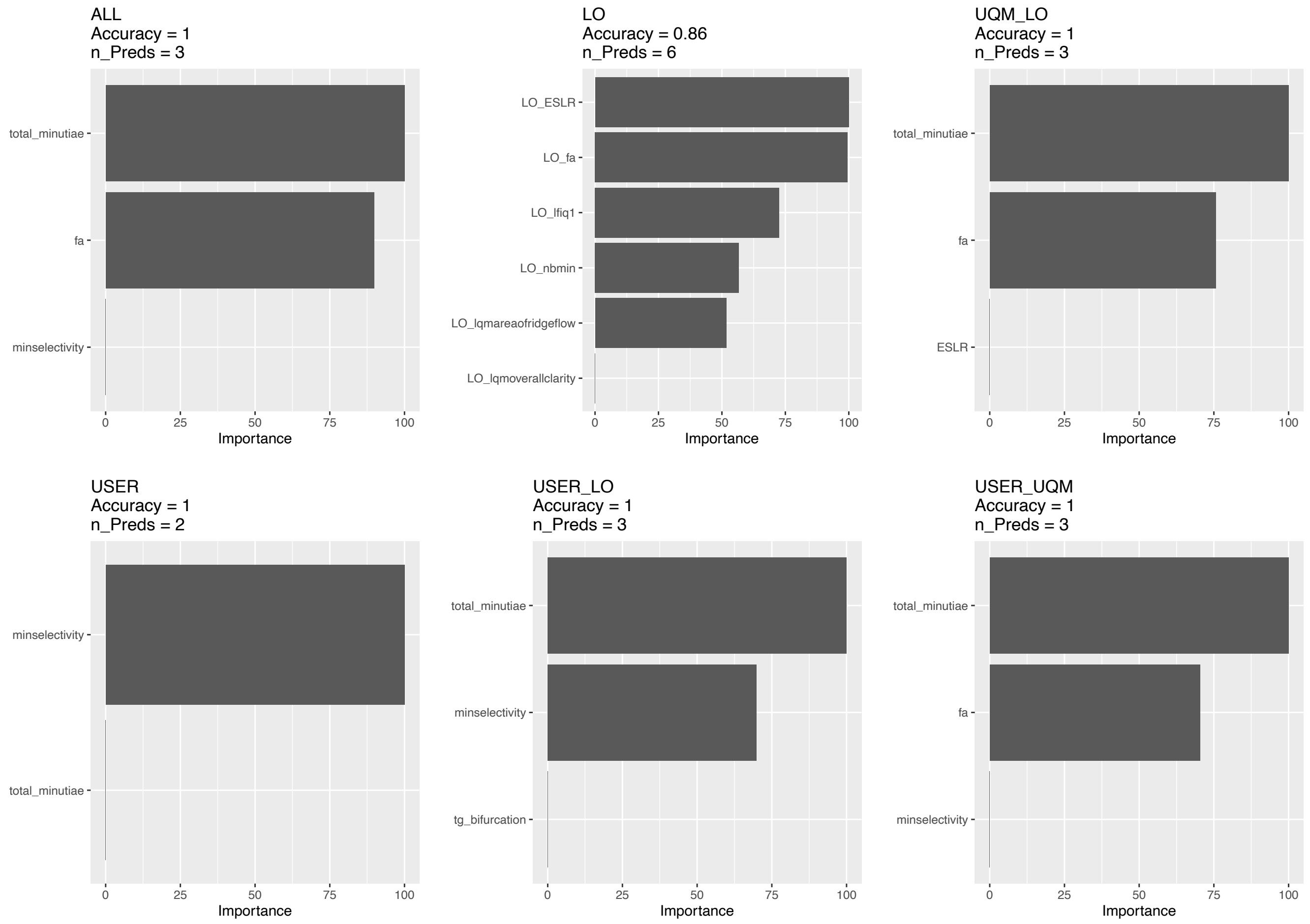


Figure 4.1 - 6 Difficulty scale: Variable importance from the most impacting (100%) to the least impacting for each set of variables.

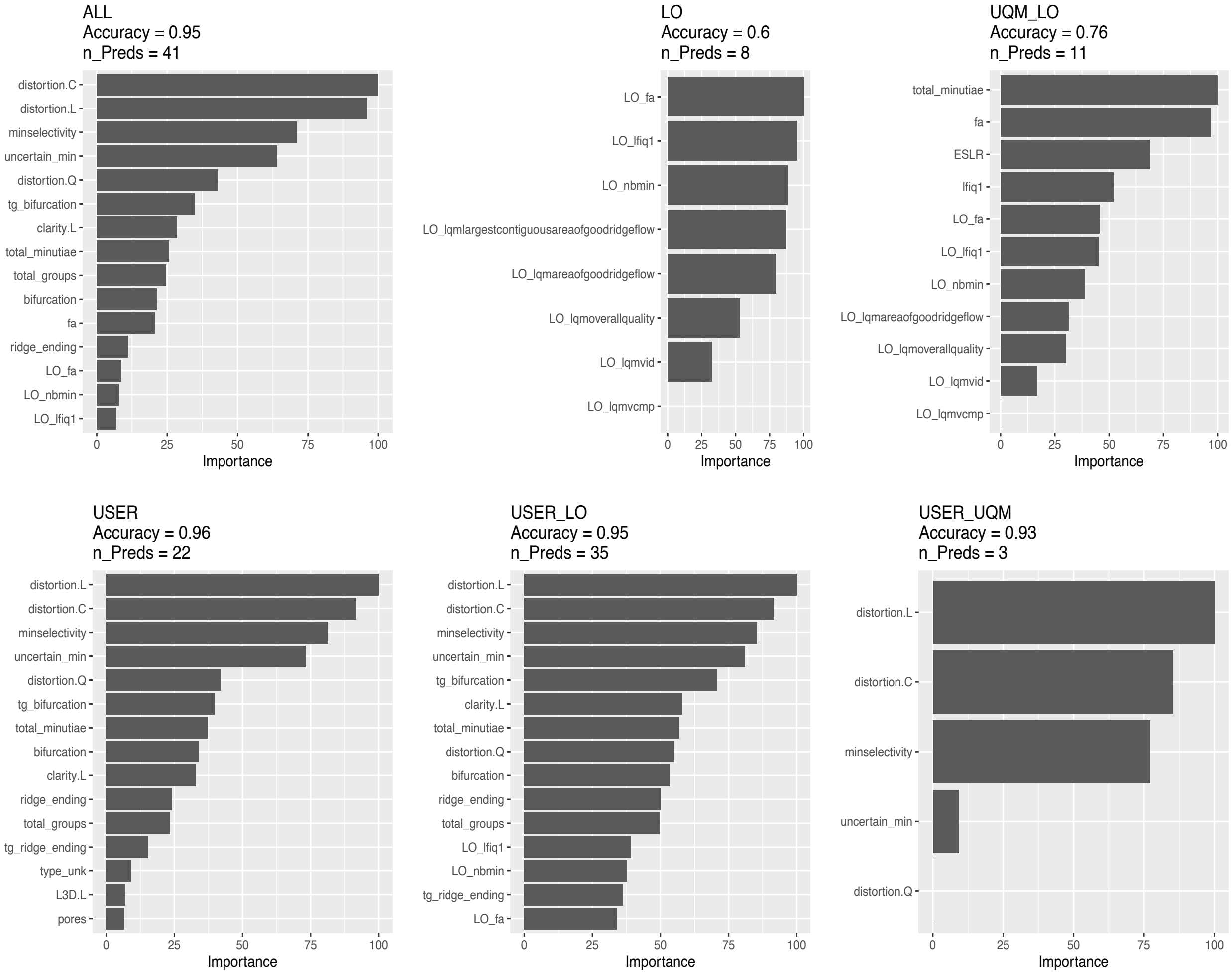


Figure 4.1 - 7 Complexity scale: Variable importance from the most impacting (100%) to the least impacting for each set of variables.

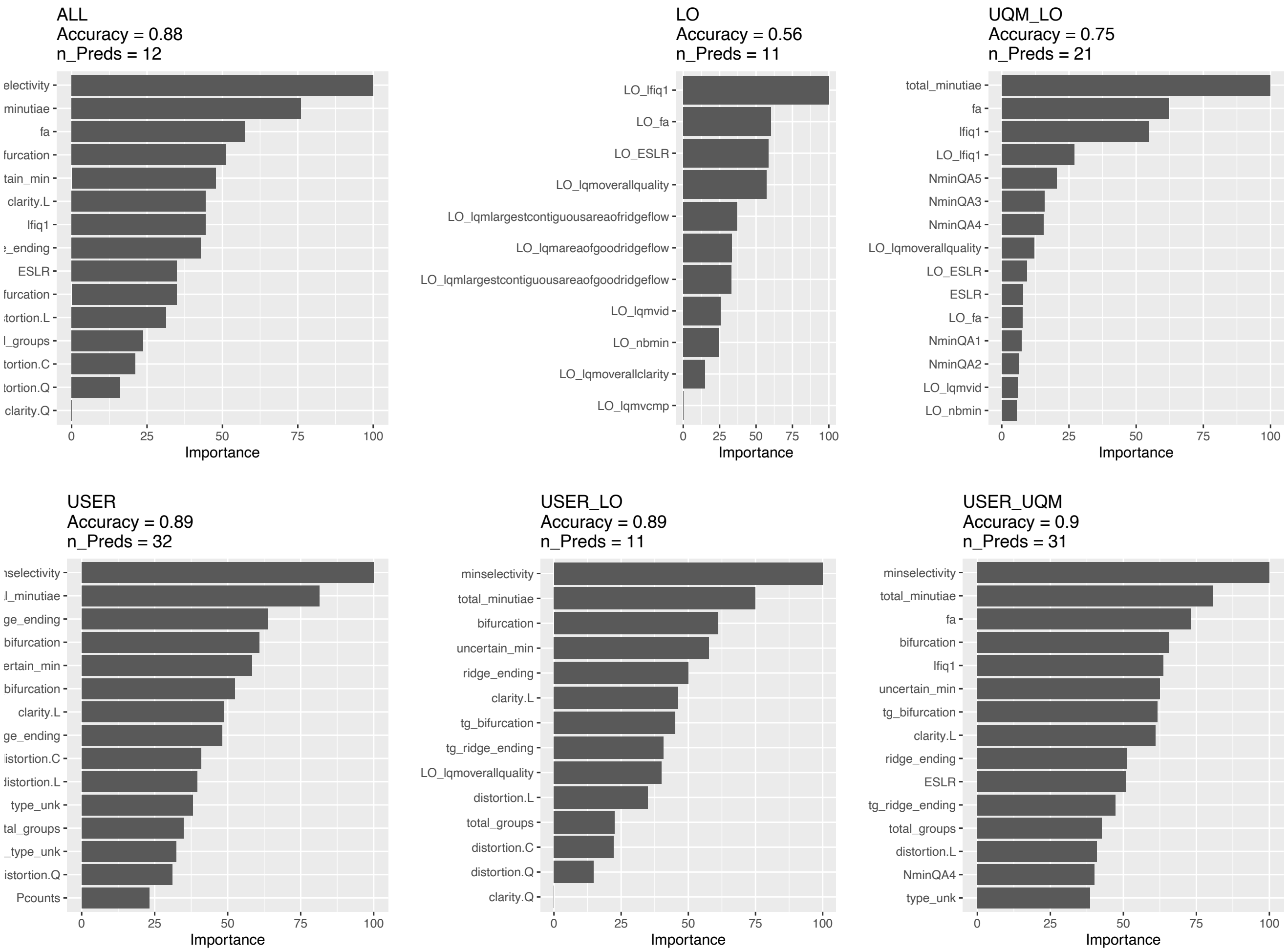
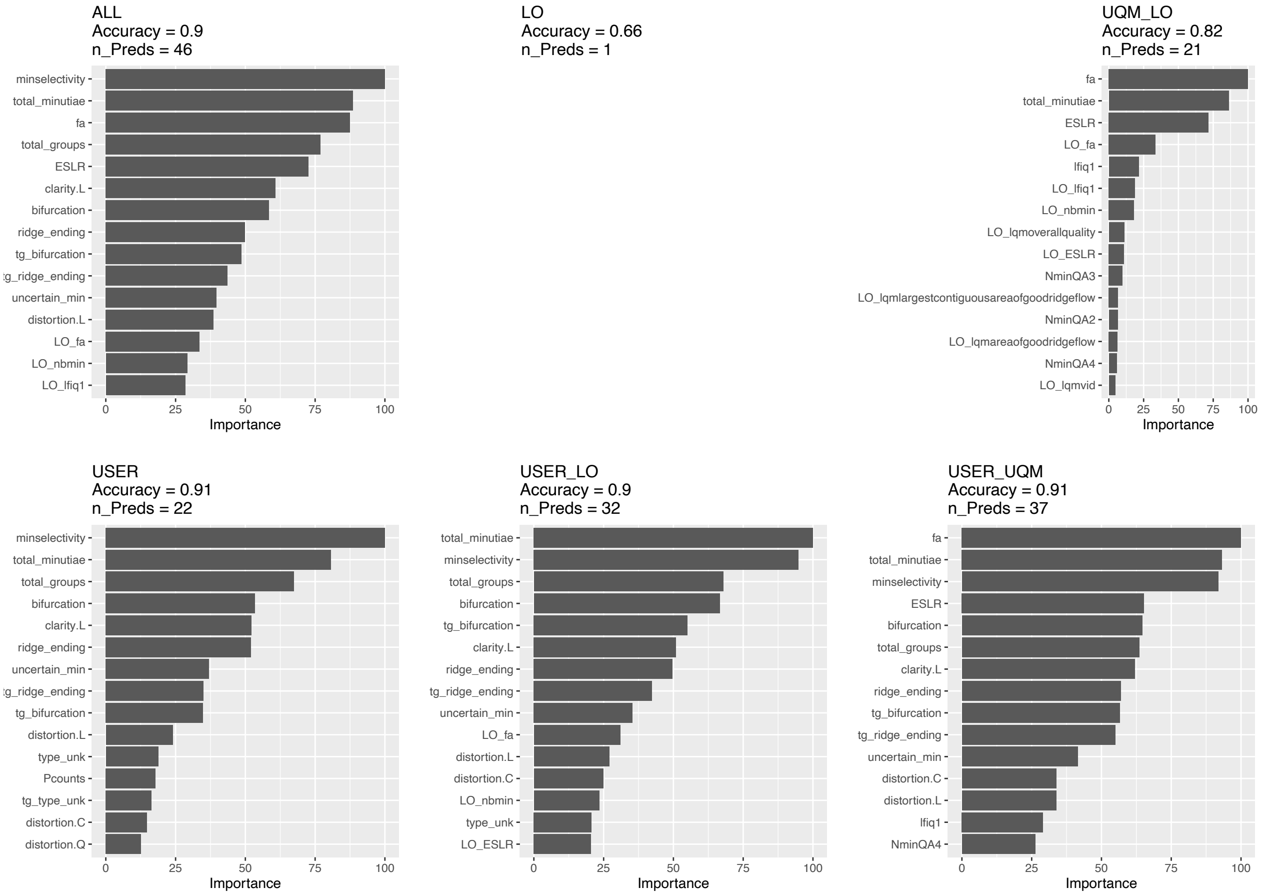


Figure 4.1 - 8 AFIS scale: Variable importance from the most impacting (100%) to the least impacting for each set of variables. The LO panel is blank because there was only 1 predictor used, thus it was at an importance of 100.



It can be seen in these figures that when the variable set included USER variables, “total_minutiae” and “minselectivity” were almost always the top two variables of importance, across variable sets and scales. The exception to this was the difficulty scale, where distortion became more important, but minselectivity still remained near the top of the list. In addition, “uncertain_min”, “distortion”, and “clarity” appeared frequently on the top variables of importance lists. When the UQM variables were included, “fa”, “lfiq1”, and “ESLR” were frequently present in the top predictors of importance.

When LO variables were included, “LO_lfiq1”, “LO_fa”, “LO_ESLR”, and “LO_nbmin” most consistently appeared on the list of variables of importance. Thus, these 12 variables were selected to bring forward to Step Two and are summarized in Table 4.1 - 4.

Table 4.1 - 4 Selected predictors

Predictors	Description
minselectivity	A measure of the perceived selectivity of annotated minutiae calculated by adding the ratio of confident to non-confident minutiae to a weighted value of the number of distinctive minutiae clusters and presence of target group
total minutiae	The total number of minutiae that were annotated by the participant
uncertain min	The number of minutiae annotated by the participant using minutiae markers that indicate uncertainty about their type or presence
distortion	The participant's global assessment of the distortion present in an impression
clarity	The participant's global assessment of the clarity of an impression
fa	The measure based on the score of the AFIS system returned after the ESLR transaction using the minutiae selected by participants
ESLR	The Expected Score-Based Likelihood Ratio, incorporating the number of minutiae selected by participants
lfiq1	The automated global quality measure incorporating the number of minutiae selected by participants
LO lfiq1	The lfiq1 measure based on the auto-encoded minutiae
LO fa	The fa measure based on the auto-encoded minutiae
LO ESLR	The ESLR measure based on the auto-encoded minutiae
LO nbmin	The number of minutiae detected by the auto-encoder

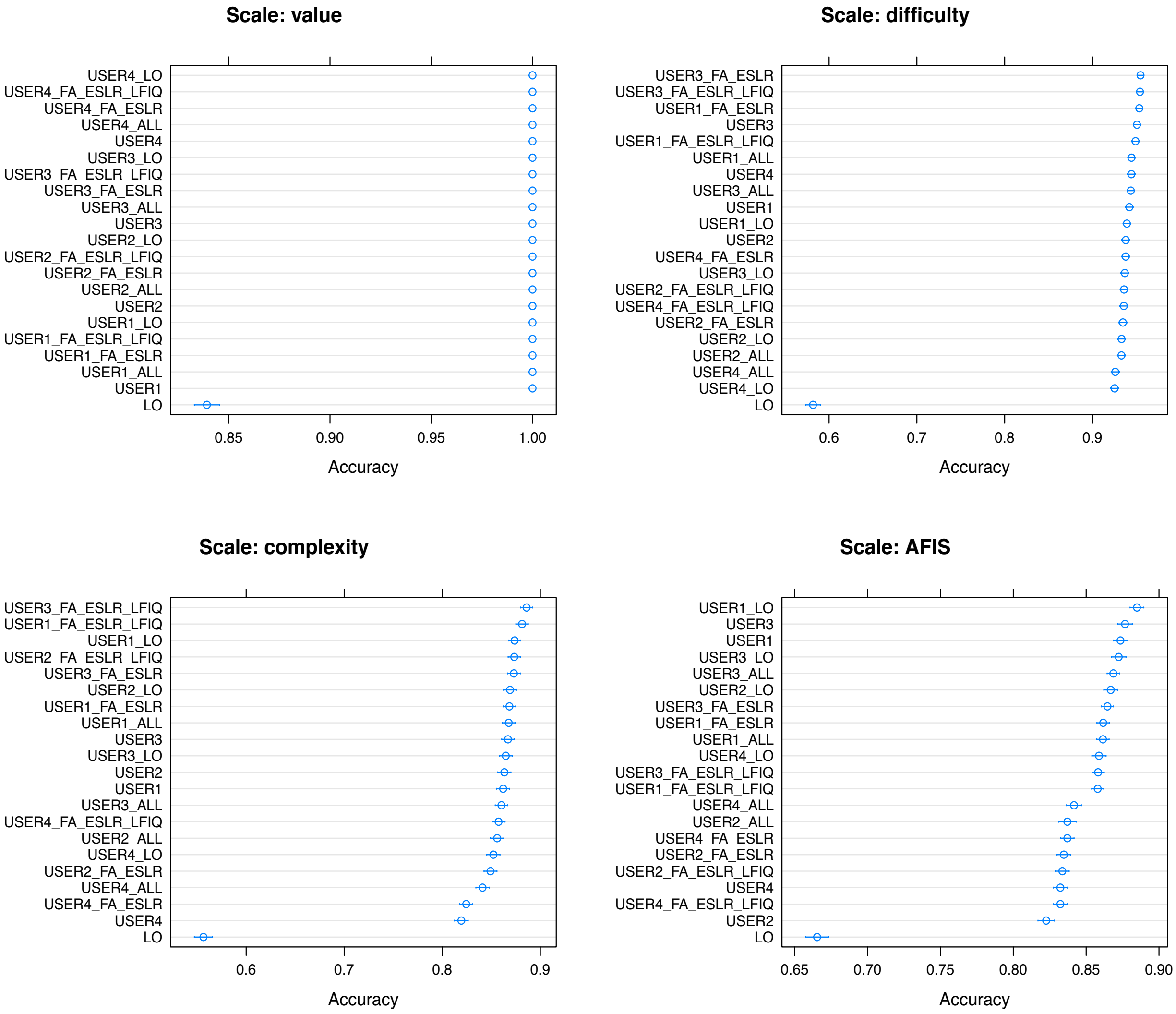
Step Two – Ideal Performance Testing

Once RF had been chosen as the MLA that would be used in the model and the list of potential predictors had been narrowed down (Step One), we set out to see how well the model could perform under ideal circumstances considering two different conditions: Optimal and Operational. When considering the Optimal condition, we assumed that we had unlimited resources of time and computation. Hence we could afford asking a lot of input from the examiners and deploy elaborated ML models.

When considering the Operational condition, we tried to maintain accuracy while removing variables that would be difficult to include in an operational model due to cost constraints involved in developing the software support to use these variables, or due to the computational load required to run the model with their inclusion. The fully optimized Operational model will represent something that could actually be deployed in working laboratories and is expected to suffer some trade-offs in accuracy for its ease and practicality of use. In both cases (Optimal and Operational), we continued to use the average user data, thus representing the best-case scenario for how these models could perform considering these two conditions. Because we were still using average user data at this stage, no decisions were made about what variables would be used in the final model; this step aimed to assess the theoretical limits of accuracy on the model's best day and which variables seemed to best support that performance.

Figure 4.1 - 9 summarizes the performance of each variable set from the combinations in Table 4.1 - 2 on each of the four scales.

Figure 4.1 - 9 Accuracy of the RF models trained on the average user data of the "superusers" who agreed with the ground truth. Accuracy is shown for each scale and for each selection of variables.



Note that for the Value scale, the only variable set for which a mean with associated error registered is the LO set, which performed comparatively poorly (Mean Accuracy for Value: 83.92%). The other variable sets all performed at accuracy of 1. The reason LO performed so poorly in relation to all other variable sets is because all the other variable sets included user data. Recall that the user data being used at this stage was the average user data of the "superusers" who agreed with the ground truth. These averaged and ideal data gave the model the best possible scenario for making a binary decision on Value, which became an easy task. The LO model did not have this advantage. There was no average data, there was only the single observation obtained from the automatic application of the algorithms. Thus, the LO model was operating at a distinct disadvantage.

Because all of the variable sets in the Value quadrant of the plot except LO performed equally, their order is irrelevant. However, for the other 3 scales, we can see trends in which variable sets tended to be the top accurate performers. For all 3 scales, variable sets that included USER1 or USER 3 tended to outperform those that included USER2 or USER4. Variable sets that included FA_ESLR and LFIQ were also frequently (but not always) better performers, particularly for the Complexity scale. Thus, if we were to consider optimal performance, the best overall performer for all but the AFIS scale would be USER3_FA_ESLR_LFIQ (Mean Accuracy for Value: 1; Difficulty: 95.41%; Complexity: 88.60%; AFIS: 85.82%). Even on the AFIS scale, this one's performance was less than 3 percentage points below USER1_LO (Mean Accuracy for AFIS: 88.49%), the top performer on that scale.

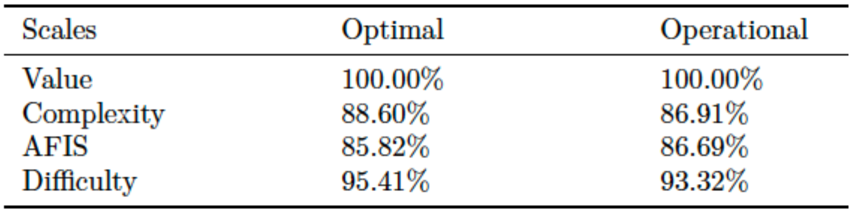
Any choice will be a balance between performance, user input and computational cost of the quality metrics. For example, fa, ESLR, and lfiq1 when used outside of the LO variable set are all calculated based upon the number of minutiae entered by the user. Software development could be done to create a dynamic interface that could wait for the user input, send it through the quality metrics, and receive a live response, but that development comes at a cost in terms of IT architecture and computing time. When disregarding any variable sets that include FA_ESLR or LFIQ, the top performing variable set across the 3 scales becomes difficult to call between USER1_LO and USER3. Both include the variable minselectivity, which turned out to be a powerfully diagnostic predictor. As was previously noted, minselectivity was a calculated variable that took into account the ratio of confident total minutiae, the number of highly diagnostic combined minutiae groups noted, and whether a target group was noted. This gave an overall measure of the perceived selectivity of the mark. Including such a variable in a user interface would require examiners to distinguish their types of minutiae and groups. As for the quality metric variables, such manual input will come at a cost here in terms of manpower.

Another finding supported by these data is that the notion of a hybrid model is a powerful one. Most of the top-performing variable set combinations included both one of the USER sets (incorporating participant input) and one of the automated variable sets (either LO or some combination of FA_ESLR and LFIQ, which are automated metrics relying on number of participant-selected minutiae). LO by itself performed extremely poorly and USER sets alone in general performed worse than when they were combined with at least one of the automated measures. This means two things: first that there is strong evidence that the human examiner still brings value to the suitability decision. These data do not support that a lights-out only assessment of suitability is superior to one made with human input for any of the 4 scales. The other is that human decision-making can be aided by automated measures.

After removing all variable set combinations that include FA_ESLR or LFIQ and any that include USER1 or USER3, the best performing model out of those remaining across the 3 non-Value scales becomes USER2_LO (Mean Accuracy for Value: 1; Difficulty: 93.32%; Complexity: 86.91%; AFIS: 86.69%).

Table 4.1 - 5 summarizes the performance of the best Optimal model and the best Operational model and shows that there is very little difference between the two. In fact, these data show that, when using the averaged data of only users who agreed with ground truth, both models perform quite well at predicting the consensus ground truth decision. This is encouraging because it suggests that (1) annotations and automated quality metrics working together can be used to predict consensus decisions; (2) consensus annotations made by examiners are a good basis for supporting suitability decisions; and (3) in theory, and under idealized circumstances, these hybrid examiner/automated models provide efficient guidance that can aid examiners, particularly in small laboratories where insufficient co-workers are available to form consensus panels and the model may serve as a proxy for the consensus.

Table 4.1 - 5 Mean accuracy obtained for the Optimal and the Operational model for the superuser.



Scales	Optimal	Operational
Value	100.00%	100.00%
Complexity	88.60%	86.91%
AFIS	85.82%	86.69%
Difficulty	95.41%	93.32%

Finally remember that at this stage the above contrasting measures of accuracy are based on our ideal superuser input. We would like to wait to make a decision regarding which predictors to retain until we have tested the performance of all sets of variables on individual users (Step Three). It may well be that the above observed trends become insignificant when we use the models accounting for all the examiners' data instead of only our superusers.

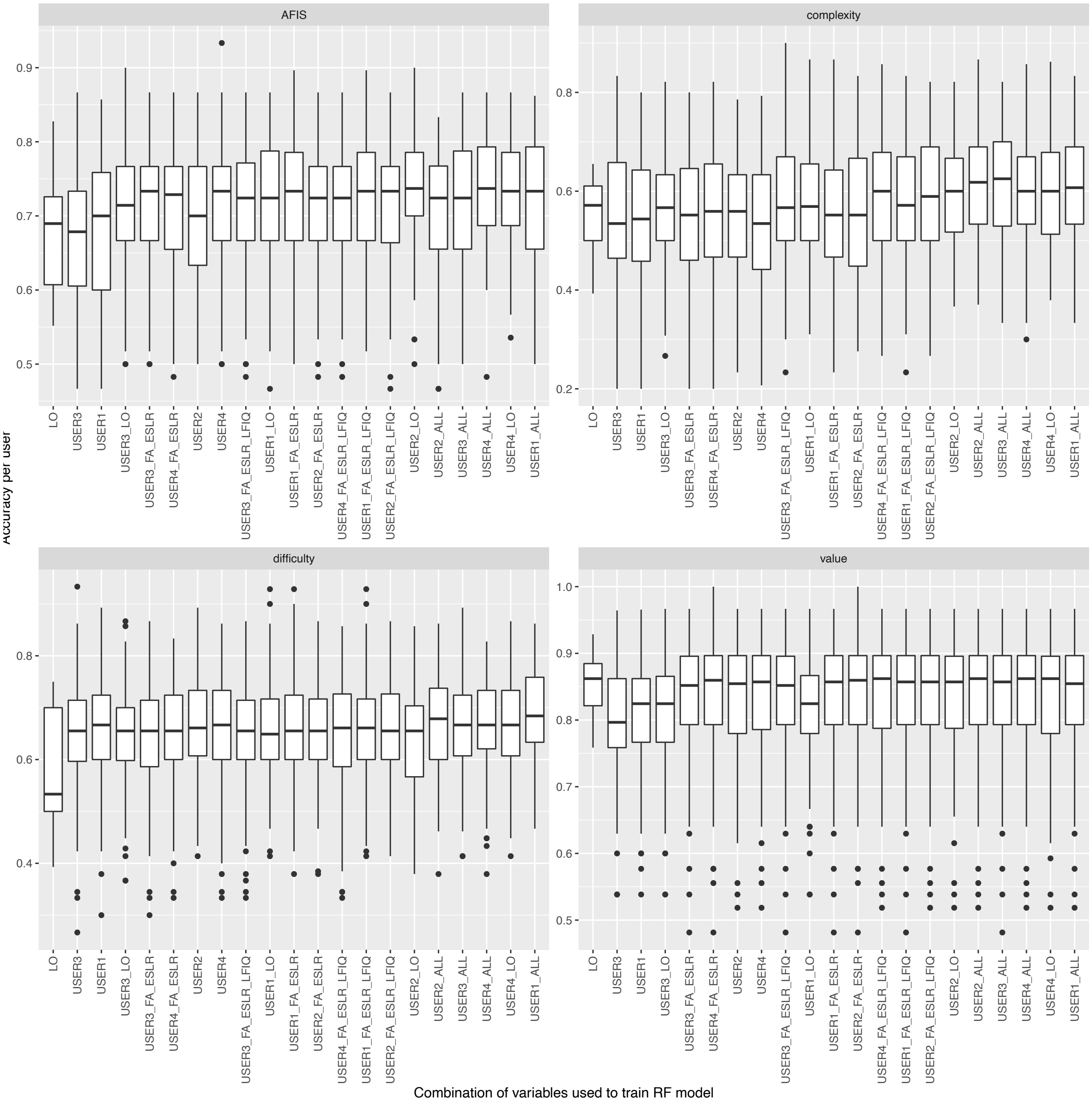
Step Three – Model Optimization in realistic applications

Although the accuracy of the model for the best-case scenario was generally quite high, demonstrating that the theoretical backbone of the exercise is sound, there is often a difference between the best-case and how things work in the real world. Additionally, the results of Step 2 took into account the performance of the model in predicting ground truth for each image, but we are really more interested in its ability to predict ground truth based on the annotations of a particular user.

Thus, we re-tested the RF models from step 2, this time using all the data in the study, meaning all user inputs from users who did not agree with the consensus ground truth as well as the all user inputs (not just averaged consensus inputs) from users who did agree with the consensus ground truth. We have adopted a leave-one-out validation scheme in the sense that the models were trained on 99 cases and tested on the remaining case with all its associated users. When all cases have been tested, we can then compute the accuracy for each user (aggregated across all of the completed trials by the user) against the declared ground truth established previously by the majority.

The models' performance per user, not per image, are shown in Figure 4.1 - 10. It gives the accuracy per user for the variable set combinations named on the x-axis used in the RF models to predict the suitability decision for each scale. Each data point used to construct the boxplot represents a single user.

Figure 4.1 - 10 Accuracy of the predictions based on user individual data for each scale and for each selection of variables.



In Figure 4.1 - 10, each scale is presented separately and the variable set combinations are ordered according to an overall mean performance score across all 4 scales, with the best overall performing being the furthest to the right.

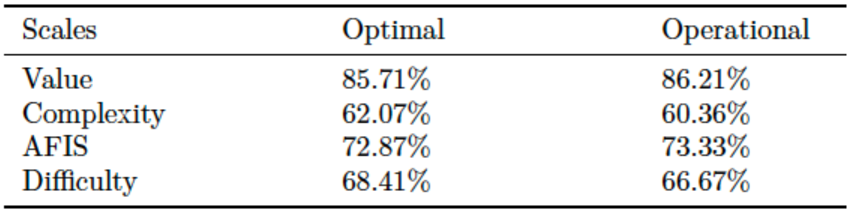
The dark line in the middle of each boxplot represents the median of the data which makes this type of data presentation more robust to outliers. For example, the lone dot at the bottom of the “USER1_FA_ESLR_LFIQ” boxplot on the Complexity scale represents a user whose annotations did not at all well assist in predicting the consensus ground truth.

Viewing these data, it is easy to see that the model using the USER1_ALL variable set performs best overall. However, the performance of the model has deteriorated after adding in all the messy, real-world user data, as expected. The median accuracy values for the USER1_ALL model on each of the four scales are: AFIS-72.87%; Complexity-62.07%; Difficulty-68.41%; and Value-85.71%. This represents the Optimal model for predicting ground truth using examiner annotations when resources are unlimited.

However, as noted under Step 2, the USER1 variable is quite demanding on the user input due to the inclusion of the minselectivity variable. The next best performing model overall-USER4_LO-does not include the minselectivity variable, or any other variables that would require quality metrics computed based on user input. The effect of choosing a simpler model is very limited on the overall accuracy. The median accuracy values for this model on each of the four scales are: AFIS-73.33%; Complexity-60.36%; Difficulty-66.67%; and Value-86.21%. Thus, this model has been selected to be taken forward to external validation.

The comparison between the median performance of the two models is shown in Table 4.1 - 6 (Table 6). These data suggest that the chosen Operational model can be used in the real world, without much loss in accuracy compared to the Optimal model.

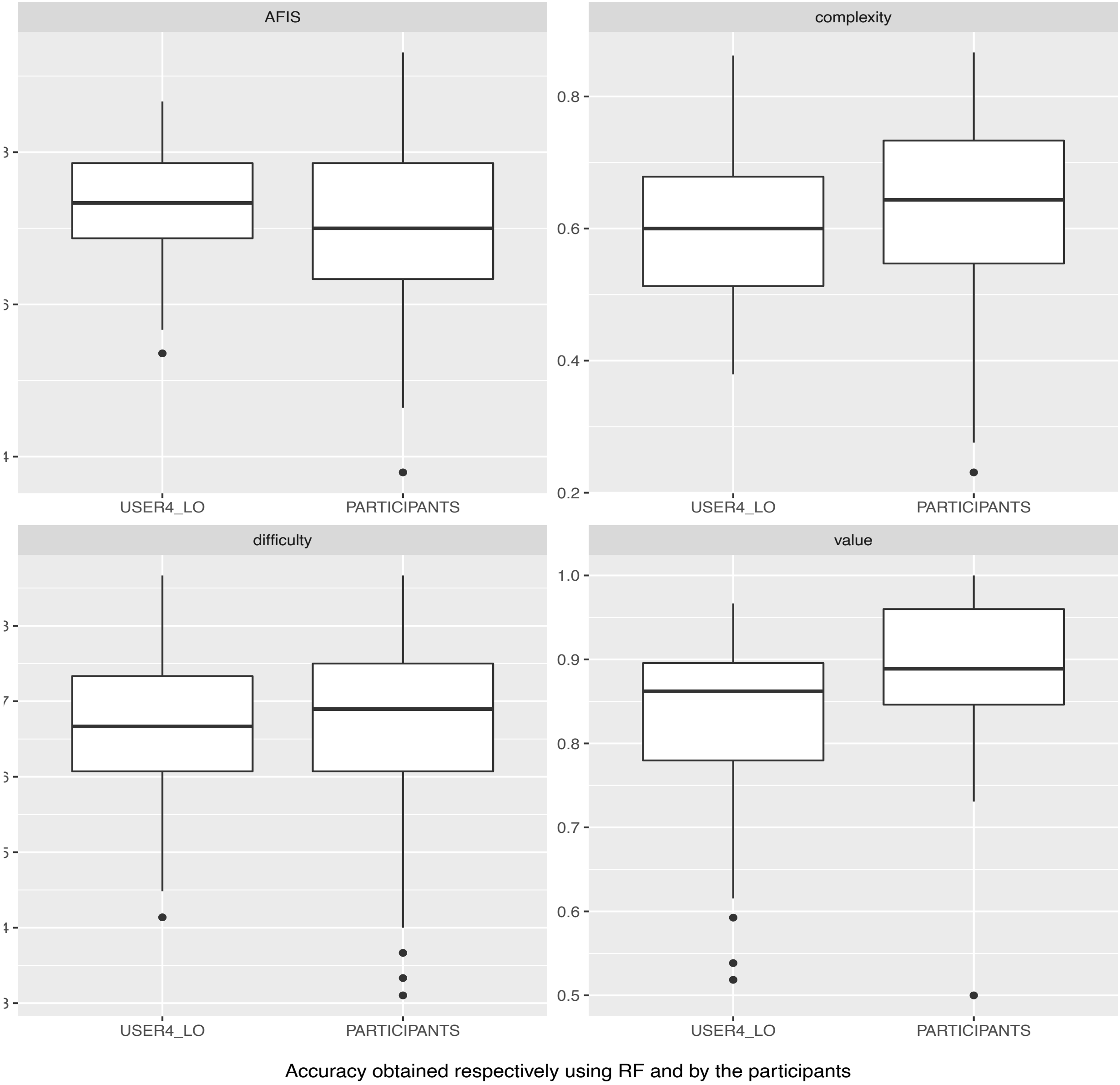
Table 4.1 - 6 Median accuracy obtained for the Optimal and the Operational model for the individual users.



Scales	Optimal	Operational
Value	85.71%	86.21%
Complexity	62.07%	60.36%
AFIS	72.87%	73.33%
Difficulty	68.41%	66.67%

Once the final model had been selected, we compared its performance to the performance of users (with their individual responses) in selecting the consensus ground truth to see which was more accurate. The results of these comparisons are shown in Figure 4.1 - 11.

Figure 4.1 - 11 Accuracy of the predictions of the RF model compared to the individual responses given by the users.



It can be seen that overall, the performance of the model was on par with the performance of the examiners, although examiners did slightly outperform the model on the value scale. These data provide evidence that a model that incorporates both examiner and automated observations can be used to assist examiners, both by streamlining the analysis process and by providing consensus guidance along 4 scales of suitability, without a loss of accuracy.

The results presented up to this point have all been based upon the prediction accuracy of the models based upon a decision threshold (50%) based on prior probabilities of each response defined by the training set. For example, for the binary value/no value decision, if the probability of a set of data leading to a GT conclusion of “Of value” was greater than 50%, the model would predict “Of value”. However, this does not present a complete picture. Many laboratories may not treat the prior probability of calling a given mark no value or of value based on these prior probabilities. For various operational, societal, fiscal, or political reasons, laboratories may wish to adjust the threshold of when they call a mark no value versus of value. For example, a very high throughput laboratory that sees many low-penalty crimes and has a huge backlog may put a priority on getting cases out the door and may only want the very best marks to be called "Of value" whereas they are happy to "miss" many marks that other laboratories might be willing to compare. Such a laboratory might wish to set a very high threshold for the value decision, such that a probability of 70 or 80 percent is required to reach an "Of value" prediction. For this reason, it is desirable to consider how the models perform across all possible probability thresholds between 0 and 1. For this purpose, a Receiver Operating Characteristic (ROC) curve is a superior evaluation tool. ROC curves evaluate the performance of a classifier by plotting the trade-off between sensitivity (True Positive Rate) and 1-specificity (False Positive Rate). The dashed diagonal line represents a baseline where the two are equal, thus, the closer to the line that a curve lies, the less accurate it is. The higher the curve is toward the upper-lefthand corner, the higher its accuracy. However, ROC curves are generally used for binary classifiers and 3 of our scales are not binary. Thus, we have used a MultiROC analysis (Wei and Wang 2018) to present the performance of the models against one another. Figure 4.1 - 12 to Figure 4.1 - 15 present the results of this analysis.

Figure 4.1 - 12 MultiROC curves associated with the Value scale.

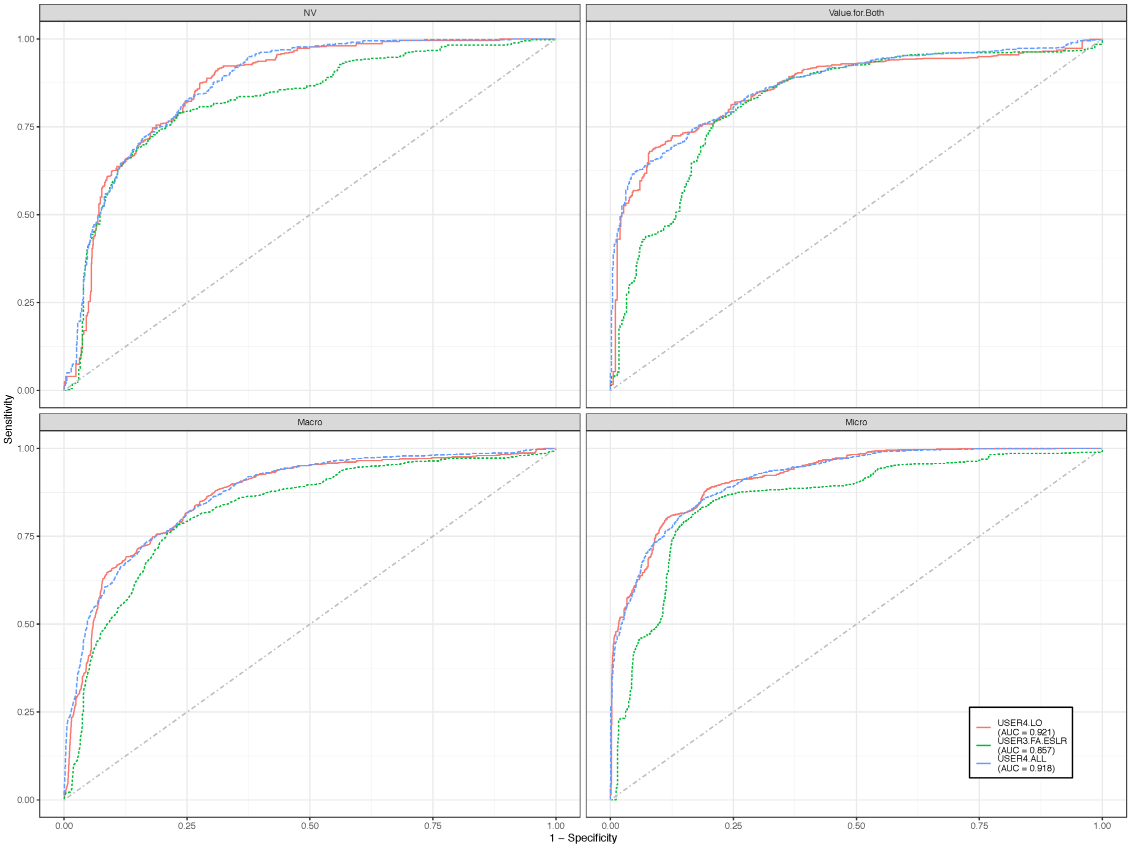


Figure 4.1 - 13 MultiROC curves associated with the Difficulty scale.

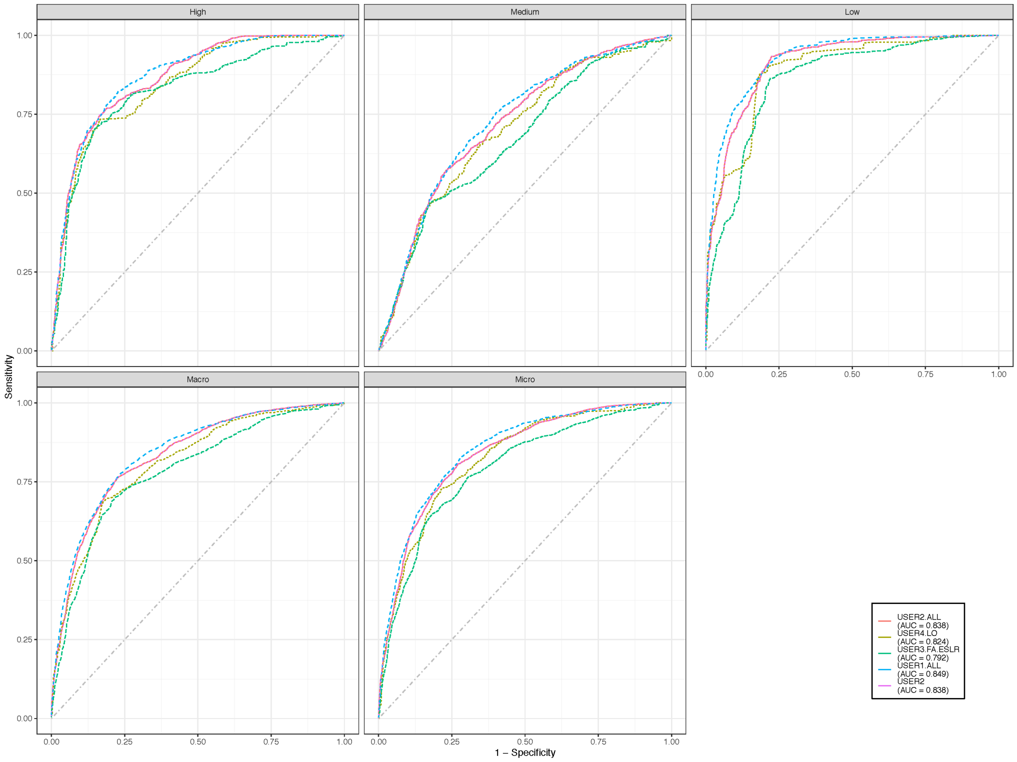


Figure 4.1 - 14 MultiROC curves associated with the Complexity scale.

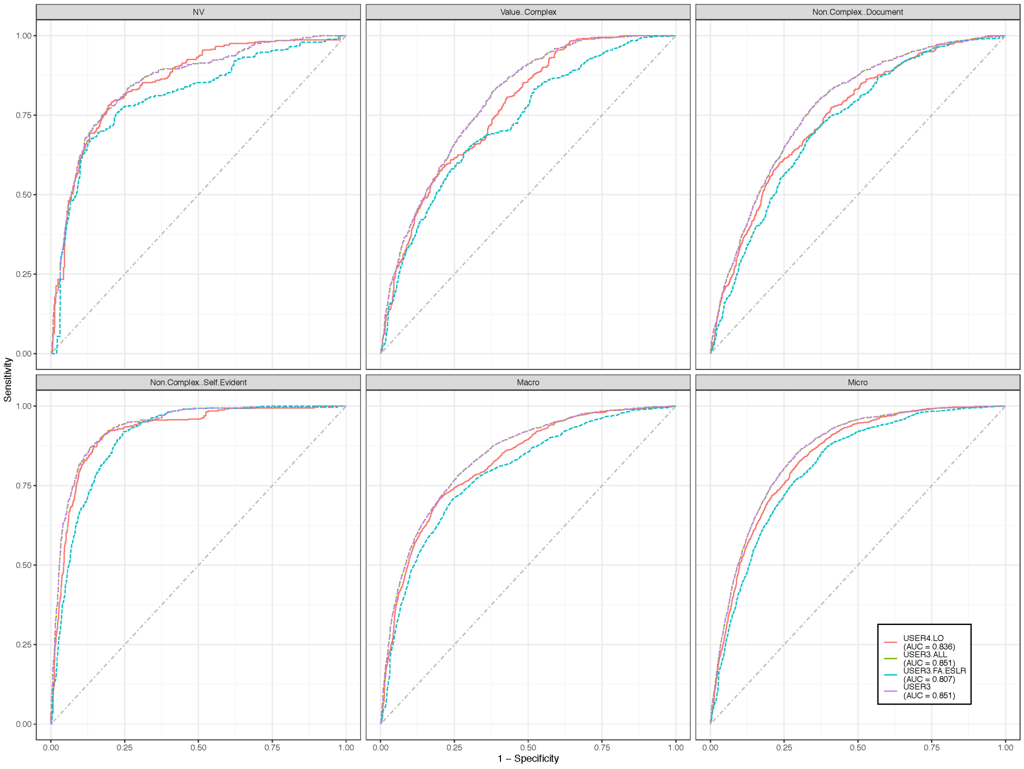
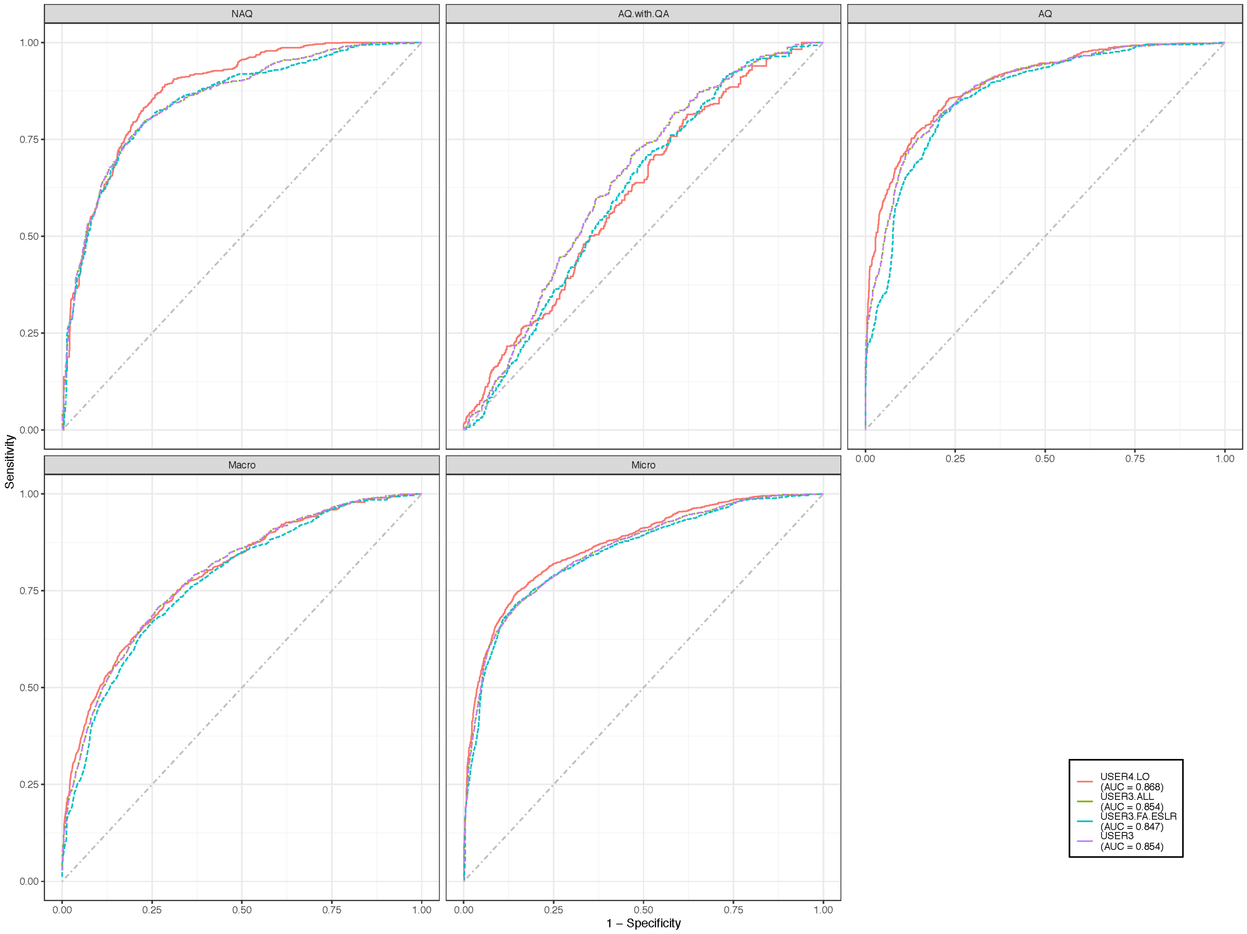


Figure 4.1 - 15 MultiROC curves associated with the AFIS scale.



Note that if we were to evaluate all 21 variable set combinations on the same ROC curve, it would be quite difficult to read. Thus, we have elected to compare only the top 2 performers, the worst performer, and our selected model (USER4_LO) for each scale. This allows us to evaluate how the USER4_LO model performs compared to the range of performance for each scale. In some cases there may be more or fewer than exactly 4 models represented if USER4_LO was one of the 2 best performers (or the worst one), or if there was a tie in the best- or worst-performer categories. Each model is presented along with its Area Under Curve (AUC), which is a measurement of the total performance of the model; generally speaking, a higher AUC indicates a more accurate model, although this may not be true for every probability threshold. Some models may perform better or worse than others at specific points along the curve.

The MultiROC output depicted in Figure 4.1 - 12 to Figure 4.1 - 15 breaks the performance of the models out along the different classifiers being evaluated in the top row of each plot. Thus, we can separately see the performance of the models at predicting each classification. For example, on the difficulty plot, we can see that all the models are better at predicting “High” and “Low” difficulty conclusions than they are at making accurate “Medium” difficulty determinations. This mirrors the behavior of human examiners, who are also better at making accurate decisions at the extremes.

In the second row of each plot, there are “Macro” and “Micro” curves presented. These are a way of measuring the aggregate performance of each model across all of the possible classifications. Because some classes are often larger than others (for example, in our data, the category “AQ with QA” has fewer examples than the other AFIS quality categories), there is a danger of the larger categories dominating the smaller categories when averaging of the model performance is done across categories. Therefore, macroaveraging and microaveraging weight the averages differently such that the effectiveness of large classes is best represented in microaveraging and the effectiveness of small classes is best represented in macroaveraging (see (Van Asch 2013) for a more in-depth discussion of micro- and macroaveraging).

We can see that the selected model, USER4_LO, generally performed quite well compared to the other models. USER4_LO had the highest AUC for both the Value and AFIS scales (0.921 and 0.868, respectively). Although it did not have the highest values for the Difficulty or Complexity scales, neither were its AUCs the lowest, and they generally compared favorably (0.824 versus 0.849 on the Difficulty scale and 0.836 versus 0.851 on the Complexity scale).

Ultimately, based upon the described evaluations of the performance of the potential models and upon the operational constraints limiting our choice of variables, the USER4_LO model was selected as the final model to take forward to external validation testing. The variables present in the USER4_LO model are: total_minutiae, clarity, distortion, LO_fa, LO_ESLR, LO_lfiq1, and LO_nbmin.

As was noted in Step 2 (using limited, idealized data), these Step 3 results using the full data from all the users still support that a hybrid model incorporating both automated quality metric data and expert-input data provides the best performance. Although the performance of the model in Step 3 was, as expected, much lower than the performance of the best models in Step 2, comparison of its effectiveness to that of human examiners making the same determinations showed that the model can be expected to perform about as well as the human examiner, making it a viable tool for both reducing the time spent on analysis of marks and for providing nuanced guidance along four scales, especially when consensus guidance is desirable and otherwise unavailable.

External Validation

Once USER4_LO was selected as the final model, we conducted one additional data-collection to test it externally using a mix of old and new images and using a fresh recruitment of participants (which may or may not have included some of the same participants as the original study, as they were all anonymous). Because we selected the model by down-selecting from all the data we initially gathered to only the variables that proved to be most diagnostic, the structure of the follow-up study was fundamentally different from that of the original study from the perspective of a user. Because they were being asked to provide different input than in the original study, this might change the way they approached the analysis, which could in turn affect the performance of the model.

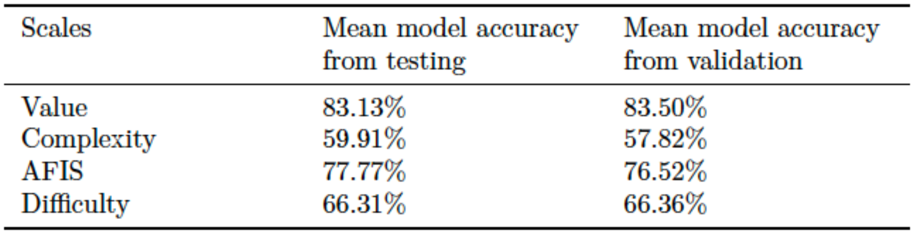
Additionally, we wanted to see whether the model performed equally well on brand-new images that were not used in the training of the model and whether it was equally able to predict consensus ground truth decisions when different users were using it. Finally, we wanted to get a sense of the usefulness of the tool-that is, whether the examiners usually agreed with the predictions of the model or whether they found it to be something that would hinder them by usually being "wrong" in their eyes. If they disagreed, we wanted to understand why the model had failed-was it a flaw in the model, a difference in opinion that was unavoidable, a difference in how thresholds or definitions were applied, or a failure of the participant to follow the study instructions?

Model performance against ground truth

We measured the performance of the model at predicting ground truth in three ways: (1) how the model performed with these follow-up data overall compared to how it performed in Step 3 using the original images only (Table 4.1 - 7); (2) how it performed on the original images (in its final form and with a new set of participants) compared to how it performed on new images it had never used before (Table 4.1 - 8); and (3) how it performed across all follow-up images compared to how well users performed at making conclusions that matched the consensus ground truth (Table 4.1 - 9).

Table 4.1 - 7 shows that there was no drastic difference between the mean performance of the model in Step 3 (development) and its performance using the follow-up validation data. These findings support that, globally at least, the model is robust to new images and new users.

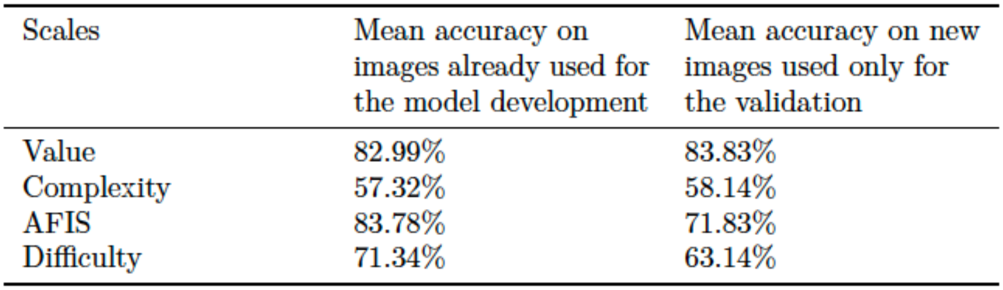
Table 4.1 - 7 Mean accuracy obtained for the chosen model on respectively the test data and on the validation data.



Scales	Mean model accuracy from testing	Mean model accuracy from validation
Value	83.13%	83.50%
Complexity	59.91%	57.82%
AFIS	77.77%	76.52%
Difficulty	66.31%	66.36%

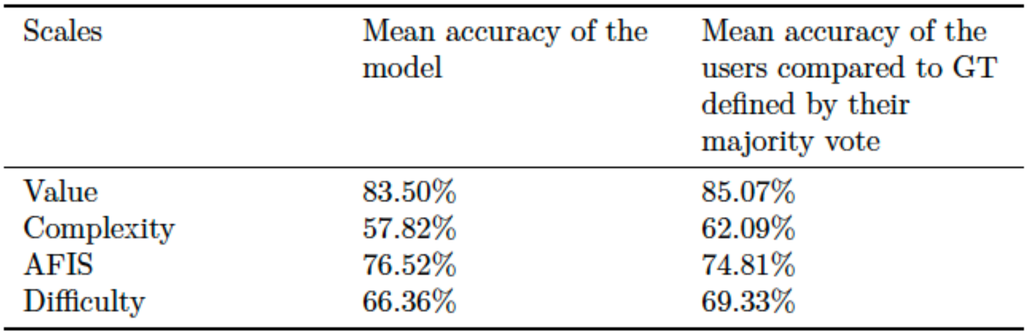
Table 4.1 - 8 shows that there was virtually no difference in mean accuracy rates between images from the original study and new images for the Value and Complexity scales. There was, however, a notable difference in accuracy between the two image sets on the AFIS and Difficulty scales. It is not clear why these two scales have been impacted by the new images whereas the Value and Complexity scales were not. It should be noted that the drop in accuracy on new images for the Difficulty scale was small when comparing it to the mean accuracy of the same scale in Step 3 (66.31% down to 63.14%). However, the drop in accuracy for the AFIS scale is larger (77.77% down to 71.83%).

Table 4.1 - 8 Mean accuracy obtained for the chosen model on respectively the images used already during the model development and new images used only for the validation.



Scales	Mean accuracy on images already used for the model development	Mean accuracy on new images used only for the validation
Value	82.99%	83.83%
Complexity	57.32%	58.14%
AFIS	83.78%	71.83%
Difficulty	71.34%	63.14%

Table 4.1 - 9 Mean accuracy obtained for the chosen model versus the mean accuracy obtained by the users when their responses were compared to the majority vote considered as ground truth (GT).



Scales	Mean accuracy of the model	Mean accuracy of the users compared to GT defined by their majority vote
Value	83.50%	85.07%
Complexity	57.82%	62.09%
AFIS	76.52%	74.81%
Difficulty	66.36%	69.33%

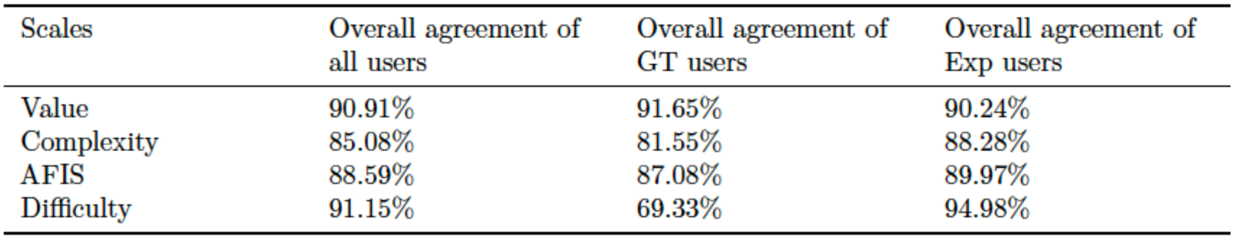
In reviewing the instances in which the prediction of the model did not match the ground truth designation on the AFIS scale, it is not apparent why the new images proved more difficult to categorize. The number of disagreements per image ranged from 0 to 82 (out of 82 total participants), with 6 images that had more than 40 disagreements being responsible for approximately 57% of the disagreement on the AFIS scale. Ground truth was AQ with QA for 4 of these images and AQ for 2. The predicted responses were mixed. In all 6 cases, the location of the core was discernible, although the images were all degraded to some degree. It may well be a simple artifact of the random sample selection that the new images happened to be slightly more difficult to classify for AFIS use.

Table 4.1 - 9 shows that the model's and the users' performance were quite close on all four scales. Users slightly outperformed the model on all except the AFIS scale. The superior performance of the model on the AFIS scale may be explained by the fact that several users left text notes stating that they don't usually make AFIS determinations in their laboratories. This lack of experience from some users may have impacted the overall average performance of the group.

<u>Model performance against examiner opinions</u>

We also took three measurements of the level of overall agreement between examiners and the model (i.e. what percentage of the time did the examiner agree with the prediction of the model): (1) the overall level of agreement across all users; (2) the level of overall agreement of users from the GT group; and (3) the level of overall agreement of users from the Exp group. These are summarized in Table 4.1 - 10.

Table 4.1 - 10 Overall agreement of respectively all users, GT users, and Exp users.



Scales	Overall agreement of all users	Overall agreement of GT users	Overall agreement of Exp users
Value	90.91%	91.65%	90.24%
Complexity	85.08%	81.55%	88.28%
AFIS	88.59%	87.08%	89.97%
Difficulty	91.15%	69.33%	94.98%

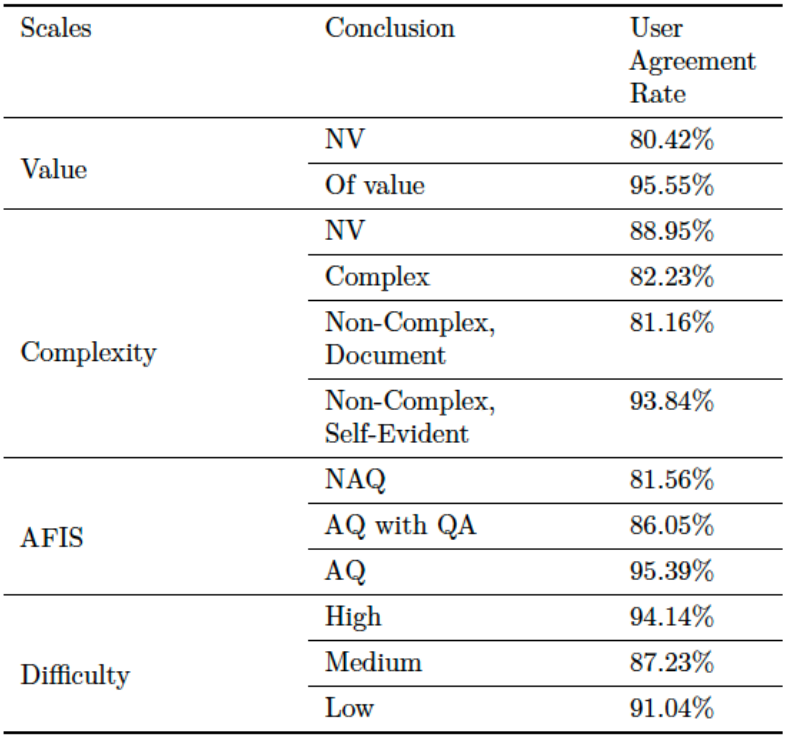
The reason we took the levels of agreement of the GT group and the Exp group separately was because we wanted to guard against a biasing effect of seeing the model prediction. Recall that in the GT group, the participants had to form and enter their own opinions on each of the four scales before they saw the model predictions. In the Exp group, participants went straight to the model prediction and were asked whether they agreed with it. We were concerned that the Exp group might be influenced by the model prediction and be more likely to simply agree with it. However, we also wanted to see whether the Exp group would be amenable to the suggestion of the model, or whether they would frequently be annoyed by the model and want to argue with it.

As Table 4.1 - 10 demonstrates, first of all, the overall level of agreement with the model was very high. Regardless of their group, examiners could accept the determination of the model most of the time. However, there was a small effect of seeing the model prediction before forming their own opinion. This effect was particularly pronounced on the Complexity and Difficulty scales, where the Exp users were much more likely to agree with the model than the GT users were to have blindly selected the same conclusion as the model. This effect was not seen on the Value scale, where agreement was consistently very high.

We also wanted to see whether there were any particular conclusions on each scale that fostered greater or less agreement than the scale as a whole. Table 4.1 - 11 presents the agreement rates of all participants for each scale, broken down by the conclusions within those scales.

Although agreement rates stay above 80% across all conclusions, there is a trend in all but the Difficulty scale that the highest-level of the scale shows the greatest agreement. This is in line with our observations in (Eldridge et al. 2020) that the only time examiners agreed with one another was at the high ends of the scales. It appears that agreement with the model follows the same trend.

Table 4.1 - 11 User agreement rate for each scale per conclusion.



Scales	Conclusion	User Agreement Rate
Value	NV	80.42%
	Of value	95.55%
Complexity	NV	88.95%
	Complex	82.23%
	Non-Complex, Document	81.16%
	Non-Complex, Self-Evident	93.84%
AFIS	NAQ	81.56%
	AQ with QA	86.05%
	AQ	95.39%
Difficulty	High	94.14%
	Medium	87.23%
	Low	91.04%

Many of the written notes where disagreement occurred on the Value scale centered around the notion of Value for Exclusion Only (VEO). In many cases in which the model predicted NV, the notes from the participant indicated that they would consider the mark to be VEO. Although the model in this follow-up study did not distinguish between VID and VEO, lumping them all into a single “Of value” category, it would appear that in some cases, it was setting the threshold for “Of value” too high and missing out on some marks that could be considered VEO. This confusion is likely due to the split in the field between examiners who consider VEO to be of value, and those who consider it to not be of value giving more attention to VID marks instead. Because ground truth was established by a majority of votes, this disparity could have caused some VEO marks to be classified as NV and some to be classified as “Of value” depending on the makeup of voters who saw each image during the initial study. These fluctuations may have impacted the machine learning as marks that might be considered VEO were not consistently classified into either NV or Of value for the algorithm to learn from.

On the Complexity scale, the disagreements largely fell into two categories: disagreements about whether a mark was NV or Complex, and disagreements about whether a non-complex mark required standard documentation or not. The first group of comments largely mirrored the Value scale in that many people were commenting that marks predicted by the model to be NV were VEO.

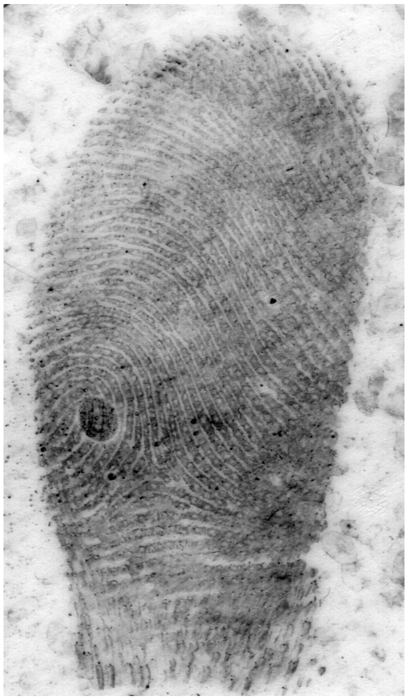


Figure 4.1 - 16 Image076 declared to be Self-evident by majority but misclassified by the model.

Although there were some instances in which the “Non-Complex, Documentation Required” predicted marks were thought to be Complex by examiners, the second group of comments was dominated by marks that were predicted as “Non-Complex, Documentation Required” but the examiner didn’t think the documentation was warranted. In some of these cases, the examiners argued that they felt the clarity was sufficient that the mark belonged in the highest category. However, many of the comments seemed to reflect either an unwillingness to document marks at all that were not complex, or a failure to understand the intent of this category. There were several comments stating that the mark should not require “additional documentation” even though the description for this category stated that it was only requiring “standard documentation” as required by laboratory protocols. Only the Complex category required “additional” documentation. For example, one examiner even explicitly stated, “I feel that no additional documentation is needed other than what I wrote in the case notes”. Of course, the definition of this category was that the mark should have the normal documentation that would be included in the case notes (as opposed to reduced documentation for the top category, or additional documentation for the Complex marks). There were some images where a large number of people agreed that the image was very high quality and belonged in the Self-evident category, where the model predicted differently. One such example is image076, which is presented in Figure 4.1 - 16.

Of some note is that there were only 36 instances (out of 624 total disagreements with the model about complexity) in which the model predicted “Non-Complex, Self-Evident” and the examiner disagreed. In nearly all of these cases, the examiner noted some small distortion that they believed should be documented.

On the AFIS scale, 76.1% of the disagreements occurred when the model predicted NAQ. In most of these cases, the examiners commented either something to the effect of “the mark is borderline, but I’d give it a shot” or stated that they thought it could be entered with additional QA measures. Only 18 (of 477 disagreements on the AFIS scale) disagreements occurred when the model predicted AQ with QA. Of the disagreements with the model’s AQ predictions, most come from examiners who urged slightly more caution with the particular mark, recommending that additional QA measures be in place. A few comments stated they would not search the mark at all, whereas some specified that their agency's minutiae threshold was not met, or that the mark was a tip, which they would not search. One limitation of this study was that information on the presence or absence of a core was not captured, which may have had an influence on AFIS decisions.

Finally, for the Difficulty scale, there was not a clearly discernible pattern for the written comments. Many were one category away from the prediction in either direction (Low and High predictions were called Medium, Medium predictions were called Low or High) but there were also a good number of High predictions where the examiner felt the mark should be Low difficulty. Interestingly, on this scale, there were a few cases where the examiner seemed to have gotten confused and marked that they disagreed with the model, yet their notes matched the model prediction.

Overall, the results of the validation have shown that the model has promise. Its mean accuracy has been comparable to the accuracy of users in every category and user agreement with the model's predictions was uniformly above 80% and frequently above 90%. The design of the model leverages both automated, data-driven assessments and a very quick, streamlined examiner input interface to produce results that are quick, transparent, and provide a consensus-based recommendation. Examiners, particularly in small laboratories, could implement this model into casework in order to save time on in-depth analysis, while being able to explain in court that the results were based on both their own expert observations and reproducible automated measures and that they represent what a consensus of experts would be likely to conclude. In addition, the nuanced guidance provided by having four scales of suitability has the flexibility to assist examiners in a variety of situations.

One weakness of the model is that no information on the presence of cores and deltas was captured from the examiners during the white box study. This may skew the results on the AFIS scale and may partially explain the relatively low accuracy of the model on this scale, because impressions such as tips may have been predicted to score high on the AFIS scale based upon their clarity and number of minutiae whereas human examiners may have reached an NAQ decision based upon not having a clear indication of core location. Similarly, it is unclear how the lack of a core or delta may have influenced human examiners in making their decisions on the other scales, but this did not enter into the model's calculus at all.

Another limitation of the model was the necessity to change the Value scale from the 5-conclusion scale that was used in the white box study to the simple binary "Of value-No Value" scale that was ultimately used to build the model. This was necessary because people's unfamiliarity with the new conclusions resulted in them not being the "consensus ground truth" option often enough to be effective in machine learning; thus the model could not effectively predict these outcomes without sufficient samples on which to train. The result of this is that we lost a great deal of nuance in the Value decision that we might otherwise have explored and we may have also lost some accuracy in the model because there developed a disparity between how human examiners and the model treated VEO marks. This disparity is reflected in the latent print examiner community, where there is a split between those who consider VEO marks to be of value and those who do not.

Conclusion

The purpose of this research was to leverage observations made by examiners that were used to support suitability decisions and automatically derived quality measures to develop a model that can be used to predict suitability. We quickly realized predicting value based only on automatic measures (in a sort of light-out mode) was not performing well and the addition of user input increased accuracy. Data from a white box study (Eldridge et al. 2020) that collected examiner observations were then combined with variables from automated quality metrics using machine learning algorithms to develop a model capable of predicting consensus suitability decisions along four scales: Value, Complexity, AFIS, and Difficulty.

The best-performing model was selected as the one that best balanced meeting operational constraints and reducing the number of user-input variables to save on analysis time with maintaining the highest achievable accuracy. This means that the model will allow examiners to reach a suitability decision very quickly, without having to perform a full analysis, yet will predict decisions with which the majority of expert examiners would agree. This would reduce the burden on latent print examiners in terms of time spent on analysis, allowing them to focus their efforts more on complex impressions. After optimization, the final model included both observations from expert examiners and variables taken from automated measures. This supports that the human examiner and automated measures both bring value to the suitability decision and furthermore that both can work together successfully. There is no argument based on these data for removing the human examiner entirely and relying solely upon automated measures for making suitability determinations.

The proposed model is accurate in making these predictions at up to 83.13% accuracy when using the full study data (as opposed to only that of examiners who agreed with the consensus) and this accuracy held steady at 83.50% on the Value scale when using new users and new images for validation. The model did not perform as well on the other three scales; however, the accuracy of the models was always on par with that of users making the same predictions.

Agreement between human examiners and the predictions of the model was generally high during the external validation study, ranging from 85.08-91.15% overall across the four scales, indicating that human experts generally would accept the prediction of the model on each of the four scales.

We have demonstrated that the hybrid examiner/automated metric model can predict with reasonable accuracy the consensus conclusions for suitability along four scales. This means that the model could be implemented successfully in operational laboratories to help streamline the analysis process, to assist in decision-making when a second opinion is sought (particularly in smaller laboratories that don't have access to sufficient experts to form consensus panels), and to provide guidance in varied applications, such as when to apply additional or reduced QA measures; how to rate the difficulty of marks for training, testing, and research; and to serve as an aid in testimony to describe the difficulty of the mark. Additionally, laboratories could adjust the probability thresholds of the model to better reflect their own priorities and operational constraints in how high they would like to set the threshold for declaring a mark of value, AFIS quality, or complex.

Beyond its usefulness for guidance in reaching suitability determinations for manual comparison, this model could also be incorporated into existing AFIS systems to assist with the AFIS suitability decision. Finally, because the model essentially predicts what the consensus opinion would be regarding the suitability of a mark, it can be used to support the decisions of individual examiners. If an examiner is challenged on why they reached a particular suitability decision, and they have used this model and found it to agree with their decision, they can claim that, within the accuracy of the model, their decision is likely to comport with that of the majority of experts. The model provides the examiner with a virtual consensus panel to which they may refer.

Acknowledgments

The authors would like to thank all the participants of this study who took the time to complete the trials.

4.2 Additional discussion on model development, optimization, and performance

This section provides additional insight and explanation of choices that were made and challenges that were encountered and resolved during the development and optimization of the model as well as discussion on the outputs of the model. These details were too in-depth for the peer-reviewed journal manuscript but are useful in understanding the evolution of the model and the learning that occurred.

4.2.1 Using ridits to average categorical responses

Early on in the analysis of the data associated with this project, we encountered the problem of what to do when we wanted to average categorical responses—such as participants' responses to the clarity question, which had 3 possible responses, or their conclusions on each of the 4 suitability scales, which ranged from 3 to 5 possible responses. How could we state what the average clarity conclusion or AFIS determination was without numbers that could be used to make calculations?

We briefly considered converting the ordered factors of the possible responses into a Likert-type scale whereby the bottom response was assigned 1, the next response 2, and so on, but this assumed that the categories were evenly spaced in weight, or importance, which was not an assumption we believed would hold, particularly for categories such as the Value scale where the 5 stops were No value, Some probative or investigative value, VEO, VIDO, and Value for Both. It was likely that there was less distance between Value for Both and VIDO, for example, than there was between others of the categories.

We found in the literature a possible solution, known as a ridit (Bross 1958). A ridit is a means of mathematically transforming an ordinal response to a numerical value based upon the distribution of a reference data set (in our case, the entire collected responses of the participants). Rather than presuming an equal distance between each category, the ridits assign a proportion to each category based upon its underlying distribution. These numerical transformations can then replace the categorical responses and be used in all standard statistical calculations as well as for ranking.

The process for calculating a ridit, taken from Bross, takes 4 steps to arrive at and represents the proportion of votes in all the categories below the one being evaluated plus ½ the proportion of the votes in that category. Essentially, it is a process by which weight is assigned to each of the categories relative to the overall distribution of the data. We built these 4 steps into a function in R and used that function to transform all our categorical data into ridits, which we then used to calculate mean values and standard deviations for those categories. The mean values were then mapped back onto categorical labels based upon where the mean value fell between the ridit-weighted categories to arrive at a consensus response for each ordinal variable.

Although conceptually, the use of ridits to determine consensus responses to ordinal variables seemed to be the way to go, they introduced some challenges in implementation and were eventually abandoned. The largest problem caused by the ridits was that we discovered later in the model development that the process of creating mathematical averages of ordinal data and mapping those averages back onto ordinal labels resulted, in some instances, in selecting a "consensus response" that nobody actually chose! This became important when we started omitting data from individuals who did not agree with the consensus ground truth, because there were instances in which NOBODY agreed with the consensus ground truth, which was not only philosophically problematic (how can it be the consensus if nobody chose it?), but also resulted in entire images being left out of the data analysis.

Once we made the decision not to use ridits, we re-assigned ground truth to all ordinal variables using a simple majority-voting scheme. This also had some disadvantages, as because there was no weighting to the categories, there could be situations in which the votes were very evenly spread across categories, yet the category that had just one more vote would be assigned as "ground truth," which might not fairly represent the breadth of participant opinions. Nonetheless, once this process was completed, we did spot-check the accuracy of the models in predicting the ridit responses versus predicting the majority-voted responses and found that they were comparable.

One other challenge with the majority-voting approach to assigning ground truth was the question of ties. In re-assigning the scale decision labels, there were ties for the most-voted conclusion 3 times on the Value scale, 4 times on the Complexity scale, 2 times on the AFIS scale, and 1 time on the Difficulty scale. These ties had to be resolved manually. It was done by looking both at the overall shape of the data (to which side did the majority of the dissenters from the two tied categories lie?) and looking at the image of the impression itself and selecting the category that seemed to encapsulate the opinions of more of the participants.

A similar process was done later with the determination of "average user data" wherein we had to determine what the consensus ordinal response was for each predictor variable from among the participants who agreed with the consensus ground truth per image/scale combination. These determinations were made for ties by again looking at the tails of the distributions and taking as consensus the one that was closer to the larger tail.

4.2.2 Unrealistically high performance and the genesis of the average user data

Early in the machine learning process, we were using all of the data collected in the white box portion of the study and randomly splitting it into a training set and a testing set. However, these runs using the automated variables only (using user-selected minutiae) were resulting in unrealistically high model accuracy—around 95% on all four scales! This didn't seem feasible and we were concerned about over-fitting of the model, so we began to look for reasons for this unexpected result.

We realized that the problem was stemming from the way the training and testing sets were being divided. Although we had 100 distinct images, each image had data from multiple users but the same image always had the same ground truth, regardless of which user's data were being considered. When the training and testing sets were being randomly defined, the same image would often end up in BOTH sets, albeit with different associated user data. This means that when the model was being tested on the testing set, it was seeing many (if not all) of the same images it had learned on in the training set, just with different user data to use as predictors. To make matters worse, because we were using the automated suite of variables, these did not change from user to user; only the variables predicated on the number of minutiae selected by the user were changing from case to case. This explained the unusually high performance of the model, and also alerted us that the training and testing sets needed to be formed differently.

Our first solution was a random selection of one user per image. In this scenario, we would randomly select the participant data from just *one* user for each of the 100 images, resulting in a data set of 100 images, each having a single user's data associated with it. One of these images would be left out as the test set and the other 99 would be used as the training set, hence adopting a leave one out cross validation approach (LOOCV). This would be done iteratively until all 100 images had had a turn being the one left out. This approach would use only one set of 100 cases (out of more than 3,000 data sets collected), which was not effectively exploiting the bulk of the data we had collected to use in training the model, even with us repeating the process of random selection and LOOCV machine learning several times to ensure the model was stable.

We also realized that by randomly selecting a single user to train the model on, not only were we throwing away a lot of usable training data, but we might randomly select a user who did not agree with the consensus ground truth for that case. It seemed nonsensical to train a model on the variables that led to a conclusion when the person who provided those variables did not provide the sought-after conclusion. Thus, the decision was made to first identify all users who did not agree with the consensus ground truth decision for each image/scale combination and remove their data from the training set entirely. This left us with a set of "super-users" who always agreed with the consensus conclusions. Thus, we could train the model to predict ground truth based on the observations of examiners who agreed with that ground truth.

However, because we were still only selecting the data from a single user who agreed with ground truth for each training iteration, we were still failing to use a lot of agreeing data to train the model. To rectify this, we decided to use *all* the data from the super-users who agreed with ground truth to train the model. We achieved this by taking their *average* responses for each predictor variable to be the single case that was used for training. Thus, we created the "average user data" set referred to in the Section 4.1 article. This both allowed us to take full advantage of all the super-user data available to us, and created a pseudo-data set that was distinct from the full data set from all these users *and* the non-agreeing users, allowing us a separate testing set that could be used later. Additionally, at this point we changed from LOOCV to a k-fold cross-validation for the machine learning.

4.2.3 The Value scale: five bins or two?

During the white box study (Chapter 3), we noticed that the Value scale data were strongly grouped in the NV and Value for Both bins. This was unsurprising as the community still largely thinks of value as a binary decision and even VEO, though more familiar, is not in common usage. At the time of the white box data analysis, it seemed encouraging that the three middle options on the Value scale were being used as much as they were—supporting the idea that these categories could be useful and one day might see widespread acceptance.

However, when we got to the machine learning portion of the research, the relative lack of use in the middle three categories became problematic. Although many participants utilized these categories, they did not do so frequently enough, or consistently enough on the same images, for them to very often be the *consensus* value decision for many images. In fact, it turned out that only 4 images of the 100 ended up being assigned a consensus ground truth decision that was *not* NV or Value for Both, and 3 of those 4 were in cases where ground truth was set manually, due to the need to break a tie!

This caused difficulties in the machine learning process. None of the models was very efficient at predicting these 3 intermediate categories, which is unsurprising considering they had no data from which to learn what a mark in one of these categories should look like. Thus, we had to make the decision to reduce the Value scale down to the only 2 categories we had data to support: NV and Value for Both.

Another potential challenge present in these Value data was that our distribution of responses was uneven. After removing the three middle categories, we had 96 images remaining with consensus responses of 17 for NV (17.7%) and 79 for VB (82.3%). When data are unbalanced like this, there is a risk in machine learning of the predictions being skewed toward the larger category simply because the algorithm sees more of it. For instance, imagine we were predicting whether someone was a professional basketball player based on their height and weight, but the training sample included only 3 basketball players with 20 non-basketball players. If we further imagined that there was overlap in the data such that some non-basketball players' height and weight fell within the ranges of basketball players, the MLA looking at these data would see more non-basketball players than basketball players in the ranges occupied by basketball players and could reasonably always predict a person was not a basketball player and be correct most of the time, according to the training data.

This phenomenon is known as having a rare event in your sample and is typically attributed to a data set in which one response happens less than 15% of the time. Because our consensus NV response occurred 17.7% of the time in our data, it is not *quite* considered a rare event, but we should nonetheless be cognizant that this imbalance could have affected the machine learning. In order to mitigate this effect, one could use the R function SMOTE (Synthetic Minority Over-sampling Technique)(Chawla et al. 2002), which synthesizes new minority class instances between the actual data of the minority class to better balance the data for machine learning.

We did not use this function in our data analysis, but the performance of the model may have been slightly improved by its use.

The need to reduce the Value scale to two categories is unfortunate for the thesis, because one of the aims of the research was to introduce these new stops on the scale (VIDO and Investigative or Probative value only) and use them as a way to advocate for policy changes. That is, it was hoped that we could demonstrate that there were marks that could be identified at the analysis stage as insufficient for a categorical conclusion, but comparable, that would be used only to generate investigative leads but would be blocked by policy from being used for the categorical conclusions. We also hoped to highlight the distinction between marks that were of value for ID but not exclusion from those that were of value for exclusion but not ID.

Although the white box use of these categories to some degree provides hope that they could be adopted one day, we found that it was premature to try to include them in modeling at this time. It appears that first, we must persuade the community of the usefulness of these categories and get them used to using them. Only after that is accomplished could we re-visit this categorization task with new images and examiners and hope to successfully build a model to predict these conclusions.

4.2.4 The importance of minutiae selectivity for suitability decisions

One major question raised by this research was whether other factors besides minutiae count factored into the suitability decision in a meaningful way. We argued in Chapter 3 that minutiae count cannot be the only determining factor, or we would see sharp thresholds at the “magical number” of 7 or 8. However, when minutiae count alone is not enough to tip the balance, what is the next most important factor that is considered? We posited that clarity must be a factor, but also that the perceived selectivity of the features must carry weight in the mind of the examiner. If a cluster of features is perceived to be highly distinctive, an examiner will give it more weight than features that are not, and that may be enough to tip the balance for a mark that is on the borderline of a decision threshold.

We collected data on the minutiae that were annotated, whether or not a target group was selected, and how many “distinctive clusters” of minutiae were annotated by our participants. However, although each of these data points gave some local information, we wanted a measure of the overall perceived selectivity of the mark. Thus, we created the variable “minselectivity,” which was described in Section 4.1 and took into account the ratio of confident total minutiae, the number of highly diagnostic combined minutiae groups noted, and whether a target group was noted. This served to combine selectivity and clarity by accounting for both distinctive features and whether the examiner had confidence that they were truly present as observed.

When the machine learning was undertaken, it became immediately apparent that this minselectivity variable was highly diagnostic. As described in Section 4.1, it was near the top of the variables of importance in nearly every model we ran in which it was included, often vying with total_minutiae for the top spot. This provided strong evidence that the distinctiveness of the features in an impression and the clarity surrounding those features are both highly diagnostic pieces of information that weigh very heavily in examiners' decision-making for all four scales, and may in fact be the key predictors of suitability decisions when minutiae count alone is not enough to make the decision clear.

Although it is unfortunate that this variable could not be tested in the validation study due to operational constraints, the fact that we have demonstrated its importance to the decision-making process is a significant contribution to our understanding of suitability. Additionally, the automated variable ESLR, which is also a measure of the rarity of the features in the mark, has been included in the validation study and the final model.

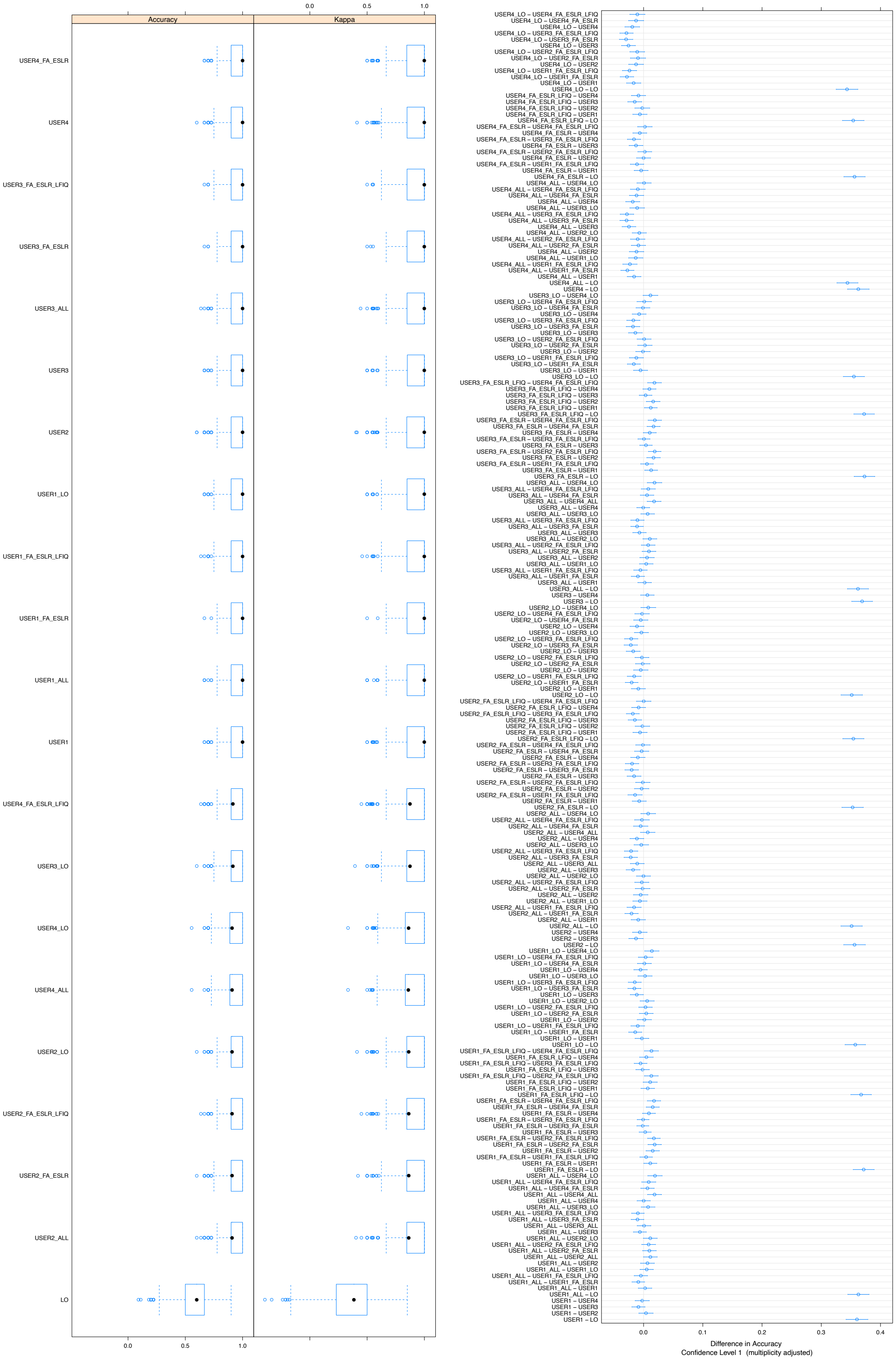
4.2.5 Significance of differences in model performance

In the manuscript presented in Section 4.1, we discuss the relative performance of different variable sets using the RF model under both idealized (Step 2 with "superuser" average user data) and real-world (Step 3 with "messy" data from all users) conditions. We summarize these data and demonstrate that there is little practical difference in the performance of these models, providing mean accuracy summaries (Step 2) and boxplots (Step 3) to compare them.

However, we also produced comparative plots for Step 2 in which differences in accuracy for each combination of variable sets were calculated, along with an evaluation of the significance of those differences. These plots were not included in the submitted manuscript due to space and readability, but an example of one (the Difficulty scale) is included here as Figure 19. In the plot, the difference in accuracy between each pair of models being compared is shown with error bars around it. The thin, grey vertical line at 0.0 indicates that there is no difference in the performance of the two compared models. If the blue dot with its error bars are far enough away from the grey line to not be touching it, that indicates a significant difference in performance between the two compared models.

Most pairwise comparisons of models do not show a significant difference in performance, with the exception of the LO models, which consistently performed poorly and are addressed in the manuscript in Section 4.1. Even for the pairings where there *is* a significant difference in performance, that difference is generally not far into the realm of significance. It was partially based upon these four plots that we made our claim in the manuscript that the performance of the models based on different variable sets (with the exception of LO) was all approximately equal.

Figure 19 Accuracy of models and pairwise comparison for the Difficulty scale.



4.2.6 Assessment of images along four scales

As was noted in Chapter 2, the idea behind the development of this utility tool was to break the suitability decision into four separate assessments because a single image could occupy different bins on each such that looking at suitability as four different scales could provide more nuanced information than just considering a value/no value decision. After developing and optimizing the model, we were in a better position to examine how this might look when applied to real-world images by working examiners.

We reviewed the model prediction results for the images used in the validation study and selected a few to illustrate the range of decisions across the four scales. Because the decisions of the model are based upon user inputs, there is no single "model prediction result" for each image; the results will vary per user. Thus, to get a single model result for each image, we utilized a majority voting scheme across the model predictions that were returned for each image, based upon the inputs of all the users who saw each image.

Naturally, there were impressions that scored in the top category across all four scales. For example, for img203 the model predicted "Of value; Non-Complex, Self-Evident; AQ; Low Difficulty" (Figure 20). For images like this, the four scales provide no additional information. Yet in a situation like this, one could argue that no model is needed at all because the high quality of the impression is so obvious. There are, however, impressions like img221 (Figure 21) for which the model predicted "Of value; Non-Complex, Document; AQ; Low Difficulty". For this image, the impression was still predicted to fall into the highest categories on three of the scales, yet fell down a category in the fourth. Looking at the two impressions together, it is clear that there is a large difference in their overall quality and quantity of information. Yet, with a simple value/no value determination, that would not be evident.

At the other end of the spectrum were the images that were of very poor quality. For the most part, such as with img008 and img009 (Figure 22 and Figure 23), the model predicted the lowest classification across all four scales. However, there were images, such as img032 (Figure 24), for which the model predicted a higher category on one of the scales. In this case, img032 has very little information, yet is relatively clear, which likely accounts for it being ranked as "Medium Difficulty" although it is still predicted not to be of value or AFIS quality

Figure 20 An impression at the high end of the suitability spectrum. The model predicted the categories for this image would be "Of value; Non-Complex, Self-Evident; AQ; Low Difficulty" as the majority predictions across all input from users who viewed the image in the validation study.

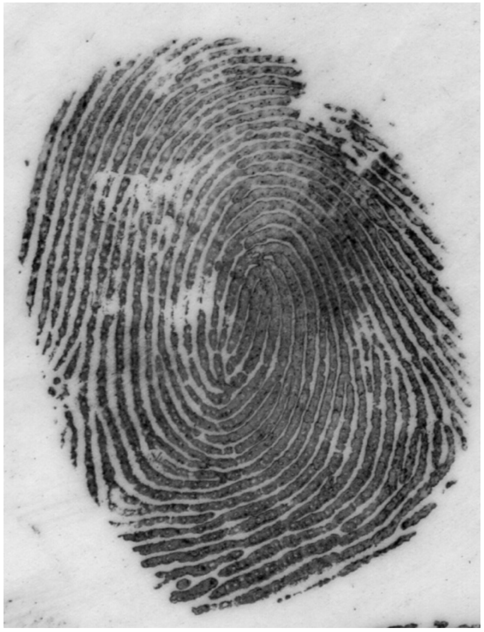


Figure 21 An impression not quite as high on the suitability spectrum. The model still predicted this impression to be "Of value; AQ; Low Difficulty" but for complexity, predicted its category to be "Non-Complex, Document" as the majority prediction across all input from users who viewed the image in the validation study.

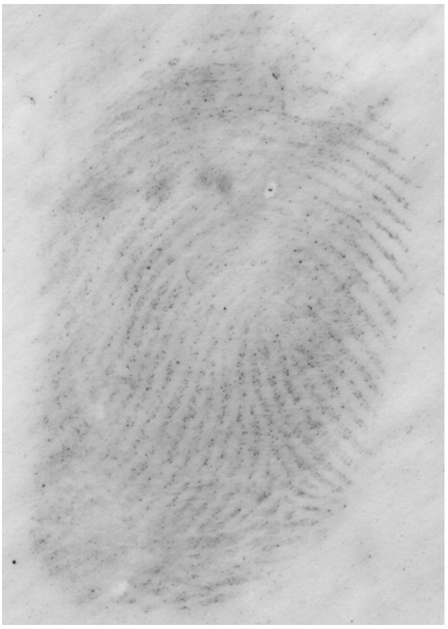


Figure 22 img008, a low-suitability impression that was predicted by the model to be in the lowest category across all four scales (No value; No value; NAQ; High Difficulty).

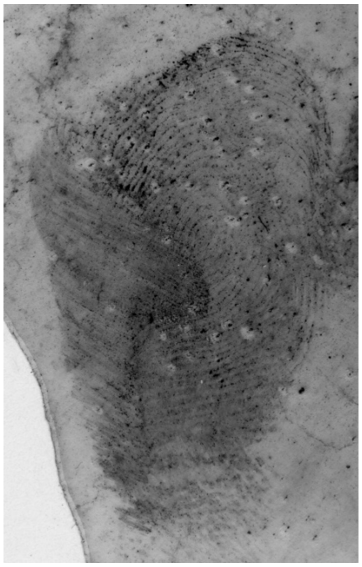


Figure 23 img009, a low-suitability impression that was predicted by the model to be in the lowest category across all four scales (No value; No value; NAQ; High Difficulty).

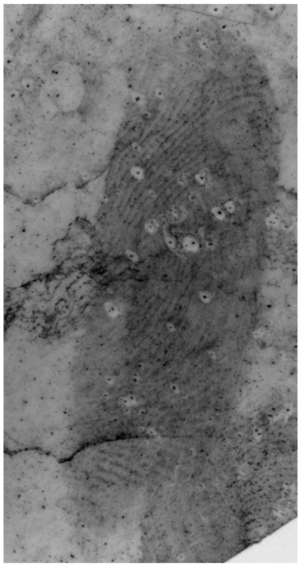


Figure 24 img032, a low-suitability impression with higher clarity that was predicted by the model to be in the lowest category across three scales (No value; No value; NAQ), but was also predicted to be categorized as Medium Difficulty.

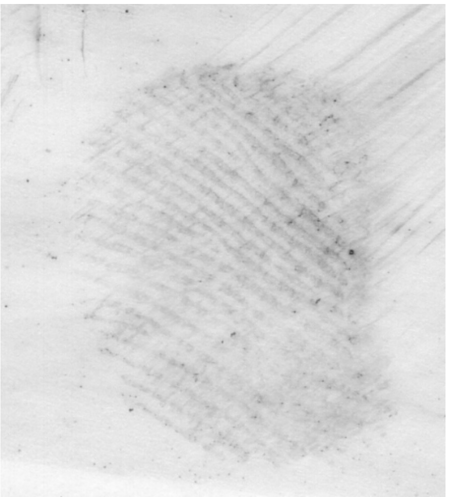
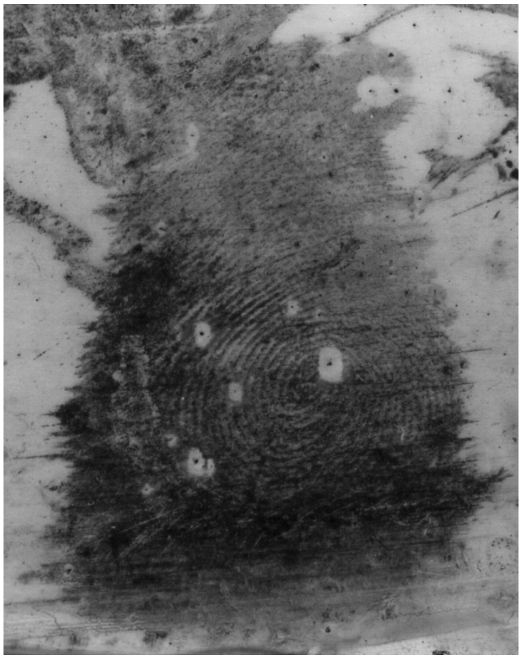


Figure 25 shows img007, another low-suitability impression that was predicted by the model to be NAQ and High Difficulty as with the impressions shown in Figure 22 and Figure 23. However, the model predicted img007 would be categorized as “Of value; Value, Complex” on the value and complexity scales. This illustrates well that marks that are more suitable for one application or another can “split” the scales, going lower on one scale and higher on another, which supports that a single suitability determination does not provide enough nuance on which to base full decision-making.

Figure 25 img007, a low-suitability impression that, despite predictions of NAQ and High Difficulty, is still predicted to be of value, albeit complex.



Toward the middle of the suitability scales, Figure 26 and Figure 27 present another pair of images for which the four-scale suitability model is able to distinguish between nuances on one scale that would be lost with a simple, binary suitability determination.

Img218 (Figure 26) was predicted by the model to be categorized as “Of value; Value, Complex; NAQ; High Difficulty” whereas img216 (Figure 27) was predicted by the model to be categorized as “Of value; Non-complex, Document; NAQ; High Difficulty.” Although it can be seen that the difference in quality between the two images is noticeable, the broadness of the categories on the value, AFIS, and difficulty scales allows the two images to be categorized the same, while one is judged as complex and the other as non-complex, but requiring documentation.

Use of the model in the validation study has demonstrated both that it is capable of making distinctions along single scales that would be lost using a binary decision model with only one dimension, and that there is value in doing so. In other words, examples have been shown of impression pairs that exhibit discernible differences in quantity and quality and that those differences have led to different utilities along one of the four scales that should be highlighted to enhance decision-making for examiners.

Figure 26 img218, an impression that was predicted by the model to be in the lowest category across three scales (No value; NAQ; High Difficulty), but was also predicted to be categorized as "Of value, Complex".

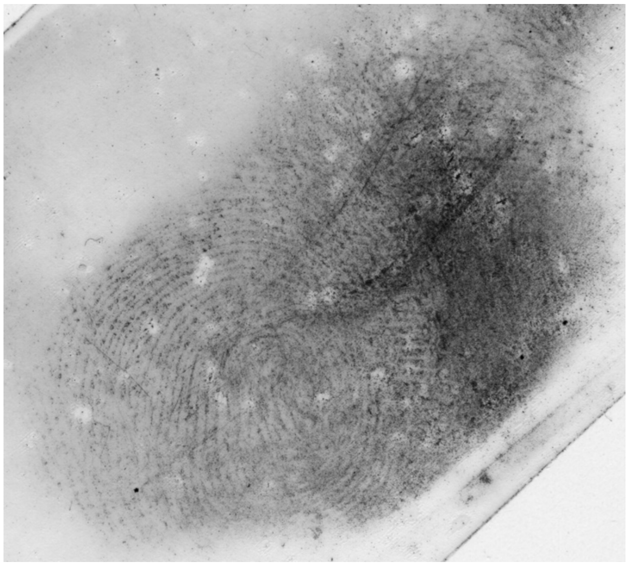


Figure 27 img216, an impression that was similarly predicted by the model to be in the lowest category for the value, AFIS, and difficulty scales, yet was predicted to be categorized as "Non-Complex; Document" on the complexity scale.

4.3 Chapter 4 summary

Chapter 4 presented a submitted article for publication that described the development, optimization, and validation of a model to predict consensus suitability decisions on each of the four previously described scales. Chapter 4 also dove deeper into the rationale behind many of the choices made during model development, including describing many challenges that were faced during development that informed those choices. Finally, Chapter 4 discussed the nuances provided by the four scales, which allow one to differentiate between marks of different qualities that may be more or less useful on one of the scales than the others.

The predictive model was developed in two stages: Model Development and Optimization, and External Validation. Stage One was comprised of three steps: Model Selection, Ideal Performance Testing, and Model Optimization in Realistic Applications.

In Step One (Model Selection), a wide range of potential MLA solutions were tested using all the variables available. These included automated variables and user input variables, both from observations of the impressions and responses to demographic and policy questions. MLAs were tested in different combinations of user input variables, automated variables, and both types of variables combined. RF was chosen as the MLA that would be optimized and used for the model. From the full range of possible predictor variables, those that showed the highest variable importance for predicting in RF were retained for optimization in Steps Two and Three.

In Step Two (Ideal Performance Testing), the RF model was evaluated with different combinations of user, automated, and combined predictor variables from those that were retained from Step One. In Step Two, the average user data were used, which was made up of averaged responses from all the “super-users” who agreed with ground truth. This represented a best-case scenario for accuracy of the model and gave us an idea of its potential for accuracy.

In Step Three (Model Optimization in Realistic Applications), the RF model was once again evaluated with different combinations of user, automated, and combined predictor variables from those that were retained from Step One. This time, all data from all users were included in the modeling, not only data from “super-users”, and no averaging of responses was done. This represented a more messy, real-world scenario where not every user will agree with the ground truth. In Step Three, we identified a set of variables that performed best under “Optimal” conditions, meaning that we put no constraints on the amount of user input or computational resources required, and a second set of variables that performed best under “Operational” conditions, where we tried to minimize the user in put necessary to achieve the highest accuracy and also tried to keep computational resources realistic.

Under these conditions, we found that the best-performing Operational model used the variable set “User4_LO,” which included the user-input variables total number of minutiae marked, distortion, and clarity and the automated variables total number of minutiae, fa, ESLR, and lfiq (all using auto-extracted minutiae) . This model achieved median predictive accuracy levels for the four scales of: AFIS-73.33%; Complexity-60.36%; Difficulty-66.67%; and Value-86.21%. These results were comparable to the accuracy of users in making suitability determinations that matched the consensus response.

These results support both that a model can be developed and optimized that is comparable to the performance of human examiners, and that a hybrid model that uses limited user input supplemented by limited automated measures is superior to either automated measures or user input alone.

For Stage Two, we performed an external validation study using new users and a mix of new and old images to test the generalizability of the optimized model. In the validation study, users input only the three user variables (total number of minutiae, and distortion and clarity assessments) and these were combined with the pre-calculated automated measures necessary for the model to arrive at predicted consensus outcomes. Participants were then asked whether they agreed with the prediction of the model on each of the four scales. Agreement between participants and the model was generally high, supporting both that the model is doing a good job at predicting the consensus opinion and that friction ridge examiners are likely to accept the conclusions of the model in their daily practice.

Real world applications and benefits of the model were discussed, including that the limited user input required can save time spent in analysis, that the ability to predict a consensus response can be useful for laboratories too small to form consensus panels, and that the additional nuances provided by the four scales can support good QA policies and practices.

Section 4.2 discussed choices that were made during the development of the model, such as the inclusion, and then discarding, of ridits to average categorical responses; the use of the average user data; and the reduction of the Value scale from 5 categories to 2 and the ramifications of that decision. It also discussed the importance of minutiae selectivity in decision-making and examined more closely the comparative performance of all the considered models before the final model was selected.

Finally, example images were provided that illustrated the nuance given by the use of the four scales of suitability. Impressions that have visibly different quality and quantity, but which would fall into the same category in a binary (value/no-value) scale were separated into different categories, depending upon which scale was being considered. These examples support that there is utility in using the four-scale model to consider different dimensions of the suitability of a mark.

5 Limitations, Potential Impact, Recommendations, and Future Research

5.1 Limitations

The research undertaken during this project spanned a period of more than 8 years from concept to publication. During this period, changes in practices in the friction ridge community, changes in the features available in the PiAnoS software, and new insights gained from initial data explorations incrementally but continually changed the course of the project. With the benefit of hindsight, there are choices that could have been made differently. There are also some limitations of the study that are simply the nature of the beast and could not have been changed under any circumstances. In this section, we briefly discuss some of the limitations of this work.

5.1.1 Ground truth is neither known, nor objective

One limitation that cannot be overcome is that there is no true ground truth for the suitability decision, in any of its facets. Because suitability is, essentially, a subjective determination, we had to develop a means of assigning the expected outcome of each decision that we could use as the standard against which to measure the performance of our candidate models. We elected to use the majority-voted response as "consensus ground truth." That decision has advantages and disadvantages.

Of course, the advantage was that for each decision, we had a conclusion that *most* examiners would agree with to use as guidance. The flipside is that often, many examiners would not. The biggest disadvantage of using consensus responses in the development of the model is that we lost a lot of nuance in examiners' observations. When a mark sat at or near the threshold of a decision, the individual observations of examiners could be illuminating in understanding what specific factors helped to push them one way or another (for example, the image presented in Figure 13 in Section 3.2.2 makes one wonder whether the 2 highly distinctive dots near the bottom of the image could outweigh the lower overall quality and quantity of minutiae). Because we were averaging examiner observations for the purposes of the modeling, these nuanced tradeoffs were lost.

Exacerbating this challenge is the fact that not all examiners are equally skilled. By using a majority vote approach, the opinions of all examiners were given equal weight. However, it is quite possible that some examiners were uniformly poor performers and the model may have been improved by their exclusion from the training data. We could have used an assessment process to identify the most trustworthy examiners to include in the voting process, while excluding the least trustworthy examiners. This could have been achieved by reviewing the annotations and responses of each participant and looking for poor documentation habits, incoherent responses in light of the annotations made, or written notes that revealed poor practices or logic. However, this process would have been extremely labor-intensive and in addition may have resulted in a detrimentally reduced set of data for training.

Additionally, there were challenges in the modelling that were caused by the majority-voting approach to assigning consensus ground truth, such as the fact that there is disparity in the community around how to treat VEO marks (as NV or of value) which could push these marks one way or another based on the luck of the draw of which examiners happened to view them. Another challenge this caused was the fact that some of the original 5 bins proposed for the Value scale were never the consensus decision, and thus could not be modeled at all. Finally, the AFIS scale results, at least in the external validation study, may have been skewed by the opinions of examiners who noted that they did not regularly perform AFIS work in their job duties.

5.1.2 Garbage in/Garbage out

A related limitation of the study that is the nature of the beast and could not be mitigated is the principle of garbage in/garbage out. One of the reasons this study was undertaken was as an attempt to reduce the variability between examiners during the analysis phase. However, that very variability limited the usefulness of the resulting model. Machine learning is only as good as the data it is provided on which to learn. Because the participants in the study were so variable in both their observations and their suitability determinations, there were limits to how well the model could learn to predict their decisions based upon their observations. The fact that it predicts as well as it does is astonishing on some level.

5.1.3 Comparison outcomes are not known

This study was focused exclusively on the suitability determination and did not consider comparison outcomes at all. Although the scope of the project was large enough without comparison entering into it, it is a limitation of the study that we were not able to correlate performance of the model with comparison outcomes. It would have been beneficial (and would make for a good follow-up study) to take marks with known ground truth pairs, use the model to determine their suitability, and then subject examiners to comparisons to see, for example, whether the marks designated as “complex” or “investigative or probative value only,” were more likely to lead to erroneous comparison conclusions.

5.1.4 The four suitability scales are new, and unfamiliar

One of the goals of this research was to introduce the idea of four different scales of suitability and also to introduce some new categories within scales that were already familiar. We have done this and the data support that these new categories were utilized and may have value to the community. However, there was a risk that by their very novelty, they may be used inconsistently because they were unfamiliar to examiners and represented a large cognitive load in keeping new definitions in mind and applying them as they performed an otherwise-familiar task. Some text comments received from participants or inconsistencies in their responses indicated that this may, in fact, have been a problem for some. It was clear that some examiners did not fully understand the definitions or intent of the newer categories.

Balanced against this is the fact that multiple previous research studies have shown that examiners are not consistent in their application of the value decision anyway—thus, the effect of introducing new categories to this decision may have been negligible and, we hope, the benefits offered by these categories outweigh the learning curve necessary for their use.

5.1.5 Observation of the cores and deltas was not made

During the initial experimental design phase of this project, the decision was made not to ask participants to record whether they observed cores or deltas in the impressions. The reasoning at the time was that we were already asking them whether they could discern the pattern type, which was the main reason you would need core and delta information, so this question was unnecessary and we were asking participants to annotate *so much* information, that we did not want to overburden them with additional questions that did not add materially to the project.

In retrospect, the project would have been strengthened by collecting these data. First, they would have allowed us to know whether participants were confident of the orientation of the mark (as would an additional question specifically targeting orientation knowledge). Second, we failed to account for the importance of the core in AFIS decisions. The model sometimes struggled to predict AFIS decisions, calling a mark of AFIS quality when it had high clarity and number of minutiae whereas many examiners would call the same mark NAQ because it was missing the core. Because the model did not have information about the presence of a core and its importance to AFIS decisions to learn on, it was unable to correctly classify in these situations.

5.1.6 Technological challenges to real-world implementation

Finally, there are technological challenges that limit the implementation and optimization of the model for use in operational laboratories. Because the model draws on several sources of information, these have had to be amalgamated across multiple machines for each of the study images. There is not currently a single standalone version of the model into which a user could simply upload an image and get a real-time response. A platform to combine all the parts of the model into a single user interface that could be utilized in working laboratories could theoretically be built, but at a cost of time and money that is beyond the scope of this project.

Similarly, because the development of the model took place over several iterations of the development of PiAnoS, some of the early features, such as the ability to annotate target groups and highly distinctive minutiae clusters, were lost. This negated our ability to include the minselectivity variable into the final model. Although the performance of the model is not much worse without this variable, minselectivity did prove to be a variable of high importance and it would have been better to include it in the final model.

5.2 Potential impact to policy and practice

The development of a utility tool for the suitability decision can impact friction ridge policy and practice in a number of ways. This section will outline the five main areas of benefit considered by this study, delineating the challenges each faces and how this tool could be used to foster improvement.

5.2.1 Research

Since the release of the NRC Report in 2009 (Edwards 2009) and the PCAST report in 2016 (President's Council of Advisors on Science & Technology 2016), the one thing that fingerprint examiners and critics alike seem to agree upon is the need for more research. These reports demanded answers to a variety of questions related to fingerprint comparison science, including questions regarding error rate, standards, proficiency, bias, and others. Researchers are developing experiments that will provide the data to answer these questions and, hopefully, will strengthen the foundation upon which friction ridge comparison science rests.

Most, if not all, of these research projects will hinge upon test subjects (typically examiners in the field, but sometimes also novices) examining some number of known and unknown impressions and reaching some conclusions about them under a variety of circumstances. The difficulty scale of this tool could be used to evaluate potential study marks prior to their being included in research test sets. This would lead to the ability to more directly and meaningfully compare the results of studies from different researchers. The use of the utility tool could provide consistency from researcher to researcher on what constitutes a "difficult" mark. The other scales could likewise be used to give researchers uniform expectations of which marks in their study are likely to be classified as VEO, VID, no value, AFIS quality, complex, etc. Thus, if a researcher were to say for example, that their study included 10 low, 10 medium, and 10 high difficulty marks, there would be a common understanding within the field of what this means.

Most often today, research focuses on the number of same source or different sources trials without any discussion at all around the difficulty of the impressions (see, e.g. (Tangen et al. 2011), which compares the performance of experts and novices, yet makes no statement about the difficulty of the images used; the sole figure presenting example images from the study (their Fig. 1) shows three quite easy comparisons), or simply a label such as "difficult" without an accompanying definition of why a particular mark might be categorized this way.

5.2.2 Proficiency Testing and Training

Similar to research, proficiency testing and training exercises would benefit from the application of a utility tool to their test samples. There is ample research in the literature on the cognitive psychology behind learning to support that people learn best when training samples start off easy and become progressively more challenging to promote learning from errors (Metcalfe 2017). The use of a tool to judge the difficulty of the images being used in training would allow the creation of increasingly difficult training samples to optimize the training progression.

Companies such as CTS sell proficiency tests to laboratories throughout the United States. One of the criticisms of these tests is that they are too easy (Max et al. 2019; Koertner and Swofford 2018) and indeed, the head of CTS has gone on public record at the National Commission on Forensic Science (NCFS) stating that the tests are crafted to be easy because that's what the customer demands (National Commission on Forensic Science (NCFS) 2016). However, these test results are often invoked in court as proof of an examiner's continued competence and expertise.

There is currently no stated standard of how many low, medium, or high difficulty marks are given on commercially available proficiency tests. Without this information, it is unclear what level of proficiency is being tested. For the results of these tests to be meaningful, proficiency test companies should be employing a standard metric when designing their tests to ensure that each year, a pre-determined number of marks from each category of difficulty is given to test-takers. Only by doing this can there be certainty that the test was of sufficient difficulty each year to warrant a claim of "proficiency" by its successful completion. Again, the difficulty scale of the utility tool developed by this research could provide a uniform way to make these assessments when constructing tests.

In addition, an improvement to current test designs would be to design tests of differing difficulty, or to award proficiency at differing levels, based upon the level of difficulty of the marks that were successfully completed on the test. This would serve a number of useful functions.

First, it could change the entire culture around the notion of "passing" a proficiency test. Rather than having a test that was pass/fail with a stigma assigned to any examiner who failed, it would be expected that most examiners would fail at some point—the purpose of the proficiency test would be to identify that breaking point. Rather than saying they passed the test, examiners would be able to state what level of proficiency they had attained—that is, what the examiner's accuracy was at each difficulty level. This approach to proficiency testing would be in line with recommendations by the National Commission on Forensic Science (NCFS) that the discipline should be testing their examiners and systems to identify the limits of their abilities, not just to see if they possess the minimum required level of competence (Bell et al. 2018).

Second, it would give both managers and juries a clearer idea of the skill level of individual examiners. For managers, this information could be used to identify examiners who would benefit from additional training, or to select the most skilled examiners for particularly difficult or high-profile cases. For jurors, it could give a sense of how concerned they need be about the likelihood of an error made in this particular case, by this particular examiner, rather than trying to extrapolate from some general error rate for the field, a challenge that has been explored by Dror (2020). This brings us to our next point.

5.2.3 Testimony

As previously noted, currently all identifications are treated as equivalent in value and certainty. However, it has been shown that more errors occur in the comparison of marks that are of marginal quality (Langenburg 2012). Furthermore, not all examiners possess equal skill in comparing complex marks. Therefore, it may be helpful to factfinders to be given information regarding both the level of difficulty of the specific mark in a given case, and the level of skill of the examiner testifying to it.

The utility tool can be useful on both counts. First, by applying it to any mark on which one is going to testify, the examiner can provide the factfinder with information regarding the difficulty of the mark and whether or not it was considered complex. The examiner may give an idea of how much concern or care is warranted in the interpretation of the results, based upon the quality of the mark. This may aid the factfinder in determining how much weight to attribute to the evidence. In the case of a poor quality mark, the examiner may be able to demonstrate why he reached his conclusion sufficiently to allay any concerns the factfinder may have had—but by presenting the quality information, they will at least know that they ought to have had concerns to begin with. Second, if annual proficiency tests have been designed to test examiners at differing levels of difficulty, as suggested above, the examiner can reassure the factfinder of her accuracy rate on proficiency test samples at the requisite difficulty level to be competent to have examined the mark at issue in a particular case.

This application of the tool is in line with the suggestions of Mnookin (2010) that testimony might be offered in conjunction with proficiency test results. She calls for the development of a metric for use in designing proficiency tests. She specifically recommends that this metric of difficulty be validated and be applied to proficiency tests in such a way that the level of proficiency of an examiner may be ascertained and compared to the level of difficulty of the prints being presented in a particular case.

Under the prevalent two-decision model of friction ridge conclusions (Identification v. Exclusion), testifying to one's conclusion has been relatively straightforward, if somewhat lacking in transparency. However, with the emergence of more nuanced decision-making, friction ridge examiners will need to re-think the way they present their findings to a jury (Cole 2011, 2014; Carter et al. 2020).

With increased documentation, it will be easier to recall and report the thought processes that went into a conclusion, as well as to increase transparency about the process. However, it is unknown what effect this information will have on a jury. Will it help to clarify matters for them, or only serve to confuse them? Furthermore, is the same level of detail in testimony necessary for all marks, or can an abbreviated explanation suffice in cases where the mark is of high quality? The utility tool can once again aid in establishing thresholds. At low difficulty or complexity levels, abbreviated explanations may be more appropriate.

With the recent push for greater transparency in testimony and greater modesty in conclusions, it seems desirable to present the jury with as much information as possible regarding how conclusions were arrived at, how strong the evidence is, and to give conclusions that do not overstate the significance of the evidence. However, research is needed to establish appropriate language to achieve these goals without simply bogging the jury down with unneeded and confusing technical information. As Lennard puts it (2013):

> “We justifiably strive for scientifically defensible means of presenting our evidence, but we arguably need an approach that better meets the needs of the court, including the jury, the judge, and the other legal practitioners involved in the process.”

In other words, just because greater technical accuracy satisfies good scientific principle, it doesn’t necessarily follow that it makes the information more understandable to a jury. We must remember that explaining our findings to the jury in a manner that aids them to understand is the ultimate goal of testimony, and that most of them are not scientists themselves (Eldridge 2012, 2019).

As it is currently unknown if information regarding quality scores, weight attributed to rarity, discipline error rates, significance of findings, and even explanations of the limitations of those findings will tend to enlighten or confuse the jury, further research in this area is clearly needed (Eldridge 2019; Langenburg 2012; Garrett et al. 2020). Recent research has however demonstrated that jurors appropriately adjust the weight given to fingerprint testimony according to hearing how well the examiner in the case has performed on proficiency tests (Mitchell and Garrett 2019). Happily, previous research on juror perceptions (Holmgren and Fordham 2011) tends to show that jurors appreciate experts’ attempts at humility, admitting errors, and appearing human, which seems to indicate that the inclusion of greater transparency and more modest claims would be welcomed and would not lead jurors to lose faith in the experts’ expertise, as some practitioners have feared.

5.2.4 Quality Assurance (QA)

Many laboratories are striving to improve their quality assurance measures and documentation, in order to provide additional safeguards against errors. One example would be to require additional documentation or review of complex marks. However, it is difficult to write policy that is predicated on terminology that has not been defined—how does one require under written policy that an additional verification be done on all complex marks, for example, when the agency has not defined what constitutes “complex”? On the other hand, marks of very high quality could be subjected to fewer QA policies, which could save laboratories time that could be better allocated to more difficult marks. The complexity scale of this utility tool is capable of categorizing marks as complex. It also separates non-complex marks into two categories to distinguish between those that should require a standard level of documentation, and those that are so superior in quality that reduced documentation can be supported.

Additionally, the utility tool could be used to review decisions made by examiners as continuing performance review and to test the sensitivity and risk appetite of employees. It can also assist during disputes between examiners as will be outlined in Section 5.2.5.

5.2.5 Providing a Consensus

The development and output of the utility tool was discussed in much greater detail in Chapter 4, but it is important to note here that the tool works by predicting the consensus response of a group of examiners, not an individual response. In other words, individual examiners may disagree with the model in many individual cases; however, the model will generally successfully predict what the consensus of a group of experts would be for a given mark along the four scales. This information can be useful in several ways.

5.2.5.1 Guidance

The tool can be used to provide guidance in borderline cases, or cases where there is disagreement between examiners. If an examiner reaches a suitability decision and finds that the tool did not agree with them, it is an indication that a panel of experts would likely not agree with them either, which is a red flag that they should reconsider their initial decision. The model is not always "right," so examiners should not use disagreement as an indication that they *should* change their initial decision—only that it may be worth further consideration.

Furthermore, if two examiners have a disagreement about the suitability of a mark for a particular use (value, AFIS entry, or whether it is complex), the model could be used as a 'tie-breaker' in the sense that whichever examiner it agrees with is likely to be the examiner a consulted panel of experts would also agree with.

5.2.5.2 Small Laboratories

Many laboratories in the United States struggle with disagreements because they have only one to three examiners, so there is often nobody else available to ask when there is a difference of opinion. Although larger laboratories can set up consensus panels to help settle disputes, small laboratories do not have this luxury. For these laboratories, the tool can function as a virtual consensus panel, providing a consensus response when there are not other examiners available to consult.

5.2.5.3 Utility Functions

Finally, the utility tool could be adjusted to meet specific agency preferences according to their utility functions. As noted by Biedermann et al. (2008), agencies and even particular examiners will have different priorities for threshold-setting based upon the makeup of their predominant workflow and their risk appetite. For example, agencies that do high volume crimes and only search through AFIS may set a higher threshold for value or AFIS quality than agencies that deal with a large number of homicides. Agencies or examiners who have recently had high-profile errors exposed might become very conservative and pull back their value and complexity thresholds to reduce the number of challenging marks they compare and increase the number of comparisons that will be subject to additional QA measures. The Dutch experts discussed by Langenburg (2012) put a high premium on consistency and set very stringent criteria for minutiae selection, with the result that fewer marks may be considered of value in comparison to other laboratory systems. And the "eagle eye" examiner in many laboratories may be willing to compare marks that most of their colleagues would consider complex or not of value at all.

Because the performance of the model is based upon a default probability threshold of how likely we want it to be that the observed data predict a particular consensus ground truth classification, this threshold can be adjusted to reflect how loose or tight an agency wants the predictions to be. For instance, if an agency wanted to minimize risk of making an error, they could set the tool to only predict "Of value" when the probability of the data leading to an "Of value" prediction surpasses some very high threshold, such as 80% or more, with the understanding that they are allowing many marks to go uncompared that could potentially have been identified. Of course, the overall accuracy of the model will change as the thresholds for classification change. The range of these values can be well-represented by ROC curves, which were presented in the paper reproduced in Section 4.1.

5.3 Recommendations for policy and practice

The results of this study support several recommendations for changes to policy and practice to reduce variability in analysis decisions and improve friction ridge analysis generally. These recommendations are outlined briefly below.

5.3.1 Do not annotate with minutiae type-specific markers

As noted in Section 3.1, our results showed that examiners are not consistent or cohesive in their use of minutiae marker types. Markers of high-confidence type were used in areas of poor quality. The same minutia was often marked by some examiners as a ridge ending and others as a bifurcation (at high confidence), even in areas of high clarity and the "votes" for each were often very close to evenly split. Examiners frequently used markers of high confidence type for clear cases of connective ambiguity. These results support that there is really no justification for designating a particular minutia as either a ridge ending or a bifurcation. It is clear that, in most cases, the examiner can't tell one from the other with certainty and therefore the designation is arbitrary. Until and unless criteria are created that clearly specify when a minutia is a ridge ending versus when it is a bifurcation, and examiners are trained to those criteria, we recommend that specific minutiae type markers not be used as they imply a level of certainty that cannot be supported.

5.3.2 Develop consensus-based standards for suitability decisions

The high degree of variability in suitability decisions demonstrated in Section 3.1 makes it clear that the decision is currently far too subjective and standards are needed to guide examiners in these decisions. There are three main points where high levels of variability can be introduced: deciding what features "count" toward the decision; deciding how much weight those features should get toward the decision; and deciding how much total information is needed to reach the various decision thresholds.

Ideally, all three of these points should be resolved through research to determine standards and thresholds that minimize both variability and risk of error. However, this research is tricky to do and it may be some time before evidence-based guidance is available. In the meantime, these decisions still need to be made and should be consensus-based, whether at the agency level, or by guidance bodies such as OSAC. There need to be written and standardized definitions for what level of clarity is necessary for a feature within that local area to "count" toward suitability, for how heavily to weigh different types of features, and for how much information is needed to cross each threshold. This is the only way that variability can be reduced and suitability determinations can be reliable rather than subject to the whim and style of the individual examiner.

5.3.3 Document analysis, including confidence level

The best way to support the suitability decision, and to allow others to review the factors that went into that decision, providing transparency, is to document the analysis of the mark. Using the model will already involve documenting the minutiae relied upon as well as the overall clarity and distortion assessments for the mark. If the recommendation made in Section 5.3.2 is followed and standardized criteria are set around definitions of features, weights, and thresholds, documentation will be the only way to review work and ensure these criteria have been met.

Additionally, the confidence the examiner has in the minutiae selected should be documented. Since the weight assigned to features should be dependent partially upon their clarity, or the confidence the examiner has in them, this confidence should be documented. First, this will provide support for the examiner's decision. Second, it holds the examiner accountable to their initial confidence in the feature. If, for example, the examiner marks a minutia at high confidence, then begins comparison to a print, fails to find that minutia, and decides to discount it and make an ID anyway, that should raise a red flag to reviewers. Examiners should not be discounting features during comparison that they were highly confident in during analysis. But without documentation of the confidence level at analysis, this would be impossible to detect. Conversely, if a minutia is initially marked with low confidence but is not found, or its position or type are changed, during comparison, this should be met with tolerance by a reviewer, knowing that the examiner was not confident of the feature in the first place.

5.3.4 Use the model as a second—or first—opinion

The performance of the model developed in this research has shown that it does a good job of predicting the consensus opinion of expert examiners along all four scales of suitability. In all cases, it performed approximately as well as examiners did on average. In addition, examiners in the external validation study had a very high rate of agreement with the conclusions of the model.

Taken together, these two findings indicate that the model's guidance can be taken as an indication of the suitability determination that *most* expert examiners would agree with as a consensus. This means that if an examiner reaches a determination that does *not* comport with the guidance offered by the model, this should serve to them as a warning that a majority of experts would likely not agree with their decision and they may need to re-evaluate that decision more closely. In fact, the model could replace a verifier in the analysis phase. An examiner could form their own analysis opinion, then use the model as a check, to provide a warning when an examiner is about to render a decision that should be more closely examined, or conversely, to reassure an examiner that the majority of experts would agree with them if they are not entirely confident in their own decision but it comports with the prediction of the model. The model could be used in this way to determine when the intervention of another examiner is needed; if the examiner and the model agree, no further action is needed, but if they disagree, the opinion of a second expert will be sought (Montani et al. 2019).

This could even be taken a step further. Not only does the model tend to reach the same decision as a consensus of experts, but it does so more quickly than a single examiner completing a full analysis on their own. Thus, we can foresee a workflow in which the examiner uses the model *first* to reach their suitability decisions (keeping in mind that the examiner has to enter 3 inputs based upon their own observations, so they will be engaged in the analysis process and will form their own opinion in so-doing) and only completes a full analysis if they do not agree with the judgment of the model, in which case they must justify why their judgment is superior by documenting specific information they used to reach their decision, to which the model did not have access (Montani et al. 2019). This scheme utilizes distributed cognition (Dror and Mnookin 2010) to share the load between examiner and automation; the examiner would see and recognize features, while the model would make decisions based on the input information. This would result in savings of both analysis time and cognitive load while reducing variability in ultimate decisions.

5.3.5 Use the new categories proposed by this research

A number of new decision categories were introduced by this research across several scales. These included Investigative or Probative Value Only, VIDO, AQ with QA, Non-Complex with standard documentation, and Non-Complex, Self-Evident (reduced QA/Documentation requirements). Each of these categories was included for a specific purpose that either encouraged a new way of thinking about suitability, or suggested QA policies to accompany it.

Although some of the Value scale categories were rarely or never chosen as the consensus response, all newly-introduced categories were nevertheless chosen with encouraging regularity by most of the participants, as detailed in Section 3.1. We recommend that laboratories consider adopting the use of these new scales and categories to expand the way examiners think about suitability and to support QA policies. In particular, we recommend the adoption and expansion of the AFIS and Complexity scales.

The new category on the AFIS scale, AQ with QA, was very frequently chosen by participants in both the original and external validation studies and was supported by many written comments from participants stating that they believed particular impressions were suitable for AFIS entry, but should have additional QA measures applied to them due to observed risk factors. Some of the comments even specifically repeated recommendations of the author, such as additional documentation or using additional minutiae that had *not* been entered into AFIS to effect an ID. It seems that there is a practical use for this category and that the community is ready to embrace it.

The two new categories on the Complexity scale both dealt with marks that were *not* complex—that is, the Non-Complex with standard documentation and Non-Complex, Self-Evident. From participants' behavior and written comments, it is clear that the distinction between these two categories was not 100% clear to them. This confusion likely stems from the fact that documentation practices currently vary so widely between laboratories that for some participants, *any* documentation felt like an extreme measure whereas for others, so much documentation is routinely done that they could not conceive of doing *more*. For these categories to be effective, laboratories will first need to adopt a policy that standardizes the expected amount of documentation for a run-of-the-mill mark. From that point, they can then require *additional* documentation for complex marks and *reduced* documentation for the self-evident marks. Making this change should reduce errors, increase transparency, and focus the laboratory's limited time and effort where it belongs—more on complex marks, somewhat on mid-range marks, and very little on the highest quality marks.

5.3.6 Separate the notions of suitability for manual comparison versus AFIS entry

In many laboratories, there is a philosophy that if a mark can be compared, it can be entered into AFIS. And as we have seen, much of the research focused on quality metrics and the suitability decision has focused on AFIS, either equating the two, or ignoring the needs of manual comparison. Yet, we have seen throughout this work that the AFIS and manual comparison suitability decisions are *not* the same. Marks that are suitable for manual comparison may not be suitable for AFIS entry. And the criteria for AFIS search may be different than those for manual comparison (e.g. wanting a core for AFIS, higher minutiae thresholds, etc). In addition, AFIS thresholds may be affected by factors such as the size of the AFIS database (a mark that may be sufficient for searching in a small database may be too risky in a large one), or the AFIS vendor an agency uses (some matchers are "better" than others and thus may carry a reduced or heightened risk of coincidental match). These variables were not directly considered in this research but may nonetheless have an impact on individual agency or examiner AFIS decision thresholds.

Thus, we recognize that these are two fundamentally separate decisions that should be treated as such, even if the results often coincide. Laboratories should consider establishing separate thresholds for AFIS entry than those they use for a suitability decision for manual comparison and may even wish to consider establishing separate AFIS thresholds for different situations.

5.4 Future Research and Development

The research presented in this thesis has provided a foundation to understanding the most diagnostic data considered by examiners in making suitability decisions and has provided a proof-of-concept that a hybrid examiner-automation model can be used to predict suitability decisions along four scales, which it has introduced, with an accuracy level commensurate with examiner abilities. Further work could be done to improve and expand upon this concept and to make it implementable in operational forensic laboratories.

First, the model could be improved as noted in Section 5.2 with the inclusion of core and delta information, with the inclusion of the tools needed to calculate the minselectivity variable, and by being built into a single, standalone system into which examiners could upload their casework images and receive real-time model predictions.

Second, the predictions produced by this model could be combined with other user input and used to develop specific thresholds and criteria to classify marks according their level of suitability. These criteria could then be tested on images with known ground truth against AFIS search ranks and other existing quality metrics to measure their improvement in ability to reliably classify impressions, particularly in the "grey area" where it is less clear whether or not the mark should be considered of value. This next step of the research is already underway.

Finally, images with known ground truth pairs could be evaluated by the model, then given to examiners for comparison. Comparison outcomes could be compared to the suitability decisions predicted by the model to evaluate whether the model correctly classified marks that were at higher risk of erroneous comparison conclusions (e.g. complex, or no value, marks).

6 Conclusion

The friction ridge comparison discipline is subject to variability at key decision points throughout the process. This variability has been observed both between and within examiners and in the analysis, comparison, and evaluation phases of an examination. This research was undertaken to closely examine the suitability decision—what suitability decisions can be made, what information do experts use to support these decisions, can the decisions of experts be predicted, and is there a way to reduce the variability in their decisions?

We proposed considering the suitability decision not as a single, binary decision—value versus no value—but as four separate dimensions of the utility of a mark for various purposes. These dimensions were represented as four proposed suitability scales: Value, Complexity, AFIS, and Difficulty. Each of these scales considers a different use to which a single impression could be put.

The Value scale considers whether or not a mark should be taken forward to a comparison. It was originally conceived as having five possible categories (but unfortunately had to be collapsed to a binary value/no value choice during machine learning):

- No Value (the mark will not be used any further)
- Investigative or Probative Value (the mark is not suitable for reaching an identification or exclusion, but could potentially provide an investigative lead or other probative information)
- VEO (the mark could be used to exclude, but not to identify)
- VIDO (the mark could be used to identify, but not to exclude)
- Value for Both (the mark is suitable for both identifications and exclusions to appropriate exemplars)

The Complexity scale considers whether a mark, once determined to be suitable for comparison, is complex or not. This decision could be used to drive QA policies, such as requiring additional documentation and review for complex marks, but allowing abbreviated documentation and review for very high quality marks.

The AFIS scale considers whether a mark is suitable for entry and search in an AFIS. Furthermore, the scale considers whether a mark that will be searched in an AFIS has characteristics that put it at higher risk of a coincidental match and should therefore be subjected to additional QA measures.

Finally, the Difficulty scale considers the overall difficulty anticipated in comparing a mark and can be used for building research, training, and testing at consistent and stratified difficulty levels. It can also be used for testimony purposes as a way of describing the difficulty level of a mark being presented in court.

To explore these four scales, we first undertook a white box study to better understand the information experts use to make their suitability decisions and to observe how they reacted to the introduction of new scales and conclusions when thinking about suitability. Volunteer experts were provided with fingermarks from casework and asked to annotate only the information they used to reach their suitability decisions, then they were asked to render those decisions on each of the four suitability scales.

Examiners were found to be variable in the features they relied upon, their perceptions of amount of clarity and distortion, and their ultimate decisions regarding suitability. Consensus observations were fairly good predictors of consensus decisions; however, the variation in the data suggested that individual examiners would not agree with the consensus opinion in many cases. Although variability will unavoidably be present to some extent in human endeavors that rely upon subjective assessment of visual cues, the discipline should nonetheless make efforts to reduce this variability to the extent possible because differences in opinion over the suitability of a mark for comparison or AFIS could have real-world consequences to the criminal justice system.

Minutiae count was the strongest driver of value decisions, whereas clarity and distortion together better explained decisions on the other scales. Overall, the variables that were most consistently relied upon to reach suitability decisions along all four scales were: total number of minutiae marked, number of confident minutiae marked (as opposed to uncertain minutiae marked), the clarity of the image, the level of distortion in the image, whether the pattern type could be determined with confidence, and the selectivity of the minutiae.

Although examiners tended to agree on which images were very high quality, there was no consensus on no value images. Also, if an image was not extremely high quality, it was likely that many examiners would assign it to the highest value category, whereas many other examiners would disagree. Many images had nearly equal votes across all value categories.

Examiners tended to express strong confidence in minutia type, even when there was connective ambiguity. More degraded images resulted in a higher use of uncertain minutia markers, but certain minutia marker types were still used. Thus, examiners should not use specific minutiae marker types unless and until specific criteria are developed to define each because the current practice lends a misleading veneer of certainty to what is often an arbitrary decision.

Furthermore, examiners should document the features and observations they relied upon to reach their suitability decision(s) because it is known that these decisions can vary widely. Without a way to substantiate *why* a particular decision was reached, the decision appears opaque and arbitrary.

The new suitability categories that were introduced along the four scales of suitability were chosen often by participants in this study. There appears to be value in expanding the notion of “suitability” of latent marks and considering different uses for which a mark may be useful as well as considering more granular conclusion options that may suggest additional quality assurance measures.

Next, we undertook to develop, optimize, and validate a predictive model using the data from the white box study to predict the consensus response on each of the suitability scales using only a few key inputs. A main question of this research was whether a model informed by user inputs alone, a fully automated model, or some combination of the two, would provide the highest accuracy predictions, thus answering the question of whether a human examiner should be involved in suitability decisions, or whether they should be fully automated.

Our results showed that, although only a few key data inputs were required, models that utilized both user input and automated measures exhibited superior predictive accuracy on all four scales to models that relied on only one or the other. Further, our external validation study showed stable predictive accuracy values with new images and new users, supporting the generalizability of the model, and also showed high levels of agreement with the model from participants, supporting the likelihood of the model being accepted for use within operational laboratories.

In Step One of the model development, 11 MLAs were tested on each of the four scales with 6 variables each, for a total of 24 sets of results. From these initial data, RF was selected as the MLA to use for all subsequent modelling and RFE was used to narrow the potential predictors to 12.

In Step Two, average user data from “super-users” who always agreed with the consensus ground truth decision were used to train and test the RF model with different combinations of variables to determine the best-case scenario of how the model could perform with idealized data. The best operational set of variables using the average user data resulted in the following mean accuracies for prediction: Value—100%; Complexity—86.91%; AFIS—86.69%; Difficulty—93.32%.

In Step Three, *all* user data were used, including data from users who did not agree with ground truth and without averaging any of the data. The same combinations of variables were tested as in Step Two. The best operational set of variables under these more real-world conditions yielded the following mean accuracies for prediction: Value—83.13%; Complexity—59.91%; AFIS—77.77%; Difficulty—66.31%. Additionally, these values were compared to the accuracy of individual users in reaching decisions that matched the consensus ground truth, with similar results. MultiROC curves were used to test the performance of the selected model against other models at all probability thresholds and the selected, operationally-realistic model performed favorably.

Finally, an external validation study was performed to test the generalizability of these results on new images and new users. The performance of the model was evaluated in four ways. First, we considered the overall mean predictive accuracy of the model during optimization (Step Three) against its overall mean predictive accuracy during the validation study. The results for

Step Three are given above. The results for the validation study were: Value—83.50%; Complexity—57.82%; AFIS—76.52%; Difficulty—66.36%, showing good stability in overall predictive accuracy for the external validation.

The second measure of performance for the final model in the validation study was comparing the mean predictive accuracy of the model on images that were used in the initial development of the model versus images that were new to the validation study. This was done to ensure against a sampling effect from the original images. These results were (old images followed by new images): Value—82.99% versus 83.33%; Complexity—57.32% v. 58.14%; AFIS—83.78% v. 71.83%; Difficulty—71.34% v. 63.14%, showing that accuracy was very stable for the Value and Complexity scales, but dropped off somewhat on the AFIS and Difficulty scales.

The third measure of performance for the final model in the validation study was comparing the overall mean predictive value of the model (given above) to the overall mean accuracy of the participants in reaching a conclusion that matched the consensus ground truth. The accuracy values of the users were: Value—85.07%; Complexity—62.09%; AFIS—74.81%; Difficulty—69.33%, which shows that even though the model was not always predicting at a very high accuracy, its performance was always comparable to the performance of users making decisions on their own.

The last measure of the performance of the final model was agreement with users. We wanted to know whether LPEs would generally accept the judgement of the model. This was examined in three ways: the overall level of agreement for all data in the validation study, the level of agreement with data from the GT user group, and the level of agreement with the data from the Exp user group. Participants in the validation study were alternately assigned to one of two groups as they signed up—GT and Exp. The GT group declared their decisions on each of the four scales prior to seeing the predictions of the model and being asked whether they agreed with them, whereas the Exp group was shown the model's predictions and asked whether they agreed with them without committing to their own decisions first. This was done both to establish consensus ground truth decisions for the new images, and to see whether there was a biasing effect of seeing the model's decision before the examiner had to form one of their own. Across all three measurements, agreement with the model was consistently high. With the exception of the GT users on the Difficulty scale, which showed only 69.33% agreement, all other scale and user combinations showed agreement ranging between 81.55% and 94.98%, indicating that there is a high likelihood that examiners would accept the predictions of the model on all four scales.

The results of our investigations have demonstrated both that there is utility in thinking about the suitability decision in terms of four separate scales, and also that the predictive model we have built is capable of predicting the consensus response on each of the four scales. The findings of this research lead us to a vision of future policy and practice as respects friction ridge suitability.

It is clear from our results that both human examiner and automated algorithm have a role to play in determining suitability, but what should be the division of labor? Who should "win" if there is a disagreement? And what, exactly, should we consider when we think about the "suitability" of a mark?

It is our belief that the suitability scale should be broken into four scales as proposed by this research, and that those scales should be used to drive policy. Suitability is not a single determination of value or no value; we have shown that there is value to considering a mark's suitability for AFIS entry as well as its complexity. Furthermore, there are multiple QA and policy decisions that can be made along different scales that can result in better efficiency in workflow, a reduction in variability, and an increase in accuracy.

Currently, guidance organizations such as OSAC are advocating for an increased number of possible comparison conclusions, including new categories of Support for Same Source and Support for Different Sources that indicate evidence that is leaning in one direction, but has not crossed a threshold for an identification or exclusion. However, nobody has previously discussed adjusting the Value scale to mirror this structure. If there are pairs of impressions that do not contain sufficient information to conclude an identification, it stands to reason that there are similarly single impressions that one could designate as insufficient to support an identification prior to even seeing the exemplar. Although the modeling portion of this research was unfortunately unable to fully explore a way to predict these conclusions, the white box portion of the research did show good use of new conclusions along the Value scale that would indicate just those sort of marks. If the friction ridge community were to begin incorporating these different Value scale conclusions and become used to them, much more research could be done in this area to result in much more nuanced Value determinations.

The Complexity scale is a necessary addition to operational workflows in order to identify those marks that should be subject to additional, or reduced, QA measures, such as documentation and review. Many friction ridge units are badly backlogged, yet mistakes can happen when people are rushing. There is a strong body of literature calling for additional caution on complex marks, but complexity has not been defined. By using the Complexity scale of this model, resources could be diverted to the marks that need the most caution and oversight (complex marks), while expending fewer resources on what is essentially "busy work" on high-quality, self-evident marks. We envision the incorporation of a Complexity scale into laboratory policy such that marks designated as "complex" will require documentation of all information located and being relied upon in a mark; those designated as "non-complex, requiring documentation" will require documentation only of the information required to meet the agency's suitability criteria policy; and those designated as "non-complex, self-evident" will require only documentation of the features relied upon to reach that complexity determination. We would further recommend that complex marks require blind verification, whereas the other two categories could enjoy standard, *open verification*.

Many agencies currently judge AFIS quality separately from value; they just haven't necessarily conceived of it as a different type of suitability decision. We do not foresee any resistance to the adoption of the AFIS scale from these agencies. However, some believe that any mark that is suitable for manual comparison is also suitable for AFIS. We hope that those agencies will consider that some marks may be suitable on the Value scale, but not be AFIS quality. For one thing, the notion of "AFIS quality" may vary according to the size of the reference database, or the quality of the particular AFIS vendor's matching algorithms. Thus, individual determinations of AFIS quality could rely on very different criteria than those used to determine value.

Furthermore, we believe the addition of the category "AFIS Quality with QA measures" is critical in light of the increased likelihood of coincidental matches in a search in a large AFIS. This category was favorably received by the participants in our study and we believe its use could greatly reduce AFIS errors by flagging marks that can be entered but pose a higher risk. Recommended additional QA measures for these marks include additional verification, blind verification, or requiring additional features *not* used in the AFIS search to be found between the two impressions before an identification may be declared.

Finally, we advocate for the use of the Difficulty scale in research, training, testing, and testimony. If the same tool were used to define the difficulty level of a mark, the results of research studies could be more directly compared to see whether performance was similar on marks of similar difficulty. Training programs could be created that were progressively more difficult. Testimony could be offered about the difficulty of marks about which an expert was testifying. And perhaps most impactful, proficiency tests could be designed with test samples at different difficulty levels. This could result in a paradigm shift in how proficiency tests are seen—rather than being a test of (very) minimal competence that every working examiner is expected to pass, they could test the *level* of proficiency an examiner possesses. The expectation could be, not that every examiner gets every comparison correct, but that the test identifies the level of proficiency at which an examiner is operating. This would allow for targeted training in areas where the examiner was weaker, and for managers to know who their strongest examiners are. There would not be a stigma around passing or failing the test, but rather a continual push for improvement, to be able to successfully complete more and more of the challenging test samples.

The challenge in deciding whether and how to implement a model such as the one we have developed lies in determining where the thresholds should lie for each category. In this research, we have predicted decisions based upon a consensus response taken at a 50% probability threshold—that is, if there is a greater than 50% probability that the consensus will vote for a particular category, that category will be chosen by the model. However, this threshold was chosen arbitrarily and another one might better represent the priorities of an agency or individual examiner. Biederman et al. (2008) have described how the threshold for a forensic decision may differ between examiners and agencies based upon their specific priorities, desires, experiences, or cost/benefit analyses. However, in our data collection, decisions were collected agnostically to all of these considerations. We did not present any scenario or framework to our participants within which to make their decisions, nor did we invoke any particular set of costs, benefits, or values. We simply showed images and asked participants to make assessments. Because each examiner will have their own values and their own agency priorities and policies, we believe that judgments based upon these can vary widely. In future research, it would be very interesting to explore how these assessments might change when particular values were imposed upon the participants while making their decisions. Nonetheless, this model is flexible in that the probability thresholds for decisions can be adjusted to suit any particular set of priorities. Because of this, we recommend that thresholds be set according to agency preference, thus removing the individual opinions and values of the examiner from the equation.

This brings us to our next point, which is: who decides—the model, or the examiner? How and when shall the model be used? Change is slow to come. We understand that. Given this reality, it is likely that were the model to be adopted, it would initially be used to support the decisions of the examiner. It could serve as a red-flag warning to an examiner if the model and the examiner did not agree; the examiner would know that they should re-visit their conclusion because it is likely a majority of examiners would not agree with it, and we would certainly advocate that in this situation, the examiner must justify why they believe their judgment to be superior to that of the model (which, we acknowledge, does not always get it right).

However, two things we have repeated throughout this dissertation bear on this argument: (1) examiners tend to be highly variable; and (2) the best-performing model incorporates *both* human and automated input. Because of these two points, we foresee a future in which the model is used *first* and is deferred to. In order to use the model, the examiner must view the mark and make 3 key inputs regarding their observations (number of minutiae, and clarity and distortion assessments). This means that the examiner has already performed a mini-analysis and has some feeling (whether made explicit at this point or not) about the mark. It also means that their input has, to some extent, been considered and incorporated by the model. By using the model first, we can reduce variability by taking the more standardized opinion. If the examiner strenuously objects to the conclusion(s) of the model, they can overrule it by documenting *why* they disagree, and what information, specifically, that they had access to and the model did not, they feel supports their decision.

This process would greatly streamline the analysis process. The mini-analysis done by the examiner in using the model would be much quicker than performing a full analysis on every mark and, in most cases, a full analysis would not be necessary. The judgment of the model could be taken as the suitability decision. If the examiner disagreed with the model, *then* a full analysis would be triggered for that mark. Similarly, verification of analysis decisions could reasonably be skipped under the argument that the decision has been "verified" by the agreement of the examiner and the virtual consensus panel represented by the model.

We see these points as the main arguments for adoption of the model—it reduces time spent in analysis and verification of analysis; it provides a built-in virtual consensus panel; it reduces variability in suitability decisions; and it expands the way suitability is conceived in ways that can drive agency policy for the improvement of efficiency and outcomes.

7 Appendix A – Terminology

Although a consistent terminology is generally a cornerstone of any professional, scientific, or academic domain, it is unfortunately true that the latent print discipline suffers from a lack of standardized terminology. Although the discipline generally agrees on a number of terms in a broad sense, closer inspection, or indeed, even conversation, will reveal that these familiar terms are often used differently, to great confusion. Additionally, there are some terms that are used interchangeably by some whereas they are separated by nuance of meaning for others. Finally, there are some terms that are well-known to many members of the profession, yet still new and puzzling to others. Thus, this appendix will describe how some of the terms throughout this work are used, particularly those that are close to one another in nature and may require some disambiguation. Note that we have grouped these terms to have those that are most closely related together and they are not necessarily presented in alphabetical order. When multiple terms are used to describe the same concept, we have selected a single term that will be used throughout the thesis.

Mark and Print — Within the US, impressions found at crime scenes, or made by chance, are typically referred to as "latents" whereas impressions taken deliberately from known sources are typically referred to as "knowns," "exemplars," or "inked prints." This work will follow the non-US convention of referring to unknown impressions as "marks" and will largely follow the same convention of referring to known impressions as "prints" although they will occasionally be referred to as "exemplars" as well.

Suitability and Sufficiency — Within the latent print community, these two terms are often used interchangeably. However, in this work, a distinction will be made between the two that is recognized by some practitioners and not by others. "Suitability" refers to the decision that is made at the end of the analysis phase and answers the question, "is there enough reliable information observed in this mark to continue using it in some way, e.g. for comparison or AFIS entry." "Sufficiency" refers to the decision that is made at the end of the evaluation phase and answers the question, "is there enough reliable information observed between two impressions to draw and support a conclusion regarding whether or not they originated from the same source."

Suitability and Value — Although "suitability" refers to the decision made at the end of the analysis phase, it is also often used interchangeably with the word "value" to describe a decision to take a mark forward to the next step in the comparison process. In this work, a distinction will be drawn between these two terms. This research examines the suitability decision in detail and proposes separating it into 4 distinct scales, one of which is the Value scale. Thus, the term "value" in this work will be limited only to the decision about whether a mark is of value or not in the context of the Value scale, whereas the term "suitability" will be used broadly to encompass any decision made on any of the four scales, or the four scales taken as a whole.

Information, Information-gathering — When we refer to "information" in this thesis, we are referring to any observed data within a mark that may be used to form and support a suitability decision. This may include minutiae presence and type, pattern type, presence of level 3 detail, clarity assessment, distortion assessment, distinctive clusters, creases, scars, or any other factors that are typically considered during analysis of a mark.

Information-gathering refers to the main function of the analysis phase, which is to observe and interpret the information present in the mark in order to reach a determination about whether it is suitable for a given purpose and will proceed further in the friction ridge comparison process.

Confidence, Reliability, and Reproducibility — These three concepts are closely related and the second two inform the first. As an examiner is gathering information from a mark, they will note various features upon which they may intend to rely during an eventual comparison. For each feature considered, they should evaluate their *confidence* in the feature – that is, how certain they are that the feature is actually present and also how certain they are that the feature is what they think it is (for example, a feature may appear to be a ridge ending when in fact it is a bifurcation on the source skin).

To determine their confidence level in a given feature, an examiner may estimate its *reliability* and *reproducibility*. Reproducibility in this context refers to how likely the examiner perceives it to be that the feature would be similarly recorded in other impressions of the same source skin. Reliability in this context refers to the degree to which the examiner trusts in the information they are observing – can it be relied upon? Reliability and confidence are very nearly synonymous, except that the perceived reliability of a feature is what *gives* the examiner confidence that it is present and being interpreted correctly. Areas of an impression that have high distortion or low clarity have low reliability and thus an examiner should have low confidence in any features in that area.

Tolerance — Tolerance can be thought of as how much "wiggle room" an examiner is willing to allow in the interpretation of a feature, and it is based upon the clarity of the local area surrounding that feature (and tied closely to the notion of reliability). If a feature is located in a low clarity area, its reliability is low and thus the examiner should be willing to accept more variability in its appearance if seeing it in another impression. Conversely, in a high-quality area, the feature has high reliability and very little variation in its appearance should be tolerated in other impressions from the same source. If the appearance of a feature is found to be "out of tolerance" during a comparison, the examiner should conclude that the features in the two impressions being compared are not the same and thus the two impressions likely did not originate from the same source. The concept of tolerance is inversely related to that of weight – a feature that has low reliability and thus is given a high tolerance should be awarded very little weight toward decision-making.

Weight — Weight is an estimate of how much a given feature counts toward a decision. Features that are high in clarity or rarity should be assigned more weight than those with low clarity or rarity because it would take fewer of them to reach and support a decision. If a decision is thought of as a scale that tips in one direction for an “of value” decision and the other direction for a “no value” decision, for example, a feature that had a lot of weight would be a large pebble on one side of the scale, whereas a feature that had little weight would be a small pebble. The determination of weight is often a tradeoff between clarity and rarity because, for example, a rare feature in a low clarity area should not be given much weight – although it would be worth a lot if it were truly there, the examiner cannot be confident that it *is* really there.

Risk — In the context of this research, risk is an assessment the examiner makes (whether implicitly or explicitly) of how likely it is that they will reach a negative outcome if they choose to rely on a particular piece of information observed in an impression. They will weigh the risk of these negative outcomes against the potential benefits of using the features when deciding whether or not to use the feature, and how much weight to give it. For example, if a highly distinctive feature, such as a trifurcation, was noted in a distorted, low-clarity area, the examiner would have to weigh the benefit of being able to use this valuable feature (which could lead to a correct ID – a very desirable outcome) against the risk of negative outcomes if the feature turned out not to be what it first appeared.

Negative Outcome — A negative outcome is a consequence of a wrong decision that the examiner would like to avoid. This can range in scope from making a suitability decision with which others would not agree (negative outcome) due to misinterpreting features to making an erroneous identification (negative outcome) due to misinterpreting features or trying to compare a mark that should not have been designated as suitable in the first place. In a more macro sense, negative outcomes can also be the consequences to the examiner or agency of these smaller negative outcomes, such as risk of reprimand, re-training, loss of public confidence, or lawsuits.

Complex, Complexity — An “of value” mark may occur anywhere along a continuum of quality from barely of value to extremely high clarity and quantity. A mark’s position along this continuum can be referred to as its “complexity” with marks at the low end of the continuum being designated as “complex” (as opposed to “non-complex”). Although one can consider either the complexity of a mark on its own, or the complexity of a *comparison* when taking into account how the mark and the print being compared to it relate to one another, this work will focus on complexity of the mark alone, because that is what is considered during the analysis phase.

Complexity as used in this research refers to the chance that two examiners will disagree about the suitability, sufficiency, or interpretation of features and distortion in the mark. The more complex a mark, the more likely it is considered that a second examiner may not agree with the first on one or more of these aspects. There is currently no set threshold or list of criteria for what makes a mark complex (although SWGFAST does provide for a complexity zone in their sufficiency graph) and this decision will vary from examiner to examiner.

VEO/VIDO/VID/VB — The latent print community is already familiar with the labels "VEO" (value for exclusion only) and "VID" (value for identification), which have been used in previous research and referenced in SWGFAST documents (Scientific Working Group on Friction Ridge Analysis Study and Technology (SWGFAST) 2011). However, with the expanded Value scale proposed by this work, additional labels are needed to describe the new options on the scale. Thus, two new terms have been introduced.

"VIDO" stands for "value for identification only" and is the mirror image of VEO; rather than describing a mark that could be excluded but not identified, VIDO describes a mark that could identified, *but not excluded*. This is in contrast to the existing VID, which refers to a mark that could be identified and *may or may not* be excludable.

"VB" stands for "value for both," which refers to a mark that the examiner believes could be *both* identified or excluded, depending on the print to which it is being compared.

Categorical Conclusion — Although technically, any decision that falls into a category is a categorical conclusion, in this work, we are using the term specifically to refer to a decision at one of the two extreme ends of the comparison conclusion scale—that is, Identification or Exclusion. "Categorical" will be used in the dictionary sense of "absolute." Less-than-absolute conclusions, such as "inconclusive," "support for same source," or "cannot exclude" would not be considered categorical conclusions by this definition.

Ground Truth and Consensus Ground Truth — Ground truth refers to a situation in which it is absolutely known what the correct and truthful conclusion should be. In latent print research, this is usually referring to a case where the researcher knows what the source of a mark was, typically because the donor of the mark was observed while creating it.

In the present research, we are focused on suitability decisions. In this case, there is really no such thing as ground truth because there is currently no objectively right or wrong answer about what a mark can and should be used for. Thus, we have taken the majority voted opinion of our participants as a proxy for ground truth so that we have an expected outcome for each decision against which the machine learning algorithms can be taught. This majority-voted expected response will be referred to as "consensus ground truth."

Utility Tool and Utility — Utility is a word that means "the state of being useful or beneficial." In the context of latent print analysis, utility would refer to the usefulness of a mark for a particular purpose. In most cases, this would be synonymous with "value"—that is, a mark has utility if it can be compared to reach a conclusion about source.

However, since this research is breaking the suitability decision apart into four scales with more than binary decisions on each, there are multiple ways and degrees to which a mark may have utility. The model being developed by this research is referred to as a "utility tool" throughout the work because the model will function as a practical tool to measure the utility of a mark along each of the four scales. This is related to, but not to be confused with, the notion of a

Utility Function as introduced by Biedermann et al. (2008), which refers to a specific calculus to determine the utility of a range of potential forensic decisions. Of course, utility functions can be incorporated into this utility tool to set thresholds that meet the operational needs and priorities of a particular laboratory or examiner, a concept that is discussed in Chapters 2 and 4 of this thesis.

Objective Data In contrast to subjective data, which are collected by asking examiners their opinions on what they observe, we are using “objective data” to denote data that are collected using automated measures, such as quality scores or expected score-based likelihood ratios (ESLR’s) generated by computer algorithms or modeling.

Brute force search — A brute force search refers to a friction ridge comparison in which there are no location and orientation clues available, or those clues are ambiguous or unreliable. When there is a reliable anchor present, an examiner may proceed with confidence to the correct corresponding area in the provided prints to conduct their comparison and if there are not corresponding features, they may exclude with confidence. When these clues are absent or unreliable, the comparison must be done painstakingly, ridge-by-ridge, and checking orientation in 360 degrees. This is time-consuming, exhausting, and at the end of it, the examiner still may not feel safe excluding unless they are positive that they have received exemplars that clearly recorded every bit of available friction ridge skin. This inelegant, time-consuming ridge-by-ridge comparison is what is meant by brute force search.

Distinctive Clusters and Target Groups — These two terms are closely related, yet distinct. A target group is a cluster of minutiae or other features that an examiner would use for their initial search in a print. Ideally, a target group should be distinctive and should be located near an anchor, but any minutia(e) could technically be used to form a search image. Thus, a good target group typically will be a distinctive cluster. However, if there is not much to choose from, a target group may not be that distinctive at all. It could be as simple as a ridge ending two ridges away from the core.

In contrast, a distinctive cluster is just that – a group of minutiae that are near one another and that together, are highly distinctive. These are minutiae clusters that catch the examiner’s attention, that they are excited to look for, or that they would give extra weight to if found because they are unusual in some way. Not every distinctive cluster makes a good target group because they may be in areas of the mark that are more difficult to search.

Anchor — An anchor refers to any stable feature of an impression that helps an examiner to know where they should focus their search. Most typically, anchors are cores or deltas, but they can also be primary creases, some secondary creases, scars, or any other occasional features that record reliably and provide location and orientation information.

Pattern force area — A pattern force area is a region of a friction ridge impression where two ridge systems are converging, forcing many ridges to end or merge. These typically occur around deltas and the outflow of loops in distal phalanx impressions. Pattern force areas are characterized by a high number of ridge endings and closing bifurcations and present an increased risk of coincidental match with a non-mated source because the fact that all marks converge in these areas means there are more likely to be repeated features there between individuals.

Orientation clues — Orientation clues are characteristics of an impression that assist the examiner to know which way a mark should be distally oriented (i.e. which way is up). Common orientation clues include patterns, cores, creases, and ridge flow.

Location clues — Location clues are characteristics of an impression that assist the examiner to know or infer the anatomical area from which an impression originated, such as from a finger or from a particular region of a palm. Patterns, creases, cores, deltas, and scars can all provide location clues, as can general ridge flow in many cases (particularly in palms and joints). Anything that provides enough anatomical information to narrow a search can be a location clue, for example a delta that cannot be oriented, because it still narrows the search to delta areas.

Level 1 Detail — Level 1 detail refers to the overall *flow* of the friction ridges (or other information, e.g. creases) in an impression. Most commonly, it is taken to mean the pattern type of a fingerprint (such as arch, loop, or whorl), but it encompasses any macro flow that can assist to locate (see *location clues*) or orient (see *orientation clues*) an impression.

Level 2 Detail — Level 2 detail refers to the ridge *paths* (or the paths of other information, such as creases or scars) observable in an impression. Typically, this includes ridge endings, bifurcations, and dots, or compound features made up of more than one of these minutiae (e.g., a ridge ending and a bifurcation in close proximity form a hook).

Level 3 Detail — Level 3 detail is commonly said to refer to ridge *shapes*. More specifically, it refers to any fine detail about the ridge or about a ridge feature. Examiners do not always agree about what information falls under the umbrella of Level 3 detail, but it is commonly accepted to include features such as pores and ridge edge shapes. Level 3 detail may also include characteristics such as ridge thickness or the angles at which bifurcations open. These are very small details that give a feature, or even a section of a ridge, a distinct "personality" that could be used to differentiate it from other, similar-looking, features or ridge sections.

Level 3 detail can be very powerful information because it is so very discriminating. However, it can also be very dangerous to rely upon *because* the detail is so fine. This detail is often not highly reproducible, meaning that it may not always record, or may not always record in the same way. Thus, examiners should use extreme caution in assigning too much weight to Level 3 detail, particularly when it is located in an area of low clarity where a high degree of interpretation is required.

Pattern type — A type of Level 1 detail, pattern type refers to the overall flow of a system of ridges, when they adhere to certain pre-defined criteria. The 3 main pattern types are the arch, loop, and whorl, and these are defined according to which side of the finger the ridge system enters and leaves on, the shape of its flow within the pattern area, and the number of deltas present. Patterns can also be found within certain areas of the palmar and plantar surfaces, and less commonly, in the medial and proximal phalanges. Pattern types are further sub-divided into sub-classification types and can be used in classification systems to narrow a search.

Manual comparison —This term refers to a comparison between two friction ridge images that is done by a human examiner, usually between a scene (unknown) mark and an exemplar (known) print. It is in contrast to automated AFIS searches. This term is used to draw a distinction between quality metric models that are designed to assess the quality of a mark for the purpose of entering it into an AFIS and those that are assessing quality of a mark for the purpose of being compared by a human examiner.

OSAC —The Organization of Scientific Area Committees. This organization, formed through a joint effort of the National Institute of Standards and Technology (NIST) and the National Institute of Justice (NIJ), gathers subject matter experts and stakeholders to identify needs and draft consensus-based standards for the forensic science community. The OSAC is broken into numerous sub-committees, each of which is dedicated to a specific forensic science discipline. Standards and Best Practice Guidelines for the friction ridge discipline are handled by the Friction Ridge Subcommittee (FRS).

Blind verification vs. Open verification — Verification is the independent review of the friction ridge examination(s)n in a case to see whether a second, qualified examiner reaches the same conclusion reached by the first examiner, and whether that conclusion is sufficiently supported. In blind verification, the verifier does this work entirely independently, without knowledge of the conclusion reached by the first examiner and without access to their notations (until after the verifier’s decision has been reached and documented). In some cases, the blind verifier does not even know the identity of the first examiner. By contrast, in open verification, the verifier reviews the notes and conclusions of the first examiner before deciding whether they agree with the decision of the first examiner.

8 References

AAAS 2017. Forensic Science Assessments: A Quality and Gap Analysis - Latent Fingerprint Examination, (Report prepared by William Thompson, John Black, Anil Jain, and Joseph Kadane), doi: 10.1126/srhrl.aag2874.

Anthonioz, A., Egli, N., Champod, C., Neumann, C., Puch-Solis, R. & Bromage-Griffiths, A. 2008. Level 3 Details and Their Role in Fingerprint Identification: A Survey Among Practitioners. *Journal of Forensic Identification,* 58**,** 562-589.

Ashbaugh, D. R. 1999. *Qualitative-Quantitative Friction Ridge Analysis – An Introduction to Basic and Advanced Ridgeology,* Boca Raton, CRC Press.

Bell, S., Sah, S., Albright, T. D., Gates, S. J., Denton, M. B. & Casadevall, A. 2018. A call for more science in forensic science. *Proceedings of the National Academy of Sciences,* 115**,** 4541-4544.

Biedermann, A., Bozza, S. & Taroni, F. 2008. Decision theoretic properties of forensic identification: Underlying logic and argumentative implications. *Forensic Science International,* 177**,** 120-132.

Bross, I. D. J. 1958. How to Use Ridit Analysis. *Biometrics,* 14**,** 18-38.

Campbell, S. A. 2011. *The Fingerprint Inquiry Report,* Edinburgh, APS Group Scotland.

Carter, K. E., Vogelsang, M. D., Vanderkolk, J. & Busey, T. 2020. The Utility of Expanded Conclusion Scales During Latent Print Examinations. *Journal of Forensic Sciences,* 65(4), 1141-1154, https://onlinelibrary.wiley.com/doi/epdf/10.1111/1556-4029.14298, Last accessed: 2/3/2021.

Champod, C. 1995. Locard, Numerical Standards and «Probable» Identification. *Journal of Forensic Identification,* 45**,** 132-159.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research,* 16**,** 321-357.

Chugh, T., Cao, K., Zhou, J., Tabassi, E. & Jain, A. K. 2018. Latent Fingerprint Value Prediction: Crowd-Based Learning. *IEEE Transactions on Information Forensics and Security,* 13**,** 20-34.

Cole, S. A. 2011. Splitting Hairs? Evaluating 'Split Testimony' as an Approach to the Problem of Forensic Expert Evidence. *Sydney Law Review,* 33**,** 459-485.

Cole, S. A. 2014. Individualization is Dead, Long Live Individualization! Reforms of Reporting Practices for Fingerprint Analysis in the United States. *Law, Probability & Risk,* 13**,** 117-150.

Danov, I., Olsen, M. A. & Busch, C. Interpretation of fingerprint image quality features extracted by self-organizing maps. 2014. 907505-907505-15.

Davis, J. 1979. Thoughts on Resolution VII. *Identification News***,** 5-6.

Dror, I. 2009. How Can Francis Bacon Help Forensic Science? The Four Idols of Human Biases. *Jurimetrics Journal,* 50**,** 93-110.

Dror, I. E. 2020. The Error in 'Error Rate': Why Error Rates Are So Needed, Yet So Elusive. *Journal of Forensic Sciences,* 65(4), 1034-1039, https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.14435, Last accessed: 2/3/2021.

Dror, I. E. & Mnookin, J. L. 2010. The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science. *Law, Probability & Risk,* 9**,** 47-67.

Dror, I. E., Péron, A., Hind, S.-L. & Charlton, D. 2005. When Emotions Get to the Better of us: The Effect of Contextual Top-Down Processing on Matching Fingerprints. *Applied Cognitive Psychology,* 19**,** 799-809.

Dror, I. E., Wertheim, K., Fraser-Mackenzie, P. & Walajtys, J. 2012. The Impact of Human–Technology Cooperation and Distributed Cognition in Forensic Science: Biasing Effects of AFIS Contextual Information on Human Experts. *Journal of Forensic Sciences,* 57**,** 343-352.

Edwards, H. T. 2009. Strengthening Forensic Science in The United States: A Path Forward. *Statement before the States Senate Committee in the Judiciary.* Washington DC.

Eldridge, H. 2012. "I am 100% certain of my conclusion" (But should the jury be certain?). *Evidence Technology Magazine***,** 8.

Eldridge, H. 2019. Juror comprehension of forensic expert testimony: A literature review and gap analysis. *Forensic Science International: Synergy,* 1**,** 24-34.

Eldridge, H., De Donno, M., Furrer, J. & Champod, C. 2020. Examining and expanding the friction ridge value decision. *Forensic Science International,* 314, 110408, https://doi.org/10.1016/j.forsciint.2020.110408**,** Last accessed: 2/3/2021.

Eldridge, H., De Donno M. & Champod C. (2021). Predicting suitability of finger marks using machine learning techniques and examiner annotations. *Forensic Science International* 320, 110712, https://doi.org/10.1016/j.forsciint.2021.110712, Last accessed: 20/5/2021.

Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J. & Bigun, J. 2005. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition,* 38**,** 777-779.

Forensic Science Regulator 2015. Codes of Practice and Conduct: Fingerprint Comparison. Birmingham. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/415108/128_FSR_fingerprint_appendix__Issue1.pdf, Last accessed: 9/14/2020.

Found, B. & Rogers, D. 1996. The Forensic Investigation of Signature Complexity. *In:* SIMNER, M. L., LEEDHAM, A. J. & THOMASSEN, W. M. (eds.) *Handwriting and Drawing Research: Basic and Applied Issues.* Amsterdam: IOS Press.

Found, B., Rogers, D., Rowe, V. & Dick, D. 1998. Statistical Modelling of Experts' Perceptions of the Ease of Signature Simulation. *Journal of Forensic Document Examination,* 11**,** 73-99.

Garrett, B. L., Crozier, W. E. & Grady, R. 2020. Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence. *Journal of Forensic Sciences,* 65(4), 1199-1209, https://onlinelibrary.wiley.com/doi/full/10.1111/1556-4029.14323, Last accessed: 2/3/2021.

Greenwell, B., Boehmke, B. & Gray, B. 2019, https://CRAN.R-project.org/package=vip. vip: Variable Importance Plots, Last accessed: 9/14/2020.

Gupta, S. R. 1968. Statistical Survey of Ridge Characteristics. *International Criminal Police Review,* 5 (218)**,** 130-134.

Hicklin, R. A., Buscaglia, J. & Roberts, M. A. 2013. Assessing the Clarity of Friction Ridge Impressions. *Forensic Science International,* 226**,** 106-117.

Hicklin, R. A., Buscaglia, J., Roberts, M. A., Meagher, S. B., Fellner, W., Burge, M. J., Monaco, M., Vera, D., Pantzer, L. R., Yeung, C. C. & Unnikumaran, T. N. 2011. Latent Fingerprint Quality: A Survey of Examiners. *Journal of Forensic Identification,* 61**,** 385-418.

Holmgren, J. A. & Fordham, J. 2011. The CSI Effect and the Canadian and the Australian Jury. *Journal of Forensic Sciences,* 56**,** S63-S71.

Kelley, S., Gardner, B. O., Murrie, D. C., Pan, K. D. H. & Kafadar, K. 2020. How do latent print examiners perceive proficiency testing? An analysis of examiner perceptions, performance, and print quality. *Science & Justice,* 60**,** 120-127.

Kellman, P. J., Mnookin, J. L., Erlikhman, G., Garrigan, P., Ghose, T., Mettler, E., Charlton, D. & Dror, I. E. 2014. Forensic Comparison and Matching of Fingerprints: Using Quantitative Image Measures for Estimating Error Rates through Understanding and Predicting Difficulty. *PLoS ONE,* 9**,** e94617.

Koertner, A. J. & Swofford, H. J. 2018. Comparison of Latent Print Proficiency Tests with Latent Prints Obtained in Routine Casework Using Automated and Objective Quality Metrics. *Journal of Forensic Identification,* 68**,** 379-388.

Kuhn, M. 2020, https://CRAN.R-project.org/package=caret. caret: Classification and Regression Training, Last accessd: 9/14/2020.

Kuhn, M. & Johnson, K. 2013. *Applied Predictive Modeling,* New York, Springer-Verlag.

Kukucka, J., Dror, I. E., Yu, M., Hall, L. & Morgan, R. M. 2020. The impact of evidence lineups on fingerprint expert decisions. *Applied Cognitive Psychology,* 34(5) 1143-1153, https://doi.org/10.1002/acp.3703, Last accessed: 2/3/2021.

Langenburg, G. 2009. A Method Performance Pilot Study: Testing the Accuracy, Precision, Repeatability, Reproducibility, and Biasability of the ACE-V Process. *Journal of Forensic Identification,* 59**,** 219-257.

Langenburg, G. 2012. A critical analysis and study of the ACE-V process, PhD. School of Criminal Justice, University of Lausanne, Lausanne, https://serval.unil.ch/en/notice/serval:BIB_D54DC636916C, Last accessed: 2/3/2021.

Lennard, C. J. 2013. Fingerprint identification: how far have we come? *Australian Journal of Forensic Sciences,* 45**,** 356-367.

Maceo, A. V. 2009. Qualitative assessment of skin deformation: a pilot study. *Journal of Forensic Identification,* 59**,** 390-440.

Max, B., Cavise, J. & Gutierrez, R. 2019. Assessing Latent Print Proficiency Tests: Lofty Aims, Straighforward Samples, and the Implications of Nonexpert Performance. *Journal of Forensic Identification,* 69**,** 281-298.

Metcalfe, J. 2017. Learning from Errors. *Annu Rev Psychol,* 68**,** 465-489.

Mitchell, G. & Garrett, B. L. 2019. The impact of proficiency testing information and error aversions on the weight given to fingerprint evidence. *Behavioral Sciences & the Law,* 37**,** 195-210.

Mnookin, J. L. 2010. The Courts, the NAS, and the Future of Forensic Science. *Brooklyn Law Review,* 75**,** 1209-1275.

Moenssens, A. 1979. Resolution VII - A Monkey On Our Back. *Identification News***,** 3-5.

Montani, I., Marquis, R., Egli Anthonioz, N. & Champod, C. 2019. Resolving differing expert opinions. *Science & Justice,* 59**,** 1-8.

Murch, R. S., Abbott, A. L., Fox, E. A., Hsiao, M. S. & Budowle, B. 2012. Establishing the Quantitative Basis for Sufficiency Thresholds and Metrics for Friction Ridge Pattern Detail and the Foundation for a Standard. Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/239049.pdf, Last accessed: 9/14/2020.

National Commission on Forensic Science (NCFS) 2016. Views of the Commission: Facilitating Research on Laboratory Performance (adopted unanimously September 13, 2016). https://www.justice.gov/ncfs/page/file/909311/download, Last accessed: 9/14/2020.

Neumann, C. 2012. Statistics and Probabilities as a Means to Support Fingerprint Examination. *In:* RAMOTOWSKI, R. S. (ed.) *Lee and Gaensslen's Advances in Fingerprint Technology.* 3rd ed. Boca Raton: CRC Press, pp. 407-452.

Neumann, C., Armstrong, D. E. & Wu, T. 2016. Determination of AFIS ‘‘sufficiency’’ in friction ridge examination. *Forensic Science International,* 263**,** 114-125.

Neumann, C., Champod, C., Yoo, M., Genessay, T. & Langenburg, G. 2013. Improving the Understanding and the Reliability of the Concept of “Sufficiency” in Friction Ridge Examination. Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf, Last accessed: 9/14/2020.

Neumann, C., Evett, I. W. & Skerrett, J. 2012. Quantifying the Weight of Evidence from a Forensic Fingerprint Comparison: a New Paradigm. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 175**,** 371-415 (with discussion).

Nill, N. B. 2007. Image Quality of Fingerprint (IQF) software application. http://www.mitre.org/tech/mtf/, Last accessed: 9/14/2020.

Nisbett, R. E. & Wilson, T. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review,* 84**,** 231-259.

Osac Friction Ridge Skin Subcommittee 2019. Standard for Friction Ridge Examination Conclusions [DRAFT]. Washington DC. https://www.nist.gov/system/files/documents/2018/07/17/standard_for_friction_ridge_examination_conclusions.pdf, Last accessed: 9/14/2020.

Osterburg, J. W., Parthasarathy, T., Raghavan, T. E. S. & Sclove, S. L. 1977. Development of a Mathematical Formula for the Calculation of Fingerprint Probabilities Based on Individual Characteristics. *Journal of the American Statistical Association,* 72**,** 772-778.

Pacheco, I., Cerchiai, B. & Stoiloff, S. 2014. Miami-Dade Research Study for the Reliability of the ACE-V Process: Accuracy & Precision in Latent Fingerprint Examinations. Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/248534.pdf, Last accessed: 9/14/2020.

President's Council of Advisors on Science & Technology 2016. Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. *In:* EXECUTIVE OFFICE OF THE PRESIDENT OF THE UNITED STATES (ed.). Washington, D.C., https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf, Last accessed: 2/3/2021.

Pulsifer, D. P., Muhlberger, S. A., Williams, S. F., Shaler, R. C. & Lakhtakia, A. 2013. An Objective Fingerprint Quality-Grading System. *Forensic Science International,* 231**,** 204-207.

R Core Team 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

R Development Core Team 2017. R: A language and environment for Statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Reneau, R. D. 2003. Unusual Latent Print Examinations. *Journal of Forensic Identification,* 53**,** 531-537.

Rstudio Team 2015. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, Inc.

Scientific Working Group on Friction Ridge Analysis Study and Technology (Swgfast). 2011. *Standards for Examining Friction Ridge Impressions and Resulting Conclusions, ver. 1.0* [Online]. Available: http://www.swgfast.org/documents/examinations-conclusions/111026_Examinations-Conclusions_1.0.pdf, Last accessed 9/14/2020.

Scientific Working Group on Friction Ridge Analysis Study and Technology (Swgfast) 2012.

Standard for the Documentation of Analysis, Comparison, Evaluation and Verification (ACE-V) (Latent), ver. 2.0, http://www.swgfast.org/documents/examinationsconclusions/111026_Examinations-Conclusions_1.0.pdf, Last accessed: 9/14/2020.

Scientific Working Group on Friction Ridge Analysis Study and Technology (Swgfast) 2013.

Standards for Examining Friction Ridge Impressions and Resulting Conclusions, ver. 2.0. http://clpex.com/swgfast/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf, Last accessed: 9/14/2020.

Stern, H. S., Angel, M., Cavanaugh, M., Lai, E. L. & Zhu, S. 2018. Assessing the complexity of handwritten signatures. *Law, Probability and Risk,* 17**,** 123-132.

Stoney, D. A., De Donno, M., Champod, C., Wertheim, P. A. & Stoney, P. L. 2020. Occurrence

and associative value of non-identifiable fingermarks. *Forensic Science International,* 309**,** 110219, https://doi.org/10.1016/j.forsciint.2020.110219, Last accessed: 2/3/2021.\.

Stoney, D. A. & Thornton, J. I. 1986. A Method for the Description of Minutiæ Pairs in Epidermal Ridge Patterns. *Journal of Forensic Sciences,* 31**,** 1217-1234.

Swofford, H. 2015. Information Paper: Use of the term "Identification" in Latent Print Technical Reports. Forest Park, GA. http://onin.com/fp/DFSC_LP_Information_Paper_Nov_2015.pdf, Last accessed: 9/14/2020.

Swofford, H., Steffan, S. M., Warner, G., Bridge, C. & Salyards, J. 2013. Inter- and Intra-Examiner Variation in the Detection of Friction Ridge Skin Minutiae. *Journal of Forensic Identification,* 63**,** 553-570.

Tabassi, E., Olsen, M. A., Makarov, A. & Busch, C. 2013. Towards NFIQ II Lite. Washington DC.

Tabassi, E., Wilson, C. L. & Watson, C. I. 2004. Fingerprint Image Quality. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=905710, Last accessed: 9/14/2020.

Tangen, J. M., Thompson, M. B. & Mccarthy, D. J. 2011. Identifying Fingerprint Expertise. *Psychological Science,* 22**,** 995-997.

The National Judicial College & Justice Speakers Institute 2019. *Science Bench Book for Judges*, The National Judicial College / Justice Speakers Institute.

Ulery, B. T., Hicklin, R. A., Buscaglia, J. & Roberts, M. A. 2011. Accuracy and Reliability of Forensic Latent Fingerprint Decisions. *Proceedings of the National Academy of Sciences, USA,* 108**,** 7733-7738.

Ulery, B. T., Hicklin, R. A., Buscaglia, J. & Roberts, M. A. 2012. Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. *PLoS ONE,* 7**,** e32800.

Ulery, B. T., Hicklin, R. A., Kiebuzinski, G. I., Roberts, M. A. & Buscaglia, J. 2013. Understanding the Sufficiency of Information for Latent Fingerprint Value Determinations. *Forensic Science International,* 230**,** 99-106.

Ulery, B. T., Hicklin, R. A., Roberts, M. A. & Buscaglia, J. 2014. Measuring What Latent Fingerprint Examiners Consider Sufficient Information for Individualization Determinations. *PLoS One,* 9**,** e110179.

Ulery, B. T., Hicklin, R. A., Roberts, M. A. & Buscaglia, J. 2015. Changes in latent fingerprint examiners' markup between analysis and comparison. *Forensic Science International,* 247**,** 54-61.

Ulery, B. T., Hicklin, R. A., Roberts, M. A. & Buscaglia, J. 2016. Interexaminer variation of minutia markup on latent fingerprints. *Forensic Science International,* 264**,** 89-99.

Ulery, B. T., Hicklin, R. A., Roberts, M. A. & Buscaglia, J. 2017. Factors associated with latent fingerprint exclusion determinations. *Forensic Science International,* 275**,** 65-75.

United States Department of Justice & Office of the Inspector General - Oversight and Review

Division 2006. A Review of the FBI's Handling of the Brandon Mayfield Case (unclassified and redacted). Washington DC, https://oig.justice.gov/sites/default/files/legacy/special/s0601/final.pdf, Last accessed: 9/14/2020.

Van Asch, V. 2013. Macro-and micro-averaged evaluation measures [[BASIC DRAFT]]. https://pdfs.semanticscholar.org/1d10/6a2730801b6210a67f7622e4d192bb309303.pdf, Last accessed: 9/14/2020.

Wang, J., Olsen, M. A. & Busch, C. Finger image quality based on singular point localization. 2014. 907503-907503-14.

Wei, R. & Wang, J. 2018. multiROC: Calculating and Visualizing ROC and PR Curves Across Multi-Class. R package version 1.1.1 ed.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Mueller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software,* 4**,** 1686.

Yen, R. & Guzman, J. 2007. Fingerprint image quality measurement algorithm. *Journal of Forensic Identification,* 57**,** 274-287.

Yoon, S., Cao, K., Liu, E. & Jain, A. K. 2013. LFIQ: Latent Fingerprint Image Quality. *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on***,** 1-8.

Yoon, S., Liu, E. & Jain, A. K. On Latent Fingerprint Image Quality. *In:* GARAIN, U. & SHAFAIT, F., eds. Proceedings of the 5th International Workshop on Computational Forensics, November 2012 Tsukuba, Japan. Springer Verlag, pp. 67-82.