2023

Enhancing the forensic comparison process of common trace materials through the development of practical and systematic methods

Meghan Nicole Prusinowski
West Virginia University, mnp0006@mix.wvu.edu

Enhancing the forensic comparison process of common trace materials through the development of practical and systematic methods

Meghan Prusinowski

Dissertation submitted
to the Eberly College of Arts and Science
at West Virginia University

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in
Forensic and Investigative Science

Tatiana Trejos, Ph.D., Chair
Aldo Romero, Ph.D.
Cedric Neumann, Ph.D.
Casper Venter, Ph.D.

Department of Forensic and Investigative Science

Morgantown, West Virginia
2023

Keywords: physical fit analysis, polymers, method development, interlaboratory studies, statistical analysis

Abstract

Enhancing the forensic comparison process of common trace materials through the development of practical and systematic methods

Meghan Prusinowski

An ongoing advancement in forensic trace evidence has driven the development of new and objective methods for comparing various materials. While many standard guides have been published for use in trace laboratories, different areas require a more comprehensive understanding of error rates and an urgent need for harmonizing methods of examination and interpretation. Two critical areas are the forensic examination of physical fits and the comparison of spectral data, which depend highly on the examiner's judgment.

The long-term goal of this study is to advance and modernize the comparative process of physical fit examinations and spectral interpretation. This goal is fulfilled through several avenues: 1) improvement of quantitative-based methods for various trace materials, 2) scrutiny of the methods through interlaboratory exercises, and 3) addressing fundamental aspects of the discipline using large experimental datasets, computational algorithms, and statistical analysis.

A substantial new body of knowledge has been established by analyzing population sets of nearly 4,000 items representative of casework evidence. First, this research identifies material-specific relevant features for duct tapes and automotive polymers. Then, this study develops reporting templates to facilitate thorough and systematic documentation of an analyst's decision-making process and minimize risks of bias. It also establishes criteria for utilizing a quantitative edge similarity score (ESS) for tapes and automotive polymers that yield relatively high accuracy (85% to 100%) and, notably, no false positives. Finally, the practicality and performance of the ESS method for duct tape physical fits are evaluated by forensic practitioners through two interlaboratory exercises. Across these studies, accuracy using the ESS method ranges between 95-99%, and again no false positives are reported. The practitioners' feedback demonstrates the method's potential to assist in training and improve peer verifications.

This research also develops and trains computational algorithms to support analysts making decisions on sample comparisons. The automated algorithms in this research show the potential to provide objective and probabilistic support for determining a physical fit and demonstrate comparative accuracy to the analyst. Furthermore, additional models are developed to extract feature edge information from the systematic comparison templates of tapes and textiles to provide insight into the relative importance of each comparison feature. A decision tree model is developed to assist physical fit examinations of duct tapes and textiles and demonstrates comparative performance to the trained analysts. The computational tools also evaluate the suitability of partial sample comparisons that simulate situations where portions of the item are lost or damaged.

Finally, an objective approach to interpreting complex spectral data is presented. A comparison metric consisting of spectral angle contrast ratios (SCAR) is used as a model to assess more than 94 different-source and 20 same-source electrical tape backings. The SCAR metric results in a discrimination power of 96% and demonstrates the capacity to capture information on the variability between different-source samples and the variability within same-source samples. Application of the random-forest model allows for the automatic detection of primary differences between samples. The developed threshold could assist analysts with making decisions on the spectral comparison of chemically similar samples.

This research provides the forensic science community with novel approaches to comparing materials commonly seen in forensic laboratories. The outcomes of this study are anticipated to offer forensic practitioners new and accessible tools for incorporation into current workflows to facilitate systematic and objective analysis and interpretation of forensic materials and support analysts' opinions.

Acknowledgments

After almost six years in the forensic science department at West Virginia University, I can only say that the experiences I have had here during my graduate education have been life-changing. I started this journey with no background in forensic science and, since then, have learned a tremendous amount about the field while growing my skills and professional network. I am so grateful for all those who have supported me during this time, and I would like to thank them all for their unwavering support.

First, I would like to thank my committee chair and mentor, Dr. Tatiana Trejos. She introduced me to forensic trace evidence analysis and pushed me out of my comfort zone to help me grow. She has been one of my biggest supporters and a truly exceptional mentor. She has paved the way for many unique and incredible opportunities over the last few years and has helped me grow professionally and personally. Words are insufficient to express my gratitude for the opportunities and support she has provided me.

I am also indebted to my other committee members. Each of my dissertation committee members has provided extensive personal and professional guidance and taught me much about scientific research and life. Dr. Casper Venter provided suggestions and support for this project. Dr. Aldo Romero introduced me to machine learning and provided substantial support as I learned about the field and how it could apply to my research. Dr. Cedric Neumann helped me develop my analytic and statistical skills and helped to sustain the high quality of the statistical foundation of this research. They all provided guidance and support throughout my project, and I am exceptionally grateful.

I would also like to thank my research group. We often joke that we are a large family, which couldn’t be more accurate. All the members have helped me throughout my research, assisting in sample collection, preparation, and data analysis. They shared my victories and were there for me in the rough and challenging times when I wondered if I could finish. I would never have made it this far without them.

And finally, thank you to my family and friends, who have endured this long process by my side, offering support and love. Completing this degree results from your endless support, and I could not be more grateful to you all.

Table of Contents

Chapter 1: Introduction

1.1. Problem Statement

Trace materials often play an important role in forensic investigations as they can provide information about potential sources and the activities that may have led to the contact and exchange of traces between a place, an object, and an individual of interest. Examining these traces often requires microscopical and instrumental analyses to characterize and compare a recovered item to a known source. These analyses usually begin with physically comparing items to determine if they could have been joined together. Without a physical fit, the practitioner compares the materials' physical, optical, and chemical features to determine if differences warrant an exclusion or if the observed similarities support an association. Finally, the examiner provides an opinion of the significance of the findings, which can include interpretation at the source or activity level to explain the relevance of the evidence under the given hypothesis.

Although most trace examination and interpretation processes enjoy scientific validity and standard methods, some areas still need to be improved regarding the perceived risk of subjectivity and bias. For example, evaluating physical fits has been a longstanding procedure in forensic science. The assumption has been that a physical fit allows for two items to be individualized to a single source. However, while the ability of analysts to determine the presence of physical fits has been assessed in research studies and proficiency exams, the examination depends only on the scientist's judgment, which is problematic in the absence of protocols with consensus-based criteria. On the other hand, some comparisons of spectral data are based solely on human-based decisions of the similarity or difference observed when two or more spectra are overlaid. Spectral overlay comparisons have been recently criticized for the level of subjectivity and inter-examiner variability in defining what constitutes an exclusionary difference.

As a result, there is a need to develop tools that can lead to more systematic and objective examinations and reports. This research aims to enhance objectivity in analyzing and interpreting polymeric materials commonly examined in trace evidence, such as tape and plastics. These goals were addressed through two main approaches: 1) the development and validation of quantitative criteria that can be used to support and inform examiners' opinions during a physical fit examination, and 2) the development of quantitative metrics for comparing spectral instrumental data.

The first objective of this study is to assess the practical application and performance of a method for the systematic comparison of polymers' physical fits. To address the need for systematic methods in physical fit examinations, this research expands upon a technique previously developed in our group for evaluating the quality of duct tape physical fits by substantially increasing the size of the dataset, conducting inter-laboratory exercises, and utilizing various statistical analyses to assess factors that can affect the quality of a fit. The method provides a similarity metric named the edge similarity score (ESS) that provides a quantitative value of the quality of a fit. The ES score method is reproducible, allows for a straightforward and transparent peer review process, and provides statistical support to the examiners' conclusions. We hypothesized that inter-laboratory collaborations could lead to fine-tuning the method, assessing practicality, evaluating sources of errors, and developing strategies for future adoption at crime labs. This study also compares the performance of human-based approaches to a computational algorithm for comparing fractured duct tape edges. We hypothesized that automated algorithms for comparing edges of fractured materials could provide a rapid analysis of a large population of samples and an objective assessment of the similarity of two separated edges. In addition, this research also includes the new development of a systematic procedure and quantitative criteria for physical fit examinations of rigid polymers, such as those typically separated from vehicles involved in accidents or criminal activities. These methods are anticipated to support analysts conducting physical fit evaluations of trace materials and serve as a foundation for future research and the development of standardized criteria.

The second objective of this study is to develop a quantitative and reproducible method for comparing micro-X-ray Fluorescence (µXRF) data collected from electrical tapes. The proposed approach introduces spectral contrast angle ratios (SCAR). The value of SCARs is demonstrated by comparing X-ray Fluorescence (XRF) data collected from electrical tape backings originating from different areas of the same rolls and various manufacturing sources and qualities of tapes. The approach has shown high discrimination power through a combination of SCARs and a two-stage approach to interpretation. It provides consistent information on data collected from various instruments and configurations. This approach offers a more widespread and transparent interpretation of spectral data.

By completing this research, the application of quantitative evaluation of similarity scores will give examiners statistical support for making decisions on physical examinations and µXRF spectral comparisons of trace materials. The assessment of error rates will serve as an experimental basis for the proposed methods' reliability. This study is anticipated to provide a means to facilitate the standardization of protocols for the identification and examination of relevant features, establish comparison criteria, transparent peer-review processes, and data-driven conclusions. The methods are intended to be easily incorporated into current protocols of trace materials and be widely applicable for various applications.

1.2. Literature Review and Background Information

Manufacturing and Use of Polymer Materials

Pressure-sensitive tapes and automotive polymers were selected as materials of interest in this research as they are commonly encountered as evidence in criminal events. The following sections describe polymer materials' main composition and manufacturing process to provide context for the methods proposed in this study.

Polymer materials are substances composed of monomers connected in chains to form large molecules. Polymers can comprise repeated monomers of the same chemical composition or molecular weight and structure. Also, most natural and synthetic polymers are made up of two or more types of monomers, known as copolymers [1]. Polymers, such as proteins, starches, cellulose, and nucleic acids, occur naturally. A particular type of polymer found in adhesives is isoprene, which is foundational to materials such as resin and rubber. Inorganic polymers that occur naturally include graphite and diamond. In addition, synthetic polymers are common materials used in everyday life. For example, polyethylene is a commonly manufactured polymer used in many items such as coatings, packaging, vehicle parts, bottles, and containers. Polyvinyl chloride (PVC) is another synthetic polymer used in films, coatings, and larger items such as pipes [2]. Polymer-based plastic products represent an extensive industry and are manufactured and used worldwide. According to Plastics Europe, in 2021, more than 390 million metric tons of plastic were produced worldwide. That value does not include polymers used to make adhesives, sealants, paints, or textiles, meaning that the number of polymer-based products is even higher [3].

Pressure-sensitive tapes, including duct, electrical, and packaging, are common consumer products that adhere to a substrate after applying light pressure. These types of tape typically contain multiple layers, with most at least including a backing and an adhesive. Additional layers may be present depending on the intended purpose of the tape. In duct tapes, the most common backing composition is polyethylene. Electrical tapes commonly contain polyvinyl chloride mixed with substances that add flexibility and flame-retardant properties [4-5]. Adhesive tape layers are usually made of a combination of natural or synthetic rubber (e.g., styrene-isoprene-styrene or styrene-butadiene). Additional components within the adhesive serve as tackifiers (terpenes or rosin derivatives), extenders (amines, phenols), colorants (titanium dioxide, zinc oxide, or carbon black), or fillers (talc, clay, calcium carbonate) [6-8]. In addition to the backing and adhesive layers, duct tape contains an additional layer to reinforce and strengthen the final product. This layer is known as the scrim, typically composed of a fiber mesh-like medical gauze. Scrim reinforcement fibers are placed between the backing and adhesive and are mostly made of cotton, polyester, rayon, or a mixture [7-8]. The composition and the weave or knit pattern can vary by manufacturer. Fibers are usually woven in a simple weave pattern in which the successive yarns are varied over and under each other. However, different patterns, such as a weft-insertion, are used [8].

The manufacturing of tape first involves making the tape's main components separately. The relevant layers (such as the backing film and reinforcement layer) are pressed together and coated with a film of liquified adhesive through a series of calendar rollers [8-10]. Some backings are plain color, while others have printed patterns of multiple colors and shapes. Electrical tapes typically are one color, while duct tapes have a wider variety of designs, colors, and patterns. Most tapes are rolled into large “jumbo” rolls for storage until the tape is cut to its final size. As such, all tape rolls originating from the same “jumbo” roll typically have very consistent physical and chemical characteristics. As a complicating factor in tracking back the manufacturing source of a tape sample, manufacturers of tape products often sell to multiple distributors. These distributors usually relabel the tape items under their own brand name, meaning that tape rolls sold under a particular brand name have the potential to share physical and chemical compositions as other rolls with different brand names [9]. Examiners must consider that in cases where the investigation requires the characterization and identification of potential manufacturers. Those investigative leads often require databases and collaborations between the forensic laboratory and industry to maintain updated exemplars.

Unlike the polymers used in films, the composition of hard plastics aims for firmness in the final product; this can incorporate a wide variety of plastic products such as interior and exterior vehicle parts, pipes used in plumbing, flooring materials, containers such as bottles and jars, packaging materials, disposable cutlery, and many more. These materials are primarily composed of polymers and composites such as acrylic or polymethyl methacrylate (PMMA), polycarbonate (PC), polyethylene (PE), polypropylene (PP), polystyrene (PS), polyethylene terephthalate (PETE or PET), or polyvinyl chloride (PVC). [2]. Many techniques are used to manufacture materials using these polymer compounds, the most common of which involves molding. Injection molding is a manufacturing process in which heated liquid material is injected into a pre-made mold and allowed to cool. The finished product conforms to the shape of the mold when removed [2]. Products made using injection molding include plastic containers (i.e., water bottles, soap containers), vehicle headlights, and taillights. Another type of injection molding is reaction injection molding (RIM), used with thermosetting polymers. This class of polymers cures as it cools, providing additional durability and stability, as they do not soften when exposed to heat after cooling [11]. Examples of plastic products made using RIM include circuit boards, packaging materials, and automotive bumpers. These polymers are common as they are durable and resistant to many substances that would otherwise damage or weaken the polymer material.

The manufacture and composition of the end-product of soft and hard polymers are quite different; as a result, analysis methods, physical characteristics, and chemical composition vary considerably. The relevance of these differences is discussed below regarding how this information contributes to forensic investigations.

Forensic Analysis of Polymer Materials

Polymer materials are ubiquitous in everyday life, so they are commonly recovered from crime scenes as potential evidence. For example, paint chips and fragments of headlights transferred to a hit-and-run victim can link the scene to a suspect vehicle; or tape illicitly used to commit a crime can be associated with a suspect roll [11]. For examination purposes, fibers, paints, tapes, and polymers are trace evidence types for which standardized analysis methods are available [12-21]. Duct tapes are found in various cases – used to gag or restrain victims in violent crimes such as homicides, kidnappings, and rapes [6]. Electrical tapes are also commonly found in crime scenes, such as those involving an improvised explosive device [7]. Tapes tend to be pliable, and their separation can be performed by hand or using a cutting instrument such as a knife, cutter, or scissors. Hard plastics are often associated with cases where the original material fractures by impact, such as hit and runs, burglaries, and vandalism [11]. Hard plastics are brittle, breaking by tension when enough force is applied [22]. Automotive plastics are generally found in relatively large-sized fragments in scenes such as hit-and-runs and car accidents.

Figure 1-1 demonstrates a typical analytical approach to the forensic analysis of polymers. Like most materials analyzed in trace evidence, the primary trend of analysis is to start with the most non-destructive technique(s) and then, as needed, move into the more destructive methods. This ensures that most evidence is left intact for additional analysis whenever possible. Some techniques provide complementary information, so the specific methods depend on the evidence type and available instruments. Many polymer materials share similar analytical techniques. While those for tapes, paints, and textiles have been widely studied and standardized, the analysis of other polymer materials is less regulated.

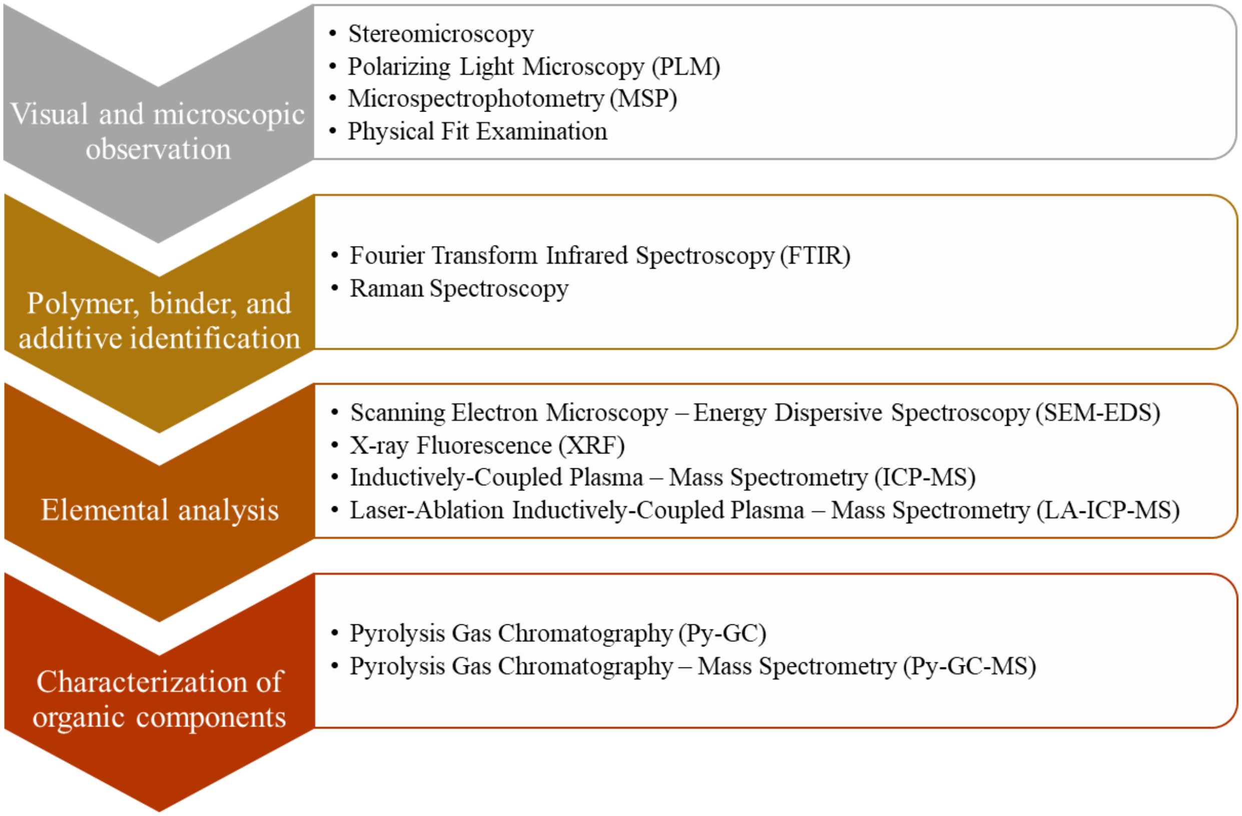

Visual and microscopic observation
• Stereomicroscopy
• Polarizing Light Microscopy (PLM)
• Microspectrophotometry (MSP)
• Physical Fit Examination
Polymer, binder, and additive identification
• Fourier Transform Infrared Spectroscopy (FTIR)
• Raman Spectroscopy
Elemental analysis
• Scanning Electron Microscopy – Energy Dispersive Spectroscopy (SEM-EDS)
• X-ray Fluorescence (XRF)
• Inductively-Coupled Plasma – Mass Spectrometry (ICP-MS)
• Laser-Ablation Inductively-Coupled Plasma – Mass Spectrometry (LA-ICP-MS)
Characterization of organic components
• Pyrolysis Gas Chromatography (Py-GC)
• Pyrolysis Gas Chromatography – Mass Spectrometry (Py-GC-MS)

Figure 1-1. Typical analytical methods used in polymer analysis. The general trend is to move from least destructive to most destructive while providing as much compositional information as possible. In instances where there are known and unknown samples, the analysis will continue until exclusionary differences are observed, at which point the source will be excluded.

When receiving trace evidence from a crime scene, the most common first step is physically examining the items. Physical characteristics are used to identify and classify samples and allow for the rapid exclusion of samples when the features as clearly distinctive, such as different colors or textures. Physical examination is also generally non-destructive, meaning the sample remains intact and usable for additional analysis techniques.

A physical examination can provide valuable information for tapes as some types are more distinguishable based on physical characteristics. Duct tapes are manufactured in various colors and designs to help compare known and unknown samples. For duct tape samples, the color and thickness of the backing and adhesive, the texture of the backing, the width of the tape, and manufacturing marks are all characteristics that can be used to distinguish between items [9, 19, 23]. In addition, the scrim layer can be evaluated. The two directions of tape scrim are known as the warp and weft, where the warp is in the direction of manufacturing along the length of the tape, and the weft direction is along the width of the tape. The scrim is characterized by counting fibers per inch in both warp and weft directions (scrim count), observing the morphology of the yarns in each direction and the weave pattern, and determining the composition of the fibers. These physical characteristics have been found to remain consistent throughout a roll of tape and provide the highest discrimination power between different duct tape rolls [9, 24]. On the other hand, electrical tapes do not have as much variability in physical characteristics. However, like duct tape, the color and thickness of backing and adhesive, texture of the backing, and width of the tape can be used as features to distinguish between samples.

Like tapes, hard plastics have a wide variety of observable physical characteristics. The intended purpose of the material influences macroscopic features on the surface. Polymers from taillights and headlights are generally transparent. They may have some degree of curvature or patterning. At the same time, plastic originating from a vehicle’s bumper is usually opaque and may have layers of paint applied over the plastic substrate. These objects can also come marked with a serial number that can be used to establish the composition of the polymer material as reported by the manufacturer. General features such as the color, thickness, hardness, presence of a coating or layers, patterning and texture, and manufacturing or spontaneous defects in the material are some characteristics that can be observed while comparing these samples. At the microscopic level, the curvature of pieces, smaller fractures radiating from the leading fractured edge, overlapping material, and distinctive fracture patterns and directionality can inform the analyst’s opinion [11, 25].

An extension to the physical examination includes a physical fit evaluation if evidence items do not demonstrate any exclusionary differences in physical features. A physical fit is an association between items based on the realignment of three-dimensional characteristics indicating they were once joined together to form a single object. The characteristics observed to determine the presence of a fit can be inherent to the material (e.g., texture, pattern, construction) or features created by the separation of the items (e.g., displacement of material, scratches, or surface damage). The recommendation for physical fit evaluations of trace materials is to, where possible, examine the items at both a macroscopic and microscopic level to evaluate all potential features and note areas of differences [26]. If no differences and characteristics indicate a physical fit, the items are considered to have the highest degree of association [27]. It is worth noting that although forensic physical fit examinations are relatively straightforward and have a wide-standing history of use in the field, analysis is heavily dependent on expert judgment, and there are limited to no consensus-based standardized methods [10].

Additional examinations can be conducted if a physical fit is not identified. Chemical characterizations of polymer materials are performed using techniques such as Fourier Transform Infrared Spectroscopy (FTIR) and Pyrolysis Gas Chromatography-Mass Spectrometry (py-GC-MS), which can provide information on the main polymer composition for the evidence item, as well as identify other compounds used in the manufacturing of the material [14, 16, 19, 20, 28, 29]. Elemental analysis can be performed using methods such as Scanning Electron Microscopy-Energy Dispersive X-Ray Spectroscopy (SEM-EDS), X-ray Fluorescence (XRF), Inductively coupled Plasma Mass Spectrometry (ICP-MS) Laser-ablation Inductively coupled Plasma Mass Spectrometry (LA-ICP-MS), and Laser-induced Breakdown Spectroscopy (LIBS) [4-6, 19, 29-31]. Unlike duct tapes, which are highly distinguishable by physical examination, electrical tapes are highly distinguishable by elemental analysis methods. Studies considering electrical tape backings have found that a variety of elemental techniques can even provide information regarding manufacturing sources [4, 5, 30, 31].

While methods for tapes, paints, and fibers are standardized, many potential polymer materials are encountered in crime scene investigations, such as latex gloves, polyurethane foam, plastic bags, condoms, and pipe materials used in explosives. Standardizing specific protocols for each polymeric item encountered is impractical. Instead, the focus has been on published standards for specific analytical techniques, including SEM-EDS and py-GC-MS for general polymeric materials [28-29]. Trained practitioners utilize these techniques and additional examinations relevant to the item.

1.3. Studies of Physical Fits of Trace Materials

Current State of Human-Based Examinations of Physical Fits

The foundation of trace evidence relies on the divisible nature of materials generated during a crime and their transfer mechanisms. It is a common belief that when fracture characteristics on separated items correspond with one another, it demonstrates they were once part of the same object, referred to as a physical fit. Due to the high probative value and non-destructive nature of physical examinations, fit comparisons are included in the first steps of evaluating physical evidence. Many materials are submitted to forensic laboratories for fracture fit examination. A survey conducted by the National Institute of Standards and Technology – Organization of Scientific Area Committees (NIST-OSAC) Trace/Materials Subcommittee in 2020 indicated that from these materials, rigid plastics and tapes are most often submitted for physical fit examinations at crime laboratories, followed by textiles, paints, and paper [32]. As described previously, while making observations at the macroscopic and microscopic level is a critical aspect of the overall examination of trace materials, the recognition and documentation of key features of a fractured edge are not currently standardized between laboratories. Only in recent years has the forensic scientific community begun to direct substantial attention to creating consensus protocols for the physical assessment of trace materials and physical fit evaluations. This area is a gap that this research project is intended to address.

For a physical fit, separated evidence items must have a three-dimensional alignment of the edges, including surface, fracture, and manufacturing or acquired features. The physical features imparted during the separation of materials depend in part on whether the material is brittle, malleable, amorphous, crystalline, or fibrous [11, 32]. The material's manufacturing process, composition, and structure are essential in the class characteristics and acquired features left on the fractured pieces. Evaluating the reliability of physical fit determinations introduces many challenges. Assessing a physical fit is inherently subjective, and analysts have no standard practices or methods to use when evaluating a potential fit. This poses a problem as examiners working in different agencies may use different protocols that can lead to inconsistent reports. Methods used in crime laboratories tend to be non-material specific, despite practitioners' acknowledgment that various materials require alternative or additional preparation for physical fit assessment [32]. There is also no consensus in the internal peer-review and verification process of results. Some laboratories only verify samples identified as fit, and some do not require a physical re-examination of the samples during the review [32]. Currently, the reliability and reproducibility of physical fit determinations between analysts are challenging to evaluate without harmonized protocols or verification processes and run the risk of introducing bias [33]. The development of quantitative criteria and determination of error rates for physical fits in trace materials has been a recognized critical research need for over a decade [34-36].

Various research studies have been conducted on physical fits in the trace evidence discipline. Most are published case studies highlighting specific physical fit identifications in casework covering a variety of materials, including paint, metals, textiles, polymers, and paper. Generally, the case reports focus on describing qualitative features or methods that contributed most to assessing the overall confidence in the fit due to more limited sample sizes or the inability to reproduce the fit experimentally [10, 37-43]. Fractography studies also provide a fundamental understanding of how materials fracture. These studies can help determine the suitability of a piece of evidence for physical fit examination, what features are relevant for physical comparison, and how to preserve features for documentation and observation. Fractography studies can address assumptions about the separation of materials, such as the assumption that fracture's features are individualizing and are not replicated randomly [37]. Finally, other studies have provided quantitative assessments of physical fit comparisons using larger sample populations to determine error rates for physical fit determinations. Nonetheless, the lack of understanding of error rates for physical fit comparisons is still a major limiting factor in using physical fit examinations during expert testimony of criminal cases. [37]

Physical Fit Studies of Polymers

As described above, numerous studies address physical fits across various materials. Among those include several publications considering polymer materials, including tapes, soft plastics (such as plastic bags), and hard plastics (such as polymer-based lenses). **Table 1-1** describes some of these studies, including the material in question, the size of the population where relevant, the statistical measures performed, and some of the main findings of each study.

The studies on pressure-sensitive tapes are particularly relevant to this research topic. Tapes are amongst the most studied materials for physical fit comparisons due to the numerous characteristics present and the relative frequency of their submission as crime scene evidence. Electrical tapes have not demonstrated strong suitability for physical fit examinations. It is generally believed they are specifically designed to be highly flexible and deformable, and therefore they do not retain distinctive edge characteristics. Nonetheless, there have not been enough studies to generalize this assumption. In 2011, Bradley et al. published a study that showed that out of a total of 106 true matches in the study, there was a 7.5% false exclusion rate and a 0.049% false positive rate [60]. While there has been some investigation into ways of reverse distortion of electrical tapes for comparison [46-47], the elemental composition of electrical tape backings has been demonstrated to offer an alternative method with highly discriminating features [4, 5, 30, 31]

Duct tapes, however, have shown feasibility for physical fit examination, and various studies have evaluated error rates of physical fit examinations. Bradley et al. in 2006 conducted a study on physical fits of duct tape, for which 92% of 81 total comparison pairs were correctly identified, with no false positives [55]. Following that, McCabe et al. published a study in 2013 that considered a large population set of tape edges, incorporating1600 tape pairs taken from 8 different kinds of tape and four separations methods (hand-torn, scissor-cut, box cutter, and an Elmendorf Tear Tester). The study reported low false positive and false negative errors, with the highest false positive rate at 3.33% and the highest false negative rate at 2.67%. This study was the first to present a potential quantitative measure - a score determined by the relative length of the fracture found to be in alignment with the total width of the edge. However, the authors did not explore how the metric corresponded with the overall performance of their method [57]. These studies formed the foundation on which our previous method was developed.

Table 1-1. A few studies that discuss physical fits of polymer materials [37]. The studies are categorized as case studies, fractography studies, or quantitative studies.

Category	Material Type	Population Size	Experimental Design	Statistical Performance Measures	Main Findings	Reference Number
Case Report	Soft plastic	1 questioned, multiple known exhibits	-Observations of physical features of the questioned and known bags -Elemental analysis via XRF -Visited the manufacturer to gain information on the production process -Determined frequency of individual bag type -Collected reference samples for determination of the period of manufacture time before the feature change -Die line slope method described by Von Bremen and Blunt was used to determine the order of manufacture	Population frequency provided	-Both questioned and known bags were the results of "J sheets" during the manufacturing process, a characteristic appearing on only 2 of 4 stock sheet rolls produced at once -A bag with the same slope as the questioned bag was produced only once every 412 bags produced -Changes in die striae and chemistry are observed in two-hour intervals, in which 254 bags of similar characteristics are produced, spread d over 16 rolls of stock film, and randomly loaded into different bag machines. -Consistency demonstrated in persistent die striae, elemental composition, tie flap offset, bag width, degree of tie-flap centering, and the presence of die flap over-tucks (due to origination from "J-sheets") between the questioned and known bags	38
Case Report	Hard plastic	2 questioned, 2 known	-Broken pieces of a wheel in the well from the scene were visually compared to the wheel well of a suspect's car	None	-Questioned pieces were found to align with known wheel well visually	39
Case Report	Metal, hard plastic	1 questioned, 1 known	-A roof located at a chop shop was compared to the roof beams of a known vehicle	None	-A physical fit was discovered due to physical examination and measurements	40
Case Report	Soft plastics	1 questioned, 1 known	-Trash bag examination for consecutive manufacture determination between questioned bags and known roll -Manufacturing plant to learn of melt pattern characteristics that can be used to associate consecutive trash bags	None	-Manufacturer-imparted, melt pattern characteristics of trash bags, such as lines and arrowheads, can be used to associate consecutive trash bags -These features can be revealed with transmitted lighting	41
Case Report	Soft plastics	4 questioned, 1 known	-Examination under the microscope revealed striations on the surface of questioned sole fragments -Examination of the soles of the suspect's boots revealed similar striations and missing portions -Voids in soles cast in Mikrosil and then compared to the fragments	None	-Direct physical fit inconclusive before casting -Fragments were concluded as having come from the suspect's soles due to alignment in striations between cast voids and sole fragments	42
Case Report	Hard plastic	1 questioned, 1 known	-Questioned blade fragments were compared visually to two known knives -Questioned sample and a section of one of the broken blade fragments were cast using Mikrosil	None	-Casts were found to have similar features -Direct comparison with reverse lighting revealed a physical fit	43

Fractography/ Qualitative	Soft plastics	-13 packages of garbage bags: 10 packages of various brands purchased from local stores; 3 retail packages obtained from 2 manufacturing plants -13 consecutively made garbage bags obtained from a manufacturing plant -7 packages of sandwich bags: 5 of various brands purchased from local stores; 2 obtained from a manufacturing plant	-Bags first examined for color, size, perforations, construction, code, pigment bands, and hairline marks presence or absence -For garbage bags, the production sequence is determined by finding the slope of a prominent marking across all bags -Bags then examined for colored striations under crossed polars, as well as individual characteristics, including fisheyes, arrowheads, streaks, and tiger stripes -Individual characteristics examined on sandwich bags include surface scratches and colored bands	None	-Knowledge from the manufacturing process can be utilized to discern the order or markings across multiple plastic bags -Bags can be thought of as consecutive when both class and individual characteristics align	44
Fractography/ Qualitative	Glass, Metal, Hard plastics	Not given	- Three different loads were used (0.98N, 2.0N, and 2.9N) for a hard indenter to reproducibly create fractures -The second part of the study included bending glass, in which a universal testing machine was used to create reproducible load distributions -The third test was with polymers using an impact "hail-stone gun." Plastic balls were discharged at polymethyl methacrylate (PMMA) sheets -Tensile tests completed on steel wires	None	-Fractures were found to have random distributions of cracks -Cracks themselves were found to have a random number, lengths, propagations, directions, shapes, and orientations -Curves and fractures made in the second study were randomly distributed -Cracks from the impact (third study) were found to be random -Curves and fracture surfaces of the wires were random and varied between the different wires -The steel wires were found to allow for a fracture match between the edges	45
Fractography/ Qualitative	Tape	Not given	-Tapes from six different manufacturers were torn by hand and observed with a comparison microscope -The edges were treated with 100 Celsius hot air for a few seconds -After treatment, the tapes were re-observed under comparison microscopy	None	-Heat treatment was found to make it easier to find the corresponding edge and improved confidence in the conclusion -The author did note, however, that applying heat treatment may destroy other evidence (DNA, fingerprints)	46
Fractography/ Qualitative	Tape	NA	-Tapes were either sheared or torn, heat-treated at 100°C with demineralized water to undo any plastic deformation occurring after fracture, cast with casting material, and each edge of the fracture cast was examined using comparison microscopy for fracture matching	None	-Each tested fracture generated an individual fracture pattern of which a cast could be taken for nearly mirror-image comparison microscopy results	47
Fractography/ Qualitative	Tape	Not given	-Tapes were torn by hand and cut with scissors to demonstrate non-reproducibility	None	-Tearing and shearing black electrical tape samples left distinct tears that were non-reproducible	48

Fractography/ Qualitative	Soft plastics	NA	-A review/recommendation for analysis of garbage bags for consecutive manufacturing identification rather than a study with actual samples -Garbage bags can be aligned according to their heat-sealed edges/ending. Transmitted light from underneath can reveal striations from the manufacturing process that can attribute to a common source	None	-Horizontal streaks in plastic bag material formed during the manufacturing process are in the following categories: 1-fisheyes (randomly distributed dark pigments) 2-arrowheads (triangular striae of dark pigment) 3-tiger stripes (horizontal striae of dark pigment) 4-die lines (become visible in the blowing and stretching process, straight horizontal lines)	49
Fractography/ Qualitative	Soft plastics	NA	-Summary of characteristics of polyethylene films that can be used for comparisons and manufacturing processes	NA	-Additives to films from manufacturing appear as striations/patterning -Extrusion marks originate from the roller -Additional scratches and surface striations come from machine wear -Dye variations come from uneven applications of dye	50
Fractography/ Qualitative	Soft plastics	NA	-A black card was cut to have ⅛ in X 6 ½ slots. Two sheets of glass were put together and placed above the grid. A 500-watt lamp illuminated the grid at a right angle -A camera was focused on putting the whole area of glass in the negative -Polyethylene piece was sandwiched between the glass sheets with the extrusion marks on the short side	NA	-The photography method was found to help visualize and document extrusion marks in polyethylene film	51
Fractography/ Qualitative	Soft plastics	NA	-This paper focuses on photographing the physical characteristics of plastic bags and film that have the potential to be used to denote matching edges or connected pieces of evidence	None	-Extrusion marks are recommended to be photographed using a secondary lens system so that the extrusion marks can be focused at any magnification -Heat marks originate from bags that are sealed together by an individual separately from the manufacturing heat seals -Secondary heat marks were often created using a soldering iron or laundry iron or by commercially made sealing machines -For sealing machines, conclusions could be made by examining the patterns left by the heat-proof fabric on the machine, by observing inclusions and irregularities created in consecutive seals made by the same machine, and by hot spots (unique areas of deformation caused by heat) -Cut edges of films could offer additional details if the instrument used to sever the edges left similar characteristics (snags, changes in cut direction t, etc.)	52

Fractography/ Qualitative	Soft plastics	NA	-Summary of a variety of methods that can be used to visualize and assess the physical properties of plastic bags and cling film -Kinds of properties that can be utilized include color and variation of die lines, polarization patterns, striations from manufacturing -Summary as well of the manufacturing of plastic bags and film: -Manufacturing: plastic bags are made by blowing polymer through a circular tube and then flattened. Cling film is also made by a blown film extrusion but forms a single sheet that is wound up -Finally, four cases mentioned in which characteristics of plastic bags were viewed to allow for matching	None	-Polarization (polarization table): used because many polymeric films are birefringent. Consecutively produced bags often have similar or consecutive colors under cross-polars, and the patterns can be compared to fit matching bags together -Shadowgraph and Schlieren imaging: shadowgraphs involve a point light source at an angle to the film, highlighting discontinuities and defects within the film. The film is photographed in front of the light. For Schlieren, the point source is directed through a convex lens or spherical mirror so that a parallel beam of light passes through the film. A matching lens or mirror catches the light and allows for photography -Incident and transmitted light microscopy: microscopes that can be adjusted to allow for visualization of inhomogeneities of the films -Four cases include an instance of printing defects showing bags produced on the same production line, a case where the polarizations colors demonstrated the bags were produced consecutively, a case where the polarization, die lines, and striations demonstrated consecutive manufacturing, and finally, a case where cling film die lines demonstrated consecutive manufacturing	53
Fractography/ Qualitative	Glass, hard plastic	60 panes of double-strength glass, 60 clear glass wine bottles, 60 polymer tail light lenses	-60 each of three sample types, two fracture methods: dynamic impact and static pressure, 30 samples each, three fracture tips (blunt, round, sharp) -Dynamic: 8x8” glass panes, wine bottles coated with RTV urethane, 5.5/8x4.1/4” plastic lens, 10 glass samples per dropping weight impact tip, 10 plastic lenses per dropping height, reassembled, imaged, and videoed for velocity measurements -Static: 8x8” sample, wine bottles coated with RTV urethane, indenter crosshead speed 10 mm/min, 10 samples per indenter tip (only wide tip used on plastic, so all 30 were the same), load vs. extension measured by Instron software, reassembled and imaged -Visual comparisons: fractures traced onto acetate and overlay one-to-one per sample at four orientations (two for bottles)	None	-Blunt fracture tips required the highest velocity (dynamic) and force (static), while sharp tips required the least -Sharp tip fracture patterns contained the fewest lines, blunt tip patterns contained the most lines Glass panes: Blunt tip created more radial and concentric fractures, and dynamic fracture patterns more simple than static Wine bottles: The number of fractures between impact tips was more evenly distributed, and fracture patterns between dynamic and static samples did not vary as much -Linear relationship expected between load and extension, curvature obtained from load profiles -In plastic lenses, velocity increased as drop height increased, causing a center crushing and edge fracturing -Plastic extension value exceeds glass values; however, the load is smaller	25,54

Quantitative	Tape	5 tests with 10 tape strips per set	-5 test sets: hand-torn from each of three rolls and scissor cut from each of the two rolls -Four examiners, individual assessments of each set. Separate sets per examiner, 20 prepared total	Performance rates	-46/50 or 92% of hand-torn end matches identified correctly -25/31 or 81% of scissor-cut end matches identified correctly -No false positives or negatives; the remaining were inconclusive -2 misidentifications occurred when examiners re-evaluated the scissor cut sets (due to lower matching percentage)	55
Quantitative	Tape	1600 torn pairs for hand-torn 200 Elmendorf-torn 200 scissor-cut 200 box cutter-cut	-4 separation methods (hand torn, Elmendorf torn, scissor cut, box cutter cut) -3 analysts, all peer-reviewing each other	-Contingency tables: inconclusive rate, accuracy rate, false-positive rate, the false-negative rate -Mean and standard deviations calculated for each analyst	Peer review results: -Hand-torn: 9 false negatives, 2 false positives, 37 inconclusive -Elmendorf-torn: 3 false negatives, 0 false positives, 11 inconclusive -Scissor-cut: 4 false positives, 0 false negatives, 1 inconclusive -Box cutter-cut: only one misidentification -Totals: Elmendorf = highest IN rates across examiners; Hand torn NGB NPB 3MGB 3MGG somewhat high; scissor-cut relatively low; box cutter-cut all 0 -Mean accuracy torn tape: 98.58 - 100.00% -Mean accuracy cut tape: 98.15 - 99.83% -Mean false positive rate torn tape: 0.00 - 0.67% -Mean false positive rate cut tape: 0.00 - 3.33% -Mean false negative rate torn tape: 0.00 - 2.67% -Mean false negative rate cut tape: 0.33%	56,57
Quantitative	Tape	11 tape sets, 200 tapes per set, 40,000 inter-comparisons, a total of 440,000 comparisons	-Sets were 200 samples each of the following fracture methods: hand torn (8 sets), Elmendorf torn (1 set), scissor cut (1 set), and box cutter (1 set) -Digital images are taken of all individual ends and fracture pair exemplars -An algorithm was developed to extract coordinates of fracture ends, thresholds set depending on image illumination and tape color, the binary image generated, noise from contamination filtered out -Similarity/distance between coordinates of a fractured pair calculated as the sum of squared residuals (SSR) value to quantify differences. Lower values indicate more similarity between items	-Frequency histograms of the true match and non-match SSR values -Box plots for SSR values among comparisons -Colored matrix plot of SSR values (shows that high and low SSRs are not random and common in certain samples) -SSR means and standard deviations between matches and non-matches	-True matching SSR values were consistently below a critical value -Majority of non-matching SSRs were orders of magnitude larger than matching -In some samples, a non-matching SSR could be even smaller than a matching SSR if fractures were somewhat similar -General grade tapes error rates with 40,000 intercomparisons: 0.0025-0.29% -General grade tapes error rate with 200 intra-comparisons: 0.5-18.50% -Professional grade tapes error rate with 40,000 intercomparisons: 0.085-0.20% -Professional grade tapes error rate with 200 intra-comparisons: 7.0-7.5%	58

Quantitative	Other	12 fracture pairs from silicon, 24 metal-coated paper samples, and 22 Perspex plates	-Fractures illuminated with oblique lighting and scanned -Two computerized systems were developed: one extracts contour representation from fracture image/scan, other compared to the database to generate the statistical probability of the match -Individual similarity scores against the databases determined by the algorithm -Correct matches were classified by human users who marked match points on the software. Pixel distances between the proposed points were then calculated -Classification process told system correct matches and non-matches for different material types and fracture line lengths. Pixel lengths between known matches and non-matches used to generate criteria for the classification of a questioned fracture -Probabilities of occurrence within generated databases used to determine optimal separation criterion for this purpose	Similarity measures between sections of fracture contour: -Difference sum of squares -Difference in standard deviation -Normalized cross-correlation -Histograms and probability density functions for correct match and populations -Likelihood ratios of a match within the material population in the database	-Correct match classification probability: 0.968 -False positive classification probability: 0.0519 -Likelihood ratio of true positive: 18.66 -Positive predictive value: 0.9491 -Bayes risk (false classifications): 0.084 -50% correct criterion positive likelihood ratio: 529 (pairs with a matching error below 0.775 will be classified as correct matches) -Probability of correct classification of a matching pair with error values between 1.05-1.15 = 0.0561 -Probability of a non-match with these error values = 0.0039 -0.93 probability of being a correct pair within these error ranges	59
Quantitative	Tape	30 test sets	-3 examiners performed end matches on 10 sets each of electrical tape fracture pairs -Each set design consisted of factor variation between tape brand, test set preparer, and mode of separation	Performance rates	-2142 end comparisons possible due to various combinations of tape ends -98/106 true matches identified -7 pairs were misidentified as inconclusive, and 1 was a false positive -A secondary reviewer also reported a false positive on the same tape pair -False positive rate was 0.049%	60
Quantitative	Tape	2280 pairs	-Tape pairs of various qualities, either hand-torn or scissor-cut -Number of areas between scrim that matched across tape edges counted (edge similarity score) and conclusion of non-match or match determined -Total population of known non-matches and matches used to evaluate score distribution and performance rates	-Performance rates -Score-based likelihood ratios	-No false positives detected in the dataset -Accuracy reported between 84-99% -ESS higher than 80% supported match, and ESS lower than 25% supported non-match	10

Quantitative	Tape	3300 pairs	-Tape pairs of 3 qualities, either hand-torn or scissor-cut -Performance rates and edge similarity score distributions determined for each set -Analysis of factors performed to see if quality, separation method, and separation have a significant influence on ESS distributions	-Performance rates -Score-based likelihood ratios -Kernel Density Functions - Logistic Regression	-No false positives detected in the dataset -Accuracy reported between 84-99% -Quality and Separation method was determined to have some effect on true fit ESS values, not on true non-fits	61
Quantitative	Tape	136 pieces of tape	- Tape from 3 qualities, all with weft-insert scrim weave -Assigned values of 0, 1, 2 to represent the type of link breakage between loops - Assigned positive or negative value to represent the distance of the next weft from the edge	-Bayesian models -Likelihood Ratios	- Overall showed high discrimination power and relatively low misleading rates of evidence - Bayesian models provided the means to calculate likelihood ratios -Limited by scrim weave type	62
Quantitative	Tape	150 pairs	- Algorithm designed in R to extract the edges of each tape sample and perform edge alignment and comparison. -Cross-correlation function (CCF) was used to assess alignment, and probability of fit -An additional exploration into the comparison of partial edges was performed	Cross-correlation function (CCF)	- Overlap of CCF values between true fit and true non-fit pairs - 48% of the true-fit scores were above the 95th quantile of the true-non-fit distribution - The algorithm could distinguish between a true fit and non-fit edges greater than 27% of the total width	63
Quantitative	Tape	21 pairs	-38 participants across two interlaboratory exercises performed end matches on 7 pairs of hand-torn duct tape -Calculated ESS values reported along with the final decision of fit or non-fit -Feedback received from the first interlaboratory was used to enhance the second	Performance rates -Z-Scores -Dunnett control charts	-Accuracy of 90-100% -Inter-participant agreements generally high -Consistency in reported ESS and conclusion improved in the second study	64

A recent paper considered loop-breaking patterns, referring to the chains created by the fracture of weft-insert scrim fibers, to quantitatively represent the comparison of two duct tape edges. In this study, van Dijk et al. [62] use these quantitative values to calculate likelihood ratios using a Bayesian network and provide evidential value to comparisons. The samples for this study consist of 136 pieces (272 edges, with a total of 127 true fit pairs) from three different quality tape sources. The authors state that when weft-insert duct tape is torn, the warp yarn usually breaks at the point where it is looped onto itself, as this is supposedly the weakest point. Where two loops intertwine, the fracture can cause at least one of these loops to break. For each of these occurrences on the edge, a value was assigned according to the appearance, as either a 0, 1, or 2 depending on if the loop was closed (or missing), open, or complex. In addition, the horizontal position of the edge fibers was denoted, assigned a positive or negative value depending on how far away the first weft fiber was from the fracture was. These values were used to generate Bayesian models and calculate likelihood ratios, which overall showed high discrimination power and relatively low misleading rates of evidence [62]. However, a limitation of this approach is that weft-insert scrim weaves are less common than simple weaves, meaning this method is less widely applicable.

Currently, only some studies address hard polymer materials for forensic physical fit examinations. One comprehensive fractography study is by Tulleners et al. on the fracture of glass and polymers. The authors broke polymer lenses, glass bottles, and glass panes under varying conditions, including different fracture tips and methods. The results indicated that sharp fracture tools required the least velocity and force to puncture the sample. They noted that an impact collision caused the polymer lens to crush instead of splinter. Of note for physical fit examinations, the authors reported that the fractures tended to follow ridges in the molding away from the impact point, potentially due to the distribution of thinner areas on the polymer lens [25, 54].

The basis of this research is our previously published systematic method for the quantitative assessment of the quality of a physical fit [10]. The quantitative metrics developed duct tape is referred to as the edge similarity score (ESS), which measures the quality of the fracture fit in alignment across the edge. The ESS has the potential to be interpreted in a variety of ways. The initial duct tape study evaluated several types of tapes and separation method and inter-analyst agreement. Over 2,000 duct tape comparisons were performed, resulting in overall accuracies between 84.9-99.8%. The quality of the duct tape affected the distribution of edge similarity scores across known true matches and known true non-match samples. The higher quality tape with thicker adhesive and more scrim fibers per square inch produced more negative misidentifications and a wider distribution of scores for the true positive samples. Despite the influence of the tape grade, no false positives were reported for the comparison sets [10]. In addition to discussion of error rates, the distribution of scores were assessed through various exploratory data analysis techniques including boxplots and histograms. The data was also evaluated using kernel density plots and score-based likelihood ratios to provide a proxy for the weight of the evidence [10].

This study serves as the foundation for this research. This dissertation further evaluated the published method through additional sample sets and interlaboratory studies and contrasted it to a computational comparison method. Furthermore, the data collected from the duct tapes were used to answer additional questions that have yet to be extensively addressed in the physical fit analysis. These include the minimum width of tape necessary to determine a physical fit reliably, identification of the most critical features for analysis, and evaluation of factors that could affect the quality of a fit. In addition, it serves as a model for developing a systematic method for comparing hard plastics. However, the characteristics, features, and comparison areas were adapted accordingly for each type of material.

Computational Methods to Aid Physical Fit Comparisons

As more studies are being conducted regarding fractography and physical fit assessments, another avenue that is being explored is using algorithms to assess edges of separated materials. Studies using physical samples and human analysts can take considerable time to complete when conducted on a statistically appropriate number of samples. Algorithms have the potential to utilize a larger population of samples to compare to questioned samples in a shorter period than can be accomplished by traditional physical comparison and provide additional support to the scientific findings. Applying algorithms to forensic evidence has become a growing area of interest as computational capacity has improved. Examples of these include the Automatic Fingerprint Identification System (AFIS), a database of forensic automotive paint samples (Paint Data Query (PDQ)), and the United States Secret Service database for ballpoint ink pens [65-67]. In these systems, data collected from an unknown sample is uploaded and compared to the stored library of samples to determine if the system returns any similar results. Often, these systems will generate a ranked list of potential sources from which an analyst can evaluate and compare the unknown sample against to make a final determination.

A few algorithms have been developed to model the fracture of various materials to learn more about the randomness of fracture patterns. Several models have also been created to associate a corresponding item with a separated fragment. Yekutieli et al. in 2012 [59] published an automated method for examining the edges of paper, silicone, and Perspex (a material like Plexiglas). This model was composed of two systems, the first performed a physical assessment on the edges, and the second collected statistics on the samples used. The input was a scan or image of the original sample, and the output was a series of segments of the material that algorithm determined to be the most likely fit. A large population of true fit and non-fit samples was introduced for the second system, which utilized the populations to determine error rates and likelihoods. Applying the algorithm to these materials resulted in combined error rates of 0.007%, 0.37%, and 0.4%, respectively [59].

Several studies have involved computational models in comparing and evaluating duct tape edges' error rates. One such algorithm was developed by Ristenpart et al. in 2017 [58], using the samples previously analyzed in McCabe's 2013 human-based study [57]. The algorithm was designed to do quantitative image analysis using scanned images of the tape backings. The algorithm used edge detection and smoothing to extract the contour of the edge of a tape backing and compare it against others in the set to try to determine what edge was the most similar. The similarity metric was a sum of square residuals, with lower values found for a true match. Their study determined that 97% of the tears result in the true match having the lowest sum of squares residual. However, they noted that scissor-cut tapes produced a substantially high false positive rate of 61.5%, likely due to straighter edges and fewer distinctive features in cut tape edges [58].

Spaulding et al. designed another recent model [63]. This study incorporated automatic edge detection and used a cross-correlation function and random match probabilities to compare duct tape fits. A total of 44850 comparisons were created, with 150 being true fits. Although there was substantial overlap in similarity scores, 48% of the true-fit scores were above the 95th quantile of the true-non-fit distribution. This study also evaluated partial duct tape edge comparisons, noting that the algorithm could distinguish between true fit and non-fit edges widths greater than 27% of the full width. Like Ristenpart et al. [58], Spaulding et al. noted limitations in the method due to distortions and artifacts introduced to the sample edges. Importantly, Spaulding et al. indicated that the results of their study demonstrated that not all fits hold the same probative value, which is consistent with the results presented in our previous study [10, 63].

It is worth noting that the algorithms described above only considered the contour of the edges. Materials also possess 3D features and surface characteristics of the material and the fracture, providing additional information. Many case studies and physical fit studies note that fracture characteristics and surface features contribute substantially to the ability to determine the presence of a fit; as a result, the performance of the algorithms was outperformed by the corresponding human-based examination. Therefore, in the model designed for this research, one of the aims is to include those features as they are expected to improve overall accuracy and reduce false positives.

1.4. Project Design and Objectives

Overall Project Goals

This research aims to address the crucial need for improving standard methods and criteria for making decisions on comparative analysis of trace evidence materials. Evaluating physical fits has been a longstanding procedure in forensic science. The assumption has been that a physical fit allows for two items to be individualized to a single source. However, while the ability of analysts to successfully determine the presence of physical fits has been assessed in research studies and proficiency exams, only some studies propose systematic and validated criteria that can be used to support decisions.

To address the need for systematic methods, this research expands upon a method previously developed for evaluating the quality of duct tape physical fits. The ESS method has been assessed using additional sample sets supporting the originally developed method for duct tapes. In addition, the results of the human examination of duct tape physical fit pairs are compared against the results of a computational algorithm designed to evaluate the same duct tape edges. Computational models are also used to extract data from the comparison templates, including critical features for comparison, and assessing the minimum width of tape necessary to perform physical fit evaluations reliably. Further verification of the duct tape ESS method has been assessed using results collected from two interlaboratory studies. The second interlaboratory exercise expands and improves the documentation and training for the ESS method based on feedback provided by the first interlaboratory. Both studies demonstrate consistent results between analysts, with high accuracy rates.

In addition, this research introduces a newly developed method for systematically comparing the physical fits of hard polymers. This method defines new comparison criteria and features for brittle plastic materials, and error rates and inter-analyst variability are determined. The similarity metrics are evaluated further through exploratory analysis. This method is anticipated to support analysts conducting physical fit evaluations of trace materials and serve as a foundation for future research and development of additional material-specific methods.

There are five main steps to the development of the systematic methods for physical fits, described below in **Figure 1-2**.

Finally, this research addresses the need for statistical comparison methods for spectral data. Electrical tape elemental data collected from X-ray Fluorescence (XRF) is compared using spectral contrast angle ratios (SCAR). The SCARs are used to evaluate thresholds by which analysts can support a decision of distinguishable or indistinguishable spectra, supporting the currently used method of spectral overlay. In addition, applying a random forest model and a two-stage approach to interpretation provides additional probative value, such as manufacturing and sourcing information.

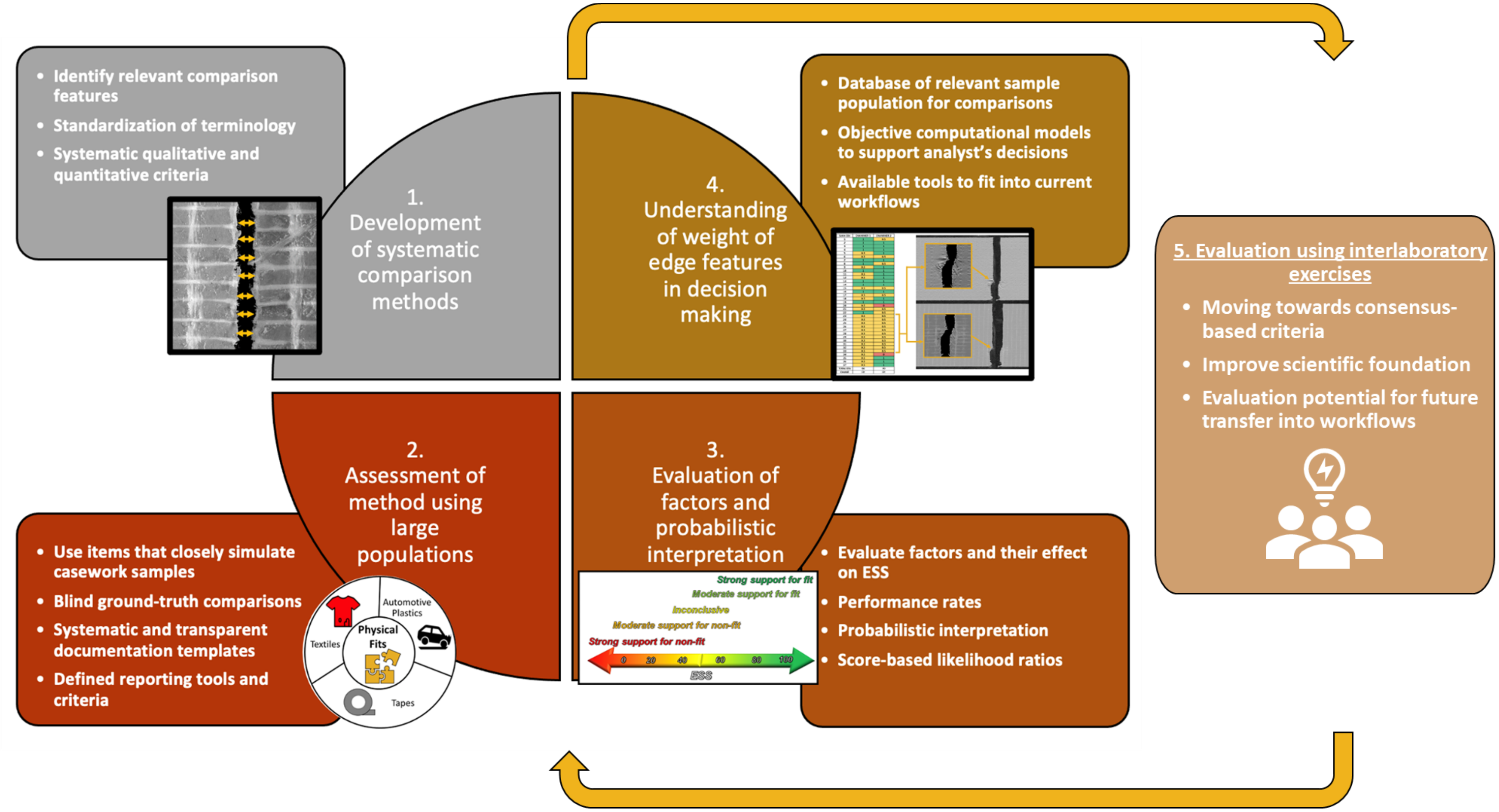

Identify relevant comparison features
Standardization of terminology
Systematic qualitative and quantitative criteria
1. Development of systematic comparison methods
4. Understanding of weight of edge features in decision making
Database of relevant sample population for comparisons
Objective computational models to support analyst's decisions
Available tools to fit into current workflows
5. Evaluation using interlaboratory exercises
Moving towards consensus-based criteria
Improve scientific foundation
Evaluation potential for future transfer into workflows
2. Assessment of method using large populations
Use items that closely simulate casework samples
Blind ground-truth comparisons
Systematic and transparent documentation templates
Defined reporting tools and criteria
Automotive Plastics
Textiles
Physical Fits
Tapes
3. Evaluation of factors and probabilistic interpretation
Strong support for fit
Moderate support for fit
Inconclusive
Moderate support for non-fit
Strong support for non-fit
0
20
40
60
80
100
ESS
Evaluate factors and their effect on ESS
Performance rates
Probabilistic interpretation
Score-based likelihood ratios

Figure 1-2. Diagram depicting the five main steps to the physical fit method development.

Objectives and Respective Tasks

Objective 1. Expand the validation of a human-based systematic comparison method for duct tape physical fits with a more extensive dataset of duct tapes and interlaboratory studies.

Task 1.1. Prepare a dataset of known samples for comparison. The sample sets consist of 3321 tape comparisons, incorporating samples from rolls of varying quality (low, medium, or high). The tape edges are either scissor-cut or hand-torn. The hand-torn subsets are also subjected to manipulation to evaluate the effect of stretching on the quality of a fit, resulting in a total of 9 subsets. The low-quality tape has 650 comparisons (250 scissor-cut, 200 hand-torn, and 200 hand-torn stretched). The high-quality has 647 comparisons (250 scissor-cut, 199 hand-torn, and 198 hand-torn stretched). The medium-quality tape has a total of 2016 comparisons (500 scissor-cut, 508 hand-torn comparisons completed by two independent analysts, and 508 hand-torn stretched). Each comparison results in an edge similarity score (ESS), a quantitative value between 0 to 100% representing the similarity between the edges [10, 61].

Task 1.2. Determine threshold values and error rates for the edge similarity score (ESS) criteria across the tape subsets. Performance rates, including sensitivity (true positive), selectivity (true negative), false positive, false negative, and accuracy, are calculated following each subset's analysis to assess the method's performance. The distribution of ESS values is also explored using boxplots and frequency histograms to identify differences in the distribution of scores between true fit and true non-fit pairs. In addition, probabilistic assessments using score-based likelihood ratios are used to provide a proxy for the probative value of a given pair.

Task 1.3. Evaluate the experimental factors that influence the quality of a physical fit. The factors of separation method (2 levels: hand-torn, scissor-cut), tape quality (3 levels: low, medium, and high), and stretching (2 levels: pristine and stretched) are evaluated using logistic regression models to assess which have a substantial influence on a duct tape physical fit examination.

Task 1.4. Create and manage an interlaboratory study on duct tape physical fits. Create a set of fractured duct tape testing kits, a training guide, and a post-survey to be distributed to interlaboratory participants. A set of physical duct tape fracture edges are created to serve as a basis to assess the duct tape ESS comparison method developed. The samples generated consist of three sets of seven pairs (a total of 21 pairs of tape), for which there are four true fits and three true non-fits per set. Pre-distribution consensus evaluations are conducted to establish expected ESS values and inter-examiner variability for each pair.

Task 1.5. Conduct inter-laboratory exercises on physical fits to assess the proposed fracture fit comparison method and evaluate the performance of the method. Thirty-eight forensic practitioners from 23 laboratories participated in the exercises, resulting in 266 comparisons.

Results are collected from each pair to evaluate error rates, reported ESS values, and the demographic results and feedback in the post-examination survey.

Task 1.6. Determine the performance of the edge similarity score method when utilized by forensic practitioners. Performance rates, including sensitivity (true positive), selectivity (true negative), false positive, false negative, and accuracy, are calculated following each interlaboratory study to assess the method's performance. The distribution of ESS values is also explored using boxplots to evaluate the distribution of ESS scores across participants for each kit and interlaboratory exercise. In addition, the reported similarity scores for the sample kits are compared against the pre-distribution consensus score using Z-scores and Dunnett control charts, which allows for identifying results significantly different from the consensus.

Objective 2. Compare the performance of a computational algorithm for comparisons of duct tape physical fits against the human-based results.

Task 2.1. Digitize a population of duct tape sample edges for use in computational modeling. Approximately 2000 tape samples are digitized using an EPSON 12000XL scanner (SilverFast 8, version 8.8.0r14 software) at a minimum of 600 dots per inch resolution. Both sides of the tape are scanned, accounting for a complete database of approximately 4000 images. The tapes in this study are the same used in the human examination [10, 61] and cover the hand-torn and scissor-cut sets from all three qualities of tape. The contrast of the tape images against the background is maximized using a black background for each image, and irrelevant artifacts and noise are reduced using Photoshop. Pre-processing of the images after digitization is performed using the ForensicFit package in python [68].

Task 2.2. Evaluate the computational algorithm's performance and compare results to those obtained through human-based examination of duct tape edges. Following the convolutional neural network training on the tape sample set, the algorithm's performance is compared to the human analysis for the same samples. Performance rates, including sensitivity (true positive), selectivity (true negative), false positive, false negative, and accuracy, are calculated for each subset to assess the algorithm's performance. These values were then compared to those of the human examination from the same tape sets. The human-based analysis provides a quantitative edge similarity score (ESS), which can be compared to the computational membership prediction (a quantitative probability of belonging to the 'fit' or 'non-fit' class). The distribution of similarity metrics can be compared as an additional assessment of the performance of the computational algorithm.

Task 2.3. Contrast the areas of the tape found to be the most important for human analysis compared to the convolutional neural network. Given pairs of tape can be extracted from the neural network and compared against the documentation of the human analysis. A method of pixel importance demonstrates which pixels of the tape image are recognized and used to decide by the neural network. These areas are observed and compared with the human comparison of the same pair, identifying features in common, such as the edge morphology, pattern of backing, or any protruding fibers.

Objective 3. Extract and interpret edge feature data of duct tape and textile physical fit examinations using computational models.

Task 3.1. Compile bin-by-bin data from the human analysis of duct tape physical fit sets as input data for the algorithm to assess the effect of tape edge size. Five subsets of duct tape pairs (three low-quality sets: scissor-cut, hand-torn, and hand-torn stretched, and two high-quality sets: scissor-cut and hand-torn) were previously analyzed and have documented bin-by-bin quantitative data for each subunit area of comparison. Each bin is assigned a value of 0, 1, or 0.5 for fit, non-fit, or inconclusive. This study accounts for 650 pairs of low-quality tape (32 bins per pair) for 20800 bins and 448 pairs of high-quality tape (55 bins per pair) for 24640 bins. The data from each bin are combined per pair and subset of tape.

Task 3.2. Extract random tape widths from bin-by-bin comparisons and calculate performance rates by bin width. The duct tape pairs' bin-by-bin alignment values (0, 0.5, or 1) are extracted, and sampling is conducted to determine the error rates based on random width segments. For example, at a given width of 10 bins, a selected tape pair has an arbitrary start point chosen, and the width of 10 consecutive bins with the corresponding alignment values would be obtained. The ESS is then calculated for this pair of tapes. This process is repeated for all possible widths for the tapes (32 bins for low-quality and 55 for high-quality). Random selection of tapes and bin widths is performed five times to account for the variation of ESS within tape samples.

Task 3.3. Estimate the minimum width of tape necessary to conduct physical fit examinations reliably. The calculated partial ESS values and resulting fit/inconclusive/non-fit decision can be compared to the ground truth, and error rates determined for each bin width. These error rates are modeled using beta regression in R to evaluate the minimum width for reliable accuracy.

Task 3.4 Extract feature information from documentation templates and determine which features are critical for decision-making. Three subsets of duct tape pairs (low-quality hand-torn stretched, high-quality scissor-cut, and high-quality hand-torn stretched) that were previously analyzed have documented bin-by-bin feature data for each subunit area of comparison following the 3-step ESS method [61]. The eight features of interest are the alignment of severed dimples, alignment of calendaring striations, alignment of warp scrim, alignment of scrim weave, corresponding protruding warp fibers and gaps, distortion explained by stretching directionality, consistent weft scrim near or at the edge of the sample, missing material. The occurrence of these features by tape subset and ground truth is calculated, and their value in deciding a fit or non-fit is assessed using mutual information gain. Mutual information gain can allow for the detection of relationships between variables, which provides information as to whether the presence of a feature can indicate a particular decision.

Task 3.5. Provide a preliminary decision-making guide for analysts using feature occurrence. Using the features highlighted by the mutual information gain, a decision tree indicates which decision (fit, non-fit, inconclusive) should be considered when observing a particular feature. To this end, an analyst could observe the decision tree, observe the features of a given tape pair, and follow the decision tree to a conclusion of fit, inconclusive, or non-fit.

Objective 4. Develop and validate a systematic method to quantify the quality of fracture fits of hard plastics.

Task 4.1. Prepare a dataset of known ground-truth polymer samples for comparison. The sample set of polymers consists of more than 500 comparisons originating from broken heads and taillights taken from vehicles. The head and taillights are broken by hand using tools including a hammer and crowbar, and fragments are then collected and reconstructed to ensure the known ground truth of each edge. Each sample is randomly labeled, and non-participating analysts document the reconstructed full lights and each pair to maintain the ground truth of the pairs. As additional information, infrared spectrums were collected from each distinct polymer type from each source light to provide compositional information.

Task 4.2. Develop a standard method for the quantitative assessment of the quality of physical fits of hard plastics and define comparison features and measurements that can be used for comparison. The comparison edge of each polymer pair is divided into five sections based on the total length of the fracture. A flexible string is used to measure the total length to account for bends or indentations in the polymer surface. A series of features were defined for use in comparison to polymer edges. The features of interest consist of the following: 3D edge alignment, surface plane/directionality alignment, edge curvature/directionality, pattern alignment, surface damage alignment, scratch alignment, extraneous material correspondence, fracture marks alignment, protruding feature correspondence, missing material. For each feature, a quantitative value referred to as a prominence value is assigned in accordance with the presence or absence of a feature, whether it indicates fit or non-fit, and if it is highly distinctive. For example, edge curvature highly indicative of fit would receive a prominence value of +2. These prominence values are summed for all five bins of a pair and are reported as a feature prominence sum (FPS). In addition, each bin of the pair is assigned a value of 0, 0.5, or 1 and reported as a relative percentage to estimate an edge similarity score (ESS).

Task 4.3. Perform blind analysis of a large population of polymer samples following a systematic comparison method. Two independent analysts blindly perform the analysis of the polymers, with the pairs prepared and assigned random numbers by a non-participating analyst to maintain blind ground truth. Five hundred comparisons are performed, with the subsets of polymers divided by the polymer type and morphology, so only pairs of similar polymer types are compared. Some pairs are repeated between the two analysts to evaluate inter-analyst variation. The analysts report a fit, non-fit, or inconclusive conclusion along with ESS and FPS for each pair.

Task 4.4. Assess the performance of the method by each polymer subset and analyst. Following analysis of the polymer subsets, performance rates of the method are calculated, including sensitivity (true positive), selectivity (true negative), false positive, false negative, and accuracy. In addition, the ESS and FPS values are explored using boxplots for true fits and non-fits to evaluate potential threshold values for determining strong support for a fit or non-fit.

Objective 5. Develop and validate a systematic method to compare XRF data of electrical tapes.

Task 5.1. Collect XRF data from known different sources and same source electrical tape backings. Spectral data from 95 black electrical tape backings are collected using XRF, with one roll as a source to evaluate intra-roll variation and the other 94 as different-source samples. Three replicates are analyzed for each sample, and additional duplicates are monitored as quality control [30]. The identified element peaks are used in the subsequent data analysis for intra-roll and inter-roll variation and calculating spectral contrast angle ratios.

Task 5.2. Calculate the spectral angle contrast ratio for each pairwise comparison. Spectral contrast angle ratios (SCAR) are calculated for each pairwise comparison (n=4,561 comparisons). Each spectrum is preprocessed to remove noise and areas with minimal information, correct baseline and normalize intensity. Spectral contrast angles are calculated for comparisons within the three replicates of each sample (within-source angles) and comparisons of each replicate from one sample to each replicate of another (between-source angles). Each sample has 3 within-source angles, and comparing two produces 9 between-source angles. The SCAR is then calculated by taking the ratio between the circular mean of the between-source angles and the circular mean of the within-source angles.

Task 5.3. Assess the potential of SCARs for classification of samples and calculating the weight of evidence. Hierarchical clustering is performed on the proximity matrix representing all pairwise SCARs between the 94 tapes, and then validated using random forest classifiers and comparing it with the manual classification. The classifiers are designed to automatically detect common and differing peaks (based on intensity and slopes) between sets of XRF spectra. Commonality and differences are assessed based on the importance of each wavelength to reach a correct classification. A two-stage approach is used to evaluate the power of the test and calculate the random match probability between given samples.

1.5. Deliverables

This project was primarily funded by the National Institute of Justice (NIJ) under award #2020-DQ-BX-0012. Objectives 1 and 3 correspond to research objectives included in the NIJ grant. All data sets and methods for data processing created during this study will be made available to interested stakeholders and archived by the NIJ. In addition, physical sample collections of approximately 3000 comparison pairs of duct tape and 500 pairs of hard plastic material will be inventoried, as well as the sample kits and templates generated for the interlaboratory studies of duct tapes. A database consisting of the digitized tape samples and a photographic inventory of the polymer samples will be delivered as part of the research. The comparison templates for tapes and hard polymers will be archived. As part of disseminating data related to this research, we have published two scientific publications in peer-review journals, shared research results at six scientific meetings in the form of posters and oral presentations, and three additional publications are in preparation.

1.6. Chapter Structure

Chapter 1 introduces the motivation of this research, the background information of relevant literature, the primary goals and tasks, and the structure of the chapters that compose this dissertation.

Chapter 2 describes the results of a joint multidisciplinary approach proposed in the National Institute of Justice award #2020-. The first part of the chapter describes an expanded evaluation of the edge similarity score (ESS) method for physical fits of duct tapes. This study expands upon the previously developed method that quantifies the similarity of the compared tape edges by systematically studying the effect of several separation methods (hand-torn and scissor-cut) and tape grades (low, medium, and high) on the quality of a fit. Over 3300 pairs of hand-torn or scissor-cut duct tapes from three different tape grades are analyzed following a three-step methodology and systematic reporting protocols. The three steps incorporate first a qualitative assessment of the overall edge alignment and description of edge pattern, second, a macroscopic evaluation of the edges' features, and finally, a bin-by-bin subunit assessment of tape edges and estimation of the edge similarity score. The first step allows for an initial examination of each sample individually and then the first side-by-side comparison. The second breaks down the sample into five macroscopic sections and evaluates several critical features for comparison. The last step involves breaking down the edges into small comparison areas by the scrim fibers, documenting comparison features, and assigning quantitative values to each bin to represent the alignment of each bin. The total values are calculated as the pair's edge similarity score and provide a metric of the fit quality that informs and supports the examiner's opinion. The method offers a visual and practical representation of the fit, non-fit, and inconclusive bins that can assist practitioners in comparing results and the decision-making process during training, technical reviews, and verifications. Statistical analysis of the separation method and tape quality reveals a potential interaction between these factors. It showed they significantly impact the edge scores for true fitting pairs but not the true non-fits' scores. The first portion of this chapter was published as article in Forensic Science International with the author of this dissertation holding primary authorship. [*Forensic Science International*. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – physical fits of torn and cut duct tapes. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567].

The second part of this chapter covers the assessment of a computational comparison algorithm for comparison of duct tape physical fits built in collaboration with Dr. Romero's group at the West Virginia University's Physics and Astronomy Department. The tape samples from the first portion of Chapter 2 are digitized, top and bottom surfaces. Then a convolutional neural network is trained on the database to perform comparisons on the same samples to predict fit or non-fit and estimate the weight of evidence. The sample sets incorporated are those from the hand-torn and scissor-cut sets of all three tape qualities. The computational model demonstrates comparable accuracy to the human examination but shows trends where the model predicts medium-quality and high-quality scissor-cut sets with higher accuracies than the hand-torn sets of the respective tape qualities. In contrast, the opposite is true for the low-quality tape. Using pixel importance, the model demonstrates that it recognizes similar features to the human examination. Overall, this model indicates that there is potential for computational models to support human analysis of physical fits by providing additional objective support.

This portion of Chapter 2 is the product of the joint effort of researchers in both the Forensic and Investigative Science Department and the Physics and Astronomy Department. The design, development, code, training, testing, and validation of the computational algorithm was the product of Pedram Tavadze, Logan Lang, and Aldo Romero. Meghan Prusinowski, Zachary Andrews, and Tatiana Trejos contributed human analysis and imaging of samples, curation of the sample image database, image adjustment, and assisted with testing and interpretation of the model output. This chapter is adapted submitted to Forensic Science International where the author of this dissertation contributed as a major co-author. [*Forensic Science International*. Tavadze P, Lang L, Prusinowski M, Andrews Z, Trejos T, Romero A. Using convolutional neural networks to support examiners in duct tape physical fit comparisons. Under review.]

Chapter 3 describes the results of two interlaboratory studies for duct tape physical fit examinations that evaluate the consistency among practitioners using the new ESS method to analyze, document, and interpret tape edge comparisons. Both studies incorporate the ESS method described in Chapter 2. However, the second interlaboratory study includes more improvements highlighted in Chapter 2, such as the three-step methodology for comparison and a more comprehensive documentation template. The two interlaboratory studies are separated by enhancements based on practitioner feedback from the first interlaboratory exercise. Both studies share a similar experimental design: three different kits containing 7 pairs of samples (4 fit, 3 non-fit). Each kit includes pairs similar in edge morphology and pre-assessed by a panel of consensus analysts. The selected pairs demonstrate less than 10% variability within the panel members. The first study incorporated results from 20 different participants and demonstrated accuracy of 95%. One of the main observations of the first exercise was that there were instances where the participant's reported conclusion did not correspond with the reported ESS value. To address these inconsistencies, the second interlaboratory exercise incorporated more thorough and standardized training and documentation templates, resulting in an overall accuracy of 99% considering the results of 18 participants. This research demonstrates the value of the ESS method standardization and transparency of physical fit examinations. It provides a means for current analysts to share their thought process transparently during training, analysis, and technical review or verification of physical fits. This chapter has been published as an article in Forensic Chemistry with the author of this dissertation holding primary authorship. [*Forensic Chemistry*. Prusinowski M, Brooks E, Neumann C, Trejos T. Forensic interlaboratory evaluations of a systematic method for examining, documenting, and interpreting duct tape physical fits. 2023. 34: 100487. https://doi.org/10.1016/j.forc.2023.100487]

Chapter 4 covers a novel extension of the duct tape ESS method, where information is extracted from the documentation templates with two primary purposes: to determine the minimal width of tape necessary for reliable physical fit examinations, and to evaluate which features are the most critical for decision making on a physical fit. To that end, the documentation templates of several tape subsets (three low quality and two high quality) and a dataset of textiles are used to compile the bin-by-bin quantitative alignment measurements (0, 1, or 0.5 to represent fit, non-fit or inconclusive). Using the collected quantitative data, simulated partial width edges are generated from random segments of the bin-by-bin documentation. With the segments and ground truth, error rates are calculated for each width, and beta regression is performed to evaluate the improvement in accuracy as bin width increases. Separately, feature data extraction allows for analysis using mutual information gain to identify features critical for comparison. The findings from the mutual information gain provide the means to inform a decision tree, by which analysts could follow the branches and make an informed decision based on the observed features of a tape or textile comparison pair. This research addresses several gaps in current practice – namely, the minimum width necessary to complete reliable physical fit examinations, and more transparency on critical decision-making features. The research presented in this chapter is part of an article in preparation that will be submitted to the Journal of Forensic Sciences. The author of this dissertation will hold primary co-authorship with Pedram Tavadze. Dr. Tavadze contributed through the development of the decision tree model, as well as performing the data extraction and statistical analysis for the mutual information. Meghan Prusinowski, Zachary Andrews, and Divyanjali Pulivendhan contributed to the human analysis that provided the bin-by-bin quantitative and qualitative data. Meghan Prusinowski, Pedram Tavadze, Zachary Andrews performed the statistical analysis and interpretation of the results under the supervision of principal investigators Tatiana Trejos and Aldo Romero, in collaboration with Cedric Neumann who provided valuable statistical support and interpretation for this study.

Chapter 5 presents a new method for comparison of hard plastics. This method broadens the previously developed ESS method of tapes into a new material commonly submitted to forensic laboratories for fit examinations. The features and observed characteristics for hard plastics vary considerably from tapes. So, it is necessary to define new comparison criteria and adjust the method to account for the material difference. The comparison method for hard plastics is a two-step method. The first is an overall assessment of the question sample, then the known, and finally a side-by-side comparison. The second step incorporates splitting the sample into five comparison bins based on the overall length of the fracture and includes the observation and documentation of nine distinct features. The feature documentation is accompanied by a prominence value, assigned based on the presence or absence of the feature, if it indicates a fit or non-fit as well as whether it is distinctive. The ESS based on the five comparison bins and the feature prominence sum are reported as final quantitative values representing the similarity between the two pieces. A total of 385 comparison pairs are generated from fracturing head and taillights to evaluate this method. The method results in 85% accuracy, and inter-analyst examinations of a subset of samples demonstrated consistent accuracy and error rates. No false positives were reported, but the rate of inconclusive and false negatives was generally higher than previously observed in other materials using the ESS method. This is hypothesized to be contributed primarily by small edge sizes, scratches and marks left after the separation event, and distortion of some edges. Overall, the preliminary results of this method indicate that with further improvements to feature definitions and documentation procedures, that the combination of ESS and feature prominence values can provide support for comparisons of hard plastics. This chapter will be part of an article in preparation, with a planned submission to Forensic Science International. The author of this dissertation will hold primary authorship. Meghan Prusinowski and Zachary Andrews provided the human analysis of the polymer samples, and worked on interpreting the results. Pedram Tavadze participated in experimental design, sample preparation, and data analysis, and the study was performed with the support of principal investigator Tatiana Trejos.

Chapter 6 addresses the final goal of this research, which is the need for statistical comparison methods for spectral data. This study compares elemental data collected from electrical tape backings using X-ray Fluorescence (XRF) using spectral contrast angle ratios (SCARs). SCARs calculated from one same source roll (n=190 comparisons) and 94 different source rolls (n=4371 comparisons) are used to determine the range of variability within and between sources. A SCAR threshold is devised to assess potential association or exclusion of sources and is optimized to minimize error rates. An automated approach to sample comparison utilizing a random forest algorithm is presented to assist in identifying the basis for similarities or differences between compared spectra. This study also presents an objective approach to reporting opinions and probabilistic determinations of spectral data that can be used as a model for other fields and materials. The SCAR metric performed well for classifying samples and making inferences about the sample's origin. This study provides a means for statistical analysis and comparison of semi-quantitative spectral data, identified as a critical research need in forensic chemical analysis. In addition, this research reinforces manufacturing and source trends observed from additional analysis of the same electrical tape samples using other instruments. This research study will be submitted as an article to Forensic Chemistry. The author of this dissertation will hold primary authorship for the data analysis and writing of this manuscript under the supervision of principal investigator Tatiana Trejos. Evie Brooks participated in data collection and analysis as well as review of the manuscript. The statistical analysis in this study was performed in collaboration with Cedric Neumann who provided valuable statistical support and interpretation.

Chapter 7 summarizes the overall conclusions of this research and proposes future work for additional expansions and improvements upon the comparison methods described in this dissertation. This chapter describes how the forensic trace evidence community benefits from transparent, systematic, and statistically supported methods that provide the means to compare samples and results. In addition, future steps, and challenges towards implementing these methods are discussed.

1.7. References

1. Subramanian M. *Basics of Polymers: Fabrication and Processing Technology*. Momentum Press, 2015. ISBN: 9781606505830.
2. Wittcoff HA, Reuben BG, Plotkin JS. *Industrial Organic Chemicals (3rd Edition), Chapter 17 – How Polymers Are Made*. John Wiley & Sons, 2012. ISBN: 978-0-470-53743-5
3. Plastics Europe. *Plastics – the Facts 2022*. October 2022. https://plasticseurope.org/wp-content/uploads/2022/12/PE-PLASTICS-THE-FACTS_FINAL_DIGITAL.pdf
4. Prusinowski M, Mehltretter AH, Martinez-Lopez C, Almirall J, Trejos T. Assessment of the utility of X-ray Fluorescence for the chemical characterization and comparison of black electrical tape backings. *Forensic Chemistry*. 2019. https://doi.org/10.1016/j.forc.2019.100146
5. Mehltretter A.H., Bradley M.J., Wright D.M. Analysis and Discrimination of Electrical Tapes: Part II. Backings. J For Sci. 2011; 56:1493-1504
6. Trejos T, Koch S, Mehltretter A. Scientific foundations and current state of trace evidence – A review. *For Chem*. 2020. https://doi.org/10.1016/j.forc.2020.100223.
7. Smith J. Forensic Examination of Pressure Sensitive Tape. From *Forensic Analysis on the Cutting Edge: New Methods for Trace Evidence Analysis.* Editor: Blackledge R. John Wiley & Sons, Inc. 2007. ISBN: 9780470166932.
8. Mehltretter A, Wright D, Dettman J, Smith M. Intra-Roll and Intra-Jumbo Variation of Duct Tapes. *J American Soc Trace Evidence Ex*. 2015; 6(1): 21-41.
9. Mehltretter AH, Bradley, M. J. Forensic Analysis and Discrimination of Duct Tapes. *J American Soc Trace Evidence Ex*. 2012; 3(1): 2-20

10. Prusinowski M, Brooks E, Trejos T. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. *Forensic Science International*. 2020. https://doi.org/10.1016/j.forsciint.2019.110103
11. Causin V. *Polymers on the Crime Scene Forensic Analysis of Polymeric Trace Evidence.* Springer. 2015. ISBN: 978-3-319-15494-7. DOI:10.1007/978-3-319-15494-7
12. ASTM International. ASTM E2225 Standard Guide for Forensic Examination of Fabrics and Cordage. 2021. DOI: 10.1520/E2225-18.
13. ASTM International. ASTM E2228 Standard Guide for Microscopical Examination of Textile Fibers. 2018. DOI: 10.1520/E2228-18
14. ASTM International. ASTM E2224 Standard Guide for Forensic Analysis of Fibers by Infrared Spectroscopy. 2018. DOI: 10.1520/E2224-18
15. ASTM International. ASTM E1610-18 Standard Guide for Forensic Paint Analysis and Comparison. 2018. DOI: 10.1520/E1610-18
16. ASTM International. ASTM E2937-18 Standard Guide for Using Infrared Spectroscopy in Forensic Paint Examinations. 2018. DOI: 10.1520/E2937-18
17. ASTM International. ASTM E2808-21a Standard Guide for Microspectrophotometry in Forensic Paint Analysis. 2021. DOI: 10.1520/E2808-21A
18. ASTM International. ASTM E3234-20 Standard Practice for Forensic Paint Analysis Training Program. 2022. DOI: 10.1520/E3234-20
19. ASTM International. ASTM E3260-21 Standard Guide for Forensic Examination and Comparison of Pressure Sensitive Tapes. 2022. DOI: 10.1520/E3260-21
20. ASTM International. ASTM E3085-17 Standard Guide for Fourier Transform Infrared Spectroscopy in Forensic Tape Examinations. 2022. DOI: 10.1520/E3085-17
21. ASTM International. ASTM E3233-20 Standard Practice for Forensic Tape Analysis Training Program. 2022. DOI: 10.1520/ E3233-20
22. Friedrich K, Almajid A. Manufacturing Aspects of Advanced Polymer Composites for Automotive Applications. *Appl Compos Mater*. 2013. 20:107-128.
23. Scientific Working Group on Materials Analysis (SWGMAT). Guideline for assessing physical characteristics in forensic tape examinations. J American Soc Trace Evidence Ex. 2014;5(1):34-50.
24. LaPorte K, Weimer R. Evaluation of Duct Tape Physical Characteristics: Part I – Within-Roll Variability. *J American Soc Trace Evidence Ex.* 2017; 7(1): 15-34
25. Tulleners FA, Thornton J, Baca AC. Determination of Unique Fracture Patterns in Glass and Glassy Polymers, Final Rep. to Natl. Inst. Justice Grant 2010-DN-BX-K219; Univ. Calif. Davis, CA. 2013. (n.d.).
26. OSAC Trace Materials Subcommittee. OSAC 2022-S-0015. *Standard Guide for Forensic Physical Fit Examination.* 2022. https://www.nist.gov/system/files/documents/2021/12/06/OSAC_2022-S-0015_Standard_Guide_for_Forensic_Physical_Fit_Examination_DRAFT_OSAC_PROPOSED.pdf

27. OSAC Trace Materials Subcommittee. OSAC 2022-S-0029 Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials. https://www.nist.gov/system/files/documents/2022/02/28/OSAC%202022-S-0029%20Standard%20Guide%20for%20Interpretation%20and%20Reporting%20in%20Forensic%20Comparisons%20of%20Trace%20OPEN%20COMMENT.pdf
28. ASTM International. ASTM E3296-22, Standard Guide for Using Pyrolysis Gas Chromatography and Pyrolysis Gas Chromatography-Mass Spectrometry in Forensic Polymer Examinations. 2022. DOI: 10.1520/ E3296-22
29. ASTM International. ASTM 2809-22 Standard Guide for Using Scanning Electron Microscopy/Energy Dispersive X-Ray Spectroscopy (SEM/EDS) in Forensic Polymer Examinations. 2022. DOI: 10.1520/2809-22
30. Brooks E, Mehltretter AH, Prusinowski M, Trejos T. Optimization and Validation for Qualitative Elemental Analysis of Electrical Tape Backings by X-ray Fluorescence. *Forensic Chemistry*. 2020. https://doi.org/10.1016/j.forc.2020.100291
31. Martinez-Lopez C, Trejos T, Mehltretter AH, Almirall JR. Elemental analysis and characterization of electrical tape backings by LA-ICP-MS. *Forensic Chemistry*. 2017. 4: 96-107.
32. Gross, S. Physical Fit Task Group – Trace Materials Subcommittee of NIST-OSAC. Survey of Physical Fit Protocols. 2019. Unpublished Survey, presented at NIST-OSAC Trace Subcommittee Meeting, March 2020, OK.
33. Dror I. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Anal. Chem.* 2020. 92:7998-8004.
34. OSAC Research Needs Trace Materials Subcommittee. Development of Quantitative Assessment and Evaluation of Error Rates in Physical Fit Determinations of Trace Materials. 2021. https://www.nist.gov/system/files/documents/2021/06/23/Trace%20Materials%20SC_R%26D%20Need_Development_of_Quantitative_Assessment_and_Evaluation_of_Error_Rates_in_Physical_Fit_Determinations_of_Trace_Materials.pdf
35. National Academy of Sciences (NAS). Strengthening Forensic Science in the United States: A Path Forward. 2009. doi: 0.17226/12589
36. President's Council of Advisors on Science and Technology (PCAST). Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/otsp/PCAST/pcast_forensic_science_report_final.pdf
37. Brooks E, Prusinowski M, Gross S, Trejos T. Forensic Physical Fits of Trace Materials: A Review. *Forensic Science International.* August 2020. Available online May 29th, 2020. https://doi.org/10.1016/j.forsciint.2020.110349
38. S. Ryland, M.M. Houck, Only circumstantial evidence, in M.M. Houck (Ed.), MuteWitn. Trace Evid. Anal., Academic Press, San Diego, CA, 2001, pp.117–137.
39. K.M. Streine, An interesting physical fracture match, AFTE J. 39 (2007) 68–69.

40. C. Caine, E. Thompson, Physical match of an automobile roof to the body section, AFTE J. 21 (1989) 632–634.
41. R.J. Kopec, C.R. Meyers, Comparative analysis of trash bags - a case history, AFTE J. 12 (1980) 23–26.
42. B. Moran, Physical match/tool mark identification involving rubber shoe sole fragments, AFTE J. 16 (1984) 126–128.
43. Y. Garcia, A fracture match in a police-involved shooting investigation, AFTE J. 44 (2012) 182–183.
44. U.G. Von Bremen, L.K.R. Blunt, Physical comparison of plastic garbage bags and sandwich bags, J. Forensic Sci. 28 (1983) 644–654, doi: http://dx.doi.org/ 10.1111/j.1365-313X.2011.04857.x.
45. H.W. Katterwe, Fracture matching and repetitive experiments: a contribution of validation, AFTE J. 37 (2005) 229–241.
46. B. Weimar, Physical match examinations of adhesive PVC-tapes: improvement of the conclusiveness by heat treatment, AFTE J. 40 (2008) 300–302.
47. B. Weimar, Physical match examination of the joint faces of adhesive PVC tapes, AFTE J. 40 (2008) 300–302, doi: http://dx.doi.org/10.1002/ cbdv.200490137/abstract.
48. N. Agron, B. Schecter, Physical comparisons and some characteristics of electrical tape, AFTE J. 18 (1986) 53–59, doi: http://dx.doi.org/10.2174/092986652566617121411007.
49. J.R. Vanderkolk, Identifying consecutively made garbage bags through manufactured characteristics, J. Forensic Identif. 45 (1995) 38–50, doi: http:// dx.doi.org/10.2174/092986652566617121411007.
50. D.S. Pierce, Identifiable markings on plastics, J. Forensic Identif. 40 (1990) 51–59.
51. S. Denton, Extrusion marks in polythene film, J. Forensic Sci. Soc. 21 (1981) 259–262.
52. K.N. Ford, The physical comparison of polythene film, J. Forensic Sci. Soc. 15 (1975) 107–113.
53. D.A. Castle, B. Gibbins, P.S. Hamer, Physical methods for examining and comparing transparent plastic bags and cling films, J. Forensic Sci. Soc. 34 (1994) 61–68.
54. A.C. Baca, J.I. Thornton, F.A. Tulleners, Determination of fracture patterns in glass and glassy polymers, J. Forensic Sci. 61 (2016) 92–101, doi: http://dx.doi.org/10.1111/1556-4029.12968.
55. M.J. Bradley, R.L. Keagy, P.C. Lowe, M.P. Rickenbach, D.M. Wright, M.A. LeBeau, A validation study for duct tape end matches, J. Forensic Sci. 51 (2006) 504– 508, doi: http://dx.doi.org/10.1111/j.1556-4029.2006.00106.x.
56. F.A. Tulleners, J. Braun, The Statistical Evaluation of Torn and Cut Duct Tape Physical End Matching, Natl. Inst. Justice 2011, July. Rep. No. 235287. (n.d.).
57. K.R. McCabe, F.A. Tulleners, J.V. Braun, G. Currie, E.N. Gorecho, A quantitative analysis of torn and cut duct tape physical end matching, J. Forensic Sci. 58 (2013) S34–S42.
58. W. Ristenpart, F.A. Tulleners, A. Alfter, Quantitative algorithm for the digital comparison of torn duct tape, Final Rep. to Natl. Inst. Justice Grant 2013-R2- CX-K009, Univ. Calif. Davis, CA, 2017.

59. Y. Yekutieli, Y. Shor, S. Wiesner, T. Tsach, Physical matching verification, Final Rep. to United States Dep. Justice Grant 2005-IJ-R-051, Natl. Crim. Justice Ref. Serv., Rockville, MD, 2012.
60. Bradley MJ, Gauntt JM, Mehltretter AM, Lowe PC, Wright DM. A Validation Study for Vinyl Electrical Tape End Matches. *J For Sci.* 2011;56(3):606-611.
61. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – physical fits of torn and cut duct tapes. *Forensic Science International*. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567.
62. van Dijk CD, van Someren A, Visser R, Sjerps M. Evidential value of duct tape comparison using loopbreaking patterns. *Forensic Sci Int*. 2022 332:111178. doi: 10.1016/j.forsciint.2022.111178.
63. Spaulding J, Picconatto G. Characterization of fracture match associations with automated image processing. *Forensic Science International*. 2023. 342. 10.1016/j.forsciint.2022.111519
64. Prusinowski M, Brooks E, Neumann C, Trejos T. Forensic interlaboratory evaluations of a systematic method for examining, documenting, and interpreting duct tape physical fits. *Forensic Chemistry*. 2023. 34: 100487. https://doi.org/10.1016/j.forc.2023.100487
65. Komarinski P. Automated Fingerprint Identification Systems (AFIS). 1st Edition. Academic Press. 2004. ISBN: 9780124183513.
66. Hodgins T, Ho A, Sandercock M. Identification of Modern Automotive Paint Systems using Paint Data Query (PDQ): A Collaborative Study. *JASTEE*. 2015. 6(3):46-63.
67. Neumann C, Margot P. New perspectives in the use of ink evidence in forensic science Part II. Development and testing of mathematical algorithms for the automatic comparison of ink samples analyzed by HPTLC. *For Sci Int*. 2009 185. 38-50. 10.1016/j.forsciint.2008.12.008.
68. Tavadze P, Lang L. romerogroup/ForensicFit: First release of ForensicFit Package. 2022. https://doi.org/10.5281/zenodo.7435058

Chapter 2: Advancement of a systematic and quantitative method for comparison of duct tape edges

2.1. Overview

This study expands upon a previously developed method that quantifies the similarity of the compared tape edges. First, we increased the tape pair datasets, introduced a standardized template to systematically document the analyst's decisions during the comparison, created a reporting tool that automatically estimates the edge similarity score (ESS) and provides a 2D visualization of the aligned bins, and studied the effect of several separation methods and tape grades on the quality of a fit. Second, we applied a computational model to analyze and compare fractured tape edges and compared the results to the human examination of the same samples.

In the human-based approach, the analysts examined more than 3300 pairs of hand-torn or scissor-cut duct tapes from three different tape grades while blind from the ground truth to minimize bias. A three-step methodology for comparing samples was followed: 1) qualitative assessment of the overall edge alignment and description of edge pattern, 2) macroscopic evaluation of the edges' features, 3) bin-by-bin subunit assessment of tape edges and estimation of the edge similarity score. A report template designed to maintain records of the decision-making process was used. Eight comparison features were defined and documented in the second and third steps using auto-populated cell options. Misidentification rates were low, with no false positives reported. Coinciding with previous research, low scores (under 20%) provided the most support for a non-fit conclusion, while high scores (80% or higher) supported a fit decision. A statistical analysis of the separation method and tape quality revealed a potential interaction between these factors and that they significantly impact the edge scores for true fitting pairs but not the true non-fits' scores. The developed comparison and documentation criteria can assist practitioners with a more straightforward, consistent, and transparent interpretation and reporting approach.

For the computer-based approach, a subset previously examined by the analysts of nearly 2,000 tapes and 4,000 images was evaluated. The tape edge images were digitized, including the top and bottom surfaces. Then a convolutional neural network was trained on the database to perform comparisons on the same samples, predict fit or non-fit, and estimate the weight of evidence. The sample sets incorporated hand-torn and scissor-cut sets of all three tape qualities. The model predicted the medium-quality and high-quality scissor-cut sets more accurately than hand-torn ones, whereas low-quality tape performed better for hand-torn separations. The computational performance was consistent with the human-based examination, and the model could identify some of the same features that the analysts observed.

The following chapter is an adaptation of a previously published article ©2023: Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations –Part I. Physical fits of torn and cut duct tapes. *Forensic Science International, 343, 111567*. Published online January 2023. https://doi.org/10.1016/j.forsciint.2023.111567.

Also, the computational approach is the product of a collaboration between Dr. Romero and Dr. Trejos’ research groups, and part of this work was submitted for publication (Using convolutional neural networks to support examiners in duct tape physical-fit comparisons. Pedram Tavadze, Logan Lang, Meghan Prusinowski, Zachary Andrews, Tatiana Trejos, and Aldo H. Romero, 2023 in review)

2.2. Introduction

Commonly performed in forensic laboratories, physical fit examinations involve comparing items to determine if there are corresponding distinctive characteristics along the fractured edges [1]. A physical fit between two items demonstrates that they were once part of the same object, as it is expected that the fracture features are not replicable [2]. Despite the probative value that a physical fit may hold for a forensic case, a scientific demonstration of the reliability and validity of these examinations is still needed. In recent years, organizations such as the Organization of Scientific Area Committees (OSAC)-Trace Materials Subcommittee have highlighted research needs that include the development of objective methods for physical fit comparisons and estimating error rates [3]. The underlying assumptions of physical fit analysis have yet to be extensively evaluated, mainly regarding which features are relevant for comparison and how the sample condition influences these features. As there are yet to be universally adopted standard methods for physical fits, these examinations are subject to errors that may undermine the evidentiary value [4,5]. Moreover, since physical fit assessments depend entirely on human judgment, it is essential to consider factors that can minimize bias and lead to standardized decision-making criteria [2].

Consequently, the field can benefit from developing consensus methodologies and data-driven criteria for analyzing and reporting physical fits. Studies that evaluate the uncertainty and error rates for physical fit examinations are necessary to support the scientific basis of this type of analysis. Such studies must account for various factors affecting samples received in crime laboratories. In response to these needs, there has been an increase in studies focusing on physical fit examinations in the last decade, some of which have addressed error rates [1]. While many materials can be examined for physical fits, the present study focuses on one of the most analyzed materials in forensic trace evidence: duct tape.

Duct tapes are well suited for physical fit examinations. They possess several layers, including a backing, an adhesive, and a reinforcement layer made of cloth (known as scrim). The scrim layer contributes to the strength of duct tape; therefore, it is less likely to distort or stretch during the events of a crime than other kinds of tape. Different rolls of duct tape can often be distinguished based on physical characteristics alone, as there are various features available for comparison between items, such as color, thickness, width, construction, and number of scrim fibers [6]. However, within a single roll of tape, the features produced by the manufacturing or construction of the tape are shown to be very consistent [7].

Studies evaluating physical fits of duct tape have been performed with relatively large populations of samples, such as the sets evaluated by Bradley et al., which reported very few misclassifications of pairs as inconclusive, and no false positives or negatives [8]. Similarly, McCabe et al. reported low misidentification rates in a study of 1600 tape pairs, at or below 3% for both hand-torn and cut tape edges [9]. More recently, quantitative methods of comparing tape edges have been proposed. McCabe et al. describe a process that calculates a similarity metric by measuring the distance of the matching area along the comparison edge and dividing it by the total length of the comparison edge [9]. Another recent article recommends using loop-breaking patterns, which refers to how the loops created by weft-insert scrim fibers can fracture. The authors employ likelihood ratios and Bayesian Networks to interpret the evidence. This method demonstrated high accuracy, and most comparisons strongly supported the respective fits or non-fits [10].

The method utilized for the present study uses an edge similarity score (ESS) to measure the quality of a fit and assess the evidence's probative value. The ESS is calculated by defining the smallest comparison unit as the areas defined between consecutive warp scrim fibers. One advantage of this method is that the number and width of scrim areas are reproducible across a single roll, thus providing a methodical comparison approach. The ESS is calculated as the percentage of comparison areas aligning across the fracture out of the total number of comparison areas. Due to its simplicity and high level of performance, our method is much more suitable to be implemented in forensic laboratories than previously proposed quantitative methods. In a previous study, the ESS was used to evaluate 2280 duct tape pairs. This initial study demonstrates high accuracy and the calculation of score-based likelihood ratios (SLRs) [5]. A mock kidnapping case illustrates the use of ESS and SLRs in assessing the evidence [5].

The goal of this study is to further enhance the scientific foundations and practicality of the quantitative method for measuring physical fits by (a) defining standardized terminology and criteria for documenting the decision-making process and (b) statistically evaluating how different factors may influence the performance of the method and the resulting similarity metric.

The main factors previously considered in the literature for duct tape include the tape's separation method and quality grade. Results concerning the separation method have had varying interpretations, depending on the methodology employed. For instance, Bradley et al. and McCabe et al. reported higher levels of misidentifications for scissor-cut sets compared to hand-torn sets, while in Prusinowski et al., the hand-torn and scissor cut sets yielded good accuracy [5, 8, 9]. These differences are attributed to the level of detail utilized in the examination. In the Prusinowski et al. study, microscopic features of the tape were noted as a critical factor contributing to the method's performance, whereas the Bradley and McCabe studies only looked at more general macroscopic features that are less discriminating in straight cuts [5, 10].

Prusinowski et al. [5] and van Dijk et al. [10] demonstrate methods of assessing the probative value of a fit through different constructions of likelihood ratios. The technique using loop-breaking patterns calculates likelihood ratios using a Bayesian network. The proposed likelihood ratio system shows that loop-breaking patterns can provide strong evidence for physical fit comparison and even provide strong support even for partial edges and tapes with reversed tearing directions [10]. In the study by Prusinowski et al., the score-based likelihood ratios are calculated from ESS to assign the value of a given fit [5]. The observed distribution of the SLRs calculated on the test sample shows that ESS values higher than 80% generally indicate strong support for a fit. In comparison, ESS values lower than 20% indicate strong support for a conclusion of non-fit [5].

In the current study, we use logistic regression [11-13] to explore the relationship between the experimental factors (e.g., separation method, tape grade) and the edge similarity score for duct tapes. A logistic regression model is a generalized linear model in which the logit link function transforms the output of a linear model from $[-\infty, \infty]$ to [0,1]. The interval [0,1] can be seen as a probability of class membership in a two class-classifier. Logistic regression is appropriate for the ESS method since ESS are always in the interval [0,1], where an ESS close to 0 is an indicator of the "non-fit class" and an ESS close to 1 is an indicator of the "fit class." One of the limitations of logistic regression models is that if one of the independent variables, or a combination thereof, completely separates the two classes, interpreting the parameters is difficult [14].

As an expanded impact of this research, we introduce an innovative approach that utilizes a computational model to compare fractured edge samples. While the results of our human-based approach are encouraging, we anticipate that the development of objective computer-based algorithms can assist practitioners in their decision-making process and improve the reliability of physical fits in testimony. The use of computational models is common in forensic science disciplines. They are commonly used in fields that require pattern recognition, such as fingerprint and footwear evidence. In addition, other types of models are used to assist in the discrimination of sources and recognize trends in data, such as that used for large-scale drug trafficking tracking [15]. Despite this, the application of computational models to physical fit examinations has been limited. A few studies have demonstrated using computational models to compare duct tape samples. Ristenpart et al. and Spaulding et al. present models that perform edge detection on tape images to outline the contour of the tape, which is then compared to other samples [16, 17]. In these studies, the computational models demonstrate strong potential for use to analyze large populations of samples in an abbreviated time frame. Despite this, both studies display higher error rates than observed in the human examination, likely due to limitations resulting from using only the backing portion of the tape. Our work has shown that there is substantial information obtained from observation of the scrim/adhesive side of duct tape that complements the features from the backing layer. Therefore, developing a model that uses both sides of the tape edges is anticipated to improve accuracy.

The model applied to our tape data utilizes a convolutional neural network. In this approach, the model's architecture is a convolutional neural network that connects to a fully connected neural network [18, 19]. Convolutional neural networks (CNN) are commonly used when performing pattern recognition on images. A CNN contains a series of layers that are used to learn from the input data (such as an image). The layers of the neural network have filters applied to recognize specific features, and the output of each layer is used as the input for the next. In addition to convolution layers, additional layers assist in computational performance, such as activation and pooling layers [20]. Following the learning using convolutional layers, the output image is carried through a fully connected network known as an artificial neural network, or ANN. The ANN takes the input and processes it through a series of hidden layers, makes decisions from the previous layer and then assesses the weight of each layer enhance the output [20]. The purpose of the two connected networks in this approach is to allow the CNN to extract the features from the tape, reduce the dimensions of the input images, and then use the fully connected network to learn from the images and make decisions on new pairs to assign probabilities of membership (i.e., fit, inconclusive, or non-fit).

In this study, the performance of the computational model is compared to the human-based examination of the same samples to evaluate how they can be complementary and provide additional objective support for a physical fit examination.

2.3. Assessing significant factors that can influence physical fit examinations - Physical fits of torn and cut duct tapes

2.3.1. Methods and Materials

The current duct tape dataset consists of 2813 physical tape comparison pairs. The tape rolls are of different grades, low quality (LQ), medium quality (MQ), and high quality (HQ). The brand name and physical characteristics of the samples are listed in **Table 2-1.**

Table 2-1. Summary of the physical characteristics of the tapes utilized in the study.

Tape Label	Scrim Count (Weft/Warp)	Weave Pattern	Color (Backing)	Width	Texture (Backing)	Color (Adhesive)
Gorilla	16/58	Simple Weave (double alternate warp scrim)	Silver	4.8 cm	Dimpled	Grey
Duck Brand	8/20	Simple Weave	Silver	4.8 cm	Dimpled	Cream
Tool Bench	8/17	Simple Weave	Silver	4.7 cm	Smooth	White

Each tape pair of a given quality originated from the respective roll. The edges are either hand-torn (HT) or scissor-cut (SC). Some subsets of hand-torn pairs are stretched (S) to evaluate against the original, non-stretched items. (**Figure 2-1**). Because two analysts independently completed the medium-quality hand-torn (MQ-HT) to evaluate inter-analyst variability, the dataset consists of

3321 comparisons. However, some samples were reused to make the stretched sets or to utilize both edges of the samples for comparisons. The 3321 comparisons were made from a total of 1998 tape samples (900 low-quality tape samples, 200 medium-quality tape samples, and 898 high-quality tape samples). Other than MQ-HT, all sets were completed by one analyst, as previously, the method has demonstrated no substantial differences in scores between analysts, and intra-examiner evaluation of tapes from a given set has an elevated risk of bias [5].

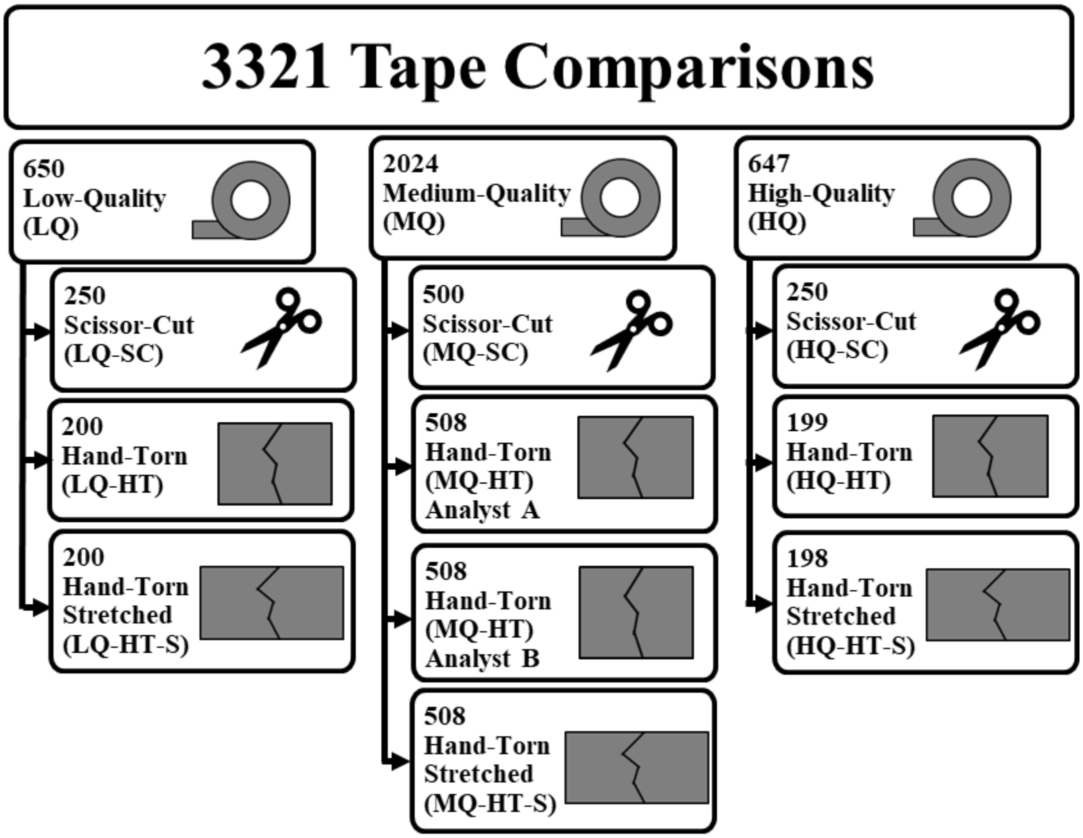


Figure 2-1. Breakdown of subsets for the tape dataset. The complete tape set consists of subsets of samples originating from each of three grades of tape, low quality (LQ), medium quality (MQ), or high quality (HQ). Edges are scissor-cut (SC), hand-torn (HT), or hand-torn with additional stretching (HT-S).

Although product labels do not explicitly describe the grade or quality of the tape, market cost and physical features are used to infer the grade range in this study. The primary distinguishing physical characteristics between the different grades include the number of scrim (reinforcement fibers) and the thickness of the backing and adhesive layers. The low-quality (LQ) tape has the thinnest adhesive thickness and the lowest scrim count per square inch (17 warp fibers, eight weft fibers). In contrast, the high-quality (HQ) tape has thicker backing and adhesive layers and a much higher ratio of warp/weft scrim fibers per square inch (58 warp, 16 weft). The medium-quality (MQ) tape has a scrim count of 20 warp to 8 weft fibers per square inch.

This study expands the previously developed method [5] by comparing samples into three steps based on feedback from practitioners in the field. **Figure 2-2** shows an examination scheme for physical fits, with the three steps incorporated into the typical workflow. In the first step, the analyst determines if the samples are suitable for a fit examination. If they are suitable, the analyst observes the overall edge morphology, including the edge pattern, denoting the edge as straight, angled, wavy, or puzzle-like (see **Figure 2-4** in the results section). The analyst compares the edge morphology of both tapes and performs an initial assessment of the level of fit based solely on edge morphology. The second step includes a macroscopic evaluation of the tape edges. This step involves splitting the tape edge into five macroscopic comparison areas and documenting the presence or absence of eight defined features (see **Figure 2-5** in the results section). The analyst also annotates if the presence or absence of the feature influences the decision of the macroscopic area as a fit or non-fit. For example, severed dimples on the backing may be present, but if only on one side of the fracture, it indicates a non-fit for that area. The third and final comparison step involves a bin-by-bin subunit comparison of the edges. The same predetermined features are again documented, but this time at the microscopic level in each comparison area (smaller bin area between two scrim fibers). The analyst then reports a score separately for each bin, assigning a value of 0, 0.5, or 1 for non-fit, inconclusive, or fit, respectively. Individual values for each comparison bin are then summed and expressed as a relative percentage of total comparison areas to estimate the similarity score (ESS).

A reporting template designed in Microsoft Excel automatically calculates the ESS and documents the decision-making process at each step, facilitating the note-taking process with built-in auto-populated cells. The bin score cells have a simple code, where if the analyst enters the number 0 (non-fit bin), the cell turns red, yellow for 0.5 (inconclusive bin), or green for 1 (fit bin). This color coding allows for rapid visualization of consecutive cells across the compared edges.

Several methods for evaluating the data are applied with the reported ESS values from each tape subset. The method’s performance is evaluated using false positive, false negative, specificity, and sensitivity rates for each subgroup. Boxplots are used to demonstrate the distribution of the ESS in ground truth datasets and provide an initial exploration of the potential relationships between the different experimental factors considered (e.g., separation method, tape quality, stretching). A logistic regression model is applied to the data to evaluate the statistical significance of the experimental factors. Finally, score-based likelihood ratios (SLR) are calculated to estimate the probative value of a pair of tapes.

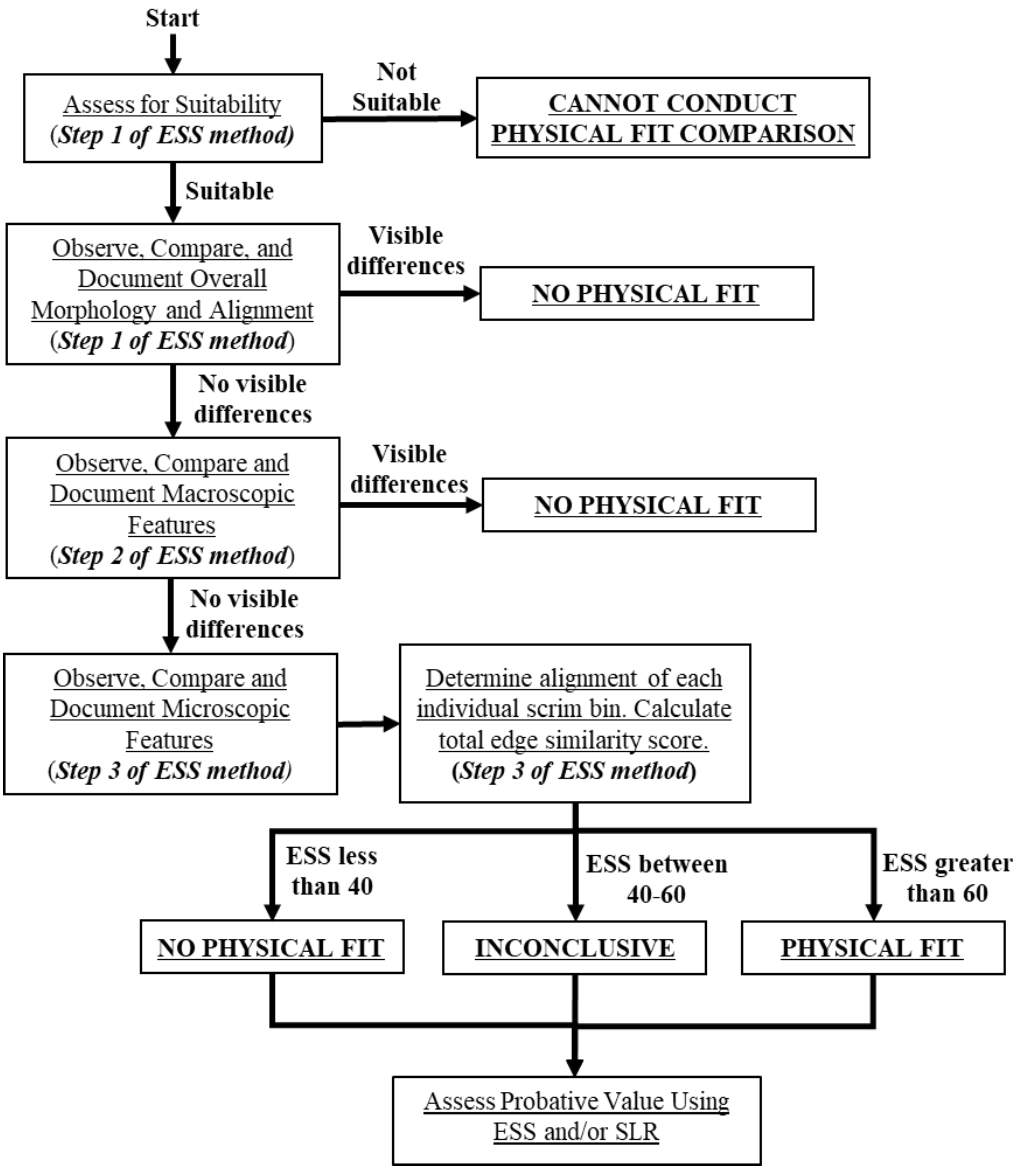


Figure 2-2. Examination scheme for physical fit comparisons. Other chemical examinations are necessary if the samples are not suitable for physical fit examination. If the sample edges demonstrate apparent differences in the comparison features at any stage, the outcome is "no physical fit." Beyond the microscopic comparison (Step 3 of ESS), the result is "no physical fit (non-fit)," "inconclusive," or "physical fit" with a description of its value. The quantitative ESS score and SLR can then be used to estimate the probative value of the outcome.

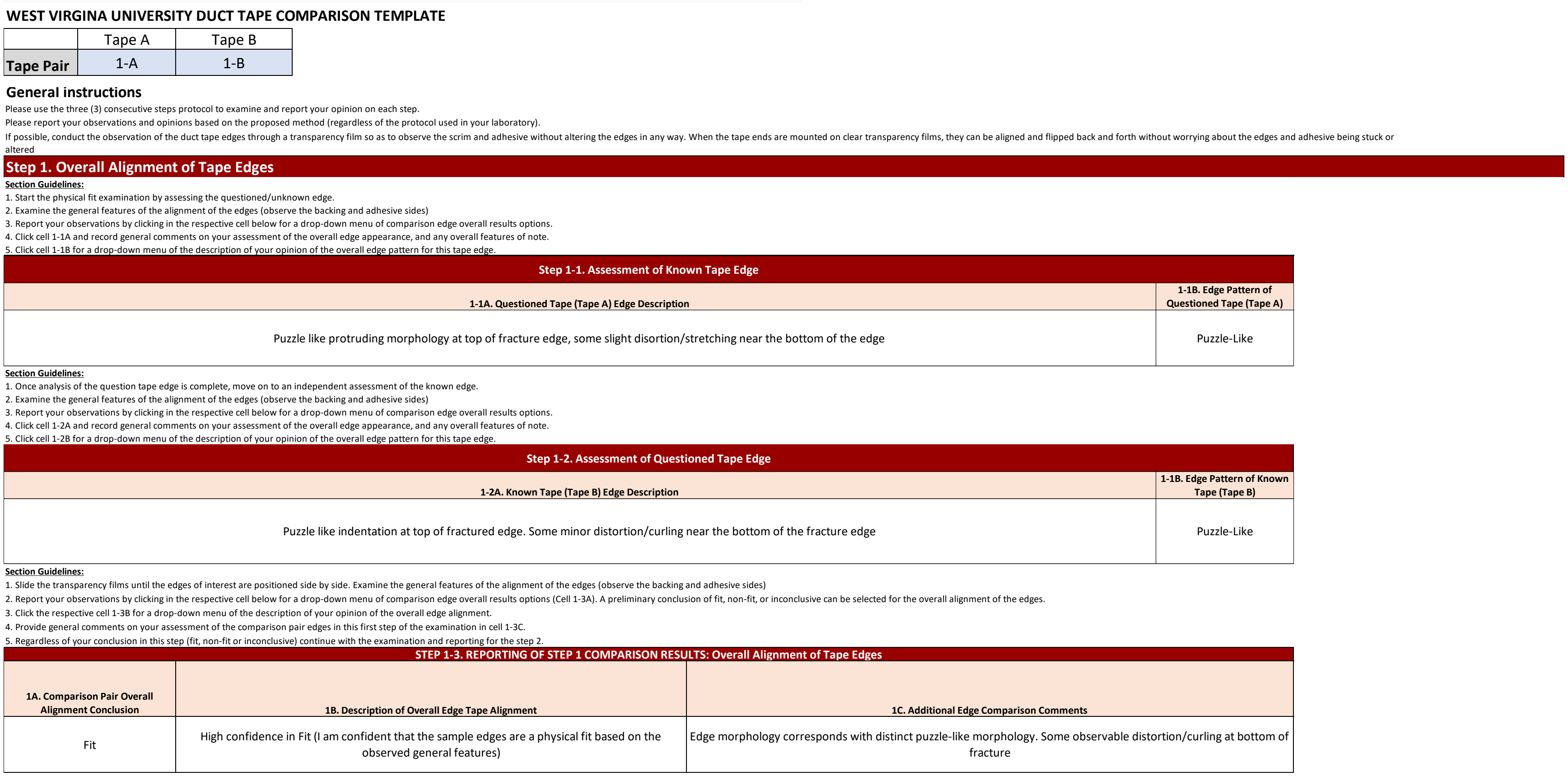
2.3.2. Results and Discussion

2.3.2.1. Development of methodical comparison and documentation criteria

This study introduces a critical improvement of the examination and documentation process of physical fits since our first publication [5], considering feedback received from practitioners who participated in an interlaboratory study. As described in the methodology section, one of the method’s novelties is creating a multi-step approach with predetermined terminology and documentation with auto-populated cells. The template is shown in **Figures 2-3, 2-4,** and **2-5**, where each step of the comparison is broken down in

In the first step, the analysts document overall observations regarding the tape morphology of each edge, along with general edge alignment characteristics. Tape edge’s standardized descriptions include one of four patterns: straight, angled, wavy, or puzzle-like [5]. Some tape features and separation methods tend to produce patterns that are more dominant than others. For example, the reduced reinforcement of the scrim in the low- and medium-quality tapes can cause irregularity and more puzzle-like edges when the tape is torn. **Figures 2-6**, **2-7,** and **2-8** display this phenomenon, particularly in the LQ-HT set. Interestingly, in **Figure 2-6**, none of the low-quality hand-torn tapes demonstrate straight or angled edges, while low-quality scissor-cut predominantly produce straight or angled edges, with only one instance of wavy or puzzle-like edges. Conversely, the high-quality hand-torn (HQ-HT) set results in very straight edges, likely due to the improved strength of the tape from the scrim and backing layers.

The few puzzle-like patterns in the HQ-HT are distinguishable from the LQ-HT set by exhibiting less prominent protrusions/cavities in the puzzle shapes. Scissor-cut edges consist of straight, angled, and wavy patterns, regardless of the tape grade. In very few cases, scissor-cut tapes produce puzzle-like patterns caused by a slight change of directionality on the cut, particularly with thicker adhesives (**Figure 2-7**).



WEST VIRGINIA UNIVERSITY DUCT TAPE COMPARISON TEMPLATE

	Tape A	Tape B
Tape Pair	1-A	1-B

General instructions

Please use the three (3) consecutive steps protocol to examine and report your opinion on each step.

Please report your observations and opinions based on the proposed method (regardless of the protocol used in your laboratory).

If possible, conduct the observation of the duct tape edges through a transparency film so as to observe the scrim and adhesive without altering the edges in any way. When the tape ends are mounted on clear transparency films, they can be aligned and flipped back and forth without worrying about the edges and adhesive being stuck or altered

Step 1. Overall Alignment of Tape Edges

Section Guidelines:

1. Start the physical fit examination by assessing the questioned/unknown edge.
2. Examine the general features of the alignment of the edges (observe the backing and adhesive sides)
3. Report your observations by clicking in the respective cell below for a drop-down menu of comparison edge overall results options.
4. Click cell 1-1A and record general comments on your assessment of the overall edge appearance, and any overall features of note.
5. Click cell 1-1B for a drop-down menu of the description of your opinion of the overall edge pattern for this tape edge.

Step 1-1. Assessment of Known Tape Edge	
1-1A. Questioned Tape (Tape A) Edge Description	1-1B. Edge Pattern of Questioned Tape (Tape A)
Puzzle like protruding morphology at top of fracture edge, some slight disortion/stretching near the bottom of the edge	Puzzle-Like

Section Guidelines:

1. Once analysis of the question tape edge is complete, move on to an independent assessment of the known edge.
2. Examine the general features of the alignment of the edges (observe the backing and adhesive sides)
3. Report your observations by clicking in the respective cell below for a drop-down menu of comparison edge overall results options.
4. Click cell 1-2A and record general comments on your assessment of the overall edge appearance, and any overall features of note.
5. Click cell 1-2B for a drop-down menu of the description of your opinion of the overall edge pattern for this tape edge.

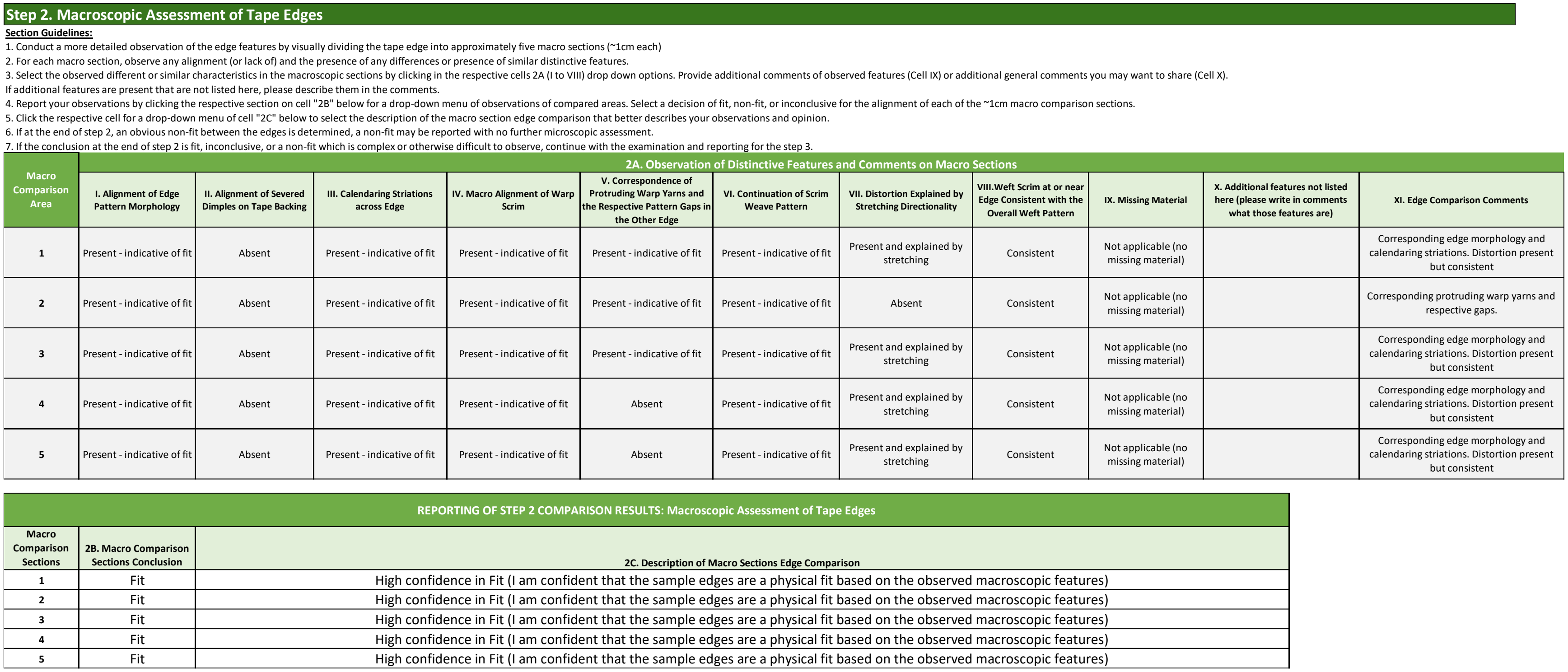
Step 1-2. Assessment of Questioned Tape Edge	
1-2A. Known Tape (Tape B) Edge Description	1-1B. Edge Pattern of Known Tape (Tape B)
Puzzle like indentation at top of fractured edge. Some minor distortion/curling near the bottom of the fracture edge	Puzzle-Like

Section Guidelines:

1. Slide the transparency films until the edges of interest are positioned side by side. Examine the general features of the alignment of the edges (observe the backing and adhesive sides)
2. Report your observations by clicking in the respective cell below for a drop-down menu of comparison edge overall results options (Cell 1-3A). A preliminary conclusion of fit, non-fit, or inconclusive can be selected for the overall alignment of the edges.
3. Click the respective cell 1-3B for a drop-down menu of the description of your opinion of the overall edge alignment.
4. Provide general comments on your assessment of the comparison pair edges in this first step of the examination in cell 1-3C.
5. Regardless of your conclusion in this step (fit, non-fit or inconclusive) continue with the examination and reporting for the step 2.

STEP 1-3. REPORTING OF STEP 1 COMPARISON RESULTS: Overall Alignment of Tape Edges		
1A. Comparison Pair Overall Alignment Conclusion	1B. Description of Overall Edge Tape Alignment	1C. Additional Edge Comparison Comments
Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed general features)	Edge morphology corresponds with distinct puzzle-like morphology. Some observable distortion/curling at bottom of fracture

Figure 2-3. Example of step 1 of the documentation template with filled out annotations. This step covers the overall assessment of each edge independently and then a side-by-side comparison, where the analyst documents whether they observe the pair as a fit or non-fit and the confidence level in that decision. The complete Excel file clean template can be accessed in **Appendix 2-A**.



Step 2. Macroscopic Assessment of Tape Edges

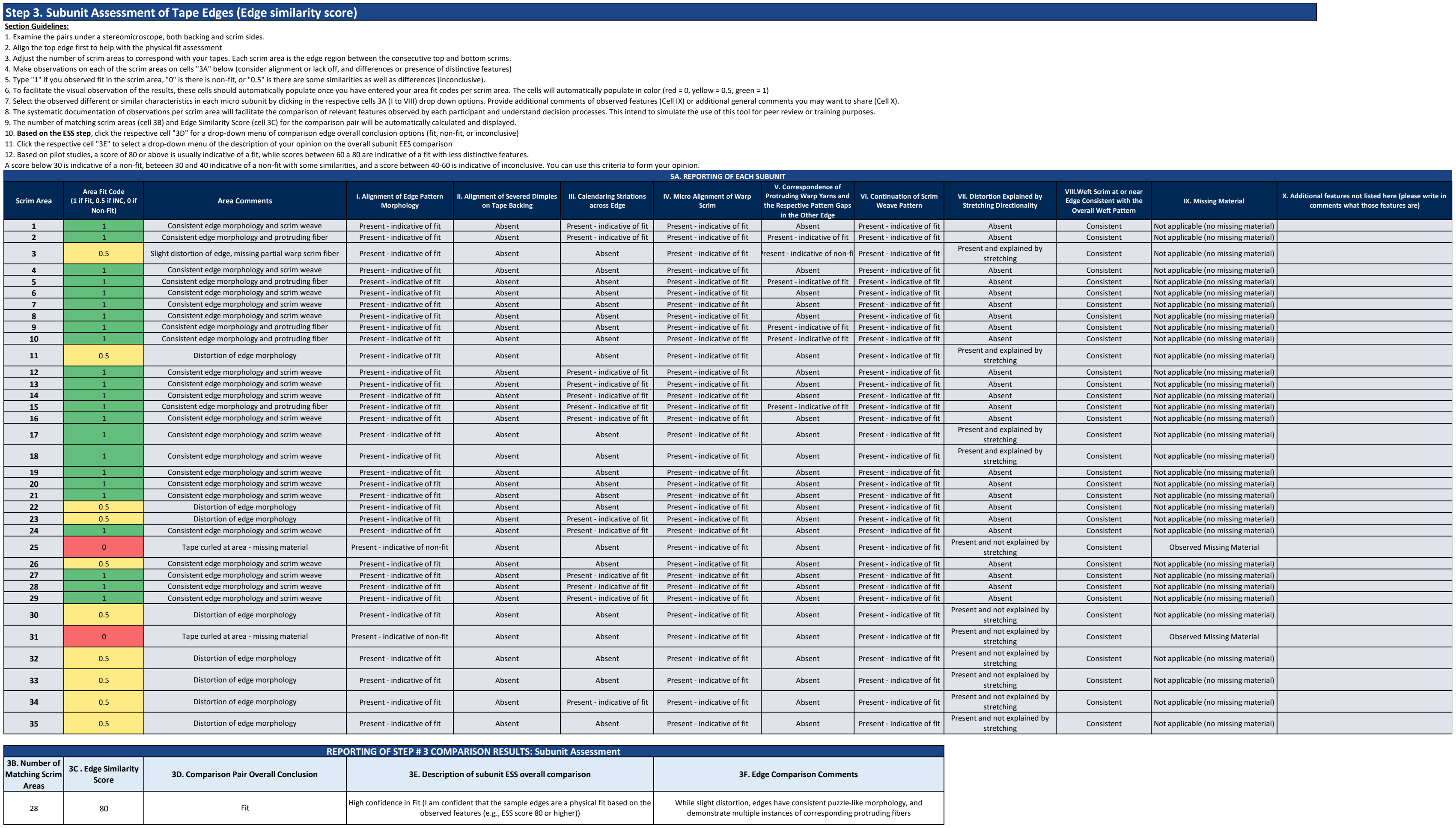
Section Guidelines:

1. Conduct a more detailed observation of the edge features by visually dividing the tape edge into approximately five macro sections (~1cm each)
2. For each macro section, observe any alignment (or lack of) and the presence of any differences or presence of similar distinctive features.
3. Select the observed different or similar characteristics in the macroscopic sections by clicking in the respective cells 2A (I to VIII) drop down options. Provide additional comments of observed features (Cell IX) or additional general comments you may want to share (Cell X).
If additional features are present that are not listed here, please describe them in the comments.
4. Report your observations by clicking the respective section on cell "2B" below for a drop-down menu of observations of compared areas. Select a decision of fit, non-fit, or inconclusive for the alignment of each of the ~1cm macro comparison sections.
5. Click the respective cell for a drop-down menu of cell "2C" below to select the description of the macro section edge comparison that better describes your observations and opinion.
6. If at the end of step 2, an obvious non-fit between the edges is determined, a non-fit may be reported with no further microscopic assessment.
7. If the conclusion at the end of step 2 is fit, inconclusive, or a non-fit which is complex or otherwise difficult to observe, continue with the examination and reporting for the step 3.

Macro Comparison Area	2A. Observation of Distinctive Features and Comments on Macro Sections										
	I. Alignment of Edge Pattern Morphology	II. Alignment of Severed Dimples on Tape Backing	III. Calendaring Striations across Edge	IV. Macro Alignment of Warp Scrim	V. Correspondence of Protruding Warp Yarns and the Respective Pattern Gaps in the Other Edge	VI. Continuation of Scrim Weave Pattern	VII. Distortion Explained by Stretching Directionality	VIII.Weft Scrim at or near Edge Consistent with the Overall Weft Pattern	IX. Missing Material	X. Additional features not listed here (please write in comments what those features are)	XI. Edge Comparison Comments
1	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)		Corresponding edge morphology and calendaring striations. Distortion present but consistent
2	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)		Corresponding protruding warp yarns and respective gaps.
3	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)		Corresponding edge morphology and calendaring striations. Distortion present but consistent
4	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)		Corresponding edge morphology and calendaring striations. Distortion present but consistent
5	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)		Corresponding edge morphology and calendaring striations. Distortion present but consistent

REPORTING OF STEP 2 COMPARISON RESULTS: Macroscopic Assessment of Tape Edges		
Macro Comparison Sections	2B. Macro Comparison Sections Conclusion	2C. Description of Macro Sections Edge Comparison
1	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features)
2	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features)
3	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features)
4	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features)
5	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features)

Figure 2-4. Example of filled out cells for step 2; this step covers the macroscopic assessment of the compared pairs of tapes. The edges are visually separated into five macroscopic sections, and nine major features are documented for their absence or presence. If present, the analyst document whether the feature indicates a fit or non-fit. In addition, each section is documented as a fit or non-fit and the confidence level in that decision. The complete Excel file clean template can be accessed in **Appendix 2-A**.



Step 3. Subunit Assessment of Tape Edges (Edge similarity score)

Section Guidelines:

1. Examine the pairs under a stereomicroscope, both backing and scrim sides.
2. Align the top edge first to help with the physical fit assessment
3. Adjust the number of scrim areas to correspond with your tapes. Each scrim area is the edge region between the consecutive top and bottom scrims.
4. Make observations on each of the scrim areas on cells "3A" below (consider alignment or lack off, and differences or presence of distinctive features)
5. Type "1" if you observed fit in the scrim area, "0" is there is non-fit, or "0.5" is there are some similarities as well as differences (inconclusive).
6. To facilitate the visual observation of the results, these cells should automatically populate once you have entered your area fit codes per scrim area. The cells will automatically populate in color (red = 0, yellow = 0.5, green = 1)
7. Select the observed different or similar characteristics in each micro subunit by clicking in the respective cells 3A (I to VIII) drop down options. Provide additional comments of observed features (Cell IX) or additional general comments you may want to share (Cell X).
8. The systematic documentation of observations per scrim area will facilitate the comparison of relevant features observed by each participant and understand decision processes. This intend to simulate the use of this tool for peer review or training purposes.
9. The number of matching scrim areas (cell 3B) and Edge Similarity Score (cell 3C) for the comparison pair will be automatically calculated and displayed.
10. **Based on the ESS step**, click the respective cell "3D" for a drop-down menu of comparison edge overall conclusion options (fit, non-fit, or inconclusive)
11. Click the respective cell "3E" to select a drop-down menu of the description of your opinion on the overall subunit EES comparison
12. Based on pilot studies, a score of 80 or above is usually indicative of a fit, while scores between 60 a 80 are indicative of a fit with less distinctive features.

A score below 30 is indicative of a non-fit, beteeen 30 and 40 indicative of a non-fit with some similarities, and a score between 40-60 is indicative of inconclusive. You can use this criteria to form your opinion.

			5A. REPORTING OF EACH SUBUNIT									
Scrim Area	**Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)**	**Area Comments**	**I. Alignment of Edge Pattern Morphology**	**II. Alignment of Severed Dimples on Tape Backing**	**III. Calendaring Striations across Edge**	**IV. Micro Alignment of Warp Scrim**	**V. Correspondence of Protruding Warp Yarns and the Respective Pattern Gaps in the Other Edge**	**VI. Continuation of Scrim Weave Pattern**	**VII. Distortion Explained by Stretching Directionality**	**VIII.Weft Scrim at or near Edge Consistent with the Overall Weft Pattern**	**IX. Missing Material**	**X. Additional features not listed here (please write in comments what those features are)**
1	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
2	1	Consistent edge morphology and protruding fiber	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
3	0.5	Slight distortion of edge, missing partial warp scrim fiber	Present - indicative of fit	Absent	Absent	Present - indicative of fit	resent - indicative of non-fi	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)	
4	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
5	1	Consistent edge morphology and protruding fiber	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
6	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
7	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
8	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
9	1	Consistent edge morphology and protruding fiber	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
10	1	Consistent edge morphology and protruding fiber	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
11	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)	
12	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
13	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
14	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
15	1	Consistent edge morphology and protruding fiber	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
16	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
17	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)	
18	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and explained by stretching	Consistent	Not applicable (no missing material)	
19	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
20	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
21	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
22	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
23	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
24	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
25	0	Tape curled at area - missing material	Present - indicative of non-fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Observed Missing Material	
26	0.5	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
27	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
28	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
29	1	Consistent edge morphology and scrim weave	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Absent	Consistent	Not applicable (no missing material)	
30	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Not applicable (no missing material)	
31	0	Tape curled at area - missing material	Present - indicative of non-fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Observed Missing Material	
32	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Not applicable (no missing material)	
33	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Not applicable (no missing material)	
34	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Present - indicative of fit	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Not applicable (no missing material)	
35	0.5	Distortion of edge morphology	Present - indicative of fit	Absent	Absent	Present - indicative of fit	Absent	Present - indicative of fit	Present and not explained by stretching	Consistent	Not applicable (no missing material)	

REPORTING OF STEP # 3 COMPARISON RESULTS: Subunit Assessment				
3B. Number of Matching Scrim Areas	**3C . Edge Similarity Score**	**3D. Comparison Pair Overall Conclusion**	**3E. Description of subunit ESS overall comparison**	**3F. Edge Comparison Comments**
28	80	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed features (e.g., ESS score 80 or higher))	While slight distortion, edges have consistent puzzle-like morphology, and demonstrate multiple instances of corresponding protruding fibers

Figure 2-5. Step 3 of the documentation template, example of a filled-out form resulting in an ESS of 80. This step covers the microscopic assessment of the compared pairs of tapes. The edges are visually separated into bins based on the number of areas between the scrim fibers. In each bin, the same nine major features from Step 2 are documented. In addition, each section is documented as a fit or non-fit and the confidence level in that decision. The analyst reports each bin as fit, non-fit, or inconclusive by coding it as 1, 0, or 0.5, which is then automatically colored and calculated as the ESS. Then, the overall conclusion is reported. The complete Excel file clean template can be accessed in **Appendix 2-A**.

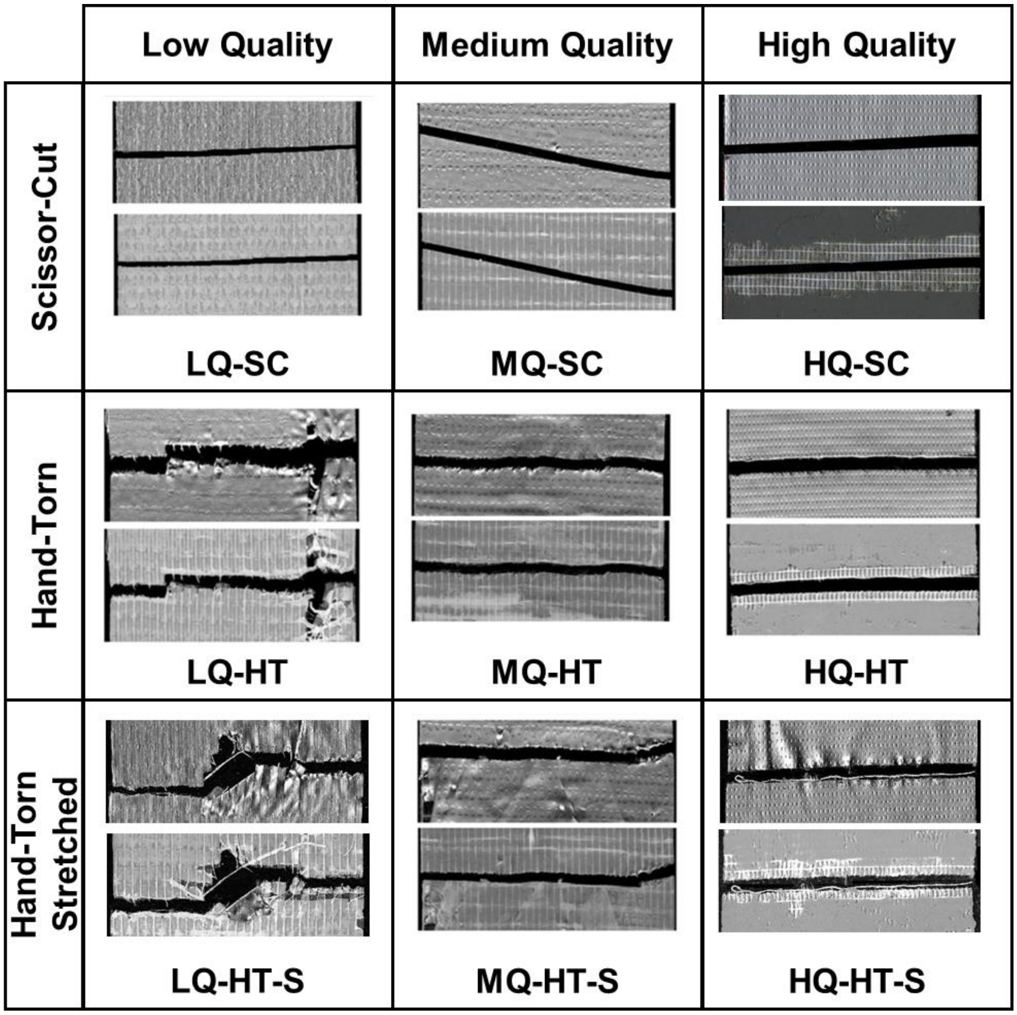


Figure 2-6. Examples of true fit pairs from the tape sets. The images show the distorted morphology observed in the MQ-HT-S, LQ-HT, and LQ-HT-S edges. Despite also being hand-torn, the edges observed in the HQ-HT set are very straight and less distinctive, even when stretched. Meanwhile, the scissor-cut sets produce much cleaner edges regardless of tape quality.

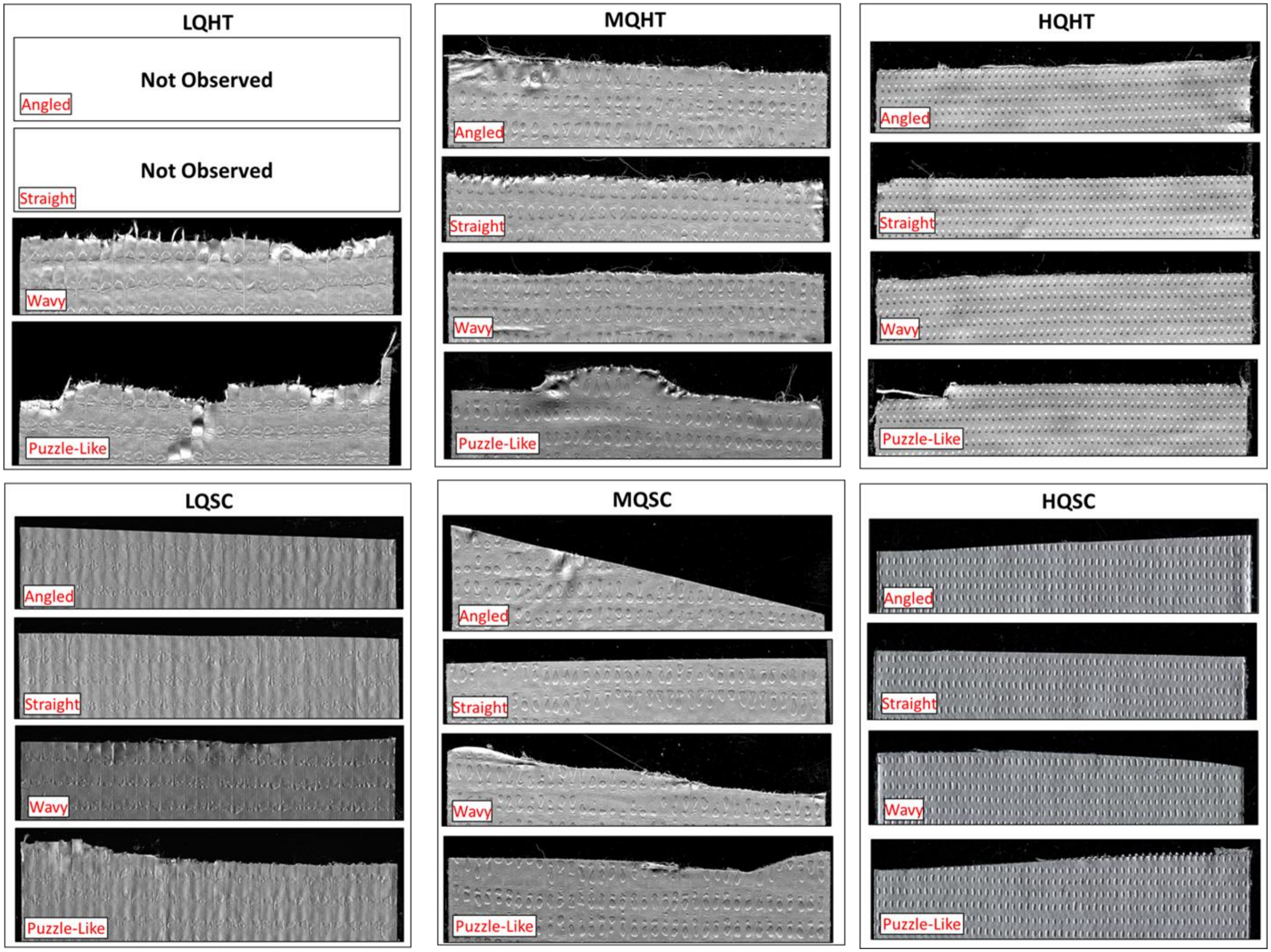


Figure 2-7. Examples of edge morphology for each sample set. Straight and angled edges were not observed in the LQ-HT set. Stretched sets shared the same edges as the non-stretched edges, so no additional examples are demonstrated here.

The medium-quality hand-torn (MQ-HT) tape has a wider variety of all four edge separation patterns, although wavy patterns are the most predominant type. The additional stretching of medium and high-quality hand-torn sets further reduces the relative occurrence of angled and straight edges (**Figure 2-8)**.

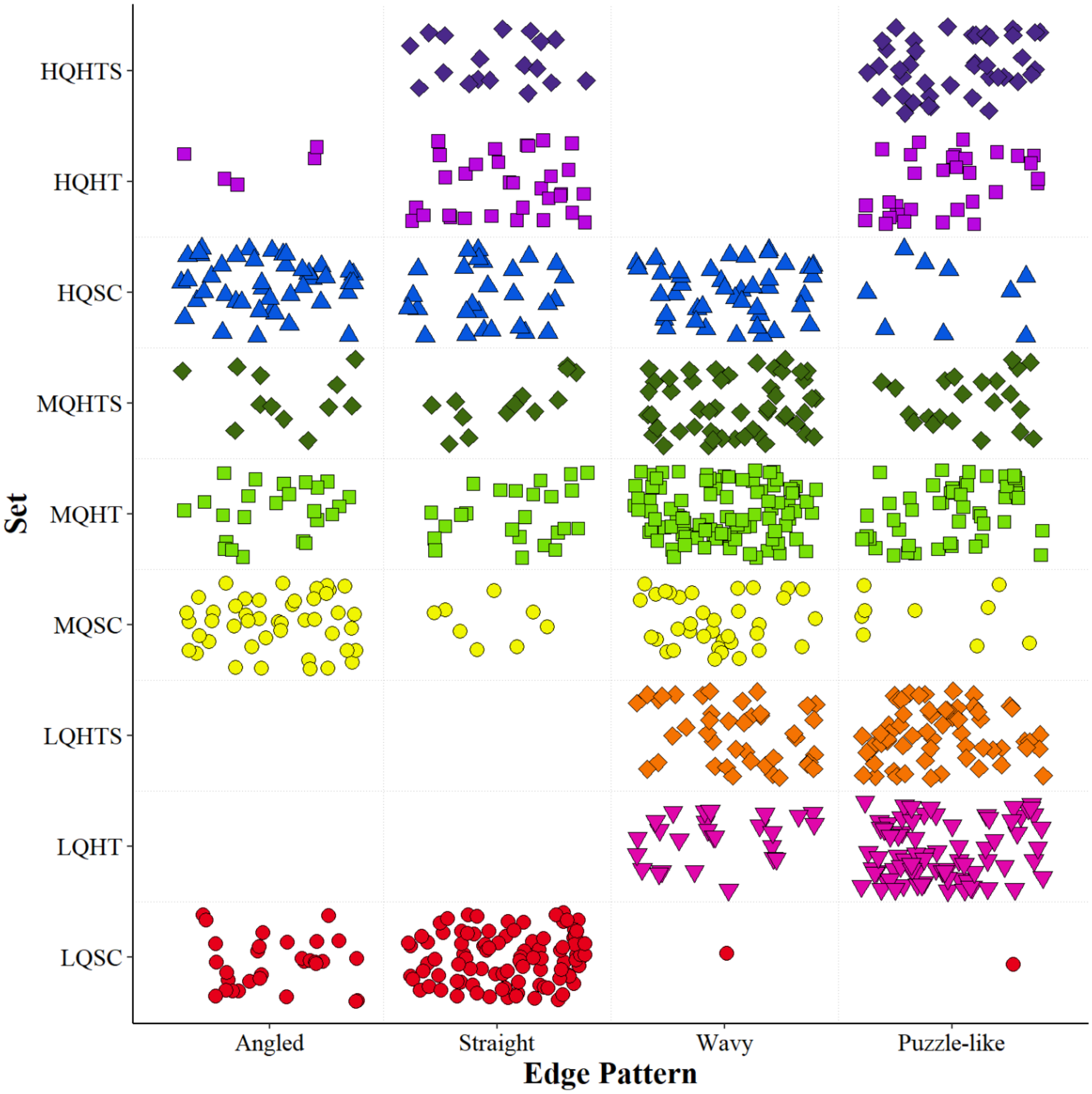
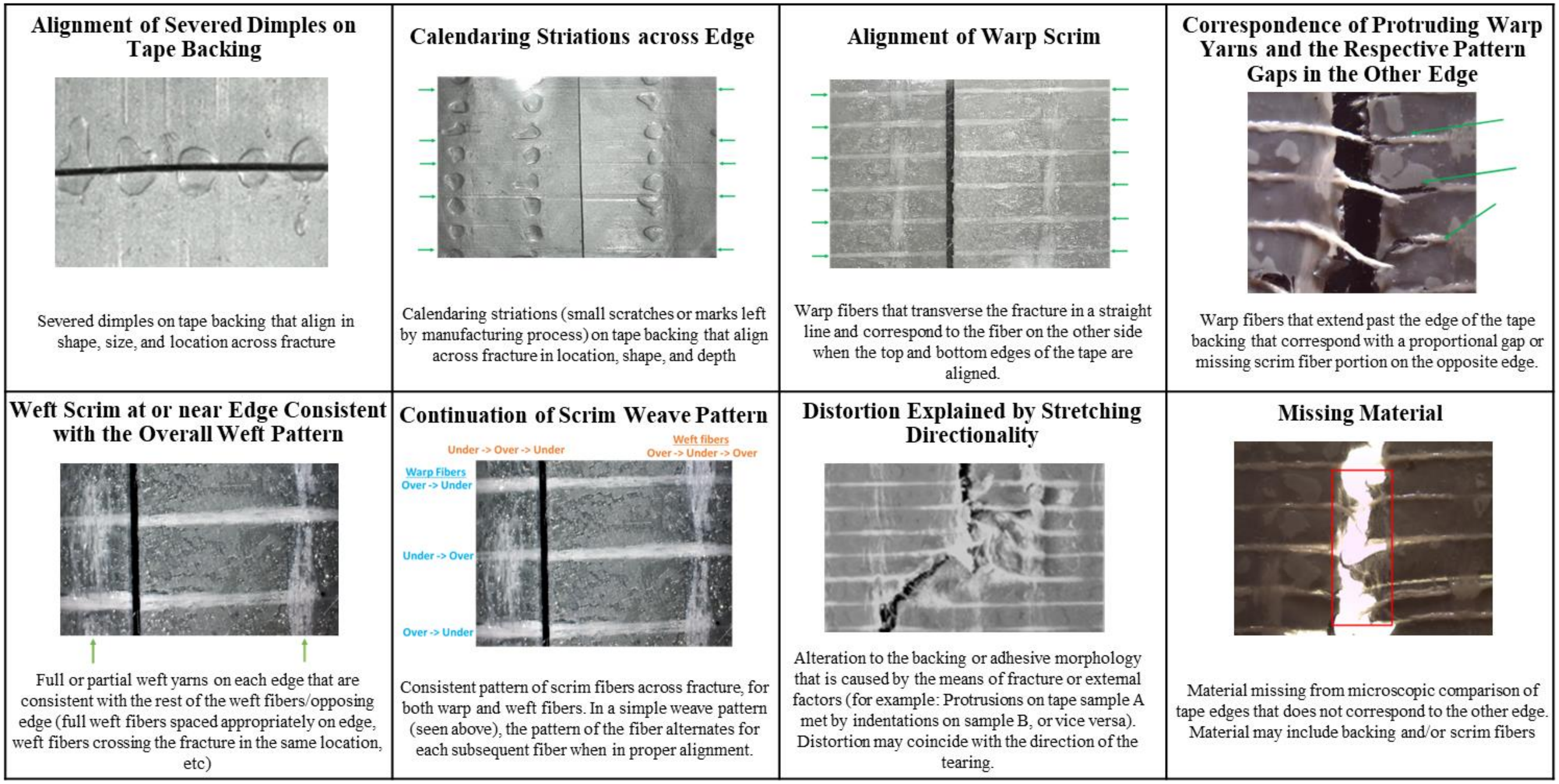


Figure 2-8. Edge pattern occurrence trends for true-fit pairs for all compared sets. Overall, puzzle-like edges are more common in hand-torn sets, while straight or angled edges are more commonly observed in scissor-cut sets.

This first step allows the examiner to document the general observations of the questioned and the known item edges separately and use this overall assessment to determine the suitability of the specimens for fit examinations. For instance, at this stage, extreme distortion or damaged edges can lead to an item not being apt for further comparisons. On the other hand, if exclusionary differences are found at this stage, they are reported, and no further physical fit examination is needed (e.g., questioned or known items with different color-shades or texture in the backing or adhesive, different scrim count). If no exclusionary differences are found in this screening stage, then the analysis continues to the second step.

In the second step, more detailed examination and documentation are conducted to identify major features in relatively large regions of the tape. One of the critical goals of this study is to provide standardized criteria for parts and terminology, which examiners can use to describe their observations consistently. Most standard operating procedures currently used by forensic laboratories for physical fit examinations are general and lack examples of material-specific features that can be observed. While some of these features may be discussed in examiner training programs, there is no consensus of defined characteristics and descriptors and therefore the consistency and trends of these features amongst different materials and examiners still need to be discovered. To reduce this gap and make documentation consistent, we recommend observing and documenting eight major features that we have found to hold the most weight in the decision process of a fit or non-fit. Examples of these features are shown in **Figure 2-9**. Some of these features, such as severed dimpling and calendaring striations, are observable on the backing side of the tape, while many are observed while looking at the adhesive side of the tape. Distortion and missing material may be viewed on either side of the tape.

These features can be observed on many separated tapes, regardless of grade quality. Some duct tapes may not have dimpling on the backing, but all duct tapes have scrim fibers that can be observed. To make the scrim fibers viewable without needing to remove the adhesive, tape pieces can be placed on transparency film. However, there may be cases where the adhesive needs to be removed to observe the scrim fibers. In those cases, such as in the high-quality roll from this study, a small strip of adhesive can be removed from the comparison edges using hexane. When such preparation is required, it is critical that examiners first observe the edges and document any shift or loss of fibers resulting from the removed adhesive, as shifted fibers may influence the documentation of features or the resulting classification of fit or non-fit.



Alignment of Severed Dimples on Tape Backing

Severed dimples on tape backing that align in shape, size, and location across fracture

Calendaring Striations across Edge

Calendaring striations (small scratches or marks left by manufacturing process) on tape backing that align across fracture in location, shape, and depth

Alignment of Warp Scrim

Warp fibers that transverse the fracture in a straight line and correspond to the fiber on the other side when the top and bottom edges of the tape are aligned.

Correspondence of Protruding Warp Yarns and the Respective Pattern Gaps in the Other Edge

Warp fibers that extend past the edge of the tape backing that correspond with a proportional gap or missing scrim fiber portion on the opposite edge.

Weft Scrim at or near Edge Consistent with the Overall Weft Pattern

Full or partial weft yarns on each edge that are consistent with the rest of the weft fibers/opposing edge (full weft fibers spaced appropriately on edge, weft fibers crossing the fracture in the same location, etc)

Continuation of Scrim Weave Pattern

Consistent pattern of scrim fibers across fracture, for both warp and weft fibers. In a simple weave pattern (seen above), the pattern of the fiber alternates for each subsequent fiber when in proper alignment.

Distortion Explained by Stretching Directionality

Alteration to the backing or adhesive morphology that is caused by the means of fracture or external factors (for example: Protrusions on tape sample A met by indentations on sample B, or vice versa). Distortion may coincide with the direction of the tearing.

Missing Material

Material missing from microscopic comparison of tape edges that does not correspond to the other edge. Material may include backing and/or scrim fibers

Figure 2-9. Examples of the eight descriptive features documented for duct tape edge comparisons. These features are some of the most observed on duct tape regardless of grade. Frequently, they help establish and document standardized criteria during sample comparisons.

Finally, the last step consists of a detailed microscopic bin-by-bin comparison, where documentation of the presence, absence, and effect of the eight defined features assists with the transparency and reproducibility of the decision-making process. Since examiners must make a data-driven assessment of each bin independently and follow specified and demonstrable criteria, we argue that this bin-by-bin methodology increases objectivity. It is in this step that ESS is calculated. This quantitative metric provides a means to evaluate and convey the quality of a fit transparently. The ESS is beneficial when conclusions are based on complex comparisons involving fewer distinctive features, distortion, or stretching. In these cases, ESS provides a transparent and explainable criterion for examiners to use when performing physical fit comparisons. Moreover, detailed documentation and a quantitative criterion can assist in comparing results from different examiners; doing so can lead to more robust training and improve the technical review and verification processes.

The use of quantitative metrics provides new alternatives to assess the quality of fit, define decision criteria, assess error rates, and understand at a deeper level how comparisons are affected by common factors observed in authentic specimens (e.g., separation method, tape grade, stretching, and distortion, to mention some).

From the research perspective, documenting all three steps is vital to generate data that can build up the current body of knowledge in this field. However, when considering the practical application and adoption of this proposed method, a detailed bin-by-bin ESS may not be necessary when non-fits are well established in the first two steps. In these cases, tapes determined to be non-fits can be further examined by physical and chemical protocols that are standardized in the laboratories to examine and compare class characteristics. Conversely, when a potential fit is found after step 2 (or when the non-fit is more complex), we strongly recommend completing step 3 due to all benefits discussed above (**Figure 2-2**).

2.3.2.2. Exploratory analysis of method performance and distributions of edge similarity scores on true-fit and true non-fit populations

Overall, the method demonstrates good performance, with accuracies for all sets at approximately 80% or higher and no false positives. While error rates are critical to evaluating the reliability of physical fit comparisons, false positives have a severe consequence for this type of examination due to the implications of the perceived probative value of a physical fit. A false positive has the potential to contribute to a miscarriage of justice, as a fit between edges is considered the strongest degree of association for trace evidence materials. As seen in **Table 2-2**, the tape subsets in this study did not result in any false positive results. This provides scientific support to the belief that the items that have been broken, torn, or separated exhibit physical features that realign in a manner that is not expected to be replicated by chance. Likewise, false negative results in forensic fit examinations should be ideally as low as possible. Tapes more prone to distortion or possessing fewer distinctive features upon separation tend to produce more false negative conclusions. For instance, high-quality hand-torn sets (HQ-HT and HQ-HT-S) exhibit relatively straight edges and limited distinctive features. For the analysts in this study, the limited features contribute to substantially higher false negative rates and more reported inconclusives than other sets. In addition, a more substantial force is required to lift the HQ tapes from a surface due to stronger adhesion and cohesion of the adhesives. Thus, the high-quality sets tend to produce additional edge distortions upon stretching than the lower-quality samples. The stretching in HQ torn tapes leads to increased false negative results and reduces the overall accuracy, as seen in **Table 2-2** (e.g., accuracy from 84.9 to 79.8 %). The low and medium-quality tapes show accuracy better than 98% (ranging from 98 to 100%, depending on the subset).

This study expands on the previous work [5] by completing two additional scissor-cut sets, one for low-quality tape, one for high-quality tape, and two other stretched sets from low and high-quality tapes (**Table 2-2**). The scissor-cut sets in this study did not contribute to higher misidentification rates. The accuracy of all three scissor-cut sets is greater than 98%, with only two false negative pairs in the LQ-SC set and one false negative in the MQ-HT set. Despite the cleaner edges observed with cut edges, the alignment of severed dimples and consistency in scrim weave (as seen in **Figure 2-9**) prove valuable for identifying true fit pairs.

The stretching aims to simulate complex sample manipulations observed in casework samples. No substantial changes in performance are observed for the low-quality hand-torn stretched set. **Table 2-2** shows that only one true fit pair from the LQ-HT-S set is reported as inconclusive, with no other misidentifications. Conversely, as discussed before, the HQ-HT-S set demonstrates a higher false negative rate than the original HQ-HT set. This is likely due to additional distortion introduced during the stretching.

Table 2-2. Summary of the method's performance rates for the duct tapes. For the low quality (LQ), medium quality (MQ), and high quality (HQ), the subsets are labeled as scissor cut (SC), hand-torn (HT), and hand-torn stretched (HT-S). Two analysts independently evaluated the MQ-HT set [7]. There are no false positives reported for any set. Inconclusive results are not included in the false positive and negative rates but are incorporated in the overall accuracy estimation.

Performance rate (%)	LQ-SC (n=250 pairs)	LQ-HT (n=200 pairs)	LQ-HT-S (n=200 pairs)	MQ-SC (n=500 pairs)	MQ-HT (Analyst A) (n=508 pairs)	MQ-HT (Analyst B) (n=508 pairs)	MQ-HT-S (n=508 pairs)	HQ-SC (n=250 pairs)	HQ-HT (n=199 pairs)	HQ-HT-S (n=198 pairs)
False positive rate (FP)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
False-negative rate (FN)	1.5	0.0	0.0	1.0	1.0	2.0	1.0	0.0	21.4	31.6
True negative rate (Specificity)	97.5	99.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
True positive rate (Sensitivity)	98.5	100.0	99.0	99.0	98.0	98.0	99.0	100.0	69.4	57.2
Accuracy	98.0	99.5	99.5	99.8	99.6	99.6	99.8	100.0	84.9	79.8

To explore the distribution of the scores in true fits and true non-fit datasets, boxplots are shown in **Figure 2-10** to document the reported ESS for each set of tapes based on the ground truth (fit or non-fit). For this series of boxplots, the one for MQ-HT combines the results from the two examiners who independently completed the set. For most sets, there is an observable separation between the ESS obtained for true-fitting and non-fitting pairs. The score distributions for all the sets for true non-fit pairs are generally consistent at 30 or below, and the majority fall below 10. Conversely, the score distributions for the true fit pairs are different for some of the sets but are predominantly higher than 80 (**Figure 2-10**). Some trends are observable in the boxplots, such as broader variability of scores and the shift in the distribution of true fitting pairs in the HQ-HT and LQ-HT to slightly lower values, resulting from the edge morphology and predominant features observed on those sets. The distribution of scores for true fits in the LQ-HT-S is also consistent with the LQ-HT set. However, these characteristics are primarily observed in the hand-torn sets. The scissor-cut sets for both low- and high-quality tape are more like the medium-quality scissor-cut set, with most fitting pairs having ESS of 90 or higher.

The distribution of scores for the true fitting pairs in HQ-HT and HQ-HT-S is wider than in any other set. The high-quality tape is prone to contribute fewer features for comparison due to the relatively straight edges and potential distortions caused by removing the thick adhesive. The additional stretching on the samples compounds the issue with the tape, indicating that the high-quality tape used in this study is not necessarily the most suitable for physical fit comparisons. While true non-fits are relatively straightforward to rule out (seen in the 100% true negative rate for both HQ-HT sets), the reduced specificity and broader distribution of scores for true fit pairs for the sets indicate that for tapes that do not contribute many comparison features, there is generally reduced quality of fits between samples. Therefore, the additional chemical analysis may be warranted even if a fit is observed.

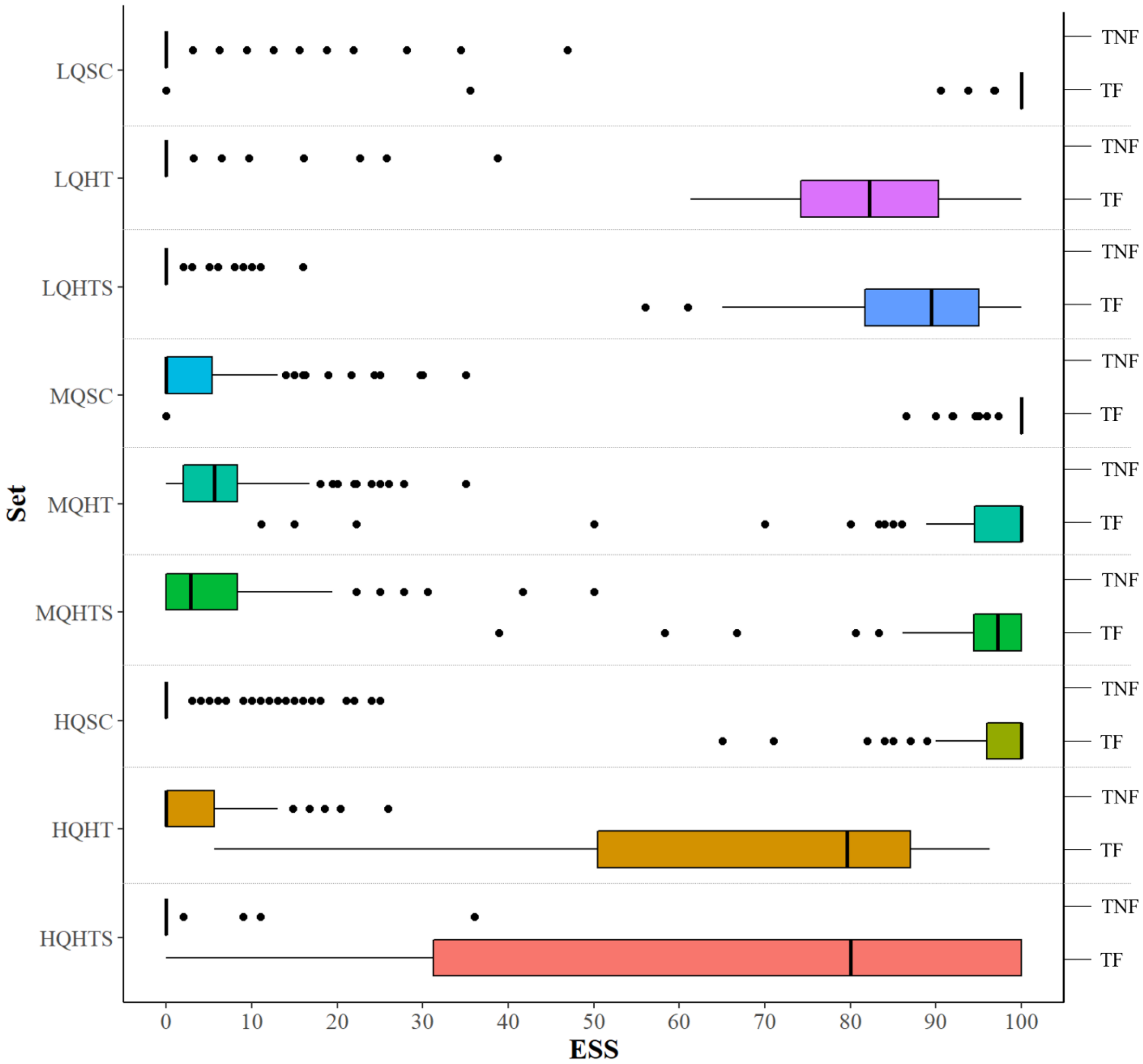


Figure 2-10. Boxplots showing the ESS distribution of each tape set, with true non-fits (TNF) displayed on the top boxplot of each set and true fits (TF) shown on the bottom boxplot of each set. Generally, there is a separation between the ESS distributions of the true fit and true non-fitting pairs, with higher scores for fits and lower scores for non-fits.

2.3.2.3. Statistical assessment of ESS values

While exploratory analysis denotes small shifts in score distributions based on the grade of tape, separation method, or stretching, it fails to determine whether these shifts are statistically significant and reveal any underlying interactions between factors. Therefore, we determine the extent of the effect of these factors on the ESS using a generalized linear mixed effect model, more specifically, a logistic regression model. Note that our model has no predictive abilities in specific cases since it only relies on general observations (such as tape quality and method of separation) made on the tape comparisons. We want to emphasize that our model is only meant to interpret the effect of the different factors by interpreting the regression coefficients.

The coefficients of our model are estimated by considering the log-odds transform of the ESS for each comparison as the dependent variable and one-hot encodings of the different levels of the different factors of interest (ground truth, quality, separation method, edge pattern) as fixed effects independent variables. Our model includes random effects to account for the replicated determination of ESS on the same samples. Initially, a random effect was added to account for the lack of independence between the ESS. Indeed, all the ESS in this study come from cross-comparisons between tape samples; thus, two ESS involving the same tape sample might be correlated. A study of this random effect showed that it was mostly negligible and has been removed in the final version of the model.

The model has been fitted using a Hamiltonian Monte-Carlo algorithm implemented in Stan 2022 [21]. Prior distributions for the coefficient values are centered on 0. We use variable selection to reduce the size of the model to select factors and interactions that are statistically significant. Variable selection relied on the widely applicable information criterion (WAIC), the standard deviation of the WAIC, and the number of parameters generated by the model with different combinations of variables and interactions. The WAIC is used to evaluate model quality as a measure of the amount of loss of information by a model as it is fit to the data set and can be used to account for both potential overfitting and underfitting of the model to the data.

The resulting model incorporates the interaction between the experimental factors of grade and the separation method, which is supported by the observations of the score distributions seen in the boxplots (**Equation 1**). In addition, there is an interaction between the ground truth and all other variables, as the ground truth of the pairs influences the ESS regardless of the other variables. Stretching is incorporated as a factor, but we found no evidence of interaction between stretching and grade or separation method.

$$\text{log odds (ESS)} \sim \text{Ground Truth} \times \ (\text{Grade} \ \times \ \text{Separation Method} + \text{Stretching} + \text{Edge Pattern}) \ (\textbf{Equation 1})$$

The posterior distributions for the value of the grade and separation method coefficients are shown in **Figure 2-11.** These distributions show the magnitude and variance of the coefficient values for each experimental factor for true-fit tape and true non-fit pairs. The distributions in **Figure 2-11** are marginalized over the remaining factors (e.g., **Figure 2-11A** shows the distributions of the coefficients for each grade of tape as a combination of the values resulting from the separation methods, stretching, and edge pattern, while **Figure 2-11B** shows the distributions of the coefficients for each separation method as a combination of the values resulting from the grade of tape, stretching and edge pattern).

Figure 2-11 shows that the value of the coefficients for the different grades of tapes and the two separation methods is significantly different from 0 for both true fits and true non-fits. Furthermore, apart from coefficients for Low and High grades of tape in non-fit pairs, all coefficient distributions are quite different. This indicates clearly that both tape quality and method of separation are factors that affect the resulting ESS.

Regarding individual levels of these factors, medium-quality tapes tend to result in the highest and lowest ESS (depending on whether the considered pairs are true (non-)fit). The more extreme ESS indicates that this type of tape contains more discriminative features. The medium-quality tape produces less distortion when torn than the other qualities of tape, and the scores are not influenced substantially by stretching (**Figure 2-8)**. Along with more occurrences of distinct-puzzle-like edges, the medium-quality tape retains many features for comparison that are not obscured by distortion or adhesive. While it could be expected that more distinctive puzzle-like edges only correspond with their true fit edge, these patterns commonly produce a greater amount of distortion in low-quality tapes, meaning that some uncertainty is caused by the degree of distortion observed.

Similarly, cut tapes tend to have more extreme ESS than pieces torn by hand, which indicates that cutting tapes does not prevent association between edges. Scissor-cut edges produce more angled edges, and if an angled edge has a sharper degree, the more likely it is to only correspond with its true edge. Certain features are observed primarily in the scissor-cut edges. Observance of severed dimples and consistent weft fibers that cross over the fracture edge is highly distinctive and strongly supports a fit conclusion. Conversely, the tapes with torn edges do not usually demonstrate these features and are more prone to stretching during separation, making it harder to associate the edges with high scores.

Counterfactual plots explore the effect of each experimental factor on the log-odds of the ESS (and, therefore, on the similarity scores). Counterfactual plots show the distributions of the expected values of the dependent variable under all combinations of levels of the different factors of a model, accounting for the uncertainty in the values of the model’s parameters. **Figure 2-12A** shows the counterfactual plot for the grade of tape. The left side of the counterfactual plot shows the distributions of the expected values of the log-odds ESS resulting from the model. In contrast, the right side of the plot shows kernel density estimated distributions of the empirical ESS data from the analyzed tape pairs.

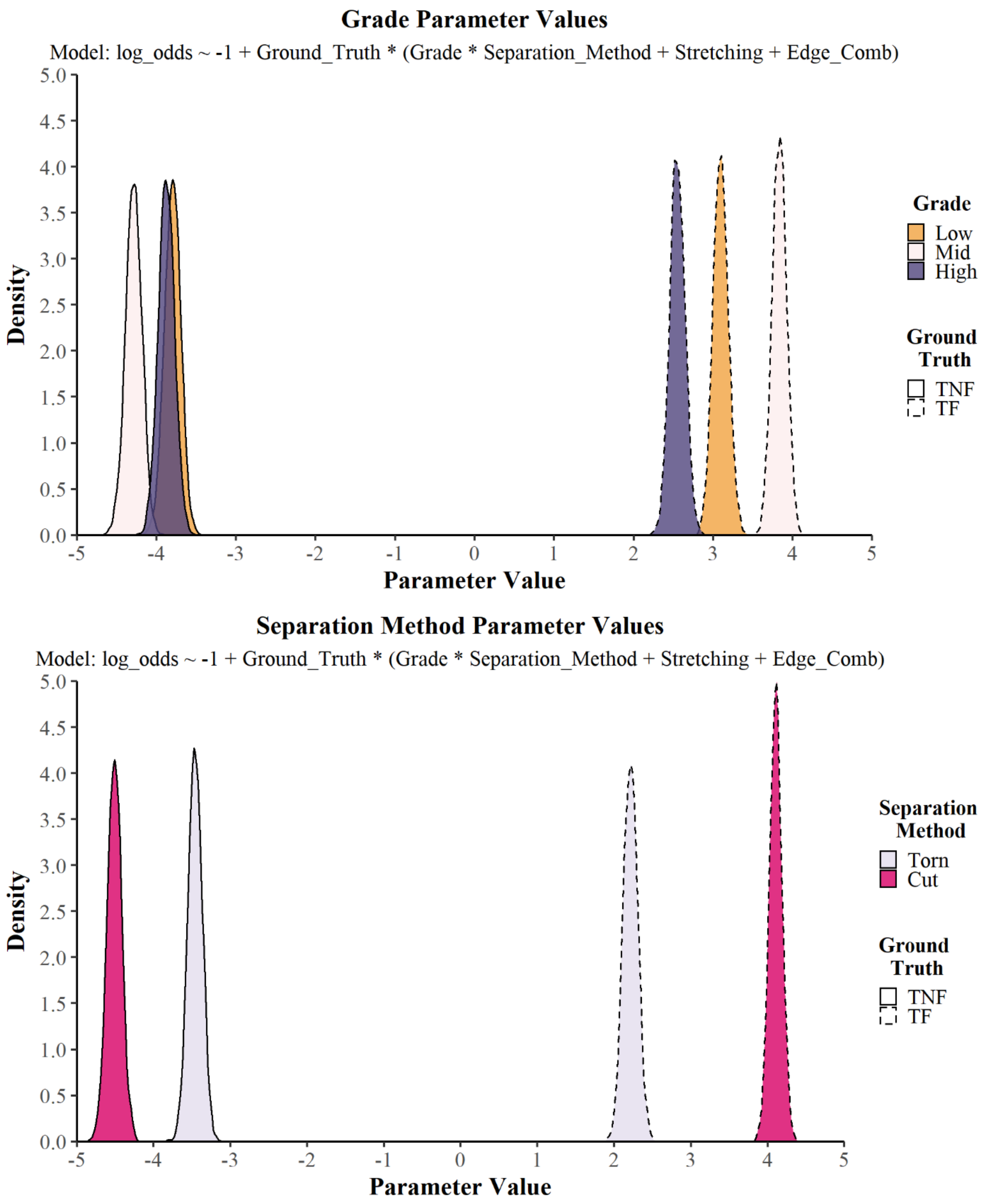


Figure 2-11. Marginalized posterior distribution of the logistic model parameters fitted to the duct tape data. **Figure 2-11A** (top) shows the distributions of the coefficients for the grade, while **Figure 2-11B** (bottom) shows the coefficients for the separation method.

First, comparing both sides of **Figure 2-12A** shows that the expected values of the log-odds ESS are similar in means but have a larger variance than empirical ESS. **Figure 2-12A** also shows that the distributions of the empirical data are multimodal. In contrast, the distributions of the expected values are Gaussian by the assumption of the model used in this study. The similarities and differences between the predicted and empirical ESS distributions allow us to conclude that (a) the model correctly captures the mean effects of the factors under investigation; however, (b) the effect(s) of some unknown factor(s) is(are) not accounted for.

Second, **Figure 2-12A** shows that while the distributions for the coefficient values for the different grades of tape are very well separated (which indicates statistically different effects for the different levels of the grade factor), the resulting impact of varying the grade factor on ESS is not particularly pronounced when accounting for the other factors and analyst variability. In addition, we observe that the model predicts a certain degree of overlap between the ESS resulting from true fits and true non-fits, just as it can be observed in the empirical data.

Figure 2-12B shows the counterfactual plot for the separation method. Similar conclusions to **Figure 2-12A** can be drawn from **Figure 2-12B**. While the distributions of the coefficients for the separation method are well separated, the predicted effect of the separation method is not as clear once other factors and analysts' variability are considered. **Figure 2-12B** confirms that cut tapes result in better separated ESS distributions than torn tape. These results indicate that despite the cleaner edges, scissor-cut edges still retain sufficient features for reliable comparisons, particularly when these observations are made at the microscopic bin sub-unit areas. Documentation of features, including severed dimples and consistent weft fibers that cross over the fracture edge, are common at the microscopic level in these samples, contributing to high ESS values.

As mentioned above, the selected logistic model for this data does not appear to account for all factors influencing the observed ESS. However, the data collected to date can still provide valuable insights about the contribution of factors to the expected ESS and quality of the fit. These trends can then be used to support the opinions of examiners conducting physical fit examinations of duct tapes and understanding how to set criteria when deciding on a fit examination outcome and associated uncertainties.

The probative value of tapes can be estimated using ESS to calculate score-based likelihood ratios [5, 22-24]. SLR values describe the likelihood of a given ESS value given the hypothesis that the pair is either a true fit or a true non-fit. The distribution of scores from each quality is estimated using kernel density estimation to calculate SLRs. These estimated distributions are shown in **Figure 2-13**. Because the counterfactual plots show that tape grade is a critical factor for determining the similarity score, the datasets have been combined by grade to generate the kernel density plots; however, expanding to larger datasets is recommended in the future.

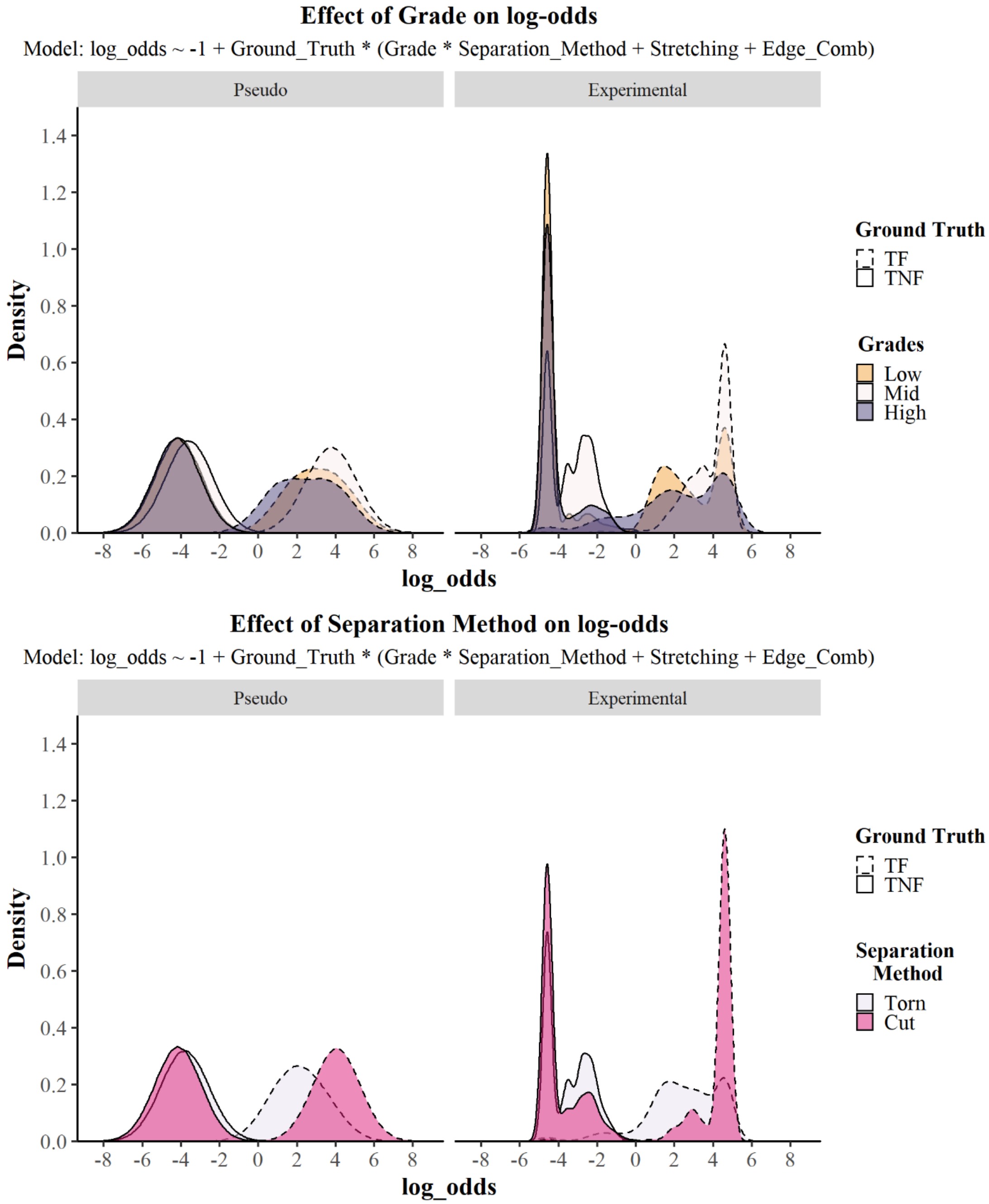


Figure 2-12. Counterfactual plot showing the distributions of the expected ESS values for duct tape data. The counterfactual plot shows the expected ESS values resulting from the model and the experimental data**. Figure 2-12A** (top) shows the distributions marginalized for the grade of tape. Medium-quality tape generally results in higher ESS than other types of tape when samples truly originated from the same tape. **Figure 2-12B** (bottom) shows the distributions marginalized for the separation method. Scissor-cut tapes generally contribute to higher ESS than torn tapes when samples truly originated from the same tape. Both grade and separation methods do not provide substantial differences in expected ESS values for true non-fit pairs.

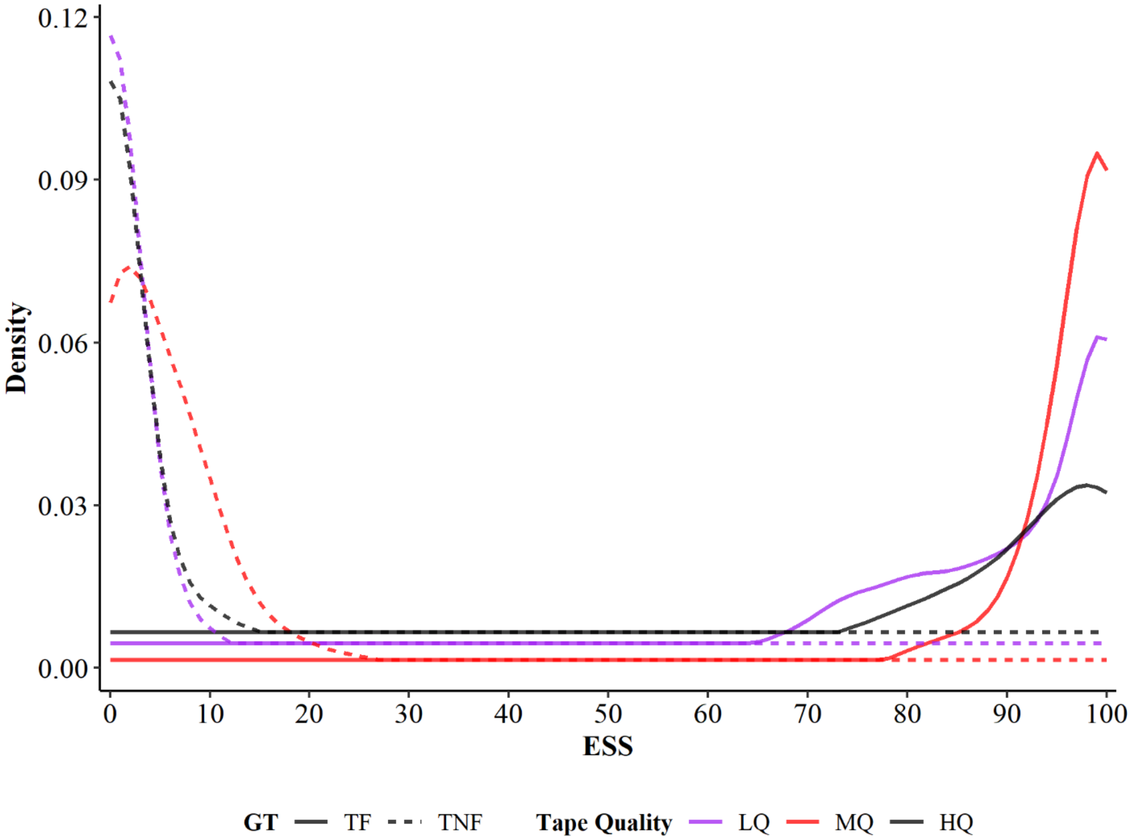


Figure 2-13. Kernel density plots depict the likelihood of obtaining a similarity score of 0 to 100 for each tape grade. The curves are separated by true fit and true non-fits. A threshold using a 3N rule [25-26] was set for each dataset to minimize extreme likelihood values resulting from a lack of data. Note that the kernel density estimates were bounded to avoid having a probability mass below 0 and above 100.

Due to the lack of overlap in the ES score distributions for true fit and true non-fit pairs, there are instances where there are no values for a given ESS value. To overcome this, a 3N [25-26] rule is applied to each dataset, in which the density of a given score is set to a threshold of $3/\mathrm{N}$, where N represents the total number of tape pairs in the dataset. For example, for medium quality, there are a combined total of 2016 ESS values in the dataset, so the density threshold is set at $3/2016$. This prevents extreme values resulting from the calculation of SLR values where the likelihood of a true fit or true non-fit is close to or equal to zero. The resulting SLR distributions are shown in **Figure 2-14**. The log of the SLR values is shown to simplify the interpretation of SLR values.

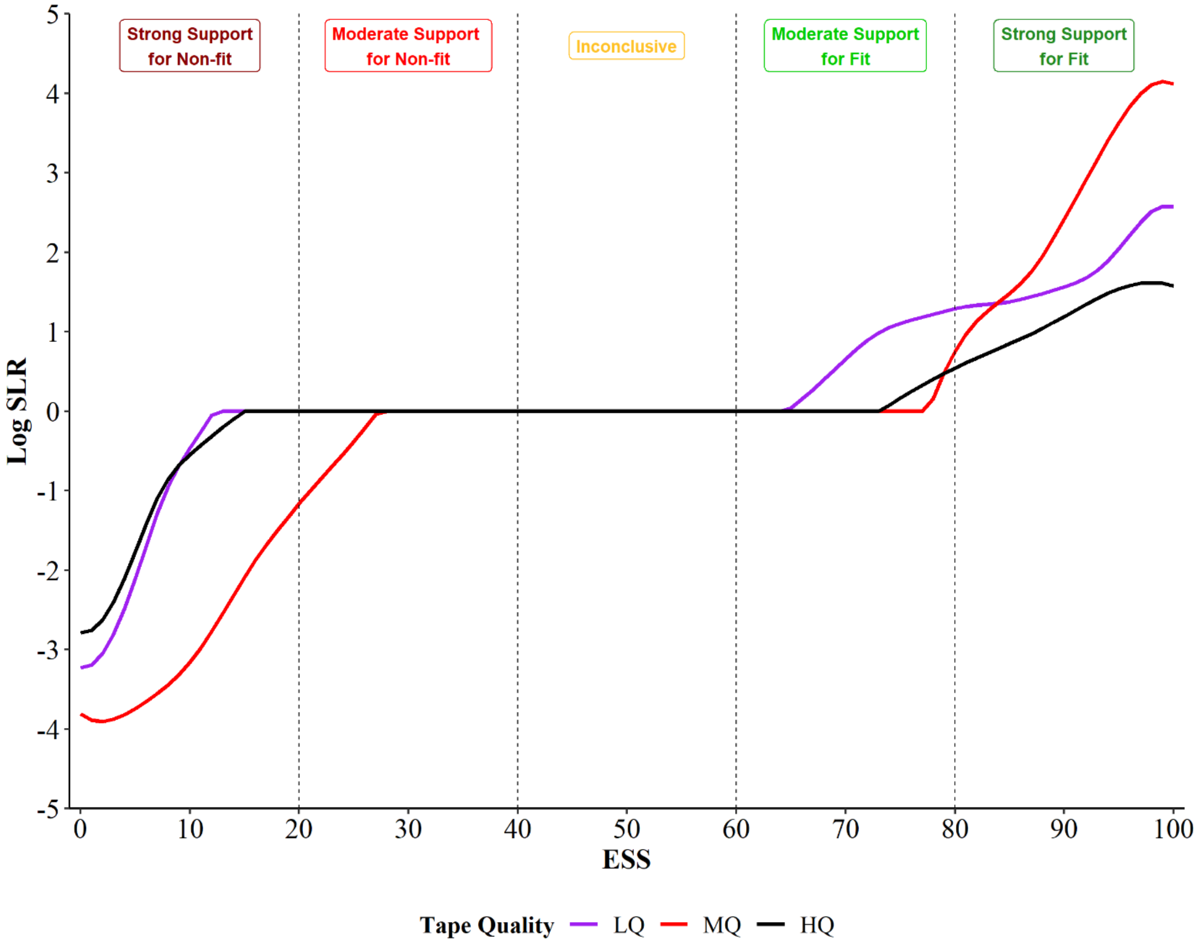


Figure 2-14. SLR distributions for each tape quality across the total range of potential ESS values. The log of the SLR values is shown to simplify the interpretation of the SLR values.

These distributions show the range of ESS values and what conclusions those values provide support for. Scores below 20 for all tape grades show the strongest support, with log SLR ranging from approximately -4 to 0 (corresponding to SLR values between 0.001 to 1). For example, a medium-quality tape pair resulting in an ESS of 0 corresponds to a log SLR of approximately -4, meaning that observing a score of 0 is 10000 times more likely for a pair that is a true non-fit than a fit. ESS values above 80 provide the strongest support for a fit, with log SLR values ranging from approximately 0.5 to around 4 (corresponding to SLR values of 3 to 10000). These values depend more on the quality, as a tape with an ESS value of 90 can result in an SLR of approximately 0.8, 1.3, or 2.4, depending on whether the tape is low, high, or medium quality. The range of 40 to 60 results in log SLR values of 0, meaning that there is no support for either a conclusion of fit or non-fit. However, data in those ESS regions may become available with larger datasets.

2.4. Application of a convolutional neural network to support physical fit comparisons of duct tape

2.4.1. Methods and Materials

2.4.1.1. Creation of a database of digitized tape samples

A subset of the tape samples previously described for the human examination was digitized for use in the computational model. The database contains images collected from the top (backing) and bottom (scrim/adhesive side) of a total of 1998 individual tapes. The images were collected using an EPSON 12000XL scanner using SilverFast 8 software (version 8.8.0r14) at a resolution of a minimum of 600 dots per inch. A background made of black poster board was placed behind each sample during scanning to improve contrast. Minor corrections to each image were made during scanning to enhance the contrast and visibility of the edges and features, such as setting the image's black point to the posterboard to ensure the background was the darkest part of the image and minimize the interference of the posterboard paper's fibers. Additional corrections to remove artifacts such as fingerprints, residual adhesive, long protruding fibers, acetate outlines, and sample labels were conducted as necessary using Adobe Photoshop (version 23.5.1).

To be used for the model, the images needed to be pre-processed. This step prepares the input (the tape images) in the most appropriate format for the model. This approach utilized a neural network, meaning the input needed to be an array of numbers with a fixed length. However, the high-resolution images required to document the tape features properly led to issues introducing the data to the model. The dimensions of the images needed to be reduced to prevent computational memory issues, but this process can result in the loss of information [27–29]. This problem was addressed in the algorithm's image collection, pre-processing, and architecture. The images were collected in as low a resolution as possible that still allowed for recognition of features (600 dpi), as well as the preprocessing of the tape images through reduction of the image width to as close to the comparison edge as possible. The process is detailed in the ForensicFit package developed by Tavadze and Lang [18, 30], available on GitHub. This python package was designed explicitly to compare forensic trace materials in physical fit evaluations. It provided the means to store and adapt raw images for pre-processing, statistical analysis, and training a machine-learning model.

2.4.1.2. Training of the computational model

To use the CNN, the input images must be all prepared in the same manner. While this may seem an easy process for a human to scan images of the samples, it is worth noting that the width of the tape roll can vary between different sources of tape, which could cause substantial issues for CNN in the learning process. To account for this potential variation, the width of the tape image (y-dimension) was cropped according to the actual width of the tape and then resized to ensure consistent inputs to the networks. While this can cause distortion if all samples in the database were treated similarly, resizing the samples would not substantially influence the output. To keep the two sides of the tape properly together as part of the decision-making process between two separate tape samples, the two images are concatenated along with those of the other sample. Two independent, identical CNN models were trained separately on the scrim and backing side of the image tapes, each of which would produce a prediction of fit or non-fit. These two outputs would then be combined, leading to a final probability that would predict a fit, non-fit, or inconclusive classification. More detail on the exact architecture of the model is described in Tavadze et al. [18].

A critical aspect of machine learning is the training of the model. Before training, two potential issues inherent to the tape material were addressed. The first considers the ratio of true fits to true non-fits. In the context of tape material used in forensic science, there are inherently more true non-fits than true fits. A given sample can only fit with one other edge but would be a true non-fit to every other edge in the database. As a result, the dataset in this study was imbalanced, which could affect the accuracy of the final model [31-33]. The model was designed with a fit to the non-fit ratio of 3:10, selected by assessing the performance using various ratios [18]. The second issue was possible misalignment or orientation of potential sides during the comparison of the tape. To prevent the model from being unable to handle random orientations of input edges, the training of the model was conducted with random mirroring, where the sides of the tapes would randomly be flipped so the model would learn not to reject edges just because they faced the same side [18]. The dataset was split into a ratio of 80:20 for training and validation. The learning rate, optimizations, and the number of iterations of the learning were optimized by trial and error [18].

2.4.2. Results and Discussion

Overall, the model demonstrates comparable performance to the human analysts. **Table 2-3**Error! Reference source not found. shows the performance rates of the model for the subset of samples compared (hand-torn and scissor-cut of all three tape qualities). The model results in lower false-negative rates for the hand-torn sets. This result is not unexpected, as hand-torn samples are more likely to produce distinct edge patterns, which is also recognized in the human examination. However, unlike in the human examination, the model did have some false positives. However, all false positive rates were below 10%.

Table 2-3. Performance of the computational model for the tape sets [18]. The tape sets referred to here are low-quality (LQ), medium-quality (MQ), or high-quality tape (HQ). The separation method for the subset is denotated as scissor-cut (SC) or hand-torn (HT).

Set	LQ-SC	LQ-HT	MQ-SC	MQ-HT	HQ-SC	HQ-HT
No. of comparison pairs	73	68	60	81	48	68
No. of fits/No. of non-fits	30/43	19/49	17/43	23/58	20/28	19/49
False-positive rate (%)	0.0	8.7	0.0	0.0	7.1	7.7
False-negative rate (%)	25.0	9.1	15.8	6.7	7.1	0.0
True-negative rate (%)	100	91.3	100	100	92.9	92.3
True-positive rate (%)	75.0	90.9	84.2	93.3	92.9	100
Accuracy (%)	72.6	76.5	88.3	85.2	81.3	71.3

It is noteworthy that the model's performance is based only on twenty percent of the dataset, as the rest is used to train the model. High-performing CNNs are often trained on datasets containing tens of thousands or millions of images, whereas this study only incorporates a database of around 4000 samples [34-36]. Despite the limitation, there are still enough images in this tape database to perform pilot testing, the feasibility of this approach, and the potential of a computational model that can perform automated quantification of the similarity/dissimilarity of duct tape edges and assist human analysis of samples.

Trends observed in the human examination are also reproduced in the computational comparison. Namely, the role of the quality of the tape shows a substantial influence on performance. In both approaches, MQ samples demonstrate high accuracy and low error rates. LQ samples perform the next best, with HQ demonstrating lower performance than the others. As observed in the human examination, likely, the more extensive distortion seen in LQ tapes and the lack of distinct features and edges observed in HQ pairs cause reduced performance and confidence in the computational examination. In addition, these samples are likely to introduce additional noise. The adhesive of the HQ samples is thick and obscuring. While removing the adhesive at the edge exposed the fibers, the residual adhesive could produce issues for the machine-learning.

Similarly, the LQ adhesive/scrim side was more transparent than HQ but not as distinctive as in the MQ, meaning that the contrast between the fibers and the bin areas could have contributed to lower performance. The separation method also has a measurable impact on performance. There is a relatively high false negative rate in all three SC sets. This is likely due to an abundance of straight edges in those sets. If the pair has strongly angled edges, non-fits are more recognizable, but not so when the edges are very straight, as it becomes more complicated to compare. We hypothesize that human analysts recognize more features in the microscopic comparison that can more definitively influence a correct decision (**Figure 2-6**). When the edge morphology is insufficient, characteristics such as alignment of dimples and calendaring striations along with the fiber weave become critical for comparison. The model is not likely to recognize these features as effectively. However, both the human-based examination and computation model demonstrate that scissor-cut edges retain sufficient features for comparison, which contrasts previous assumptions about scissor-cut edges [8].

The relationship between the quantitative metrics from the human-based and computational approaches is shown in **Figures 2-15** and **2-16**. The boxplots in **Figure 2-15** compare the distributions of ESS estimated by the analyst to the fit membership probability assigned by the CNNs. The results of each approach are separated by the respective subsets of tapes and the known ground truth (i.e., true fit or true non-fit). Notably, true fit and true non-fit pairs are generally well separated using either metric. True non-fit pairs typically receive ESS of 30 or less, with most below 10. Most true non-fits had low probabilities (generally less than 0.10) assigned by the model for the "fit-membership." True fits typically receive high values (ESS of 80 or higher or probability of fit-membership greater than 0.7 for the CNN). The hand-torn sets span a wider range of probability outputs and ESS for true fits compared to scissor-cut samples. It is also observable where the computer-based model yields false positive pairs in various subsets, where there are none in the human analysis. Despite this, it is worth noting that when considering the model's output, some of the true fit tape pairs with lower probabilities are more likely to be reported inconclusive due to a disagreement in membership probabilities reported between the scrim and the backing CNNs.

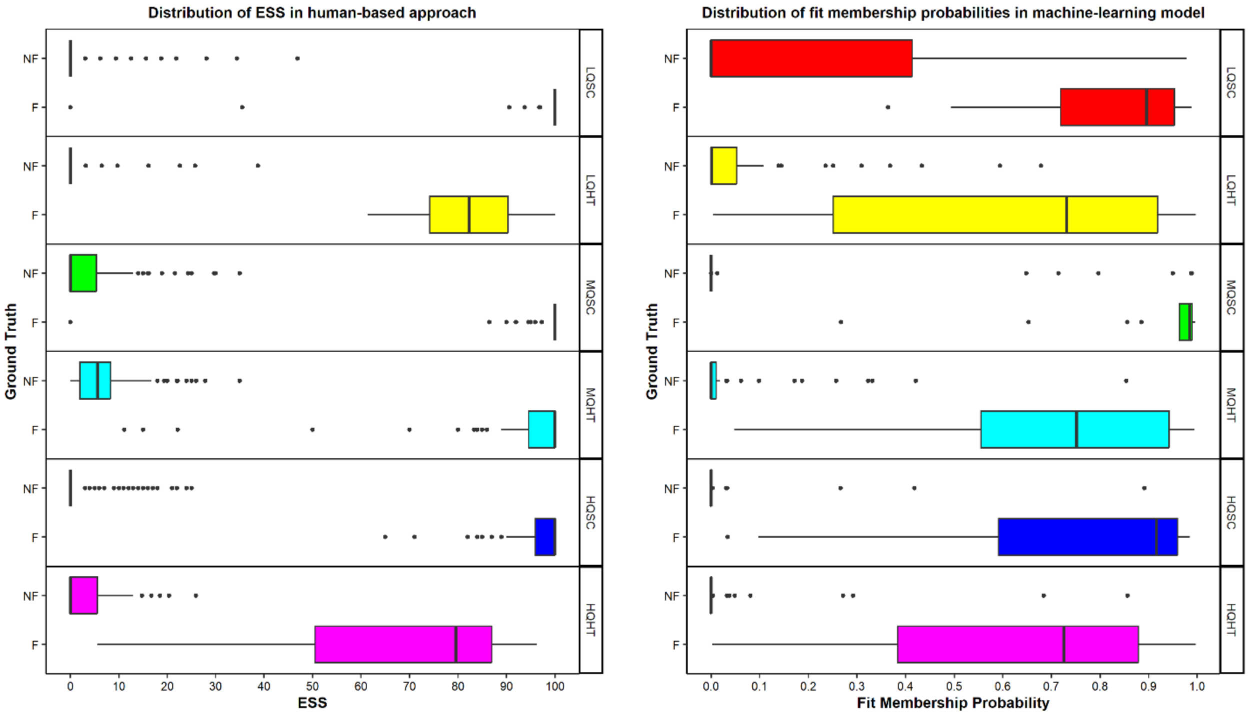
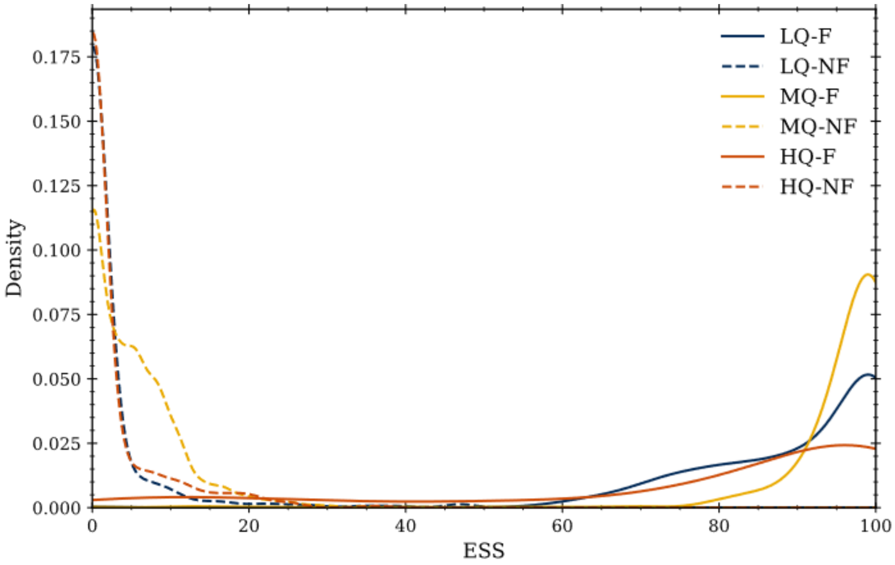


Figure 2-15. Boxplots of edge similarity score for all the sets from human examination (left). The right image shows the boxplots of the probabilities assigned by the model. The data was obtained by combining the output of the two convolutional neural networks (scrim and backing).

The edge similarity scores are obtained from all the tapes in the subsets, while the probabilities are obtained only from 20% of the dataset (validation set).

An alternative comparison of the two approaches is shown in **Figure 2-16**, where the distribution of ESS and CNN probabilities are estimated using kernel density estimation. The datasets have been combined by quality in both approaches to generate the kernel density plots. In both methods, MQ tapes more commonly receive high scores or probabilities for true fits, followed by LQ and HQ. This result is consistent across both approaches but is less pronounced in the computational model. A similar trend is observed for true non-fits, where most tape pairs have low values (ESS 20 or less, CNN fit membership probabilities less than 0.3). In addition, there is a limited overlap between true fits and true non-fits in both cases, indicating good discrimination power. More importantly, both ESS and CNN outputs can be used as the basis to estimate likelihood ratios to estimate the respective probative value or weight of the evidence [5, 37]. This provides the opportunity to use a probabilistic interpretation of the probative value of the evidence to support and inform an analyst’s conclusion.



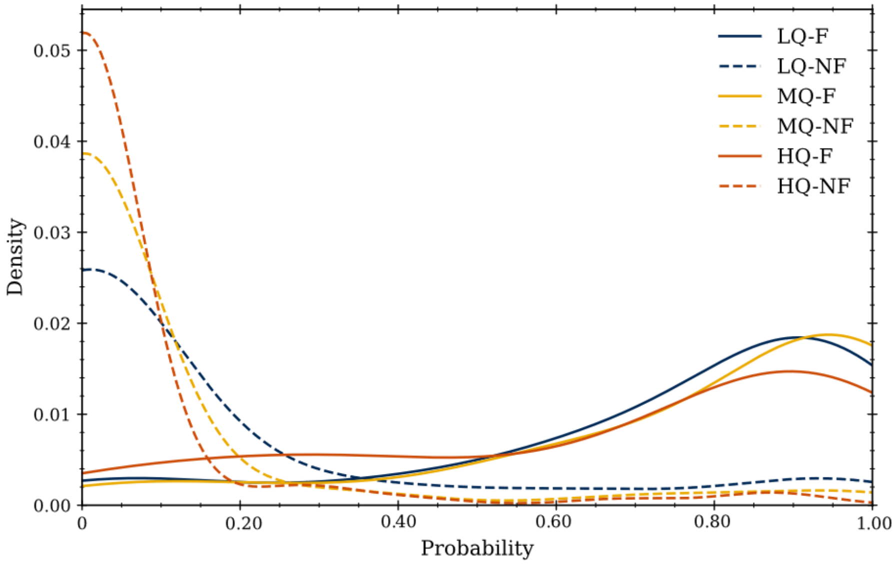
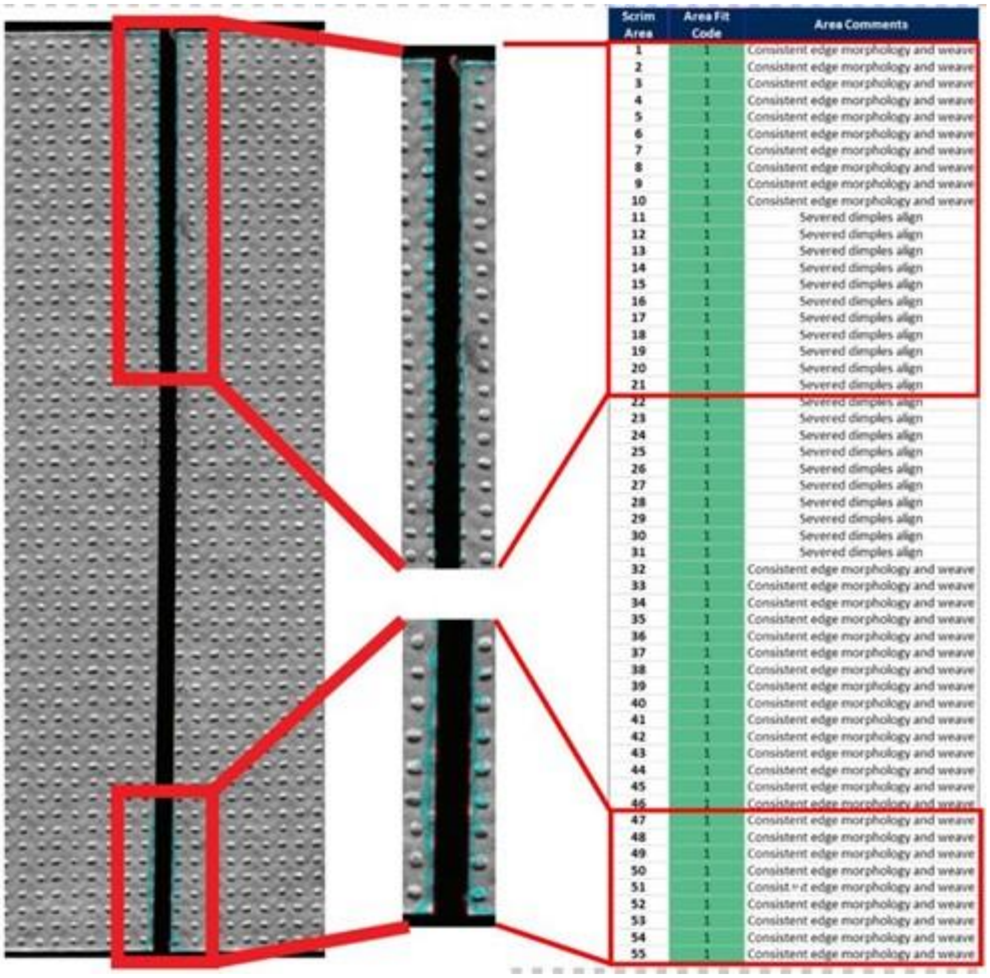


Figure 2-16. Kernel density plots demonstrating the spread of the different datasets for both the human-based approach and the computational mode. The top kernel density represents the edge similarity score estimated by the analyst's examination (top). The bottom kernel density shows the estimates of the fit membership probabilities calculated by the computational method for the tape pairs [18].

These results indicate good potential for the model in addition to the human examination; a concern of using neural networks in forensic science is the need for more transparency in how decisions are made. While the model's architecture may be known, the exact features the model is tracking and using for decision-making can be challenging to demonstrate. This is particularly important for future implementation of computational methods, as forensic practitioners who use the model are likely to be questioned during testimony on the model's inner workings. As one of the intended goals of this model is to allow it to be more accessible, it is critical to understand what the model is using for the predictions. As a solution to this concern, this technique shows the critical areas for comparison to the model by Layer-wise Relevance Propagation (LRP) [38], where the important pixels for the prediction decision are highlighted.

Figure 2-17 shows an example pair with LRP highlighting important pixels in cyan blue. This is a true fit pair from the HQ-SC subset, resulting in a model prediction of a fit with a probability of 0.87 and its conclusion with an ESS of 100. According to the LRP, the most important pixels are at the edges of the tape, where the model recognizes both edge morphology and the texture of the backing. As an example, the top portion of the pair has consistent severed dimples close to the edge highlighted by the model. The notes from the human examination also recognize these dimples as critical for the fit decision for the same area of the tape. While not severed at the bottom of the tape pair, the dimples are still recognized as important by the algorithm. This feature would also be useful during an examination by a human analyst. While more variable than other features, the spacing between dimples could still be valuable when determining non-fits if the spacing is substantially different from the rest of the observable backing.



Scrim Area	Area Fit Code	Area Comments
1	1	Consistent edge morphology and weave
2	1	Consistent edge morphology and weave
3	1	Consistent edge morphology and weave
4	1	Consistent edge morphology and weave
5	1	Consistent edge morphology and weave
6	1	Consistent edge morphology and weave
7	1	Consistent edge morphology and weave
8	1	Consistent edge morphology and weave
9	1	Consistent edge morphology and weave
10	1	Consistent edge morphology and weave
11	1	Severed dimples align
12	1	Severed dimples align
13	1	Severed dimples align
14	1	Severed dimples align
15	1	Severed dimples align
16	1	Severed dimples align
17	1	Severed dimples align
18	1	Severed dimples align
19	1	Severed dimples align
20	1	Severed dimples align
21	1	Severed dimples align
22	1	Severed dimples align
23	1	Severed dimples align
24	1	Severed dimples align
25	1	Severed dimples align
26	1	Severed dimples align
27	1	Severed dimples align
28	1	Severed dimples align
29	1	Severed dimples align
30	1	Severed dimples align
31	1	Severed dimples align
32	1	Consistent edge morphology and weave
33	1	Consistent edge morphology and weave
34	1	Consistent edge morphology and weave
35	1	Consistent edge morphology and weave
36	1	Consistent edge morphology and weave
37	1	Consistent edge morphology and weave
38	1	Consistent edge morphology and weave
39	1	Consistent edge morphology and weave
40	1	Consistent edge morphology and weave
41	1	Consistent edge morphology and weave
42	1	Consistent edge morphology and weave
43	1	Consistent edge morphology and weave
44	1	Consistent edge morphology and weave
45	1	Consistent edge morphology and weave
46	1	Consistent edge morphology and weave
47	1	Consistent edge morphology and weave
48	1	Consistent edge morphology and weave
49	1	Consistent edge morphology and weave
50	1	Consistent edge morphology and weave
51	1	Consistent edge morphology and weave
52	1	Consistent edge morphology and weave
53	1	Consistent edge morphology and weave
54	1	Consistent edge morphology and weave
55	1	Consistent edge morphology and weave

Figure 2-17. Layer-wise Relevance Propagation (LRP) analysis of a true fit HQ-SC pair compared to human documentation of the same pair. LRP identifies the most important features at the tape's top and bottom, especially the severed dimples and edge morphology [18].

Expansion of this database is expected to improve the confidence of the CNN model and address issues related to more realistic casework samples (samples with extensive artifacts and distortions). In the meantime, these approaches provide the means to estimate rates of misleading evidence.

2.5. Conclusions

The edge similarity score method provides good overall performance for physical fit examinations. The addition of the two scissor cut sets of low and high quality coincides with the accuracy seen in the previous medium-quality cut set, with accuracy over 98% and no false positives. Stretching of hand-torn tapes adds some variability to the ESS but does not significantly influence the overall conclusion. Generally, edge similarity scores for true fits are consistently greater than 80. Scores of true fit pairs are more widely distributed for the low- and high-quality hand-torn sets. The use of the systematic method provides consistent results between examiners. The defined comparison criteria assist with understanding the reported conclusion for a given pair and allow for understanding what features are most common in the different qualities and separation methods. For example, severed dimples primarily occur on scissor-cut edges and are rare on hand-torn edges. This is most often due to the relative placement of the weft fibers, which on all three tapes tend to fall underneath the dimples. Hand-torn tapes generally do not tear across a single weft fiber and so commonly do not show severed dimples. Conversely, hand-torn edges display more distortion and puzzle-like patterns due to stretching, as the scissor-cut edges do not generally present with distorted edges. An exception is the high-quality hand-torn sets, which produce less distortion upon separation and few distinctive puzzle-like patterns.

This study shows that different variables have varying impacts on the edge similarity score when evaluating the statistical effect of the experimental factors of interest. For non-fits, the influence of the separation method and tape quality on the ESS values is negligible, and the ESS trend towards low values regardless. Generally, scissor-cut tapes result in higher scores for true fit pairs than hand-torn pairs. Regarding tape quality, in true-fit pairs, medium-quality tapes receive higher scores overall, followed by low-quality and high-quality. The calculated score-based likelihood ratios based on ESS values show that the likelihood of observing an edge similarity score below 20 is much more likely if the edges are true non-fits.

Meanwhile, tapes with scores above 80 are much more likely if the samples originated from true fitting pairs. While the effects of these variables are not overt, the observed effect seen in the counterfactual plot coincides with the exploratory data analysis from the different tape sets. As such, examiners must consider the suitability of the tape material during physical examinations. The construction and physical properties of the tape may result in fewer informative features or contribute to more uncertainty in the decision-making process. While it may be challenging to determine the separation method for tapes, the findings of this study show that examiners should not assume that a potential cut edge will not possess distinctive features for comparison. Conversely, this study indicates that scissor-cut edges result in the slightest variation of scores for all three grades, meaning the examiners consistently note enough features to have a higher degree of confidence in their reported conclusion.

Establishing a set of standard features and terminology for comparison of tape edges provides more straightforward documentation of the criteria used in examinations of physical fits. The proposed methods are anticipated to assist examiners in evaluating the edges through documentation and observation in a systematic manner, reporting standardized comparison criteria, understanding potential sources of variations, and presentation of error rates. The ESS method is valuable for training and peer-review processes. The three-step comparison process allows for a more nuanced understanding of the quality of the edge from macroscopic and microscopic observations and defined criteria.

Additionally, this research provides a functional computational model for tape physical fits. The open-source python package, ForensicFit, developed by our partners in this research [18, 30], provides a means for pre-processing images for forensic physical fit examination. The package has provided data for machine learning to train two independent convolutional neural networks for the respective tape sides (backing and scrim), with analysis performed on the combined probabilities and membership assignment. The proposed computational model performs well with comparable high accuracies and low error rates to the human examination. This study confirms the previous findings observed by the analysts through physical and microscopic examination, such as the value of scissor-cut edges for comparison despite cleaner edges. As such, it demonstrates the potential of machine learning models to provide statistical support to analyst conclusions. The application of LRP allowed the computational model to recognize the most critical features and demonstrates that they are consistent with the features identified through human examination. The results also indicate that the approach has room for improvement, particularly with the addition of larger datasets in the future, which will assist the learning process of the model.

This study demonstrates that not every physical fit conclusion holds the same confidence. Therefore, we recommend accompanying a reported physical fit with an assessment of its probative value. Overall, the ESS approach demonstrates that observing a physical fit on two duct tape pieces that were not joined together is unlikely, as no false positives are observed in the various populations evaluated. It is worth noting that both studies result in errors, reinforcing that physical fit examinations are not error-free, whether conducted by an analyst or by a computer algorithm. Therefore, it is critical to raise awareness of potential error sources and factors influencing the accuracy of physical fit examinations. Even though this study uses a relatively small dataset (~4000 images) compared to what is typically used to train CNNs for other applications, the model still performs comparably with the human examination. As such, it shows the potential to help provide additional support to physical fit examinations. The model presented here allows the forensic community to collaborate and build databases of other materials of forensic interest.

2.6. References

1. Brooks E, Prusinowski M, Gross S, Trejos T. Forensic physical fits in the trace evidence discipline: A review. Forensic Sci. Int. 2020. 313: 110349.
2. Gross S. 2020 Physical Fits Survey. 2020.
3. Organization of Scientific Area Committees. Development of Quantitative Assessment and Evaluation of Error Rates in Physical Fit Determinations of Trace Materials. Natl. Inst. Sci. Technol. 2018.
4. Dror IE. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. Anal. Chem. 2020. 92: 7998–8004.

5. Prusinowski, M., Brooks, E. & Trejos, T. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. Forensic Sci. Int. 2020. 307: 110103.
6. Mehltretter AH, Bradley MJ. Forensic analysis and discrimination of duct tapes. JASTEE. 2012. 3(1):2–20.
7. LaPorte K, Weimer R. Evaluation of duct tape physical characteristics: part I – within-roll variability. JASTEE. 2017. 7(1): 15–34.
8. McCabe KR, Tulleners FA, Braun JV, Currie G, Gorecho EN. A quantitative analysis of torn and cut duct tape physical end matching. J. For. Sci. 2013. 58: 34–42.
9. Bradley MJ, Keagy RL, Lowe PC, Rickenbach MP, Wright DM, LeBeau MA. A validation study for duct tape end matches. J. For. Sci. 2006. 51: 504–508.
10. van Dijk CD, van Someren A, Visser R, Sjerps M. Evidential value of duct tape comparison using loopbreaking patterns. Forensic Sci. Int. 2022. 332: 111178.
11. Morrison G. Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio. Australian Journal of Forensic Sciences. 2013. 45(2): 173-197. DOI: 10.1080/00450618.2012.733025.
12. Sharma V, Bains M, Verma R, Verma N, Kumar R. Novel use of logistic regression and likelihood ratios for the estimation of gender of the writer from a database of handwriting features. Australian Journal of Forensic Sciences. 2021. https://doi.org/10.1080/00450618.2021.1956587
13. Biosa G, Giurghita D, Alladio E, Vincenti M, Neocleous T. Evaluation of Forensic Data Using Logistic Regression-Based Classification Methods and an R Shiny Implementation. Front. Chem. 2020. https://doi.org/10.3389/fchem.2020.00738
14. Curran J. Introduction to Data Analysis with R for Forensic Scientists. Taylor & Francis Group. Boca Raton, FL. 2010.
15. Franke K, Srihari SN. Computational Forensics: An Overview, in: Comput. Forensics. Springer Berlin Heidelberg, Berlin, Heidelberg, n.d.: pp. 1–10. https://doi.org/10.1007/978-3-540-85303-9_1.
16. Ristenpart W, Tulleners F, Alfter A. Quantitative algorithm for the digital comparison of torn and cut duct tape; Final Report to the National Institute of Justice Grant 2013-R2-CXK009. University of California at Davis. 2017.
17. Spaulding J, Picconatto G, Characterization of fracture match associations with automated image processing, Forensic Sci. Int. 2023. 342:111519. https://doi.org/10.1016/j.forsciint.2022.111519.
18. Tavadze P, Lang L, Prusinowski M, Andrews Z, Trejos T, Romero A. Using convolutional neural networks to support examiners in duct tape physical fit comparisons. Forensic Science International. 2023. Submitted, under initial review.
19. Abadi M et al. {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems. 2015. https://www.tensorflow.org/.
20. B. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. *arXiv*. 2015.

https://doi.org/10.48550/arXiv.1511.08458

21. Stan Development Team. 2022. Stan Modeling Language Users Guide and Reference Manual, v2.30. https://mc-stan.orgI
22. Leegwater AJ, Meuwly D, Sjerps M, Vergeer P, Alberink I. Performance study of a score-based likelihood ratio system for forensic fingermark comparison. J. For. Sci. 2017. 62(3):626–640.
23. Hepler A, Saunders C, Davis L, Buscaglia J. Score-based likelihood ratios for handwriting evidence. For. Sci. Int. 2012. 219:129–140.
24. Zadora G, D. Ramos D. Evaluation of glass samples for forensic purposes – an application of likelihood ratios and an informational-theoretical approach, Chemom. Intell. Lab. Syst. 2010. 102:63–83.
25. Eypasch E., Lefering R., Kum C.K., Troidl H. Probability of adverse events that have not yet occurred: A statistical reminder. BMJ. 1995. 311:619–620. doi: 10.1136/bmj.311.7005.619.
26. Möller S, Ahrenfeldt LJ. Estimating Relative Risk When Observing Zero Events-Frequentist Inference and Bayesian Credibility Intervals. Int J Environ Res Public Health. 2021. 18(11):5527. Doi: 10.3390/ijerph18115527
27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning, Springer, New York, New York, NY. 2009. https://doi.org/10.1007/978-0-387-84858-7.
28. Bellman RE. Dynamic programming. 1967. https://doi.org/10.1016/S0076-5392(08)61063-2.
29. Page ES, Bellman R. Adaptive Control Processes: A Guided Tour. 1962. https://doi.org/10.2307/2343225.
30. Tavadze P, Lang L. romerogroup/ForensicFit: First release of ForensicFit Package. 2022. https://doi.org/10.5281/zenodo.7435058.
31. Kaur H, Pannu HS, Malhi AK. A Systematic Review on Imbalanced Data Challenges in Machine Learning, ACM Comput. Surv. 2020. 52: 1–36. https://doi.org/10.1145/3343440.
32. Van Hulse J, Khoshgoftaar TM, Napolitano A. Experimental perspectives on learning from imbalanced data. in: Proc. 24th Int. Conf. Mach. Learn. - ICML ’07, ACM Press, New York, New York, USA. 2007. pp. 935–942. https://doi.org/10.1145/1273496.1273614.
33. Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. in: Lect. Notes Electr. Eng., 2020. 209–221. https://doi.org/10.1007/978-3-030-29407-6_17
34. Deng L. The MNIST database of handwritten digit images for machine learning research, IEEE Signal Process. Mag. 2012. 29: 141–142. https://doi.org/10.1109/MSP.2012.2211477
35. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8693 LNCS 2014. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

36. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009. pp. 248–255. https://doi.org/10.1109/cvpr.2009.5206848

37. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – Part I. physical fits of torn and cut duct tapes. Forensic Science International. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567

38. Binder A, Montavon G, Lapuschkin S, Müller KS, Samek W. Layer-wise relevance propagation for neural networks with local renormalization layers. Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 9887 LNCS. 2016: 63–71. https://doi.org/10.1007/978-3-319-44781-0_8

Appendix 2-A

The following object links to a spreadsheet containing the template for the duct tape comparisons.

WEST VIRGINA
UNIVERSITY DUCT T

Chapter 3: Interlaboratory evaluations of a systematic method for assessing duct tape physical fits

3.1. Overview

This research applies the method we developed to quantify compared duct tape edges through evaluation of the method by forensic practitioners. We evaluated the method through two interlaboratory exercises, the first with twenty (20) participants and the second exercise having eighteen (18) current forensic trace evidence practitioners. With each participant comparing seven samples, the total number of comparisons across both studies reaches 266.

The goal of the first exercise was to assess method performance, variation, and determine what aspects of the method needs improvement. The first study results show a 97% accuracy according to the reported conclusions made by the practitioners. However, upon further inspection of the documentation templates, there are numerous instances of participants reporting a conclusion that does not correspond with the reported edge similarity score (ESS). For example, a pair with a reported ESS of 20 may have been reported as a fit with high confidence. As such, the results are evaluated using ESS values, which instead indicates an accuracy of 95%. The discrepancies observed in the documentation are hypothesized to be due to misunderstanding of the methods, lack of standardized training, and participants relying on their own criteria to make decisions. As such, for the second exercise, more formalized training, more extensive documentation, and changes based on feedback received in the first study are incorporated to determine if performance and consistency in responses improved. The second exercise demonstrates a slightly improved accuracy (99%), and no discrepancies between the reported conclusion and ESS values. Furthermore, the enhanced documentation allowed for visual demonstration of results between participants and allowed for transparency of the thought process of the analysts for what they consider the most important features for making decisions. The formalized training ensures that all participants are interpreting the method, the documentation, and the ESS value in a consistent manner.

In addition to the performance of the method across both interlaboratory exercises, the ESS values provided by the participants are assessed using Dunnett control charts and z-scores. These two approaches indicate that most participants' results were satisfactory, with a few cautionary and insufficient results in the first exercise, and fewer cautionary results in the second exercise.

Overall, the inter-laboratory performances indicate that the proposed qualitative and quantitative criteria assist practitioners in arriving at reproducible assessments of the quality of a fit. While there are still improvements that can be made to the training and documentation template, the ESS method for duct tapes provides thorough records of physical fit evaluations and allows for a straightforward and transparent comparison of results among independent analysts, paving a way to facilitate blind verification of physical fit samples in forensic laboratories.

The following chapter is an adaptation of an article submitted for publication ©2023: Prusinowski M, Brooks E, Neumann C, Trejos T. Forensic interlaboratory evaluations of a systematic method for examining, documenting, and interpreting duct tape physical fits. Forensic Chemistry.

3.2. Introduction

While physical fit examinations have a long history of use in forensic science; there is an increased interest in the forensic community to demonstrate its reliability and to develop standardized methods for examination and interpretation [1-5].

Although previously published studies demonstrate that analysts have relatively high accuracy rates when examining physical fits, they also demonstrate that physical fits are not exempt from errors [2]. Therefore, it is of the foremost importance to develop tools to detect, evaluate and minimize potential sources of errors and bias in physical fit examinations. One material that has received attention in this field is duct tape, due to the prevalence of this type of evidence in casework [5-10].

The studies performed by Prusinowski et al. demonstrate low error rates using a method that reports a similarity score for a duct tape physical fit pair [8, 9]. The proposed method has evolved from its initial development and assessment of performance rates through large datasets of known ground truth items (> 3000 duct tape comparisons) [8], to the identification and development of standardized reporting criteria and assessment of significant factors that can influence physical fit examinations [9], and finally, its evaluation and refinement through interlaboratory studies. The examinations in the two initial studies from Prusinowski et al. are performed solely by a few researcher analysts. The present study is designed to identify the capabilities and limitations of the proposed approach when the examinations are performed by practitioners with experience in physical fit examinations but new to the proposed ESS and reporting criteria.

A critical step in the implementation of new methodologies in forensic science involves establishing the performance and robustness of the method. Accreditation standards require that forensic laboratories utilize methods that have substantial testing to verify the utility, validity, reliability, and reproducibility across laboratories and independent analysts. One of the most effective means of evaluating a new method is through interlaboratory studies where multiple practitioners analyze the same sample (or samples with similar compositions/origins) under set conditions [11]. The study is often led by a coordination body, who creates the experimental design and evaluates it with an expert panel prior to distribution to participants [12]. The coordination body also maintains the anonymity of participants, verifies the data, conducts analysis and interpretation of results and incorporates the findings and participant feedback to implement improvements.

In this paper, we report the results of two interlaboratory studies designed to assess the proposed method for examining, documenting, and interpreting physical fits of duct tapes using standardized qualitative descriptors and quantitative metrics. The second interlaboratory study applies feedback from the first to refine the proposed method, and the training provided to the participants. An overarching goal of these exercises is to incorporate this feedback to develop consensus protocols that can be practical and likely to be implemented by the scientific community.

3.3. Methods and Materials

3.3.1. Sample preparation

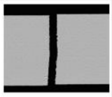
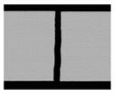
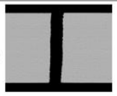
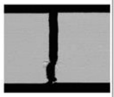
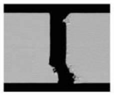
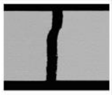
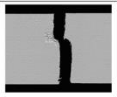
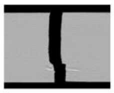
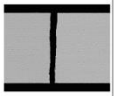
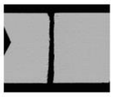
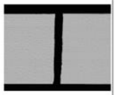
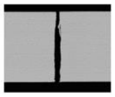
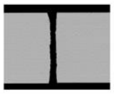
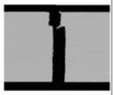
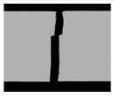
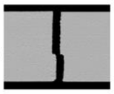
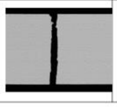
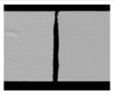
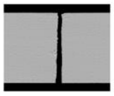
The tape samples utilized in these exercises originate from medium-quality grade duct tape (Duck Brand Electrician's Grade Gray Duct Tape, ShurTech Brands, Avon, OH) [8, 9]. The tape roll has approximately 4.0 mils backing thickness, 2.5 mils adhesive thickness, and 20/8 warp/weft scrim count. Each sample consists of a hand-torn, 6-8 cm long strip of the roll. Samples are placed on clear acetate sheets following fracture to allow for the backing and adhesive side of the tape to be examined without the need to remove the adhesive with solvent. The samples are labeled with unique identification numbers that are traceable to the coordination body but maintain the ground truth unavailable to participants. Pairs of tape samples are created, with some pairs including consecutive strips of the roll, and some pairs including non-continuous strips.

Initial tape pairings are analyzed by four independent analysts using a blind process, where the ground truth was unknown by the panel. The panel analysis consists in assessing each pair for alignment features on both the backing and adhesive sides under a stereomicroscope and following the interpretation criteria and reporting template. Twenty-one (21) pairs resulting in inter-participant ESS relative standard deviations lower than 10% ESS are selected from the samples set and rearranged into seven groups of three similar pairs each, to prepare 3 kits of 7 comparison pairs each. Classification of the seven optimized pairs includes three fits of ESS greater than 80% (F+), one fit of ESS below 80% (F-), and three non-fits of ESS lower than 40% (NF+). The sets also include a variety of edge morphology (straight, wavy, puzzle-like) as described in **Table 3-1**.

3.3.2. Design of Studies

Following discussions with practitioners, the interlaboratory study is designed to allow for direct examination and handling of the samples in a manner that closely resembles casework. To facilitate simultaneous distribution to various participants and maintain practical turn-around times, three kits containing 7 pairs of tape are used. **Table 3-1** shows that the chosen pairs share similar characteristics and ES scores to the corresponding pair in the other kits (e.g., Pair 1 and 3 of Kit A is the same ground truth, similar in edge morphology and has similar ESS values to Pair 1 and 3 of Kits B and C). Refer to **Appendix 3A** for detailed images of the pairs.

Table 3-1. Description of the seven tape pairs selected for each of the three interlaboratory kits, and respective ground truth and consensus values obtained by the pre-distribution examination panel.

Pair ID (ILS#1/ILS#2)	Ground Truth	Kit A (Kit 1 and 4) Consensus Mean and Pair Image	Kit B (Kit 2 and 4) Consensus Mean and Pair Image	Kit C (Kit 3 and 6) Consensus Mean and Pair Image	Overall Pre-Distribution Mean ESS
I (1/3)	Fit (F+)	97 ± 4 	99 ± 3 	97 ± 4 	97 ± 3
II (2/7)	Fit (F-)	77 ± 5 	70 ± 2 	75 ± 4	74 ± 5
III (3/2)	Fit (F+)	88 ± 3 	86 ± 2 	89 ± 2 	88 ± 2
IV (4/1)	Non-Fit (NF+)	11 ± 3 	10 ± 4 	10 ± 3	11 ± 3
V (5/4)	Non-Fit (NF+)	2 ± 3 	0 ± 0 	0 ± 0 	1 ± 2
VI (6/5)	Fit (F+)	95 ± 2 	96 ± 3 	92 ± 4 	94 ± 3
VII (7/6)	Non-Fit (NF+)	5 ± 4 	3 ± 3 	5 ± 4 	4 ± 3

To minimize potential bias, each participant independently receives, processes, and returns the kit and documentation to the coordination body while remaining blind to the ground truth. Study kits are distributed so that each kit returns to the coordination body before re-distribution to the next participant. The experimental design of both exercises aims for seven participants per kit (21 participants per interlaboratory study) with a turnaround time of 3-4 weeks, for approximately 7-8 months to complete all interlaboratory distributions. However, due to uncontrolled circumstances, several participants took longer to submit results, and some could not complete the exercise, so each interlaboratory study took approximately 11 months. A total of 252 examinations from 36 participants are reported from 20 and 18 participants in the first and second interlaboratory study, respectively, with five individuals participating in both studies. However, the experiments are designed to ensure that they receive different kits on each exercise to prevent bias. In the event of multiple participants from the same laboratory, instructions are provided to work independently and report individual answers.

In the first interlaboratory study, participants are provided training material containing examples of the feature descriptions and an explanation of the ESS method. An optional virtual question-

and-answer session is offered for any participant who requests additional assistance in understanding the methodology. In the sample kit, participants receive the seven pairs of tape in separate envelopes with labels. In addition, they receive physical and electronic versions of the study instructions and the reporting template. The electronic version has a page for each comparison, with cells for reporting the alignment of each bin, and an automated calculation for the ESS. The final page of the template contains a survey, including information on how long it took the participants to complete the kit, information on accreditation and proficiency testing in physical fits, and questions regarding the utility of the ESS method. An example of the instructions and reporting template is included in **Appendix B**.

Following the completion of the first interlaboratory study and a meeting with participants to discuss the outcome of the exercise, a second interlaboratory study is developed to address some of the feedback provided in the first study. Among the most important modifications derived from these discussions are 1) the creation of more detailed documentation of the bin-by-bin observations of the edge features that lead to a decision, 2) inclusion of a metric for an inconclusive decision on a bin,, 3) addition of documentation of the comparison process before to the microscopic bin-by-bin assessment, and 4) more comprehensive instructions and familiarization with the proposed method and reporting tools.

In the second study, two training videos are prepared to describe the method and procedure for the interlaboratory study and guidance for reporting qualitative feature descriptors, the quantitative bin-by-bin scores, overall ESS, and the interpretation criteria. The training consists of two videos, each approximately 20-25 minutes. The training videos are included in **Appendix C**.

The first video describes the foundation of the comparison method, including results collected from initial tests of the method on a sample set of around 3000 comparisons.[8,9] In addition, this training video describes some of the feedback and lessons learned from the first interlaboratory study. The second video details the procedure for the exercise. It highlights the contents of the sample kit, explains the template and characterized features, and gives an example of how a tape edge can be compared and documented. This video is used in place of the Microsoft PowerPoint slides included in the first interlaboratory study, as it is hypothesized that a video with demonstrable steps and a live question-and-answer session will increase the understanding of the method and comparison features. All participants in the second interlaboratory study attend a training session where the two videos are shown followed by one-on-one discussion. The live sessions are offered for each participant at their own convenience and are available on-demand for reviewing the videos at any time after training.

In addition, the reporting template is expanded to include a more thorough documentation of the physical fit examination process. The method is divided into three main steps. The first step is to observe and document the overall edge alignment and description of edge pattern morphology, the second step is to examine and document alignment and feature occurrence at a macroscopic level, and the final step is to examine and document bin by bin the features and edge alignment at a microscopic level, estimate an ESS value, and report a conclusion. In the second and third steps, eight comparison features are documented using pre-defined options, and a score of 1 (fit), 0.5 (inconclusive), or 0 (non-fit) is decided for each bin. An automated calculation then displays the ESS once the participant fills out the bin comparisons. Supported by previous research, the participants are instructed that low scores (under 20%) provided the most support for a non-fit conclusion, while high scores (80% or higher) support a fit conclusion. The updated instructions and template used in the second interlaboratory exercise is included in **Appendix B**.

Other than these adjustments, the design of the interlaboratory study is consistent with the first – three sample kits containing seven pairs of tape that prior to distribution had been documented to share similar ESS and edge morphology characteristics. The same kit specimens are used for both interlaboratory exercises to determine if the modifications made to the method improve the participants' performance and decrease inter-participant variability. To minimize potential bias, the kits and samples are renamed, the order of the comparison pairs presented to a participant is randomly reorganized, and in cases when an individual or laboratory participated in the first study, a different kit is assigned in the second exercise. Performance rates and statistical analysis are performed in Microsoft Excel (Version 19.08), JMP Pro 16 (v.2021, SAS Institute Inc., NC), and open access R (version 4.2.2, R studio version 2022.07.2+576).

3.4. Results

3.4.1. Selection of samples and the consensus values prior to experiment

The original pool of physical tape comparison pairs includes a variety of edge patterns typically observed in hand-torn separations. To select the optimal pairs to be used in the interlaboratory, the sample pairs are further divided into seven groups of three similar pairs each, to prepare three distribution kits. The grouping criteria is defined for each kit to contain seven pairs, three of which have scores representing a high-confidence fit (F+ ranging from 86% to 99%), one has ESS representing a more challenging fit (F-, ESS ranging from 70 to 77%), and three have ESS representing high-confidence non-fits (NF+, ESS ranging from 0 to 11%). Also, each of the three inter-kit pairs contains similar edge morphology while representing various edge patterns across the seven pairs. Comparison pairs not meeting this criterion are excluded from the sample pool set as a potential inter-laboratory kit item.

After this preliminary selection process, a panel of four analysts independently examine the samples while maintaining no knowledge of the ground truth or anticipated scores. For this study, a consensus of ESS between panel members is defined as all answers being within a 10 ESS range. The annotations made by the panel members on a reporting template during their examinations are used to evaluate agreement in the decision process. The reporting templates allow for a more transparent discussion and review process when comparing panel members' results to assess which samples resulted in the highest consensus in their ESS results. Finally, the samples are assigned to the kits with the goal to keep the edge morphology and consensus ESS as close as possible and minimize the variability between kits. **Table 3-1** presents the samples assigned to each kit.

3.4.2. Performance rates of the first interlaboratory study

The performance of the ESS method has been assessed based on the ground truth or consensus ESS compared to 1) the final overall conclusion the participants report for a given pair, and 2) the ESS reported by the participants [8, 9]. For the ESS criteria, an ESS value of 60% or higher is considered support for a fit, with scores higher than 80% demonstrating higher confidence in a fit. ESS values of 40% or lower support a non-fit opinion, with pairs with scores below 20% demonstrating the most support in a non-fit result. Inconclusive results occur when the ESS is between 40 to 60%, as in those instances there are substantial similarities and differences between the samples to prevent a decision of fit or non-fit. In this study, we refer to these category ranges as the ESS thresholds.

Performance rates are calculated to evaluate the accuracy, sensitivity (true positive rate; TPR), specificity (true negative rate; TNR), false positive rate (FPR) and false negative rate (FNR) for each kit [8]. The performance of the method using the participant-reported conclusions are listed in **Table 3-2**. Initial analysis of the results demonstrated high overall accuracy for the method, ranging between 95 and 100%, with only a few samples reported as inconclusive, and none as false positive or false negative. Kits 1 and 2 include pairs reported as inconclusive by the participant for true fit samples (3 total pairs, 1 from Kit 1 and 2 from Kit 2).

Table 3-2. Performance of the first interlaboratory study using the participant-reported conclusion of fit, non-fit, or inconclusive. Inconclusive conclusions are not included in sensitivity and specificity rates but are included in the overall accuracy calculation. (n=number of participants).

Performance rates (%)	Kit 1 (n=7)	Kit 2 (n=6)	Kit 3 (n=7)	Overall (n=20)
Sensitivity	96	92	100	95
Specificity	100	100	100	100
False Positive Rate	0	0	0	0
False Negative Rate	0	0	0	0
Accuracy	98	95	100	97

However, in this first study, some participants report conclusions that do not always correspond with their calculated ESS. For example, one true fit tape pair is reported as a fit with high confidence by a participant who gave an ESS of 38, while another participant reported a fit with low confidence for a pair while giving an ESS of 19. When evaluating the rationale for the discrepancy, several participants still used their own criteria in decision-making as they did not feel comfortable interpreting the ESS score. Correspondingly, the performance rates based on the participant's reported conclusion vary from those estimated from the reported ES score. **Table 3-3** reports the performance based on the ESS threshold. Utilizing the ESS approach, accuracy rates range between 90 and 100%, and the false negative rate of Kit 1 increases to 18%. In Kit 1, five pairs with ESS scores reported below 60% are considered false negatives despite being reported by the participant as a fit.

Table 3-3. Performance of the first interlaboratory study based on the participant-reported ESS and using ESS thresholds to support a fit or non-fit.

Performance rates (%)	Kit 1 (n=7)	Kit 2 (n=6)	Kit 3 (n=7)	Overall (n=20)
Sensitivity	82	92	100	94
Specificity	100	100	100	100
False Positive Rate	0	0	0	0
False Negative Rate	18	8	0	6
Accuracy	90	100	100	95

After evaluating the data and participants' feedback, the discrepancy between some of the participants' reported ESS and their conclusion is considered to have originated from a misinterpretation of the use of the ESS, and a lack of formal training in the method. Interestingly, all the instances of discrepancy in score and conclusion originate with participants who did not participate in training, and so the second study includes some improvements to minimize this source of variance; a required training session for all participants and more clear instructions on how to report conclusions based on the ESS and not on the current criteria used at their agencies.

3.4.3. Performance rates of the second interlaboratory study

Two main adjustments are implemented in the second interlaboratory exercise based on the participants feedback and the data evaluation from the first study: 1) provide a more thorough one-on-one training and requiring participants to complete it before the study, 2) update the methodology and reporting template to simulate typical casework workflow.

Table 3-4 lists the main improvements that were implemented based on the first interlaboratory study. One of the main aspects that greatly facilitated the standardization and consistency of analysts' conclusions is splitting the examination into three sequential steps, including one step for the general assessment of the comparison pair and two steps that require a more detailed examination and documentation to identify major features in relatively large regions of the tape (step 2) and in each bin (step 3).

One of the aims of this upgraded process is to provide practitioners with standardized criteria and features for documenting their observations and thought processes. To that end, the reporting of the presence, absence, and effect in the decision of a fit or non-fit is documented on each comparison region for eight relevant features. The terminology and criteria for each of these features are demonstrated in the instructions and training of the second exercise. The eight features that have been shown to influence a duct tape physical fit examination are: 1) Alignment of severed dimples on tape backing, 2) Calendaring striations across edge, 3) Macro alignment of warp scrim, 4) Correspondence of protruding warp yarns and the respective pattern gaps in the other edge, 5) Continuation of scrim weave pattern, 6) Distortion explained by stretching directionality, 7) Weft scrim at or near edge consistent with the overall weft pattern, and 8) Missing material [9].

Another improvement implemented from the participant's feedback is the addition of an inconclusive option on each comparison bin. As a result, the updated template contains an auto-populated cells that are coded with a value of 0 for a non-fit, 0.5 for inconclusive, and 1 for a fit. Moreover, we decided to boost the visual assessment of the fit examination by including a macro in the excel template that automatically colors each bin cell with green for a fit, red for a non-fit, and yellow for an inconclusive. Observers can then immediately identify the areas within the tape edges that have corresponding features. This feature also proved helpful during the comparison of the reported results across participants. Moving forward, this green-red-yellow visualization tool has the potential to facilitate peer-review and presentation of results in the courtroom. After the second interlaboratory test, another improvement that will be incorporated in the template is to conduct the examination of the questioned sample in step one, independently, and before observing and documenting the known sample, to further minimize potential bias. ISO/IEC 17025 and studies on contextual bias recommends that where possible, questioned samples should be examined before known samples to help reduce the influence of contextual bias, such as looking for features observed in the known in the questioned sample instead of focusing on finding differences [13-17]. In the context of physical fit examinations, because samples must be compared side-by-side, this aspect is incorporated into step one, where each sample can be examined independently before comparison begins.

Since physical fits depend on the analysts' judgment, the risk of subjectivity and bias is inherent [14-16]. While some risk factors can be controlled by laboratory procedures (such as managing case information through linear sequential unmasking) [14], the proposed approach addresses many areas of potential bias specific to physical fit comparisons. For example, creating standardized methods and defined criteria can help ensure that practitioners consistently observe and interpret material-relevant features. Since the analyst documents specific criteria to assign a score to each bin independently, the decision is less likely to be influenced by previous bin observations. In addition, several laboratories have previously reported that they only perform verification on samples for which a fit is identified, leading to a potential bias that verifying practitioners are more likely to agree with the original decision [17]. The method evaluated here provides a path to facilitate blind verification by independent examiners that transparently document the criteria and judgment to arrive at a fit or non-fit conclusion. For instance, the initial examiner and the verifier can systematically document their thought process and provide a tangible comparison of the rationale of their conclusions through the standardized reporting template and the respective qualitative features and quantitative scores. Implementing an independent examination and documentation system allows transparent documentation of agreement and discrepancy rates between examiners and can facilitate a mechanism for introducing verifications of various conclusion types and complexities (i.e., establish policies to verify fits, no-fits, and inconclusive conclusions).

The findings show that the overall performance improves in the second interlaboratory with the adjustments made. Improved training and documentation in the templates have helped to provide more consistency between participants. In the new template, the additional standardized terminology and steps help to add transparency and make an independent review process more straightforward. In the second study, there are no discrepancies between the ESS and conclusions reported by participants. Only 1 comparison is a false negative, and there are no false positives. As such, the accuracy ranges between 98% to 100%, both for the ESS threshold approach and the conclusion reported by the participant (**Table 3-5**).

Table 3-4. Comparison of the evolution of improvements from the method and reporting templates used in the first and second interlaboratory study.

Feature	Interlaboratory #1 Reporting Template	Interlaboratory #2 Reporting Template
Bin by bin comparisons	Yes	Yes
Bin by bin score codes	Yes, 1(fit), 0 (non-fit)	Yes, added inconclusive option: 1(fit), 0 (non-fit), 0.5 (inconclusive)
Auto-populated estimation of ESS metric	Yes	Yes
Color-coded visualization tools	No	Yes
Overall comparison of edge qualifiers, drop-down	Yes, including the following terminology: M+ = Match with high confidence M- = Match with low confidence INC = Inconclusive NM- = Non-match with low confidence NM+ = Non-match with high confidence	Yes, adjusted to new terminology: F+: High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features) F-: High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed macroscopic features) IN: Inconclusive (I believe there are similarities and differences throughout the edges that do not let me support a decision of either fit or non-fit) NF-: Low confidence in Non-fit (I believe that generally the edges do not fit, but there are some areas of alignment or distortion that could be obscuring potential alignment) NF+: High confidence in Non-Fit (I am confident that the sample edges are not a physical fit based on the observed macroscopic features)

Overall conclusion drop-down	Yes,
	M = Match
	INC = Inconclusive
	NM= Non-match
		Yes, adjusted terminology
		F = Fit
		INC = Inconclusive
		NF= Non-fit

Multistep examination and documentation	No, single step. Subunit assessment of tape edges (Edge Similarity Score)	Yes, three steps 1) Overall alignment of tape edges 2) Macroscopic assessment of tape edges 3) Subunit assessment of tape edges (Edge Similarity Score)
Documentation of features relevant for the decision process	No, although the 8 features were explained in the instructions, they were not explicitly documented in the template	Yes, documentation of 8 features included in steps 2 (macro-assessment) and step 3 (bin-by-bin)

Table 3-5. Performance of the second interlaboratory study using the participant-reported conclusions and ESS values.

Performance rates (%)	Kit 4 (n=7)	Kit 5 (n=6)	Kit 6 (n=5)	Overall (n=18)
Sensitivity	100	96	100	99
Specificity	100	100	100	100
False Positive Rate	0	0	0	0
False Negative Rate	0	4	0	1
Accuracy	100	98	100	99

A reported accuracy of 100% does not mean that there is no variation in reported scores and bin areas across the participants. Rather, the variation lies within minor differences in interpretation of the tape edges features, representing less than 20% ESS variation between participants and not affecting the final report of the fit or non-fit as per the ground truth. Some variation of ESS between analysts is not unexpected, based on level of experience in physical fit examination, the material of interest, the sample morphology, and distortion level. Probabilistic interpretation has shown that values within +/- 20 ESS can still provide similar support for a conclusion [8, 9]. These findings further demonstrate that the proposed protocol aids in the standardization of the physical fit examination process.

3.4.4. Comparison of results across both interlaboratory studies

The distribution of scores across the sample pairs and kits are evaluated to compare the results of both interlaboratory studies. The pairs in the kits consist of three true non-fit pairs (labeled as NF), and four true fit pairs (labeled as F+). One of the true fit pairs has been specifically included as a more challenging comparison, with an expected score that is lower than the other true fit pairs (labeled as F-). All reported ESS values across both interlaboratory exercises can be seen in **Appendix D.** As seen in **Figure 3-1**, the boxplots representing the score distributions for all kits across both studies generally fall in the expected range. The first study shows that there was a much wider range of scores for both the F- and F+ pairs, and even some instances of false negatives. In addition, there are a few ESS that fall into the range of 40-60%, which are considered inconclusive. Meanwhile, the distribution of scores in the second interlaboratory study are less variable overall, and the fewer outliers were still within the expected thresholds. Only one false negative, and no inconclusive results or false positives are observed across all 126 examinations in the second exercise.

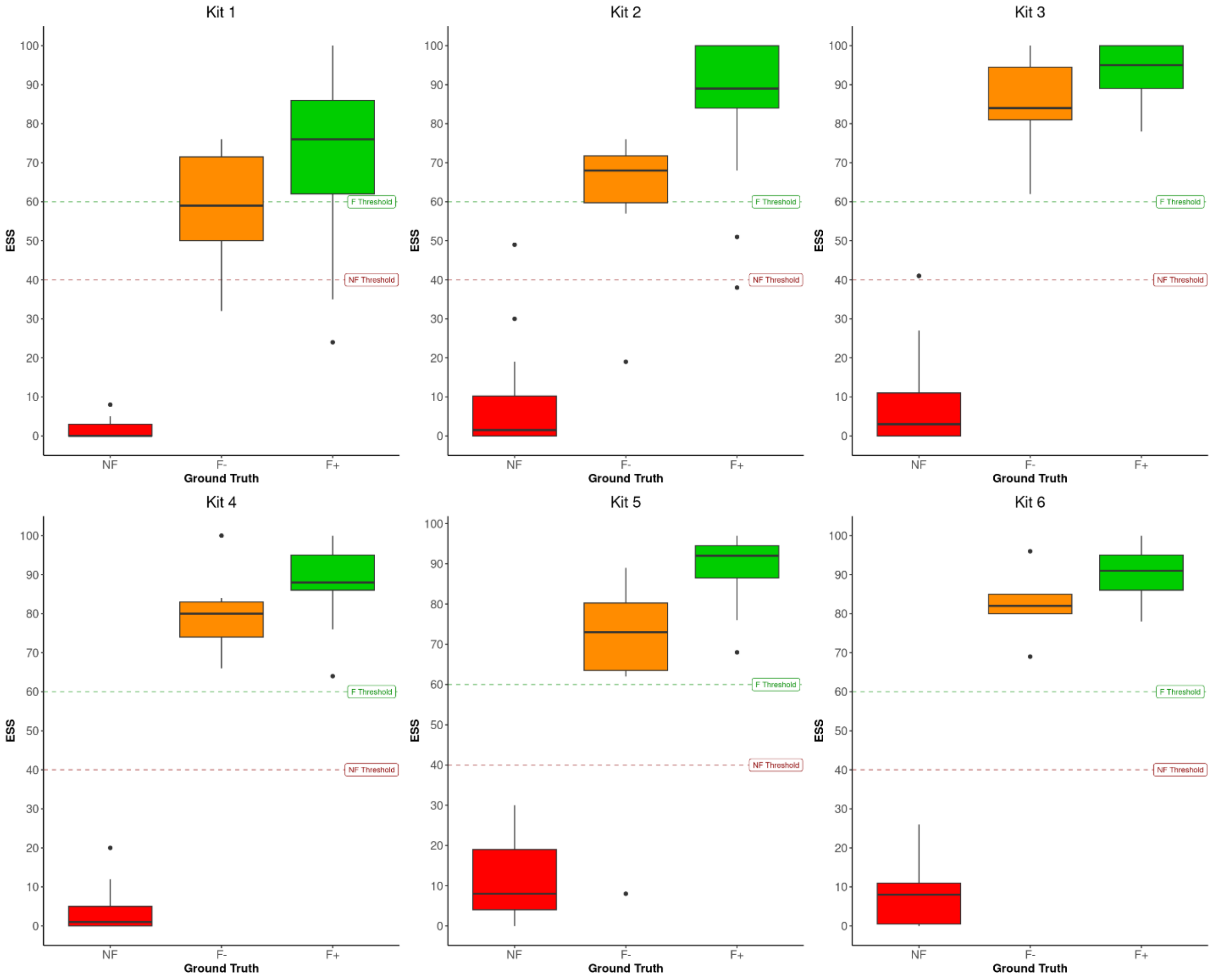


Figure 3-1. Boxplots showing distribution of scores for true non-fits (NF+), true fits (F+), and challenging true fit (F-) pairs for each of the sample kits. The thresholds for fits and non-fits have been set at 60% and 40% respectively. The variability of scores for the second interlaboratory study (ILS 2 – Kits 4 through 6) shows narrower quantiles and whiskers in comparison to ILS 1 (Kits 1 through 3).

In addition to the overall ESS distribution, the proximity of reported ESS for each participant is also evaluated according to statistical significance testing by analysis of variance with the Dunnett's test at a 95% confidence interval. Using this test, about 74.3% (104 of 140 comparisons) of the reported ESS in the first study show no significant differences from the respective pre-distribution consensus mean scores. The Dunnett's control chart in **Figure 3-2** shows the interval of reported scores for each ground-truth in comparison to the consensus mean. Participants who reported scores within the decision limits did not demonstrate significant differences against the consensus mean. Interestingly, most participants from the first interlaboratory exercise who report significantly different ESS scores from the consensus range opt out formal instructional training. Most of the variation among the eight participants that significantly differ from the consensus ESS come from the participants that did not receive the in-person training. Four of those eight participants that have deviations from the consensus report results that do not correspond to a change in the expected conclusion. For example, reporting an ESS of 40 for a non-fit would be a bit higher than expected for the non-fit, but it does not preclude the sample from being rightfully considered a non-fit. As expected, the F- pairs shows more variation in ESS scores according to Dunnett's due to the increase in the complexity of the comparison (**Figure 3-2**).

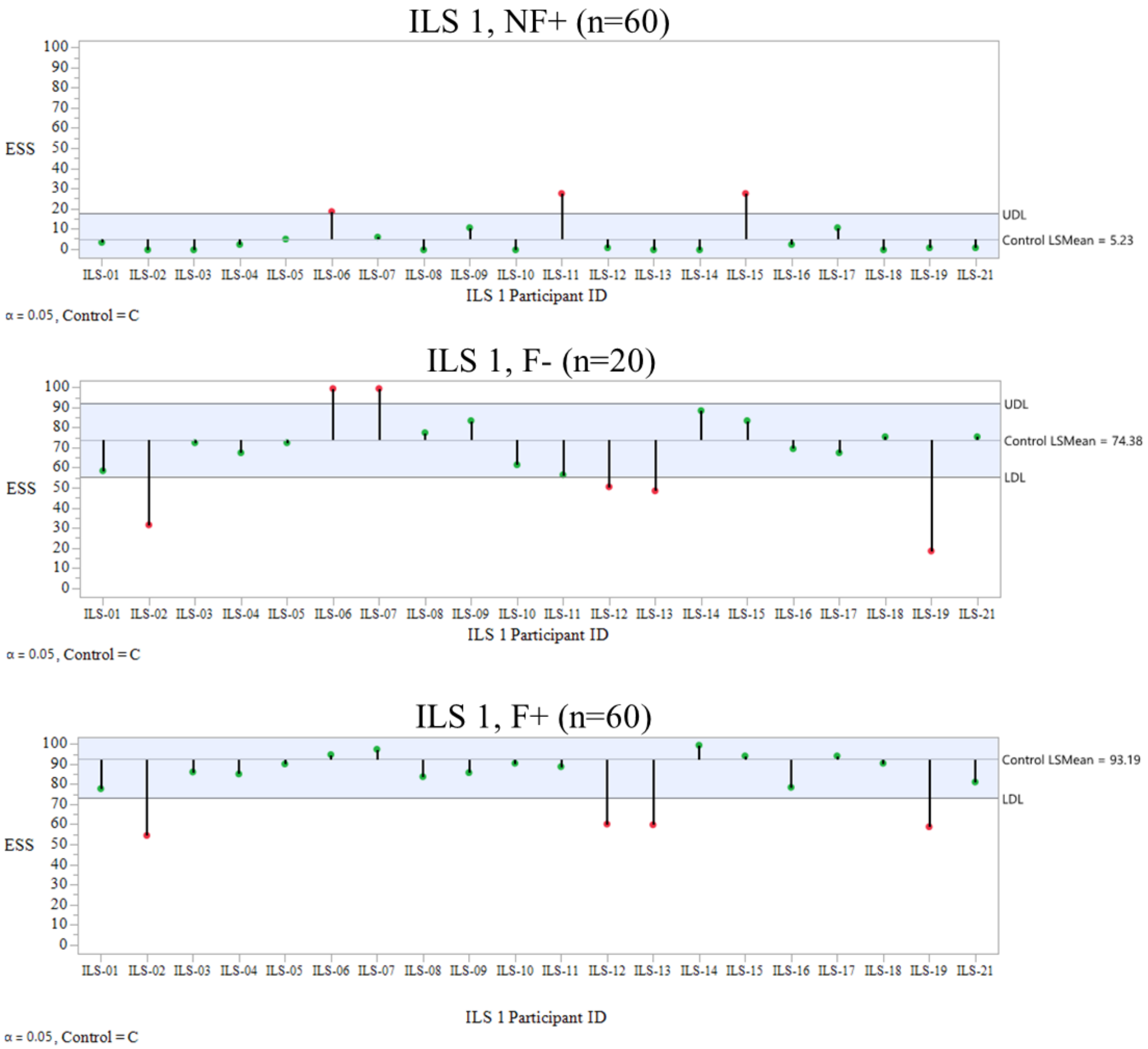


Figure 3-2. Dunnett's control plot for the three ground truth difficulty sets (F+, F-, NF+) for the 20 participants of the first study (ILS 1). The participant IDs are independent of the IDs used in ILS 2. The plots for the NF+ and F+ account for all samples with those respective categories (3 samples each). While most participants fell within the decision limits, several participants reported scores outside the expected range.

There are ten instances where the participants of ILS 1 report an overall conclusion that is inconsistent with the reported ESS. Several of those are where participants report a fit with high confidence for a low ESS, but in one instance, a participant reports a non-fit for a pair that has a high ESS. This trend is likely a result of the participants following their own criteria for making decisions instead of using the ESS to support their overall conclusion. For example, there is one instance of a pair reported as a non-fit while having an ESS higher than 80. In that comparison, the participant notes “Overall there are no individualizing characteristics, therefore would not be considered a fracture match at our laboratory. The pattern is good with minor distortions but there are no specific areas that indicate that this fracture match is specific to this roll and this tear and that it couldn't be reproduced.” Finally, other participants note in the comments that the edges show distortion, such as one participant who notes “Overall shape of fractured edges appear to correspond (scrim areas 25-37) but because those areas were wrinkled and/or folded over on the acetate, it was difficult to accurately assess the features and make a conclusion. Those areas were graded as "0 non-match".” As such, in ILS 1, there is indication that some participants revert to criteria they use in their own laboratories’ method instead of concluding based on the ESS. To that end, the more thorough documentation and training implemented in ILS 2 addresses this deficit.

In comparison, 81.7% (103 of 126) of the reported ESS in the second study show no significant differences from the pre-distribution consensus means. The Dunnett’s control chart in **Figure 3-3** shows the interval of reported scores for each ground-truth for ILS 2. Once again, the F- pairs shows more variation in ESS scores, but the extent of the variation is lower than in the first study. In addition, the F+ pairs demonstrate only two participants who are at the cusp of the lower the decision limit of the Dunnett control mean, and fewer outside the limits for both NF+ and the F- pairs. Converse to the results seen in the first study, all ESS assignments in the second ILS were paired with a consistent conclusion of fit or non-fit. This improved over the trends seen in the first study, where it is more common to see conclusions that do not correspond with the reported ESS. In other words, except for the false negative result, all instances outside of the decision limits in ILS 2 were close enough to not create a change in the decision qualifier (i.e., they correctly remain as F- (ESS >80), F- (ESS 60-80), or NF+ (ESS <20), based on the expected ESS threshold). Even for the false negative, the participant reported the overall conclusion as a non-fit.

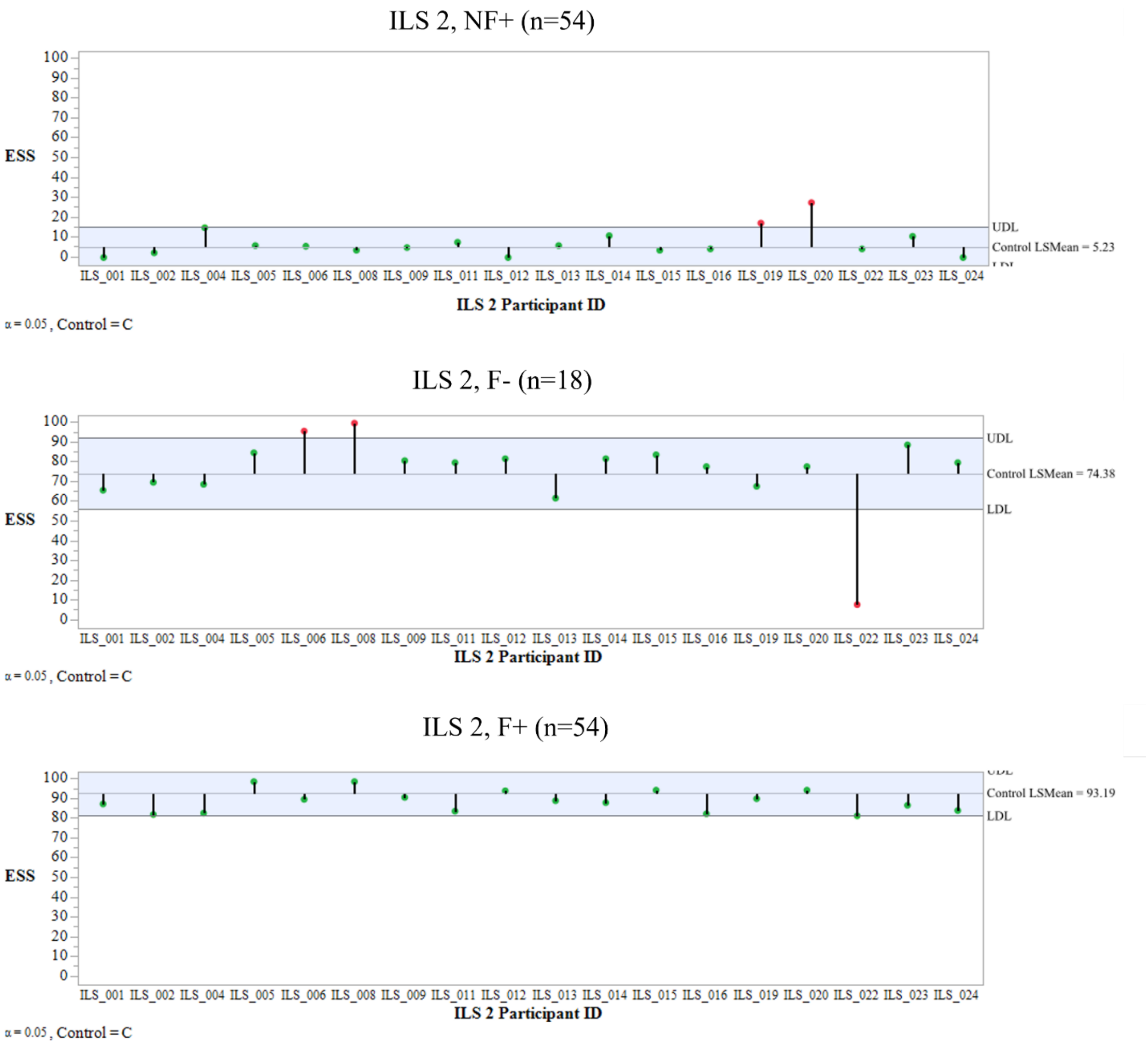


Figure 3-3. Dunnett's control plot for the three ground truth difficulties for the second interlaboratory study (ILS 2). The participant IDs are independent of the IDs used in ILS 1. As expected, there is more variance in the F- pair, however, the variation is not as extensive as in ILS 1. In addition, fewer participants reported scores outside the expected range for all three categories.

Dunnett's test is helpful when using a reference value as the control group; in this case, we use the consensus ESS established by the expert panel in lieu of a "ground truth." However, the consensus ESS is estimated from only four analysts that are already very familiar with the ESS method. A complementary metric to evaluate the inter-examiner agreement is using z-scores that consider the deviation from a participant from the study mean. Therefore, to further evaluate performance using reported ESS, z-scores are calculated for each comparison across the practitioners. Z-scores are a common method used for interlaboratory exercises, which describe the standardized deviation of an individual result from the mean value of the study. The study mean ESS and standard deviations are calculated using all results from each respective ILS per sample, regardless of kit. Outliers are evaluated but not removed from the Z-score calculations, as analysis shows that while there are a few (3 or less) outliers in each ILS across all comparisons, removing the outliers from the study mean did not contribute to a substantial improvement in performance. Most importantly, maintaining all data points allows the identification of laboratories and samples that were outside expected study limits.

In this context, an individual comparison is considered satisfactory if the absolute value of its corresponding z score is less than 2. The comparison is considered cautionary if the absolute value of the z score lies between 2 and 3 and insufficient if the absolute value of the z score is greater than 3 [11]. **Figures 3-4 and 3-5** show the calculated z-scores for each sample in both interlaboratory exercises. Negative z-values indicate that the reported ESS was less than the study mean, while positive z-values indicate the reported ESS was higher than the mean. The z-values have been color-coded for visualization of which comparisons were considered satisfactory or too far outside the z-score interval. Values marked in yellow or red represent those that are cautionary (yellow) or insufficient (red).

In the first interlaboratory exercise, 130 of 140 comparisons (93%) are considered satisfactory, while 8 (6%) are considered cautionary and 2 (1%) are insufficient. In the second exercise, out of the 126 comparisons, only 4 are considered cautionary (3%), 1 is insufficient (0.8%), and the rest are satisfactory (96%). The individual comparisons identified as outliers across both studies were marked either as warnings or unsatisfactory using this criterion. The few red and yellow flags for unsatisfactory or cautionary results are observed in non-fits as well as fits of different complexity (F- and F+), indicating that there are no trends for a given comparison type. In the first interlaboratory exercise, there is at least one participant who participated in the additional informational training who still has responses marked as cautionary, meaning that the difference in training between participants does not fully exclude the possibility of variances between participants. The same is true considering the second interlaboratory study, where all participants trained in the same way, still results in cautionary results as well, although there are fewer than ILS 1. All calculated Z-scores are available in **Appendix D**.

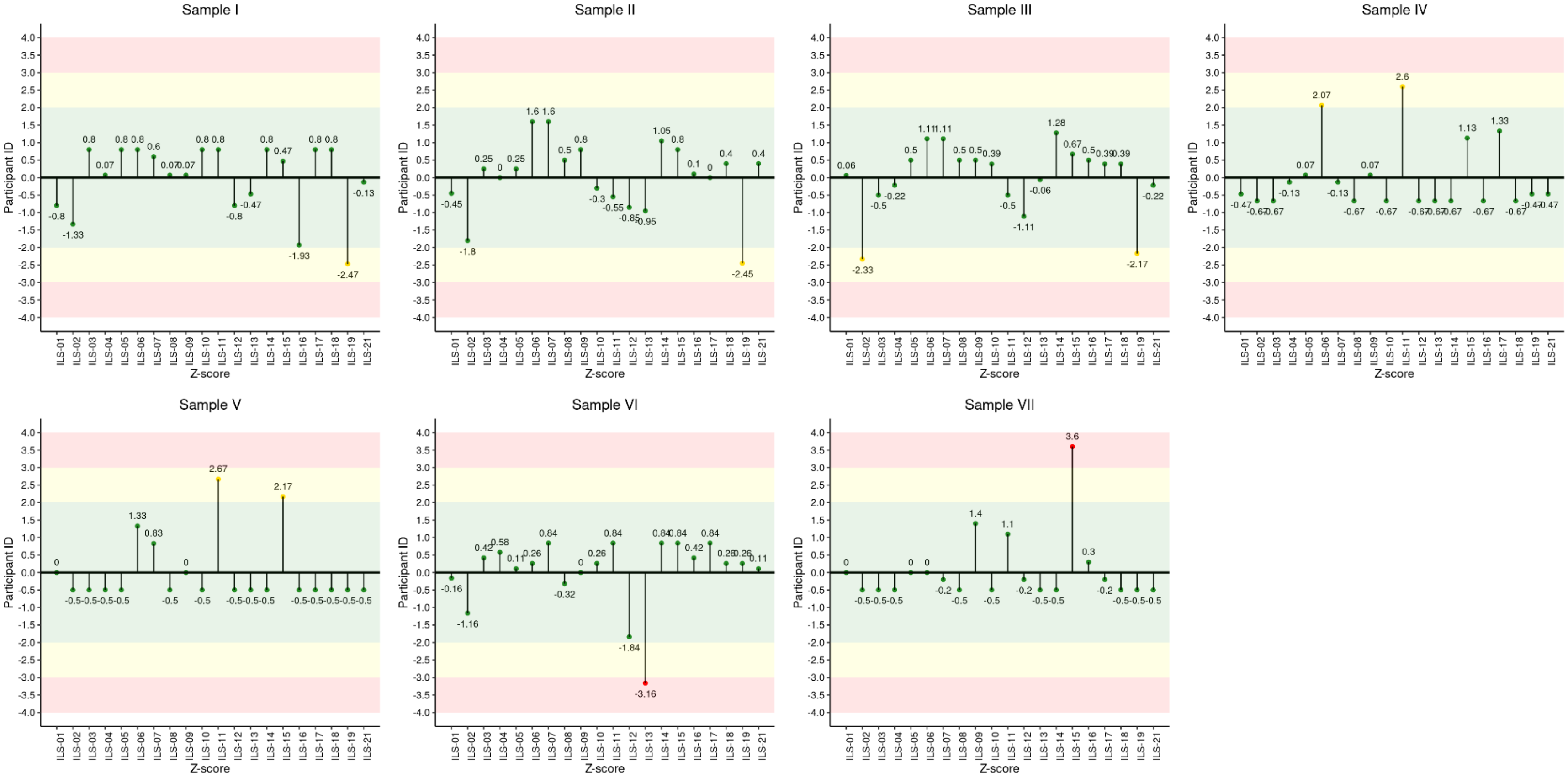


Figure 3-4. Z-scores of the reported ESS values for ILS 1 for each participant. The participant IDs are independent of the IDs used in ILS 2. The z-values have been color-coded for visualization. Green bars are considered satisfactory, yellow bars are considered cautionary, and red bars are considered insufficient (too far outside the confidence interval). Ground truth of the samples is as follows: Sample I (F+), Sample II (F-), Sample III (F+), Sample IV(NF+), Sample V (NF+), Sample VI (F+), and Sample VII (NF+).

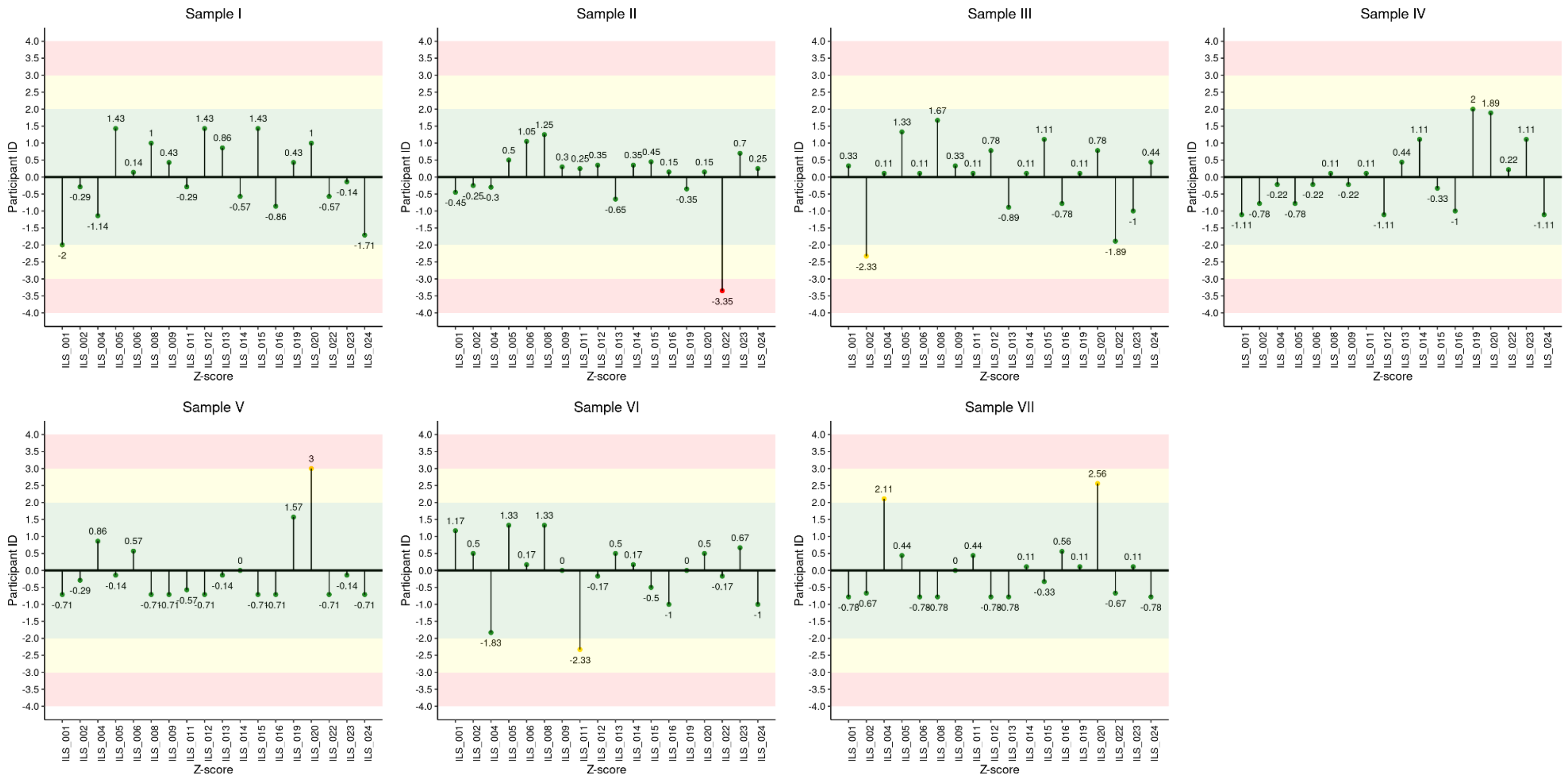


Figure 3-5. Z-scores of the reported ESS values for ILS 2 for each participant. The participant IDs are independent of the IDs used in ILS 1. The z-values have been color-coded for visualization. Green bars are considered satisfactory, yellow bars are considered cautionary, and red bars are considered insufficient (too far outside the confidence interval). Ground truth of the samples is as follows: Sample I (F+), Sample II (F-), Sample III (F+), Sample IV(NF+), Sample V (NF+), Sample VI (F+), and Sample VII (NF+).

As explained before, the improvement in performance and reduced variation of ES scores from the first to second exercise is attributed to improved training and understanding of the ESS method in the second study, as well as more consistent documentation, interpretation, and application of the ESS value to support the participants' opinion for a given pair. For the five repeat participants, good performance is consistent between four participants on the two exercises, while improvements are seen in one participant on the consistency of reported conclusions to ESS. This shows that the first interlaboratory feedback enhanced the introduction of the method to new participants and shows good progress toward future implementation procedures.

3.4.5. Survey Results

The final tab of the sample template incorporates a survey that evaluates each participant's experience level in physical fits, the occurrence of duct tape physical fits in their laboratory, and their overall opinion on the application of the duct tape ESS methodology.

Overall, the 36 survey responses across both studies show that all participants work at accredited forensic laboratories. Only two participants across both studies report not having completed proficiency testing in either duct tape or physical (fracture) matches. One participant reports having no casework experience with physical fit examinations, and four report having physical fit experience in other materials besides duct tape. The majority state having at least 1-10 cases involving physical fits of tapes, with six participants indicating they have had more than 20 cases. As such, this study benefits from participants that are familiar with physical fit examination prior to introduction to the ESS method, and therefore the participants provide substantial feedback.

In the first study, 94% of participants note that analysis of the sample kit took more than 90 minutes. In the second, the modified question asks how many hours it took the participants to complete the kits. The average during the second study was 15 hours to complete the 7 comparison pairs in the set, with at least four participants reporting 24 hours or longer for completion (**Figure 3-6**). While this may seem high, many participants also state that a typical physical fit examination takes them at least one to three days, and some report spending a week on a typical exam. This timeframe seems reasonable for the completion of seven pairs, as unlike standard physical fit examinations, the samples in this set were already prepared, and participants were not required to document the overall physical characteristics and image the samples, unlike standard practice in casework.

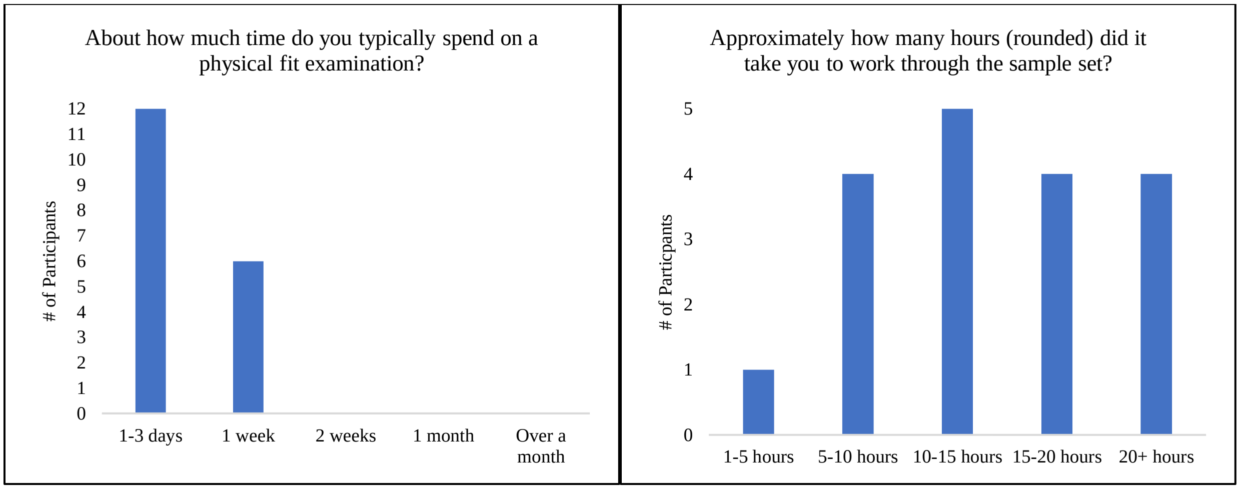
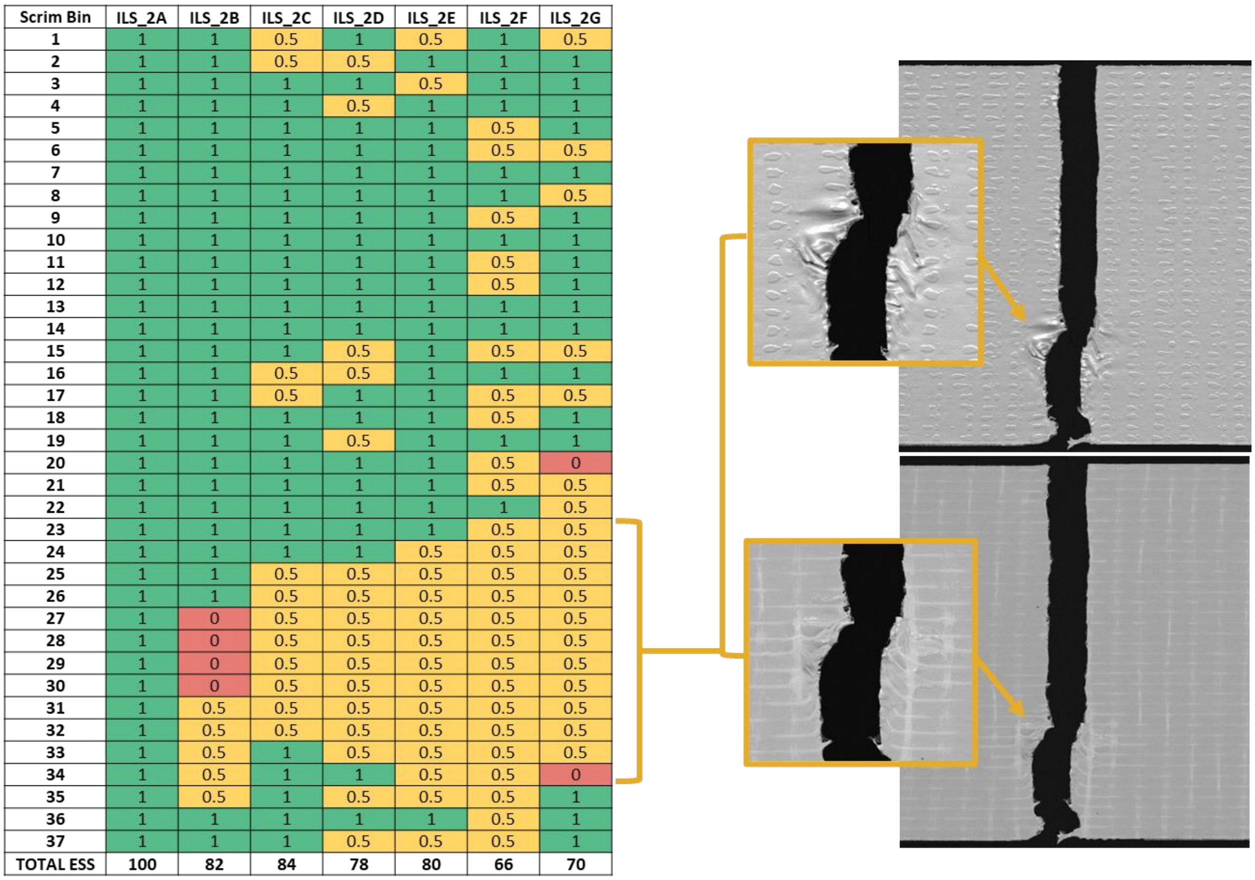


Figure 3-6. Distribution of responses to the second ILS survey questions. **Figure 3-6A** (left) shows that most respondents typically spend 1-3 days on a typical physical fit examination. **Figure 3-6B** (right) shows that most participants spent between 5-20 hours on the sample kit.

In the first study, approximately half of the participants indicate they did not find the approach helpful. Based on the feedback provided, this is due to several reasons. Those that do not find the method useful tended to report that it was due to 1) concerns about areas that had indicated both fit and non-fit, and how there was not an option to report inconclusive, 2) that they rely more on overall observation or macroscopic assessments when making decisions, 3) that they feel that the ESS could be misleading if it doesn't indicate a 100% fit or a 0% non-fit, and/or 4) laboratory specific policies that differ from what this method requests. Specific feedback is listed in **Appendix E**. For the first two areas of concern, we incorporate and address this feedback in the second exercise by adding the option to report bins as inconclusive (0.5), as well as introduce the three-step methodology that follows an overall, macroscopic, and then microscopic comparison process. As for the concern over the ESS being misleading, previous studies have shown that not all pairs have the same degree of similarity, even if they are true fits.[9] This is particularly true when the samples have been stretched or damaged. A critical goal of this method is to provide analysts with the means to transparently document the features and criteria they use to make decisions. Finally, regarding conflicting laboratory practices, one of the key motivations for the development of the ESS method is that there is currently no consensus-based method for physical fit comparisons of any material in forensic trace evidence. As a result, there is no consensus for what should be considered an individualizing feature, or when a tape pair should undergo more analytical evaluation. This is what the ESS method is proposed to help address with further research and feedback from current practitioners.

To help address these areas of feedback, the second study incorporates more thorough documentation, and more standardized training to address. The main improvements include incorporating the 3-step methodology, specific feature documentation at the macroscopic and microscopic level, the ability to report inconclusive bins, and ensuring all participants of the second study attended a virtual training with the authors. The training ensures that all participants have the method explained in a clear and transparent manner and can clarify any questions they have before they start comparisons. As a result, 11 out of 18 participants (61%) in the second exercise report that the ESS method was useful for documenting their thought process for a physical fit examination. Of the seven who disagreed, few state why they felt the method is not helpful. The main observation made is that while the macroscopic assessment is helpful, going all the way to do the microscopic ESS documentation becomes too rigid, and the overall assessment becomes lost in the detailed documentation process. Another participant notes that they feel the template has too much detail and causes the most critical features to be misplaced. Specific feedback from the second interlaboratory studies is available in the **Appendix E**. Despite these concerns, the results from the study indicate that the microscopic ESS assessment leads to inter-examiner agreement and more tangible support of the initial macro-assessment.

On the other hand, the value of the ESS method for training and technical review is seen favorably by most participants across both studies. One of the critical aspects of this method is to provide practitioners with a metric and visual representation of the similarity or differences between the compared samples. As such, two practitioners can independently complete their examination of a pair of samples, and quickly compare results to determine if there are discrepancies. An example application of inter-examiner comparisons is shown in **Figure 3-7** for one of the more challenging F- pairs, with several examples of participants' opinions on the same tape pair. Although there are slight variations in reported values in the scrim bins, the overall ESS value and the documentation of the areas of distortion are consistent. As represented by the color-coded bins in yellow (inconclusive) or red (non-fit), most of the participants report lower scores on bins 25 to 35, where tape distortions are present (see **Figure 3-7** insert images)



Scrim Bin	ILS_2A	ILS_2B	ILS_2C	ILS_2D	ILS_2E	ILS_2F	ILS_2G
1	1	1	0.5	1	0.5	1	0.5
2	1	1	0.5	0.5	1	1	1
3	1	1	1	1	0.5	1	1
4	1	1	1	0.5	1	1	1
5	1	1	1	1	1	0.5	1
6	1	1	1	1	1	0.5	0.5
7	1	1	1	1	1	1	1
8	1	1	1	1	1	1	0.5
9	1	1	1	1	1	0.5	1
10	1	1	1	1	1	1	1
11	1	1	1	1	1	0.5	1
12	1	1	1	1	1	0.5	1
13	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1
15	1	1	1	0.5	1	0.5	0.5
16	1	1	0.5	0.5	1	1	1
17	1	1	0.5	1	1	0.5	0.5
18	1	1	1	1	1	0.5	1
19	1	1	1	0.5	1	1	1
20	1	1	1	1	1	0.5	0
21	1	1	1	1	1	0.5	0.5
22	1	1	1	1	1	1	0.5
23	1	1	1	1	1	0.5	0.5
24	1	1	1	1	0.5	0.5	0.5
25	1	1	0.5	0.5	0.5	0.5	0.5
26	1	1	0.5	0.5	0.5	0.5	0.5
27	1	0	0.5	0.5	0.5	0.5	0.5
28	1	0	0.5	0.5	0.5	0.5	0.5
29	1	0	0.5	0.5	0.5	0.5	0.5
30	1	0	0.5	0.5	0.5	0.5	0.5
31	1	0.5	0.5	0.5	0.5	0.5	0.5
32	1	0.5	0.5	0.5	0.5	0.5	0.5
33	1	0.5	1	0.5	0.5	0.5	0.5
34	1	0.5	1	1	0.5	0.5	0
35	1	0.5	1	0.5	0.5	0.5	1
36	1	1	1	1	1	0.5	1
37	1	1	1	0.5	0.5	0.5	1
TOTAL ESS	100	82	84	78	80	66	70

Figure 3-7. Example of application of documentation template to a duct tape physical fit examination for the pair F- of Kit 1. The participants in this example have slightly different reported ESS, but the overall conclusion of fit is consistent, and most of the participants report the area of distortion consistently.

Moreover, when looking at their reporting templates, it becomes evident that the reasons for the lower confidence in the yellow areas are caused by feature distortion, such as the reported misalignment of warp fibers, correspondence of protruding warp fibers, and the respective gaps across the edges, and particularly stretching observed but explained by the distortion of the edge. This is particularly critical considering that all participants from the second study report that their laboratory completes verification on 100% of duct tape physical fits. According to a larger survey conducting regarding physical fits across forensic laboratories, most laboratories only complete verification on samples reported as fits, but others perform verification on all pairs [17]. In the same survey, no participant reported any discrepancy during verification of physical fit cases in the last five years (**Figure 3-8**). This is problematic as it can introduce bias by creating general expectations of agreeing with peers. All participants report that in a case of discrepancy, a discussion will be held between the original analyst and the verifying analyst to determine where the discrepancy lies. To facilitate blind verification, this method endorses systematic documentation by an independent examiner, and for the first time, provides a transparent and quantifiable way of comparing the results of two or more examiners in a way that thoroughly demonstrates the thought process and observations made by each examiner. These tools and evaluation mechanisms can motivate managers to implement verification processes that are not limited to "fit decisions" and can lead to monitoring performance trends.

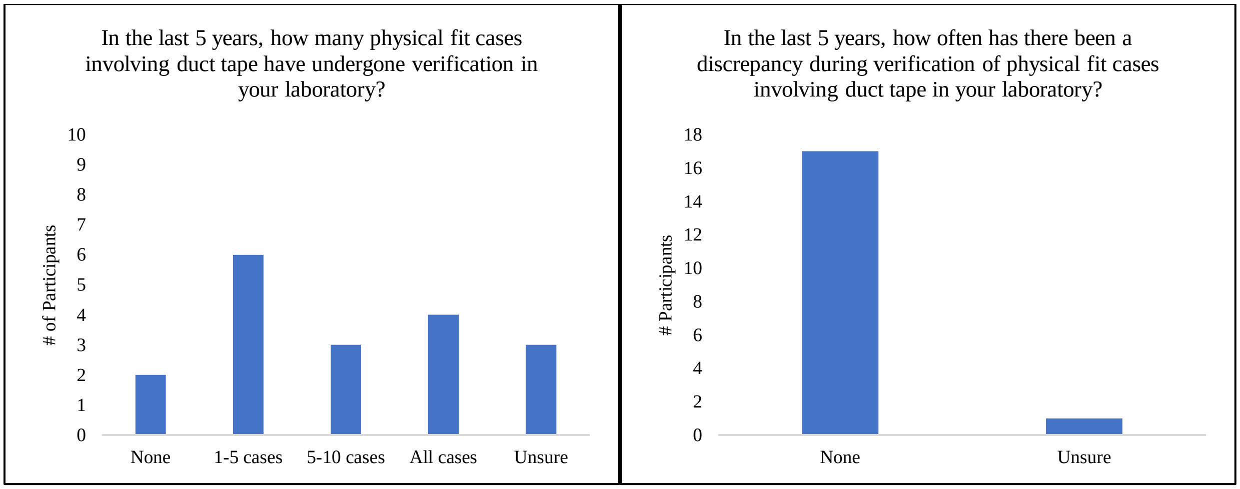


Figure 3-8. Distribution of responses to the second ILS survey questions. **Figure 3-8A** (left) shows that distribution of the number of duct tape physical fit cases reported to have undergone verification in the last five years. **Figure 3-8B** (right) shows that no participants reported any instances of a discrepancy during verification. One participant was not currently performing casework, and so was unsure.

The survey reveals some concerns expressed about the implementation of the ESS method which can be addressed as more training and interlaboratory collaborations are conducted. One of the most common comments across the two studies is that some participants feel that ESS has the potential to misinform juries and attorneys if it is not a 100% fit. However, the ESS method is not intended to be the only decision criteria for physical fit examinations. It is intended to support and inform the participant's opinion. Not every fit between samples in casework holds the same probative value, and since a 100% perfect fit is not always possible, it is important to maintain that transparency. Like other disciplines, the trier of fact does not necessarily expect a 100% certainty; what is important is to be able to express what a given quality of a fit means based on the scientific foundations and error rates reported in the literature for the assessment of the proposed methods. One possibility to alleviate those concerns is to provide more training and develop consensus-based standards about how to communicate the results to the jury.

3.5. Conclusions

This study compares results from two interlaboratory studies to evaluate the performance of the proposed method for the examination, documentation, and interpretation of duct tape physical fits. Thirty-eight (38) participants across 23 laboratories examined the sample kits, resulting in 266 total examinations. The overall accuracies are high (90-100%) and inter-participant agreement is generally high, with most reporting ESS scores falling within a 95% confidence interval of the mean consensus values or the inter-participant means (study means). The inter-participant agreement and accuracy improved from the first to the second study based on the practitioners' feedback, accounting for the improvements made to training and documentation of the examination process. Several participants in the initial study provided qualifiers that are unreasonable for the respective ESS ranges, which is addressed in the second study, leading to more consistent reports of ESS.

When looking at performance rates based on conclusions and ESS thresholds, discrepancies observed between some participant's conclusion and ESS were resolved in the second trial, displaying an improvement in accuracy from 95 to 100% across all three kits, for the first and second study, respectively.

When using z-score criteria, in the first study, only 10 out of 140 total comparisons across the 20 participants are outside the expected range according to z-score comparisons. Conversely, only 5 out of 126 total comparisons across the 18 participants of the second study are outside the expected range of consensus ESS means, and only one is considered unsatisfactory.

Overall, the two ILS demonstrate that the proposed ESS method provides support to participant conclusions, demonstrates scientific reliability with low error rates and high accuracies, and offers analysts systematic and transparent documentation criteria. This method provides a quantitative value representing the quality of a fit, and the interlaboratory studies have allowed for an estimation of inter-participant error rates. Moreover, the design of standard terminology and reporting of relevant features in a systematic template provides valuable means to harmonize practice. Participants generally agree that the method has the potential to facilitate systematic and transparent peer review and training. The feedback incorporated across both studies has led to improvements in documentation templates, training, and communication of the method that is anticipated to assist with future implementation into casework. Additional work in this area will continue to address the concerns and feedback provided by practitioners in physical fit examination, with the goal of continuing to improve and expand the application, communication, and interpretation of the ESS method on duct tapes and other materials.

In summary, the lessons learned in the studies serve as important benchmarks to provide criteria that assist with standardization and transparency of the examination and interpretation, and a mechanism to demonstrate the thought process during training, examination, technical review or verification of physical fits. These findings are anticipated to offer a path forward to the forensic examination of physical fits and facilitate incorporation into current guidelines. The proposed method aligns with ongoing standard guides being developed in the field for the examination of physical fits and is anticipated to easily adapt to current workflows.

3.6. References

1. OSAC. Oct 27, 2022. Lexicon web. Retrieved November 28, 2022, from https://lexicon.forensicosac.org/
2. Brooks E, Prusinowski M, Gross S, Trejos T. Forensic physical fits in the trace evidence discipline: A review. *Forensic Sci. Int.* 2020. 313: 110349.
3. President's Council of Advisors on Science and Technology. Report to the President - Forensic Science in Criminal Courts: Ensuring Scientific Validity. *Exec. Off. Pres. Pres. Counc. Advis. Sci. Technol.* 2016. 1: 1–160.
4. National Academy of Sciences. *Strengthening forensic science in the United States: A path forward*. *Strengthening Forensic Science in the United States: A Path Forward*. 2009. doi:10.17226/12589.
5. OSAC 2022-S-0015, Standard Guide for Forensic Physical Fit Examination. Trace Materials Subcommittee. November 2022.
6. McCabe KR, Tulleners FA, Braun JV, Currie G, Gorecho EN. A quantitative analysis of torn and cut duct tape physical end matching. *J. For. Sci*. 2013. 58: 34–42.
7. Bradley MJ, Keagy RL, Lowe PC, Rickenbach MP, Wright DM, LeBeau MA. A validation study for duct tape end matches. *J. For. Sci*. 2006. 51: 504–508.
8. Prusinowski, M., Brooks, E. & Trejos, T. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. *Forensic Sci. Int.* 2020. 307: 110103.
9. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – Part I. Physical fits of torn and cut duct tapes. For. Sci. Int. 2023. 343: 111567. DOI: https://doi.org/10.1016/j.forsciint.2023.111567
10. van Dijk CD, van Someren A, Visser R, Sjerps M. Evidential value of duct tape comparison using loopbreaking patterns. *Forensic Sci. Int.* 2022. 332: 111178.
11. ISO/IEC 17043:2010 Conformity assessment - General requirements for proficiency testing. 2010.
12. ASTM International. ASTM E691 - 19e1 Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method. 2019:1–26. doi:10.1080/00224065.1993.11979478
13. ISO/IEC 17025:2017. General requirements for the competence of testing and calibration laboratories. 2017.
14. Quigley-McBride A, Dror IE, Roy T, Garrett BL, Kukucka J. A Practical Tool for Information Management in Forensic Decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in Casework. *For. Sci. International: Synergy*. 2022. 4: 100216.
15. Dror IE. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Anal. Chem.* 2020. 92: 7998–8004.
16. Kassin SM, Dror IE, Kukucka J. The forensic confirmation bias: problems, perspectives, and proposed solutions. *J. Appl. Res. Mem. Cogn*. 2013. 2: 42-52. 10.1016/j.jarmac.2013.02.003
17. Gross S. NIST-OSAC Physical Fits Task Group. *Physical Fits Survey*. 2020.

Appendix 3-A

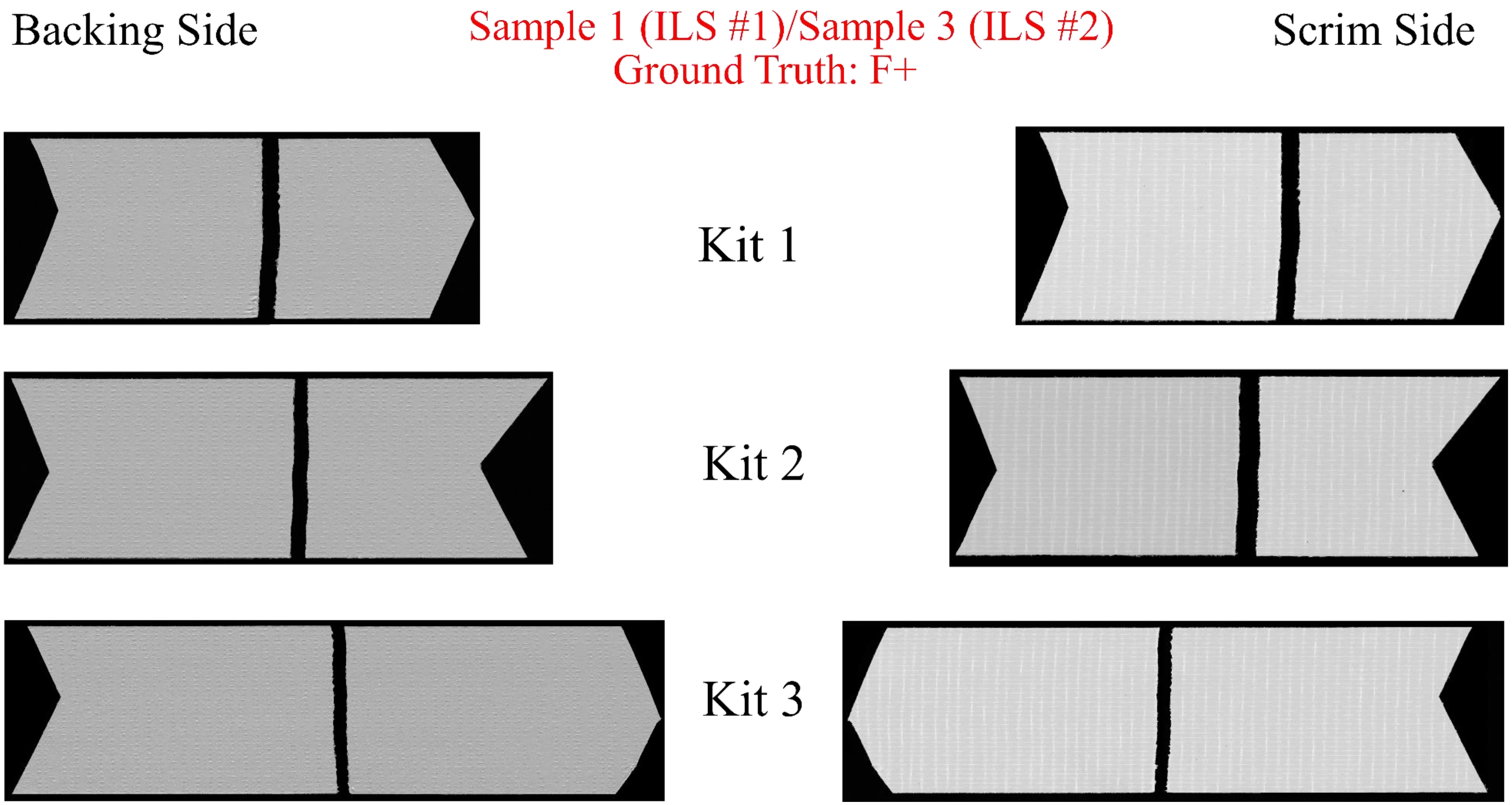


Figure 3A-1. Images of each Sample 1 pair from the three analysis kits used in ILS 1 and 2. Sample 1 in ILS 1 was labeled as Sample 3 in ILS 2. The ground truth of this sample pair was a true fit, intended to have high confidence.

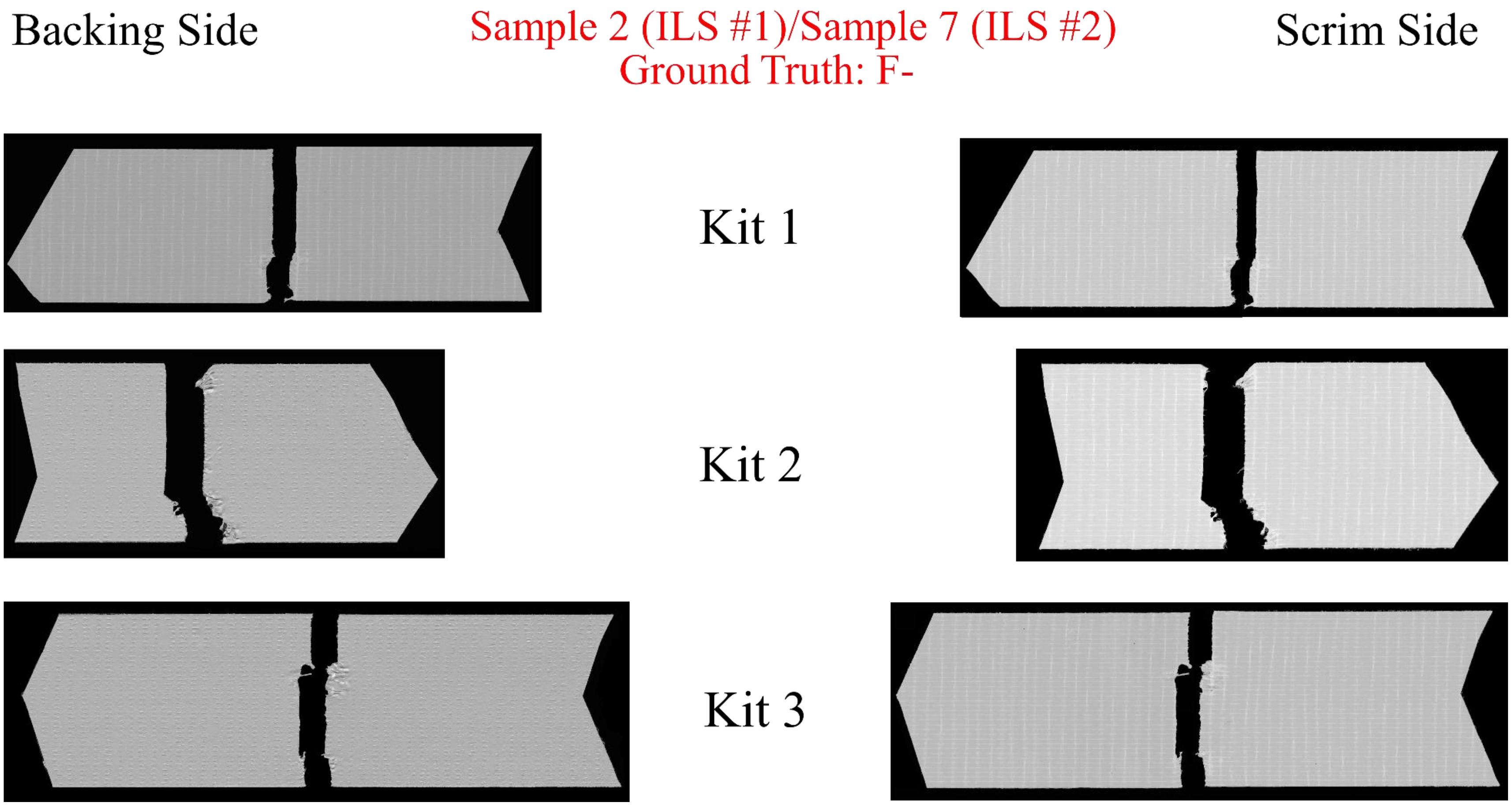


Figure 3A-2. Images of each Sample 2 pair from the three analysis kits used in ILS 1 and 2. Sample 2 in ILS 1 was labeled as Sample 7 in ILS 2. The ground truth of this sample pair was a true fit, intended to have low confidence, meaning it was expected to have lower reported ESS.

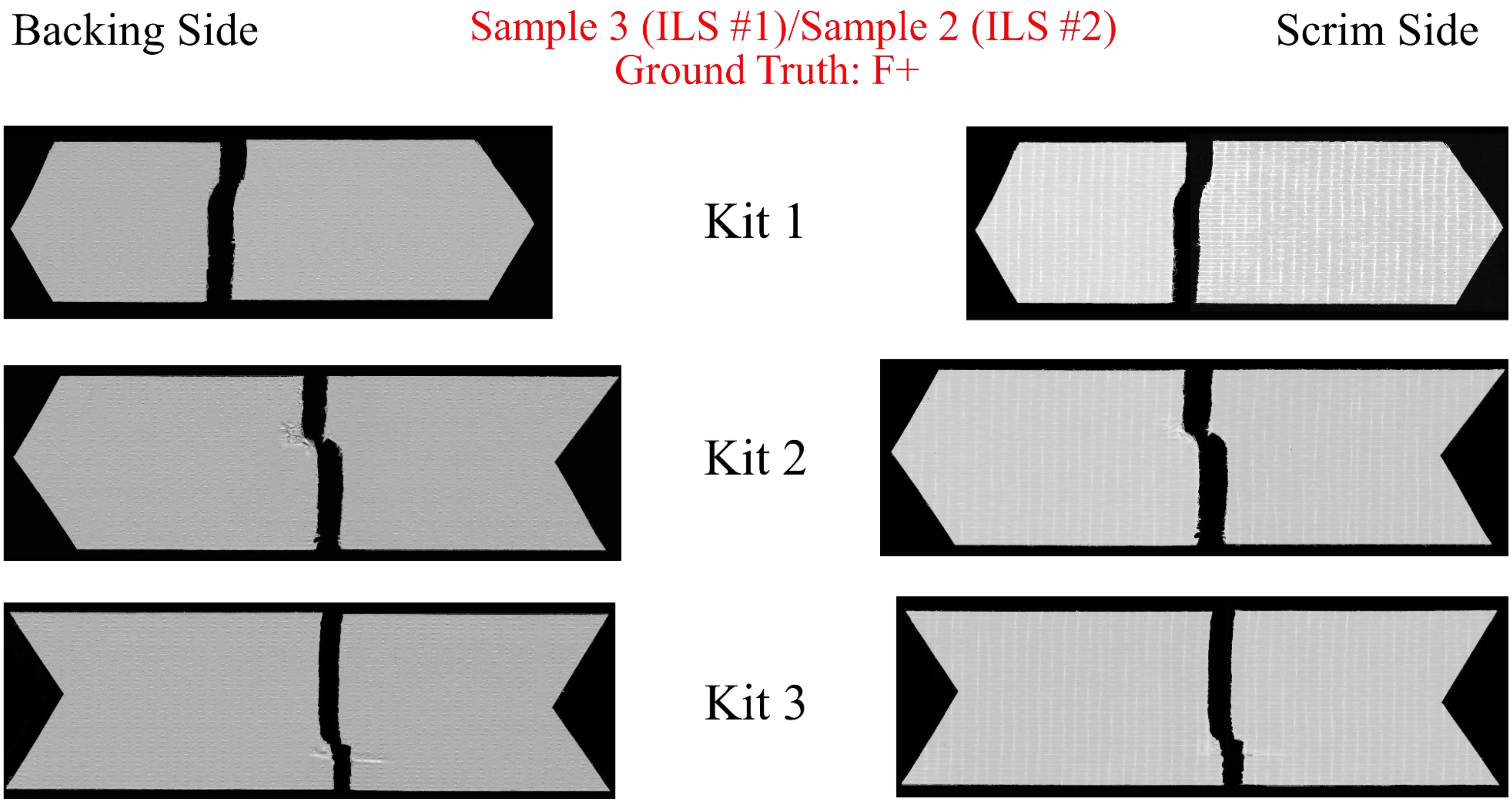


Figure 3A-3. Images of each Sample 3 pair from the three analysis kits used in ILS 1 and 2. Sample 3 in ILS 1 was labeled as Sample 2 in ILS 2. The ground truth of this sample pair was a true fit, intended to have high confidence.

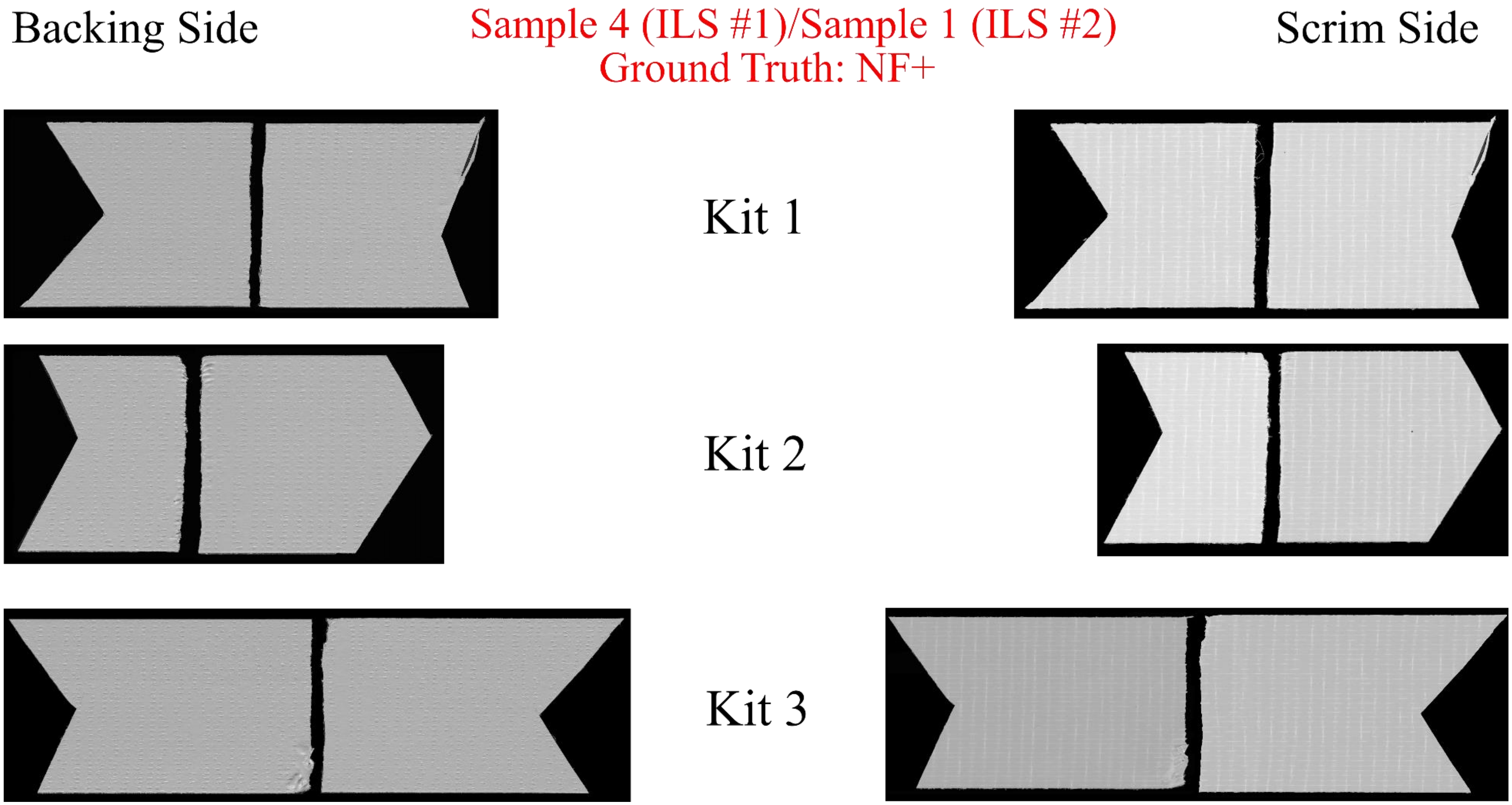


Figure 3A-4. Images of each Sample 4 pair from the three analysis kits used in ILS 1 and 2. Sample 4 in ILS 1 was labeled as Sample 1 in ILS 2. The ground truth of this sample pair was a true non-fit, intended to have high confidence.

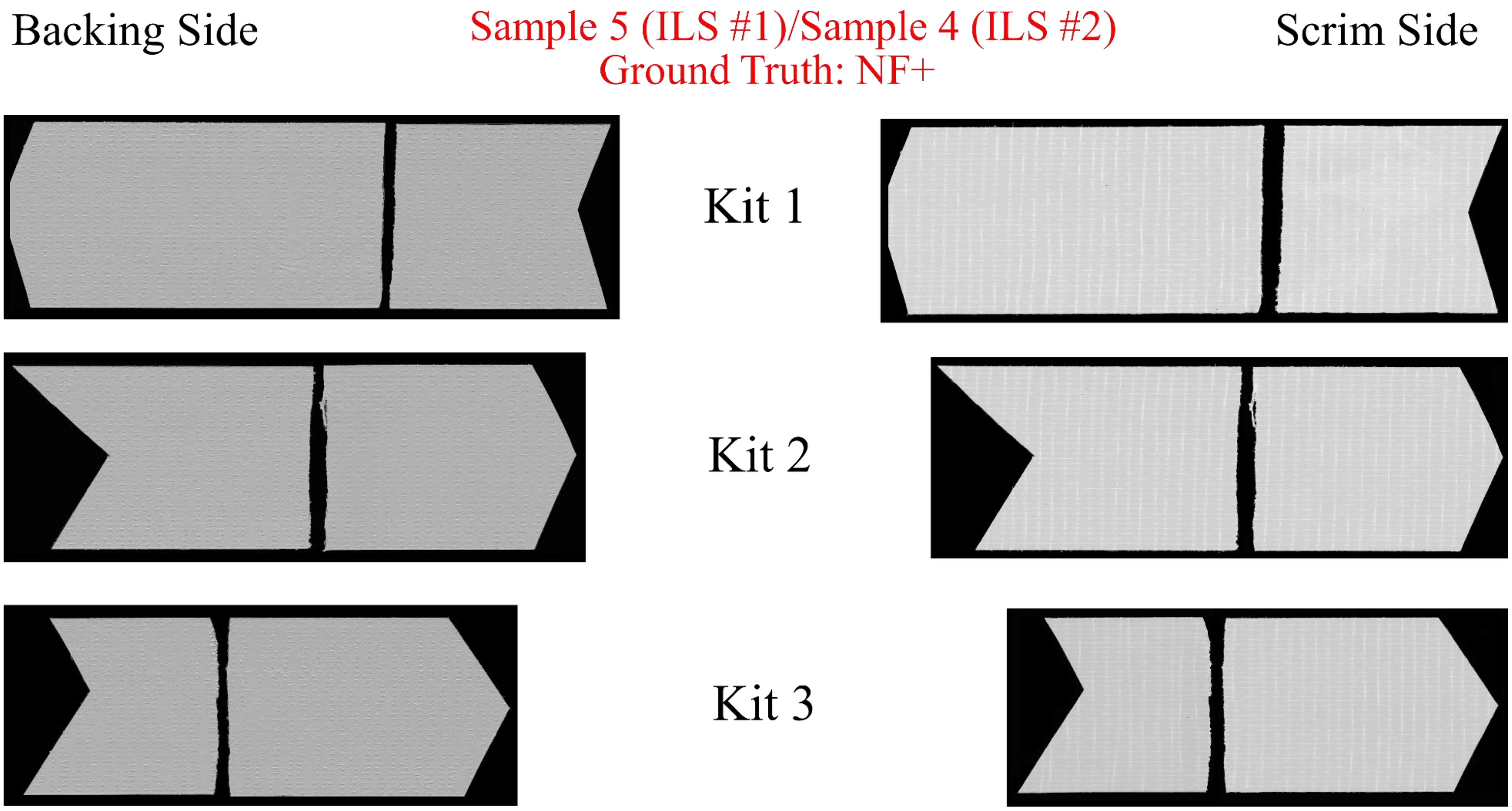


Figure 3A-5. Images of each Sample 5 pair from the three analysis kits used in ILS 1 and 2. Sample 5 in ILS 1 was labeled as Sample 4 in ILS 2. The ground truth of this sample pair was a true non-fit, intended to have high confidence.

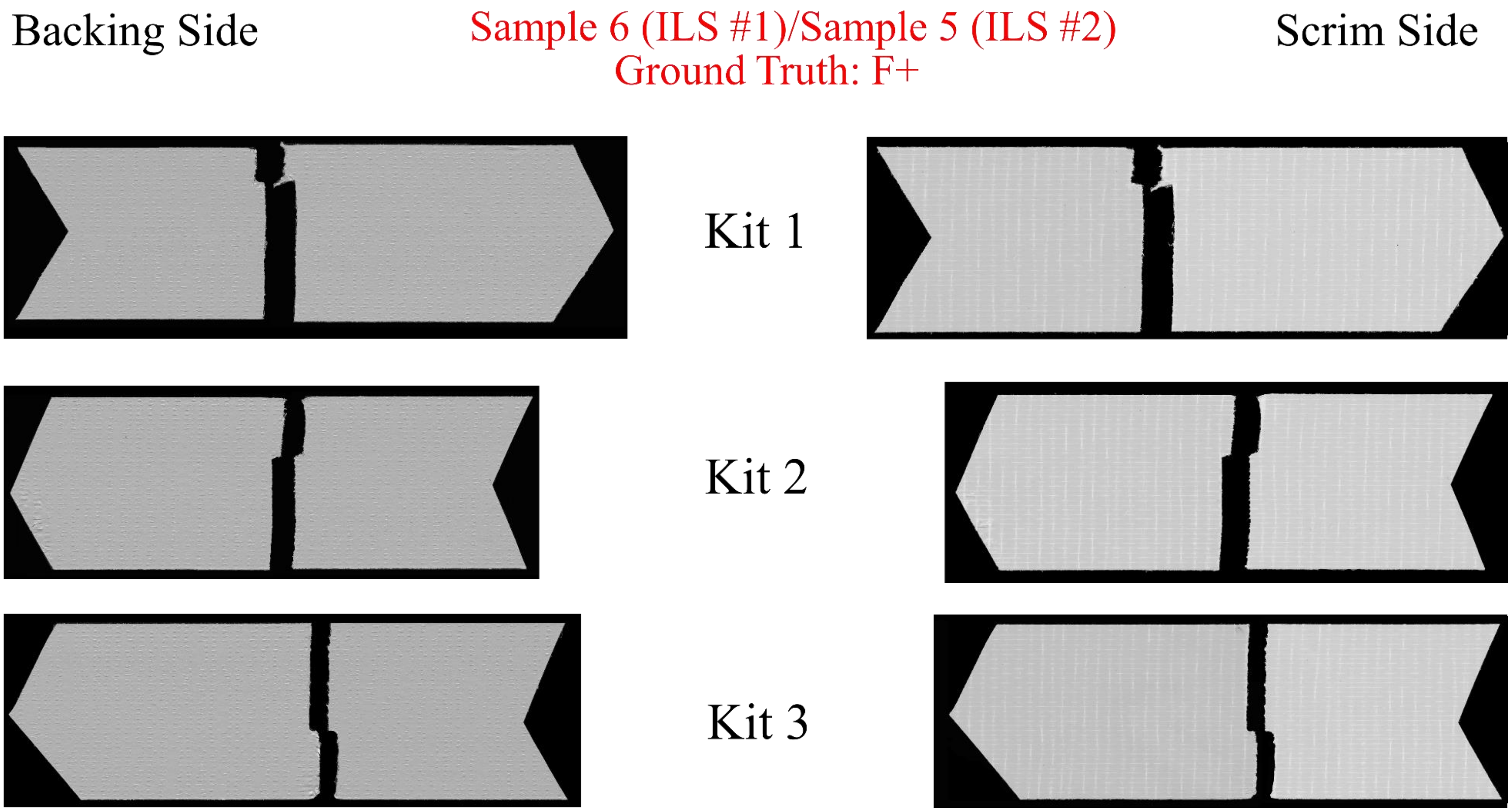


Figure 3A-6. Images of each Sample 6 pair from the three analysis kits used in ILS 1 and 2. Sample 6 in ILS 1 was labeled as Sample 5 in ILS 2. The ground truth of this sample pair was a true fit, intended to have high confidence.

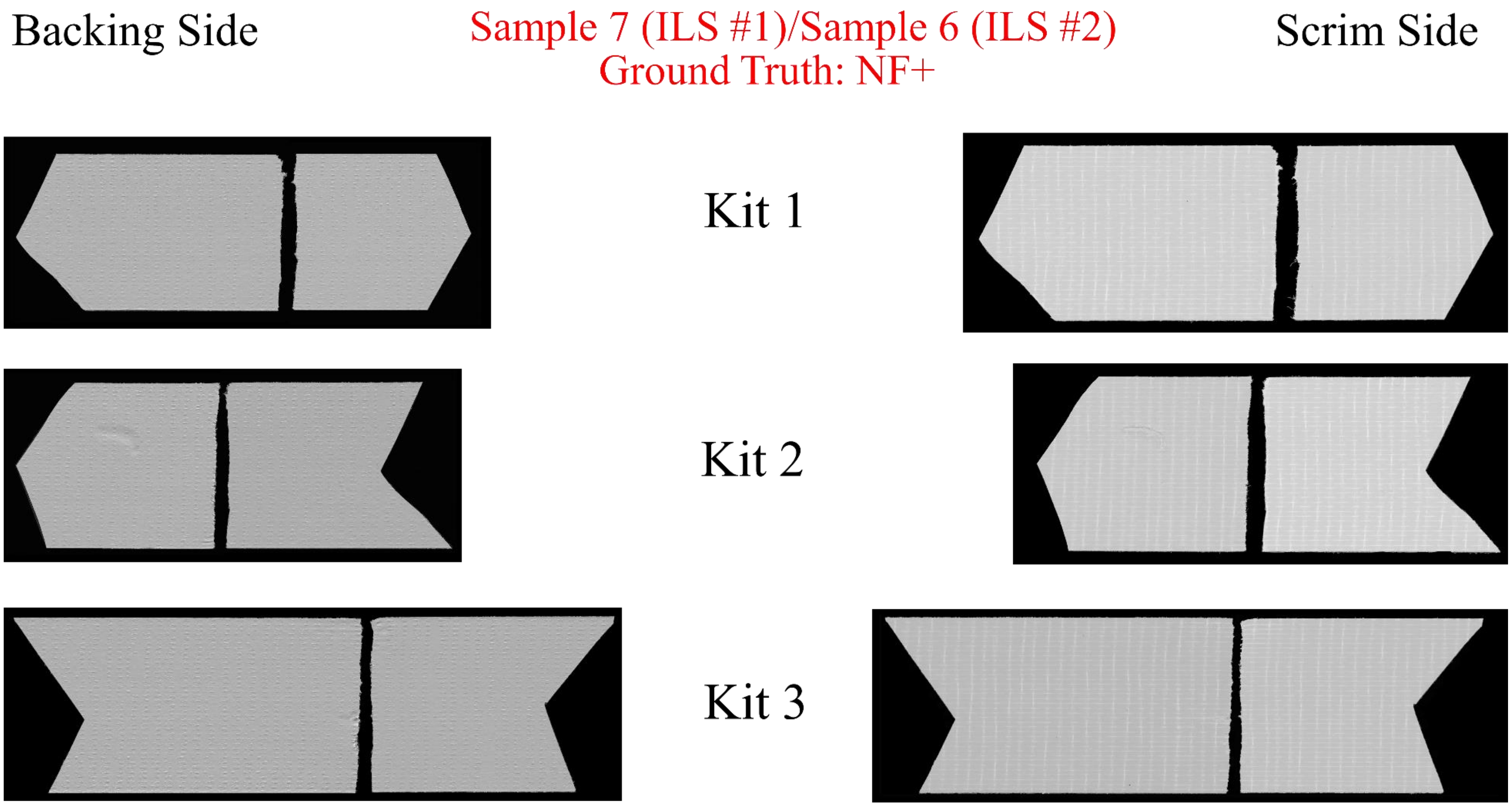


Figure 3A-7. Images of each Sample 7 pair from the three analysis kits used in ILS 1 and 2. Sample 7 in ILS 1 was labeled as Sample 6 in ILS 2. The ground truth of this sample pair was a true non-fit, intended to have high confidence.

Appendix 3-B

The following object links to a spreadsheet containing the instructions, templates, and surveys for both interlaboratory exercises (ILS1 and ILS2).

ILS_Instructions_Templates_Surveys.xlsx

Appendix 3-C

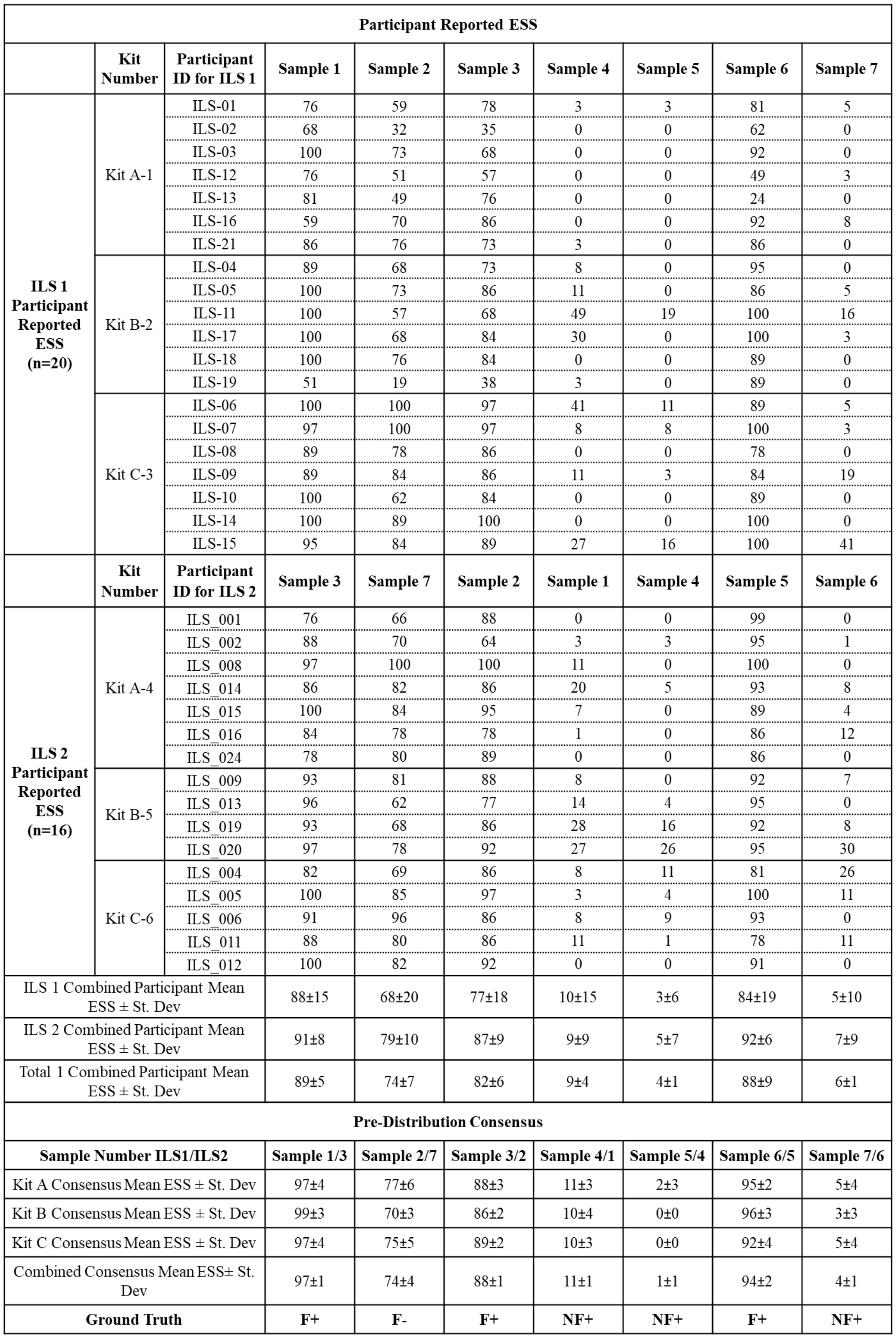
The following two links are the two training videos utilized in the second interlaboratory exercise (ILS 2). The first (Part 1) covers the foundation of the research on duct tape physical fits, as well as describing the results of the first interlaboratory exercise (ILS 1), while the second video (Part 2) explains the protocol and documentation template for the second interlaboratory exercise (ILS 2).

Part 1. Introduction to Physical Fit Research (WVU)

Part 2. WVU Round Robin 2 Protocol Video

Appendix 3-D

Table 3D-1. This table demonstrates all reported ESS values for each participant across the interlaboratory studies. The ID of each participant is listed, which kit they completed, and what the reported ESS was for each comparison they completed.



Participant Reported ESS									
	Kit Number	**Participant ID for ILS 1**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
ILS 1 Participant Reported ESS (n=20)	Kit A-1	ILS-01	76	59	78	3	3	81	5
		ILS-02	68	32	35	0	0	62	0
		ILS-03	100	73	68	0	0	92	0
		ILS-12	76	51	57	0	0	49	3
		ILS-13	81	49	76	0	0	24	0
		ILS-16	59	70	86	0	0	92	8
		ILS-21	86	76	73	3	0	86	0
	Kit B-2	ILS-04	89	68	73	8	0	95	0
		ILS-05	100	73	86	11	0	86	5
		ILS-11	100	57	68	49	19	100	16
		ILS-17	100	68	84	30	0	100	3
		ILS-18	100	76	84	0	0	89	0
		ILS-19	51	19	38	3	0	89	0
	Kit C-3	ILS-06	100	100	97	41	11	89	5
		ILS-07	97	100	97	8	8	100	3
		ILS-08	89	78	86	0	0	78	0
		ILS-09	89	84	86	11	3	84	19
		ILS-10	100	62	84	0	0	89	0
		ILS-14	100	89	100	0	0	100	0
		ILS-15	95	84	89	27	16	100	41
	Kit Number	**Participant ID for ILS 2**	**Sample 3**	**Sample 7**	**Sample 2**	**Sample 1**	**Sample 4**	**Sample 5**	**Sample 6**
ILS 2 Participant Reported ESS (n=16)	Kit A-4	ILS_001	76	66	88	0	0	99	0
		ILS_002	88	70	64	3	3	95	1
		ILS_008	97	100	100	11	0	100	0
		ILS_014	86	82	86	20	5	93	8
		ILS_015	100	84	95	7	0	89	4
		ILS_016	84	78	78	1	0	86	12
		ILS_024	78	80	89	0	0	86	0
	Kit B-5	ILS_009	93	81	88	8	0	92	7
		ILS_013	96	62	77	14	4	95	0
		ILS_019	93	68	86	28	16	92	8
		ILS_020	97	78	92	27	26	95	30
	Kit C-6	ILS_004	82	69	86	8	11	81	26
		ILS_005	100	85	97	3	4	100	11
		ILS_006	91	96	86	8	9	93	0
		ILS_011	88	80	86	11	1	78	11
		ILS_012	100	82	92	0	0	91	0
ILS 1 Combined Participant Mean ESS ± St. Dev			88±15	68±20	77±18	10±15	3±6	84±19	5±10
ILS 2 Combined Participant Mean ESS ± St. Dev			91±8	79±10	87±9	9±9	5±7	92±6	7±9
Total 1 Combined Participant Mean ESS ± St. Dev			89±5	74±7	82±6	9±4	4±1	88±9	6±1
Pre-Distribution Consensus									
Sample Number ILS1/ILS2			**Sample 1/3**	**Sample 2/7**	**Sample 3/2**	**Sample 4/1**	**Sample 5/4**	**Sample 6/5**	**Sample 7/6**
Kit A Consensus Mean ESS ± St. Dev			97±4	77±6	88±3	11±3	2±3	95±2	5±4
Kit B Consensus Mean ESS ± St. Dev			99±3	70±3	86±2	10±4	0±0	96±3	3±3
Kit C Consensus Mean ESS ± St. Dev			97±4	75±5	89±2	10±3	0±0	92±4	5±4
Combined Consensus Mean ESS± St. Dev			97±1	74±4	88±1	11±1	1±1	94±2	4±1
Ground Truth			**F+**	**F-**	**F+**	**NF+**	**NF+**	**F+**	**NF+**

Table 3D-2. This table demonstrates all calculated Z-scores for each participant across the two interlaboratory studies. The ID of each participant is listed, which kit they completed, and what their respective Z-score was for a given comparison. Z-scores are calculated using the results of each laboratory exercise independently.

Participant Z-scores									
	Kit Number	**Participant ID for ILS 1**	**Sample 1**	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**
ILS 1 Participant Z-scores (n=20)	Kit A-1	ILS-01	-0.82	-0.44	0.06	-0.47	-0.05	-0.17	0.00
		ILS-02	-1.36	-1.77	-2.39	-0.65	-0.51	-1.16	-0.55
		ILS-03	0.82	0.23	-0.55	-0.65	-0.51	0.40	-0.55
		ILS-12	-0.82	-0.84	-1.16	-0.65	-0.51	-1.86	-0.28
		ILS-13	-0.45	-0.97	-0.09	-0.65	-0.51	-3.13	-0.55
		ILS-16	-1.91	0.09	0.52	-0.65	-0.51	0.40	0.27
		ILS-21	-0.09	0.36	-0.24	-0.47	-0.51	0.11	-0.55
	Kit B-2	ILS-04	0.09	-0.04	-0.24	-0.10	-0.51	0.54	-0.55
		ILS-05	0.82	0.23	0.52	0.08	-0.51	0.11	0.00
		ILS-11	0.82	-0.57	-0.55	2.64	2.74	0.82	1.09
		ILS-17	0.82	-0.04	0.37	1.36	-0.51	0.82	-0.27
		ILS-18	0.82	0.36	0.37	-0.65	-0.51	0.25	-0.55
		ILS-19	-2.45	-2.44	-2.24	-0.47	-0.51	0.25	-0.55
	Kit C-3	ILS-06	0.82	1.56	1.14	2.09	1.35	0.25	0.00
		ILS-07	0.64	1.56	1.14	-0.10	0.88	0.82	-0.27
		ILS-08	0.09	0.49	0.50	-0.65	-0.51	-0.31	-0.55
		ILS-09	0.09	0.76	0.52	0.08	-0.05	-0.03	1.37
		ILS-10	0.82	-0.31	0.37	-0.65	-0.51	0.25	-0.55
		ILS-14	0.82	1.03	1.29	-0.65	-0.51	0.82	-0.55
		ILS-15	0.45	0.76	0.68	1.18	2.28	0.82	3.56
	Kit Number	**Participant ID for ILS 2**	**Sample 3**	**Sample 7**	**Sample 2**	**Sample 1**	**Sample 4**	**Sample 5**	**Sample 6**
ILS 2 Participant Z-scores (n=18)	Kit A-4	ILS_001	-2.00	-0.45	0.33	-1.11	-0.71	1.17	-0.78
		ILS_002	-0.29	0.25	-2.33	-0.78	-0.29	0.50	-0.67
		ILS_008	1.00	1.25	1.67	0.11	-0.71	1.33	-0.78
		ILS_014	-0.57	0.35	0.11	1.11	0,00	0.17	0.11
		ILS_015	1.53	0.45	1.11	-0.33	-0.71	-0.5	-0.33
		ILS_016	-0.86	0.15	-0.78	-1	-0.71	-1	0.56
		ILS_024	-1.71	0.25	0.44	-1.11	-0.71	-1	-0.78
	Kit B-5	ILS_009	0.43	0.3	0.33	-0.22	-0.71	0	0
		ILS_013	0.86	-0.65	-0.89	0.44	-0.14	0.5	-0.78
		ILS_019	0.43	-0.35	0.11	2.00	1.57	0	0.11
		ILS_020	1.00	0.15	0.78	1.89	3	0,50	2.56
		ILS_022	-0.57	-3.35	-1.89	0.22	-0.71	-0.17	-0.67
		ILS_023	-0.14	0.7	-1.00	1.11	-0.14	0.67	0.11
	Kit C-6	ILS_004	-1.14	-0.3	0.11	-0.22	0.86	-1.83	2.11
		ILS_005	1.43	0.5	1.33	-0.78	-0.14	1.33	0.44
		ILS_006	0.14	1.05	0.11	-0.22	0.57	0.17	-0.78
		ILS_011	-0.29	0.25	0.11	0.11	-0.57	-2.33	0.44
		ILS_012	1.43	0.35	0.78	-1.11	-0.71	-0.17	-0.78
Ground Truth			**F+**	**F-**	**F+**	**NF+**	**NF+**	**F+**	**NF+**

Appendix 3-E

Table 3E-1. Results of question seven of ILS 1 survey. Question was: Did you find the edge similarity score (ESS) approach easy to follow for duct tape end comparisons? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS-01	Difficult	I found myself consistently losing count of the scrim areas and having to go back to the top edge and count again to figure out where I was. I also found it difficult to count the scrim areas when looking at the backing side of the tape (even when using transmitted light).
ILS-02	Easy	No Comment
ILS-03	Easy	No Comment
ILS-04	Easy	No Comment
ILS-05	Easy	No Comment
ILS-06	Average	I think that using 1 for significant correspondence of features, 0 for featureless correspondence in an area, and -1 for significant differences in an area would be more in line with how an evaluation is really done.
ILS-07	Easy	Some edges were featureless and could be considered inconclusive but that selection was not available. Featureless edges were scored as a match.
ILS-08	Easy	I wish there was a sliding scale to align a number with each scrim area. I would turn to fill out paper and get lost midway down the tape and forget which area number I had done, so a lot of recounting. This opens an opportunity for error.
ILS-09	Difficult	there were several times that the edges of a scrim area could be made to align but only if you ignored the fact that the area directly above or below was VERY misaligned. I wasn't sure how to score that. There were also times that the weft yarn ends didn't align and it was hard to figure out if that should apply to the scrim area above or below the yarns, or to both.
ILS-10	Easy	Under the current system, with every area getting a score regardless, it generates an arbitrary number that could be misleading. If the tape is determined to be a physical match or not, the number generated should not give the impression that there is less than certainty of the result, either positive or negative.
ILS-11	Average	No Comment
ILS-12	Easy	No Comment
ILS-13	Average	It was more time consuming than difficult; especially when trying to keep track of the # for the scrim area that was being compared. If it were my own case I would have written on the acetate.
ILS-14	Average	The basic idea is easy to follow but did struggle trying to keep the edges lined up and counting down the areas. It would have been a little easier if both pieces of tape had been on the same acetate/been able to manipulate them some like in a real case. Difficult to view calendaring and dimple marks when using transmitted light so still had to flip back and forth, having to realign the tape and count areas again took longer.
ILS-15	Average	It was hard at times to flip the duct tape over and make sure I was looking at the same area on both backing and adhesive sides. It also took a lot more time than a typical examination.
ILS-16	Easy	No Comment
ILS-17	Very Easy	I think that scoring each of the 37 edge areas was a systematic manner of comparing each tape set. Caused me to slow down and really inspect each you the smallest of details. I also liked making notes of the striae that appear on the backing relative to each of the 37 areas. Demonstrates how much you saw and where.
ILS-18	Difficult	The layout of the ESS seemed to black/white in making a determination. In our laboratory we look at the overall characteristics but in order to form an opinion we are looking for the presence of individualizing characteristics. An overall pattern could be present but if there no individualizing characteristics are present then a physical match cannot be called and the sample will proceed to full manufacturing characteristic comparison.
ILS-19	Average	There were often several features on these samples that would lead me to conclude a PM; a few individual characteristics were "enough" to call a PM on these samples. I did not think the ESS score greatly influenced my decision.
ILS-21	Difficult	I struggled with it - I take an approach of big picture and then go smaller. It was hard to not be biased when I had seen that the samples didn't align with the contour of the tear.

Table 3E-2. Results of question eight of ILS 1 survey. Question was: Did you find the edge similarity score metric useful to inform/support your opinion? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS-01	Not Useful	Ideally, I would first examine each scrim area to generate an edge similarity score. Then the combination of my observations and the ESS would lead me to a conclusion. However, I couldn't help myself from making an initial conclusion about the physical fit almost as soon as I saw the tape. Once I had it in my head that a physical fit existed or didn't exist, I was basically only looking at the scrim areas to confirm my existing opinion. I felt like my exam of the scrim areas was biased because I had already jumped to a physical fit conclusion before calculating the ESS. This may be because I am an experienced physical fit/tape examiner. In my opinion, this method could be extremely useful when training a new examiner, to boost their confidence in their opinion before they gain the experience necessary to make physical fit calls independently.
ILS-02	Not Useful	This is a good way to document what is observed by the examiner for the review process, I don't feel the score would be anything that would push my opinion one way or the other. A physical fit exam is wholistic and the conclusion is based upon all of the data. I think it is dangerous to rely on a value that can be used to separate a match from a non-match. Also, we verify all PM exams by a second scientist.
ILS-03	Useful	No Comment
ILS-04	Not Useful	I understand that stats are more beneficial with more points of comparison; however, I observed that the features I used to associate or discriminate were present over 4-6 areas rather than each area individually. Therefore, this approach would make more sense to me if the evaluation was binned every 5 segments rather than as 37 individual segments.
ILS-05	Useful	Somewhat useful. I found that I had an opinion already before I evaluated each scrim area, and looking at the scrim areas individually affected my certainty (made me less certain). Having an ESS that highly corresponded with my initial overall opinion would ultimately support that opinion. For those ESSs that weren't on the extreme ends (near 1 or 0), I was less certain. In casework with dirty samples, I would be concerned the ESS metric might make my overall opinion less supported and I may be inclined to report more inconclusives.
ILS-06	Not Useful	No Comment
ILS-07	Useful	No Comment
ILS-08	Useful	Yes, so long as there is context. Does there have to be a number below which you should conclude a non-match or inconclusive? Is there a particular number or percentage of corresponding areas that must align to support the match?
ILS-09	Not Useful	would have been more helpful to have an inconclusive option as opposed to just match/non-match. For example, when both edges are both so distorted that you can't even see an edge to see if it matches or not. Or for edges that are roughly linear- they technically "match" but don't have any unique features that I would use in casework to determine a physical match.
ILS-10	Not Useful	Under the current system, with every area getting a score regardless, it generates an arbitrary number that could be misleading. If the tape is determined to be a physical match or not, the number generated should not give the impression that there is less than certainty of the result, either positive or negative.
ILS-11	Useful	No Comment
ILS-12	Useful	No Comment
ILS-13	Very useful	Yes and no. There were some samples where the overall shape appeared to match but the individual characteristics of the edges did not fit neatly back together. Therefore, the score metric might appear to detract from the decision to call it a match.
ILS-14	Not Useful	It could be a valuable tool for court purposes, but I find our current method of documenting the overall fracture contour and individualizing characteristics to be sufficient.
ILS-15	Not Useful	Not really useful for informing my opinion because before using the edge similarity score, I could already see whether they matched or not. It could potentially support an opinion if the numbers correlate, but typically in court an image would probably demonstrate a match / non match better.
ILS-16	Useful	No Comment
ILS-17	Useful	No Comment
ILS-18	Not Useful	Needs to be more specific as to if samples have non-individualizing characteristics and therefore no opinion can be rendered in that area.
ILS-19	Not Useful	The Similarity score metric I don't think would greatly influence my overall conclusion, given all the details documented in the sample.
ILS-21	Not Useful	I don't think it was useful for my decision, but I can see how it might be useful to support it when there is a newer scientist. My other concern would be that there isn't room for explainable differences - in the edges that are stretched and don't align, the stretching can explain why it doesn't match there.

Table 3E-3. Results of question nine of ILS 1 survey. Question was: If you were to implement the ESS approach in your examinations, would you find the report templates for the score metric useful for a peer-review process? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS-01	Very Useful	All the templates were very well designed and easy to use.
ILS-02	Useful	No Comment
ILS-03	Useful	No Comment
ILS-04	Useful	No Comment
ILS-05	Very Useful	No Comment
ILS-06	Not Useful	This approach is useful in validated physical matches in tape. But it is limited because it can only be used on certain types of tapes. Not all tapes have scrim patterns and not all duct tapes have horizontal scrim threads. Since this practice cannot be universally applied to physical matching of tapes, it is of limited use in actual casework.
ILS-07	Useful	No Comment
ILS-08	Useful	Maybe? Verifications are performed in my lab of all positive and probative physical fits. So would need to consider if the verifier would also need to complete one of these. The verification is done by looking at the original evidence, so the verifier could see different characteristics and put different weight on them than original examiner. I think it would help support but would not eliminate the need of photographs and would only add to the documentation requirements.
ILS-09	Not Useful	Would have been more helpful to have an inconclusive option as opposed to just match/non-match. For example, when both edges are both so distorted that you can't even see an edge to see if it matches or not. Or for edges that are roughly linear- they technically "match" but don't have any unique features that I would use in casework to determine a physical match.
ILS-10	Very Useful	No Comment
ILS-11	Useful	No Comment
ILS-12	Not Useful	No Comment
ILS-13	Useful	Our laboratory conducts verifications (blind whenever possibly) of every physical match case, so it could be a useful tool during this step.
ILS-14	Useful	No Comment
ILS-15	Useful	No Comment
ILS-16	Useful	No Comment
ILS-17	Very Useful	Especially if one were to go back, as a reviewer, the re-check by looking at the same samples yourself vs images or drawings. Images can only show so much but this sheet would pint me to the critical features and simplify a review.
ILS-18	Useful	With certain modifications to indicate areas that are generally non-individualizing the implementation could be useful in review and in court.
ILS-19	Useful	The report template was easy to follow.
ILS-21	Very Useful	No Comment

Table 3E-4. Results of question twelve of ILS 2 survey. Question was: Did you find the edge similarity score (ESS) approach easy to follow for duct tape end comparisons? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS_001	Agree	Found it challenging to score for distorted areas that overall fit but had to score as .5 since edges did not align. However, did not seem accurate to score as 0.
ILS_002	Somewhat disagree	No Comment
ILS_004	Somewhat Agree	No Comment
ILS_005	Somewhat Agree	No Comment
ILS_006	Agree	As long as you know you have the entire fractured ends from both sides of the tape this process could be useful. The ESS could give false numbers if section(s) of the fractured edges are missing.
ILS_008	Somewhat Agree	No Comment
ILS_009	Somewhat Agree	No Comment
ILS_011	Agree	No Comment
ILS_012	Agree	I would have liked a NA response in some cases
ILS_013	Somewhat Agree	It was easy to determine the result from the edge similarity score, however the process of analyzing each individual scrim made the determination more time-consuming.
ILS_014	Somewhat Agree	No Comment
ILS_015	Agree	No Comment
ILS_016	Somewhat Agree	It's easy to follow but can be very difficult to keep track of where you are, especially with the inner most sections.
ILS_019	Somewhat Agree	It is difficult to get my brain to wrap around treating each scrim bin as its own physical fit with its own characteristics. During training, we are taught that it is the totality of the edge features that allows one to determine a physical fit or not. So it was hard to let that go.
ILS_020	Somewhat Agree	It was generally easy to follow however a lot of what it asked seemed to be redundant information with the samples we had
ILS_024	Agree	No Comment

Table 3E-5. Results of question thirteen of ILS 2 survey. Question was: Did you find the edge similarity score metric steps useful to document your thought process in a systematic manner that would be easy to describe to other examiners? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS_001	Agree	Generally, agree. Video somewhat helpful to understand what's being asked for each feature, however the one example shown was straightforward. Suggest including more explanations of what is meant by each feature (column) and each option (drop-down) for several scrim areas.
ILS_002	Completely Disagree	I would agree document my thought process after lawyers agree to document theirs
ILS_004	Somewhat Agree	I found the macroscopic level useful, to make sure all aspects of the duct tape were considered, and to somewhat break it down. When I broke it down to the scrim level, however, I felt that I was questioning myself too much with the extreme detail and that I was losing the "overall picture"
ILS_005	Somewhat Agree	No Comment
ILS_006	Somewhat Agree	The ESS number could indicate the likelihood of fracture match no hard conclusions could be made from that number
ILS_008	Somewhat disagree	I initially thought this would be a useful evaluation tool for me (after the 1st study and as the instructions were presented). But I found that in practiced it is too granular. I can see the benefit of the macro assessment where the width is segmented into sections; however, I found the further delineation to be too prescriptive as I was evaluating the overall area for distortion or concordance of multiple features (warp, weft, backing) that existed over a larger area than the distance between two warp yarns.
ILS_009	Somewhat Agree	No Comment
ILS_011	Somewhat Agree	No Comment
ILS_012	Neither agree nor disagree	I think there is a lot of room to interpret some of the characteristics. i.e. how much stretching is needed to mark stretching is present.
ILS_013	Agree	From the edge similarity score approach, it is easy to identify in exactly which scrim there is differences, which can make it easier to allow others to review.
ILS_014	Agree	No Comment
ILS_015	Agree	No Comment
ILS_016	Neither agree nor disagree	It would be useful to document the thought process used when using this method, but this would likely have differences from the thought process I would typically be using. I also feel like sometimes you can get too "zoomed in" and areas that are truly a match have too much distortion on a microscopic level. Also, in a typical case currently, we would be removing adhesive at the ends to ensure all warp yarns line up as you can't really see all the ends with the adhesive still there.
ILS_019	Somewhat Agree	It is a different way of approaching a PF examination that takes some time to adjust to. I think it is relatively easy to explain, but maybe not so easy to put into practice.
ILS_020	Disagree	The is not the process that I normally take when approaching a physical fit, yes, I do a macro and then see if there are any features that traverse both edges but rarely do I look at each individual part and evaluate that for a fit or not
ILS_024	Somewhat Agree	No Comment

Table 3E-6. Results of question fourteen of ILS 2 survey. Question was: If you were to implement the ESS approach in your examinations, would you find the report templates for the score metric useful for a peer-review process or as a training tool? The participant's response to the question, along with any comments they left, are shown below.

Participant	Answer	Comments
ILS_001	Somewhat Agree	Due to distortion, two of the tape comparisons scored in between fit and non-fit. However, based on the totality of the correspondences, I felt confident to call them fits. These discrepancies between the ESS and my determination are why I'm responding this way, as then the final opinion is still subjective. And the ESS approach is less useful in these circumstances.
ILS_002	Disagree	No Comment
ILS_004	Somewhat Agree	I think the macroscopic level was useful, both for review and for training.
ILS_005	Neither agree nor disagree	No Comment
ILS_006	Somewhat Agree	I would probably not implement this approach due to it only being applicable for use with duct tape. I would look into implementing a model that could be used for facture matches with various substates. As for a training tool, this could be useful for a beginner of fracture matches. I don't believe it would be useful in a peer review process. The number is based on the initial examiners findings which is what the peer review is checking, not a number based off of those findings.
ILS_008	Neither agree nor disagree	No Comment
ILS_009	Neither agree nor disagree	No Comment
ILS_011	Agree	No Comment
ILS_012	Disagree	Too difficult to follow in the report templates. I think that using the numbering system is valuable to point out the specific area features are located. However, the key features are lost in the table of data. I would prefer a table that assumes consistent unless noted for a match and assumes not consistent unless noted for a non-match. Only include areas of significance in table and remove all other data.
ILS_013	Agree	I would find it useful as a peer-review process, since you can easily identify the areas with discrepancies, as well as for a training tool, since it can show trainees what examiners look for when performing physical fit comparisons for duct tapes.
ILS_014	Somewhat Agree	Somewhat agree with both applications.
ILS_015	Completely agree	No Comment
ILS_016	Agree	Yes, the templates are good for the ESS approach. I do think that some categories for the microscopic subunit section are not necessary - like continuation of scrim weave pattern and weft scrim at or near edge. Those can just be run through once in section 2 and should not change.
ILS_019	Agree	The report templates were very easy to follow - only suggestion would be to clarify that Parts 1 and 2 can be done with a stereoscope (I originally thought they both had to be done visually only).
ILS_020	Disagree	No Comment
ILS_024	Somewhat Agree	No Comment

Table 3E-7. Results of question fifteen of ILS 2 survey. Question was: With those considerations, do you think the ESS step can be used to support your opinion by providing additional comparative data and systematic criteria? The participant's response to the question, along with any comments they left, are shown below.

Participant	**Answer**	**Comments**
ILS_001	Agree	No Comment
ILS_002	Disagree	No Comment
ILS_004	Somewhat disagree	Again, I think the macroscopic assessment was useful, but the ESS was too minute as to make myself question my results more. I think finding significant, unique points of comparison is more useful. Also, if you had a duct tape that had a perfect match on half of it, and the other half was missing, the ESS score would not be an accurate reflection of the fact that the tapes are a match.
ILS_005	Somewhat Agree	No Comment
ILS_006	Disagree	no, based on the notes above the ESS number is only "valid" if it supports your conclusion.
ILS_008	Somewhat disagree	I think the macro segments approach has greater utility than the ESS approach.
ILS_009	Somewhat Agree	No Comment
ILS_011	Agree	No Comment
ILS_012	Agree	I would never include the number in a report, but I would use it in my notes to reflect the areas with similarities and differences
ILS_013	Agree	I do think this method could help when there is confusion or uncertainty on what result to report, since it has a specified range for each result.
ILS_014	Somewhat Agree	No Comment
ILS_015	Agree	No Comment
ILS_016	Somewhat Agree	It is possible it could be used to support my opinion, but my concern is that I might have too many non-matching or inconclusive areas in some fracture matches in which I am confident of a fit. Sometimes when there is too much distortion, no conclusion can be made about each individual area, when in fact they do fit as a whole.
ILS_019	Neither agree nor disagree	Even if this information isn't included in the report, the notes are still discoverable. Although there is all sorts of other documentation gathered during a PF examination, I'm worried that there is a real chance that a jury or defense attorney will misinterpret the ESS score as a type of confidence score (e.g., "they said it was a physical fit but the ESS is only 82%!") - particularly during closing arguments when we aren't there to clarify [hair analysis has previously fallen victim to this and has been called "junk science" because of it].
ILS_020	Disagree	No Comment
ILS_024	Somewhat Agree	No Comment

Chapter 4: Experimental results on data analysis algorithms for extracting and interpreting edge feature data for duct tape and textile physical fit examinations

4.1. Overview

This study proposes various computer-based methods to support the analyst’s decisions in physical fit examinations of duct tapes and textiles. First, the study evaluates the error rates associated with complex case situations that simulate the recovery of partial samples. Experimental thresholds of minimum sample size are estimated as a function of relative missing portions of the textile or tape’s width on the comparison edges. Second, a machine learning algorithm extracts and assesses the importance of edge feature information from analysts’ reporting templates. Then, a decision tree model is presented to add objectivity to the analysts' conclusions.

The duct tape dataset includes 1098 pairs originating from low and high-quality rolls. The samples are either hand-torn or scissor-cut, and several sets undergo stretching. The textile dataset consists of 600 samples taken from clothing items made of 100% cotton and fractured by stabbing the item or tearing it by hand. In the comparison templates for each material, the analysts document comparison features for each edge comparison bin and quantitative values (0, 0.5, and 1) that denote each bin decision (non-fit, inconclusive, fit). The partial edge analysis of duct tapes demonstrates that accurate physical fit comparisons are feasible with at least 35% of the edge width, while textiles are feasible with at least 40% of the edge. However, the uncertainty increases with less sample available for comparison. These results reveal that, regardless of material, accuracy for partial width comparisons suffers for more complex or distorted samples, such as high-quality hand-torn tapes, where partial sample examinations are not recommended.

The extraction and analysis of feature information show that certain features hold different weights in the decision depending on the separation methods and tape’s qualities. For example, the alignment of severed dimples is one of the most influential features of scissor-cut backings, but not other separation methods. Similarly, the importance of a feature such as a scrim weave pattern is superior to high-quality tapes than other grades. While the importance of the features observed in textiles is not as divergent as in tapes, there are still noticeable trends, such as that the print/pattern and construction alignment hold more value for stabbed samples, while the yarn alignment is more informative in hand-torn samples. This information is then used to train decision tree models, which provide comparable performance to the human analysis, and demonstrate the value of incorporating objective computational models to support the analyst’s conclusions.

4.2. Introduction

Growing attention is being paid to forensic physical fit examinations, and several studies have been published to address gaps in this area [1-6]. While most standard operating procedures in forensic laboratories are generalized [4], the features and characteristics used to inform an analyst’s opinions depend on the material. For example, the fibers’ features observed in duct tape samples differ in importance in a physical fit comparison compared to textile fabrics [7-9]. As such, one of the most critical aspects to address in physical fit examinations is identifying which features are relevant to a given material during the comparison steps and why.

With the advent of machine learning, big data, and powerful computational tools, there is a growing opportunity to narrow those gaps and better use underlying information [10]. While neural networks have become increasingly popular in this field, other statistical analysis techniques are often overlooked despite their potential advantages. For example, statistical tools such as mutual information can be handy when dealing with smaller datasets [11]. They can help with feature ranking and extraction without requiring extensive data. Here, these methods are used to understand better the significance of different features in a physical fit examination, which can improve the data curation process and design of the examination approach.

Incorporating a range of statistical analysis techniques and machine learning tools can provide valuable insights into the data generated by forensic institutions. By leveraging these insights, the examination methods can be refined to improve the training for forensic scientists and peer review processes. This approach can lead to more accurate and reliable forensic analyses, significantly impacting criminal investigations and justice outcomes. Forensic physical fit examinations are inherently subjective, which makes them vulnerable to legal challenges in court. However, by quantifying the importance of different features experienced examiners consider, new decision-making tools can be developed to increase scientific foundations, objectivity, and reliability.

Several studies already use computational modeling to provide a quantitative assessment of physical fits to address current questions in physical fit analysis. Two include studies by van Dijk et al. and Spaulding et al. on duct tapes [12-13]. van Dijk et al. use what they refer to as 'loop-breaking patterns' created by the separation of weft-insert yarns to assess the similarity of two edges quantitatively. These values, created from a comparison set of 272 tape edges, are used to calculate likelihood ratios using a Bayesian network and assess the evidential value to comparisons. The authors assign a value according to the appearance of each loop along with width as either a 0, 1, or 2 depending on if the loop was closed (or missing), open, or complex. Overall, this approach shows high discrimination power and generally low rates of misidentified evidence [12]. Spaulding et al. use a model that uses automatic edge detection to calculate cross-correlation functions and random match probabilities to compare duct tape fits. From a set of 150 end-matching pairs, this study evaluates partial duct tape edge comparisons, with the algorithm distinguishing between true fit and non-fit edges when the width is at least 27% of the full width [13].

Our research group has also developed several methods in this area based on an edge similarity score (ESS) principle. The ESS metric is intended to provide analysts with a quantitative value to denote the similarity between two edges based on the features present and the analyst's interpretation of the edges. The ESS method has been developed and applied to several materials thus far, including duct tape and textiles [7-9]. In a study of more than 3000 duct tape comparisons, the ESS method results in 84.5-99.8% accuracy, depending on the tape quality and the separation method used. In addition, when tapes are stretched, this study shows that though the variability of ESS scores increases, there is no substantial influence on overall accuracy [7, 8]. For textiles, in a population of 600 samples, accuracies of the ESS method range between 98-100% based on the construction and separation method [9]. In addition, interlaboratory studies conducted using the ESS methods for both tape and textiles have led to improved documentation of features for both materials. The interlaboratory studies result in an overall accuracy of 95-99% for the duct tapes and 89% for the textiles. [14-15].

Beyond the application to physical samples, a machine learning model [16] has also been developed to assist in the physical fit analysis of duct tapes. The developed model consists of two independent convolutional neural networks (one for each side of a tape: backing and scrim) that output combined probabilities for each network and then assign membership to either fit or non-fit for a given pair. Using a dataset of more than 4000 images of duct tapes, the computational model results in 71-88% accuracy, specifically reproducing some trends observed in the human analysis of the same samples [16].

As a new approach to the value of machine learning for physical fit examination, this study utilizes the methods developed in our group. It introduces a novel approach that uses mutual information and decision trees as fundamental data analysis tools. First, the potential minimum length of the tape necessary to complete a physical fit evaluation is assessed by evaluating the accuracy of the ESS method from randomly selected portions of tapes in the dataset. Following that, mutual information is applied to quantify the importance of features in the eyes of the analysts. Finally, a decision trees model is created as a decision assistant for physical fit examinations of duct tapes. The outcome of this study is expected to aid forensic examiners in conducting more objective physical fit examinations.

4.3.Materials and Methods

4.3.1. Sample data

The data in this study originate from samples initially studied by Prusinowski et al. [7, 8]. The datasets are from low-quality (LQ) and high-quality (HQ) tape. Specifically, data is from the low-quality scissor-cut (LQ-SC), low-quality hand-torn (LQ-HT), low-quality hand-torn stretched (LQ-HT-S), high-quality scissor-cut (HQ-SC), and high-quality hand-torn stretched (HQ-HT-S) sets. The samples are placed on pieces of transparency film to allow the adhesive and scrim fibers to be observed. For the high-quality samples, a thin strip (~3mm) of adhesive is removed from the tape edge to allow the scrim fibers to be seen, as the adhesive is dark and opaque. The dataset in this study accounts for a total of 1098 edge comparisons [8]. The textile data used in this study is derived from a larger population of textile physical fit comparison pairs explored by Andrews et al. [9]. The data includes 600 comparison pairs generated from fifteen 100% cotton garments of varying construction (knit or weave). These pairs consist of 301 stabbed pairs and 299 torn pairs.

For both materials, the pairs are examined by a human analyst using a standardized documentation spreadsheet to record the presence or absence of pre-defined comparison features and document the overall conclusions regarding each comparison pair. All comparisons are performed blindly, meaning the analysts is unaware during the comparison of the ground truth of the sample pair.

4.3.2. Standardized Reporting Template

Each comparison is broken into three main steps: an overall assessment of the edges and general morphology. The second is a macroscopic assessment of the edge performed by visually separating the sample pair into five comparison areas and documenting each section's main features and alignment. The final step is a microscopic assessment of the tape edges, where a comparison bin is defined as the area between the scrim fibers (hereby referred to as 'scrim bins'). Each scrim bin is independently observed, documented for comparison, and assigned a quantitative value representing the alignment of the individual bin (hereby referred to as the area bin code). The values are 0 for a non-aligning bin (non-fit), 0.5 for a bin with both aligning and misaligning features (inconclusive), and 1 for a bin demonstrating aligning features (fit). The template automatically calculates the pair's edge similarity score (ESS), representing the overall alignment's quantitative value as a relative percentage of the number of scrim bins aligning across the tape width [8].

A template is filled out for every pairwise comparison of tape samples. The features documented are described in **Table 4-1**. Each comparison area has a total of 8 main features, and the template has drop-down options for the analyst to select and denote the feature and the rationale for their choice in the interpretation of each bin. For most, the feature can be described as absent or present, and if present, whether it indicates a fit or non-fit. Also, the template incorporates a simple visual aid that color-codes the comparison bins to identify at a glance the fitting (green), non-fitting (red), and inconclusive bins along the overall edge width.

The documentation template for textile comparisons is similar to that of duct tape but only consists of two primary steps. The first step is an overall assessment of the textile edges, and the second step is a microscopic assessment of the edges. For this second step, the comparison edge is divided into ten areas, or bins, of equal size. Unlike tapes, the bin size on the textile fragments is based on the overall length of the fracture, divided by ten, whereas the tape is divided into bins based on the inherent number of scrim fibers. The rest of the template is very similar to the one for duct tapes: color-coded comparison areas indicating the area fit code, seven pre-defined comparison features with response options to document the presence or absence in each bin, and automatic calculation of the overall ESS as a relative percentage of the ten bins. The specific features documented for textiles are described in **Table 4-2**.

4.3.3. Data Extraction

4.3.3.1. Minimum Width on Partial Comparison Items

To evaluate partial comparisons of duct tapes and textiles, the area fit codes data of all the available tape, and textile samples in the reporting template are extracted to assess the minimum width needed for reliable physical fit examinations. Following this, partial widths simulating the recovery of only a portion of the edge are defined based on the number of scrim areas. A randomly selected starting point among the edge is chosen for a given sample pair, and the corresponding number of consecutive bins is recorded. From there, the ESS of the partial width is calculated, and an overall conclusion is assigned. If the ESS of the partial width is less than 40, the edge comparison is considered a non-fit, an ESS of 40-60 is considered inconclusive, and an ESS of 60 or higher is considered a fit. The recorded outcome of the partial width comparison is evaluated versus the known ground truth (i.e., known true fit, known true non-fit).

Table 4-1. Table of features documented for tapes and the respective response options for observation of the features at the macroscopic and microscopic level.

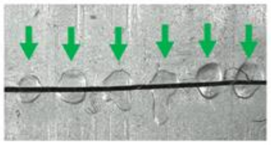
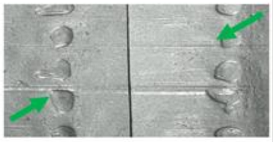
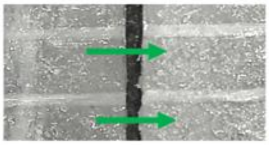
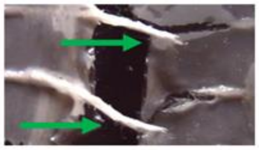
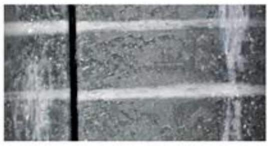
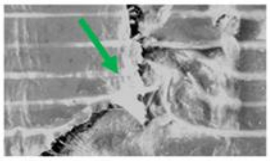
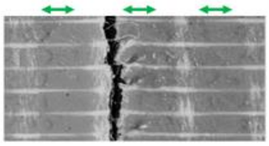
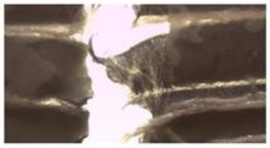
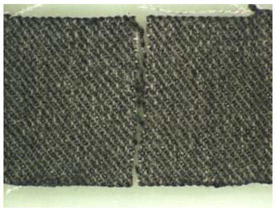
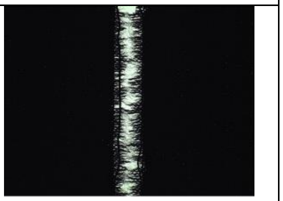
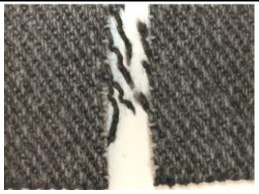
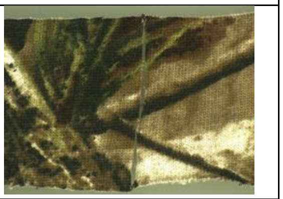
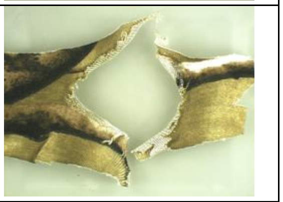
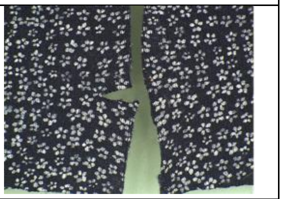
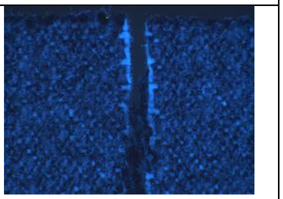
Feature	Description	Response Options	Example Image
I. Alignment of Severed Dimples on Backing	Severed dimples on tape backing that align in shape, size, and location across fracture	• Absent • Present – indicative of fit • Present – indicative of non-fit	
II. Alignment of Calendaring Striations on Backing	Calendaring striations (small scratches/marks left by manufacturing process) on tape backing that align across fracture in location, shape, and depth	• Absent • Present – indicative of fit • Present – indicative of non-fit	
III. Alignment of Warp Scrim	Warp fibers that transverse the fracture in a straight line and correspond to the fiber on the other side when the top and bottom edges of the tape are aligned.	• Absent • Present – indicative of fit • Present – indicative of non-fit	
IV. Corresponding Protruding Warp Yarns and Gaps	Warp fibers that extend past the edge of the tape backing that correspond with a proportional gap or missing scrim fiber portion on the opposite edge	• Absent • Present – indicative of fit • Present – indicative of non-fit	
V. Continuation of Scrim Weave Pattern	Consistent pattern of scrim fibers across fracture, for both warp and weft fibers where applicable. In a simple weave pattern (seen above), the pattern of the fiber alternates for each subsequent fiber when in proper alignment	• Absent • Present – indicative of fit • Present – indicative of non-fit	
VI. Distortion Explained by Observed Stretching Directionality	Alteration to the backing or adhesive morphology that is caused by the means of fracture or external factors (for example: Protrusions on tape sample A met by indentations on sample B, or vice versa). Distortion may coincide with the direction of the tearing	• Absent • Present and explained by stretching • Present and not explained by stretching	
VII. Consistent Weft Fibers at or Near the Edge	Full or partial weft yarns on each edge that are consistent with the rest of the weft fibers/opposing edge (full weft fibers spaced appropriately on edge, weft fibers crossing the fracture in the same location, etc)	• Consistent • Not Consistent	
VIII. Missing Material	Material missing from microscopic comparison of tape edges that does not correspond to the other edge. Material may include backing and/or scrim fibers	• Not applicable (no missing material) • Observed missing material	

Table 4-2. Table of features that are documented for textile comparisons, including a description of the feature, an image of the feature, and the response options available to select for each feature.

Feature	Description	Options	Image
Construction Alignment	Consistency and alignment of construction, including type (weave/knit) and direction, between two textile fragments. Consistency in the thread or yarn count between the two fragments is also considered.	• Consistent • Inconclusive • Inconsistent • Cannot be assessed	
Gap Alignment	Alignment of yarns from one fragment into corresponding gaps observed in another fragment along the comparison edge.	• Consistent • Inconclusive • Inconsistent • Cannot be assessed	
Yarn Alignment	Alignment of yarns that have been pulled out of the fracture edge between two textile fragments.	• Present - Indicative of fit • Present - Indicative of non-fit • Inconclusive • Absent	
Design Alignment	Consistency and alignment of yarn color and pattern between two textile fragments.	• Present - Indicative of fit • Present - Indicative of non-fit • Inconclusive • Absent	
Distortion	Force applied during the fracture event causes distortion that can mask other features.	• Present - Indicative of fit • Present - Indicative of non-fit • Inconclusive • Absent	
Secondary Tearing	A secondary, perpendicular tear that is not the primary fracture that is being compared.	• Present - Indicative of fit • Present - Indicative of non-fit • Inconclusive • Absent	
Fluorescence	Fluorescence of individual yarns can aid in the identification of a physical fit.	• Consistent • Inconclusive • Inconsistent • Cannot be assessed	

This process is repeated for all potential lengths across all samples in the dataset. Five iterations of random selections of widths and starting points are performed to evaluate the variability in performance across the datasets. Following the calculation of the performance rates for each partial width across all five iterations of the model, beta regression is applied to the performance rates. This allows for an evaluation of the average performance rate at each partial width and an assessment of the variability at each point. Where relevant, the data is separated by material type, quality, and separation method.

4.3.3.2. Mutual Information Theory

In 1948, Claude Shannon introduced the concept of mutual information in his seminal paper "A Mathematical Theory of Communication" [17]. Alongside entropy and channel capacity, mutual information is considered one of the foundational concepts in information theory. This measure quantifies the amount of information shared by two random variables and is used to assess their mutual dependence. It serves as a tool for quantifying the information shared between the variables.

Suppose X and Y are two random variables with corresponding probability distributions $p(X)$ and $p(Y)$, respectively, and let their joint distribution be represented as $p(X,Y)$. Mutual information, denoted as $\mathbb{I}(X;Y)$, measures the degree to which the joint distribution $p(X,Y)$ is similar to the factored distribution $p(X)p(Y)$. The mathematical formula for mutual information is given as follows [17]:

$$\mathbb{I}(X;\ Y) = \sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

If X and Y are independent, then their joint probability distribution $p(x,y)$ is equal to their factored probability distributions $p(x)p(y)$, resulting in a mutual information value of zero.

This concept can be better understood by comparing the mutual information formula and recognizing its similarities to the concept of entropy from information theory. Entropy $\mathbb{H}(X)$ in information theory is represented by the following equation:

$$\mathbb{H}(X) = -\sum_{x} p(x)\log p(x) \tag{2}$$

Entropy, in information theory, measures the amount of uncertainty (*i.e.,* randomness) in an information source. This definition is similar to entropy in statistical mechanics, which measures the degrees of disorder (*i.e.*, randomness), differing only a factor of k_B the Boltzmann factor. In other words, mutual information measures the reduction in the uncertainty of one variable (X) after knowing the value of the other variable (Y). For example, what is the uncertainty about a pair of duct tape's physical fit after knowing they come from the same brand? Please refer to **Appendix A** for a more in-depth discussion of the relationship between mutual information and entropy.

4.3.3.3. Data Preparation

Mutual information (MI) is used to analyze data from previous studies by Prusinowski et al. [8] and Andrews et al. [9]. The data is stored in tables within an Excel template, as explained in section 2.2. Each table represents a comparison between two samples, with each row corresponding to a bin on the comparison edge and each column representing observable features. One column, "Area fit code," indicates the analyst’s decision (0, 0.5, or 1) regarding whether the two samples fit together. The other columns describe the major features for comparison used by the analysts and the analyst’s opinion regarding the influence of that feature on the bin decision.

To determine the importance of each feature in the examiner's decision-making process, all comparison tables are concatenated into a single table and mapped by text values to numbers using a lookup table (see **Appendix Table 4A-1** for details). The mutual information of each feature is then calculated concerning the area fit code, which represents the decision the analyst made during the process. **Figure 4-1** provides a schematic of the process followed. To assess and calculate the mutual information, the "mutual_info_classif" function from the scikit-learn Python package is used [18].

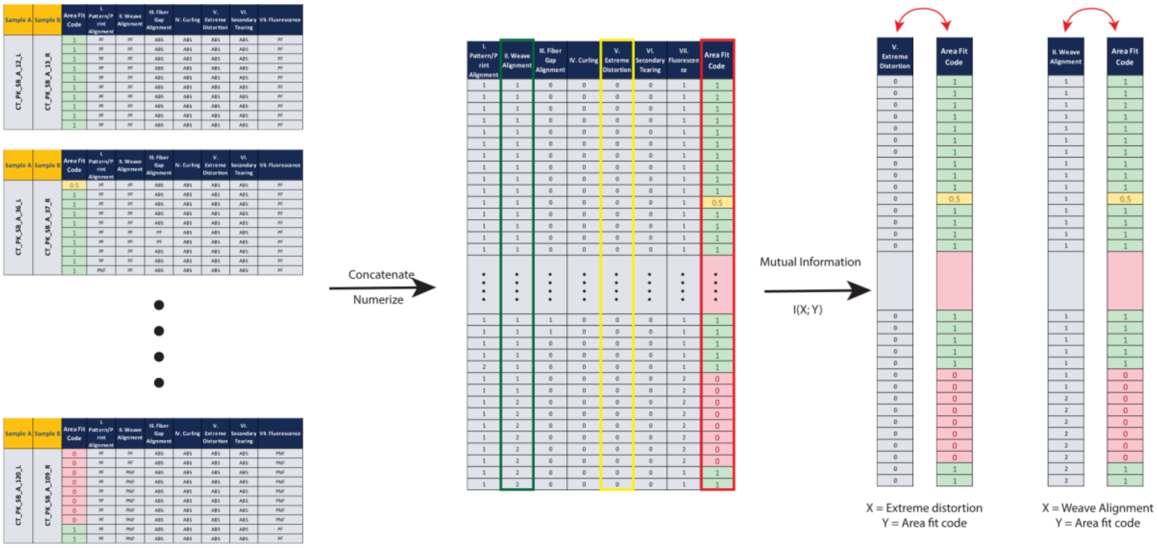


Figure 4-2 Process of data manipulation for the calculation of mutual information.

4.3.4. Decision Tree

Decision trees, also called classification and regression trees (CART) are a type of supervised machine-learning algorithms. The decision tree breaks down a complex decision-making process into smaller, more manageable steps. This is done by recursively partitioning the feature space into a set of rectangles and assigning a constant (e.g. fit or non-fit) to each. This can be represented by one tree with multiple leaves each representing one region in the feature space [10]. When dealing with two variables it is easy to represent the classification with rectangular divisions, however, in larger dimensions with many features this task becomes impossible and the feature space partitioning can be fully described by a single tree. Creating the decision tree is equivalent to finding the optimum partitioning for an n-dimensional feature space. The advantage of using a decision tree in this study is that it allows for breaking down a complex decision-making process into simpler, more manageable steps, making it easier to analyze and understand the evidence.

Finding the optimal partitioning of the feature space is shown to be a nondeterministic polynomial-time complete (NP-complete), which it is a type of computational problem that no efficient solution algorithm has been found to solve [19], and so scientists use different approaches to find locally optimal partitioning. The methods to quantify the quality of each split include misclassification rate, entropy, and Gini index [10, 20]. This study uses the `DecisionTreeClassifier` function implemented in the Scikit learn package [18]. Entropy is used as the criterion for growing of the decision tree. 80% of the data is used for the training stage but the final performance metrics are calculated using the entirety of the data.

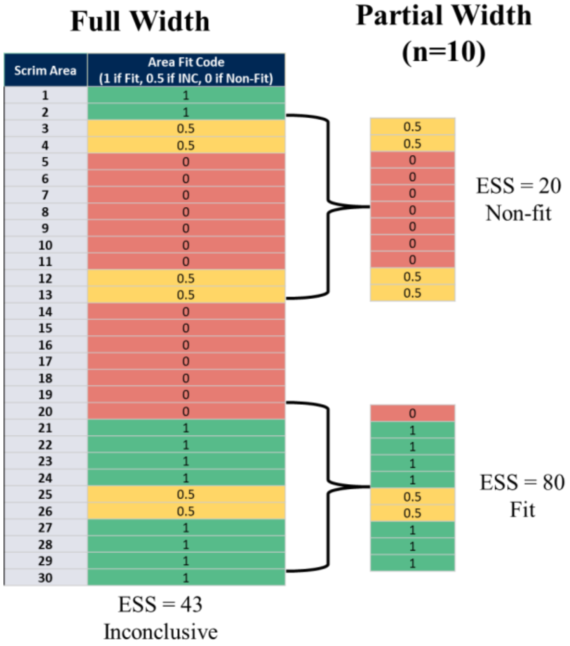
4.4.Results and Discussion

4.4.1. Minimum width on Partial Comparison Items

As described previously, it is commonplace in casework to have samples submitted for physical fit examination that are only partial fragments. When this occurs, it is often left up to the analyst to determine if enough of the fragment length is available and if there are enough distinct features to make a reasonable decision. With a lack of foundation in this area or guidelines to support analysts, this study aims to estimate the edge length necessary to reliably conduct a physical fit analysis for materials for which there is a substantial amount of available quantitative data: tape and textiles.

A critical aspect of the data extraction for the partial width comparison analysis is to ensure that despite a random starting point on the sample, the bins following are consecutive. This is intended to reflect realistic casework samples. The partial width extraction is performed five times per material type to provide additional data and insight into the variability in edge alignment of the material datasets. **Figure 4-2** describes an example of a partial width selection conducted by the model in this study, with the number of bins being 10, out of the 30 bins corresponding to the complete edge. In this example, while the full-width pair is classified as inconclusive, the two separate partial-width selections correspond to a non-fit or fit decision, depending on the starting point selection.

It has previously been reported that the construction, manufacturing, and method of separation of a sample can impact the quality of the sample edge for comparison and what features are left behind. In tapes, the edge morphology, features left during separation, and comparison performance are influenced most heavily by the quality (grade) of the roll [8]. As such, the tape sets evaluated in this study are also separated by grade. The previous research reports that scissor-cut pairs generally retain sufficient features for comparison, while hand-torn edges tend to result in more distorted but distinctive edge patterns. Generally, stretching deforms some of the features, impacts the ESS, and causes increased variability, but not to the extent of producing significant changes in the overall accuracy of the method [8]. In textiles, separating and constructing the fabric samples contributed the most towards differences in ESS values. However, the impact is less substantial than in tapes [9]. For this study, the sample set is divided by separation method, as in the original research, the separation method contributed specific features critical for observation and decision-making in physical fit analysis.



Scrim Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)
1	1
2	1
3	0.5
4	0.5
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0.5
13	0.5
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	1
22	1
23	1
24	1
25	0.5
26	0.5
27	1
28	1
29	1
30	1

0.5
0.5
0
0
0
0
0
0
0
0.5
0.5

0
1
1
1
1
0.5
0.5
1
1
1

Figure 4-2. Diagram depicting the random selection and calculation of performance of the ESS method applied to a partial sample width. Selection of two different starting points on the sample pair edge results in significantly different outcomes. The bins contain the overall bin code, colored green (fit, 1), yellow (inconclusive, 0.5), or red (non-fit, 0). The ESS for full width is 43, while the ESS of two randomly selected edge portions (33% each) lead to different ESS outcomes (20 and 80, respectively).

Starting with the duct tape samples, following the calculation of error rates of all possible partial width lengths, the trends were observed compared to the full-width comparisons of the tape samples. With LQ samples, the accuracy of the entire width of tapes by human examination is relatively high, ranging above 90% [8]. The results by partial width yield similar accuracies for the LQ scissor-cut set. However, it is notable that for the hand-torn sets, the method's accuracy decreases and has a wider spread of variability, which becomes more evident when the length of the tape is less than 30-40% of the width (**Figure 4-3**). This is likely due to the increased distortion observed in these hand-torn separated samples. **Figure 4-3** shows the accuracy of the ESS method when applied to portions of the low-quality tape subsets. The solid line represents the beta-regression model of the accuracy from the five iterations performed for data extraction. The shaded portions represent the accuracy variability for each percentage of the tape width. When considering these variability ranges, the results indicate that to maintain an accuracy above 90% for this dataset, the sample must contain at least 35% of the original width.

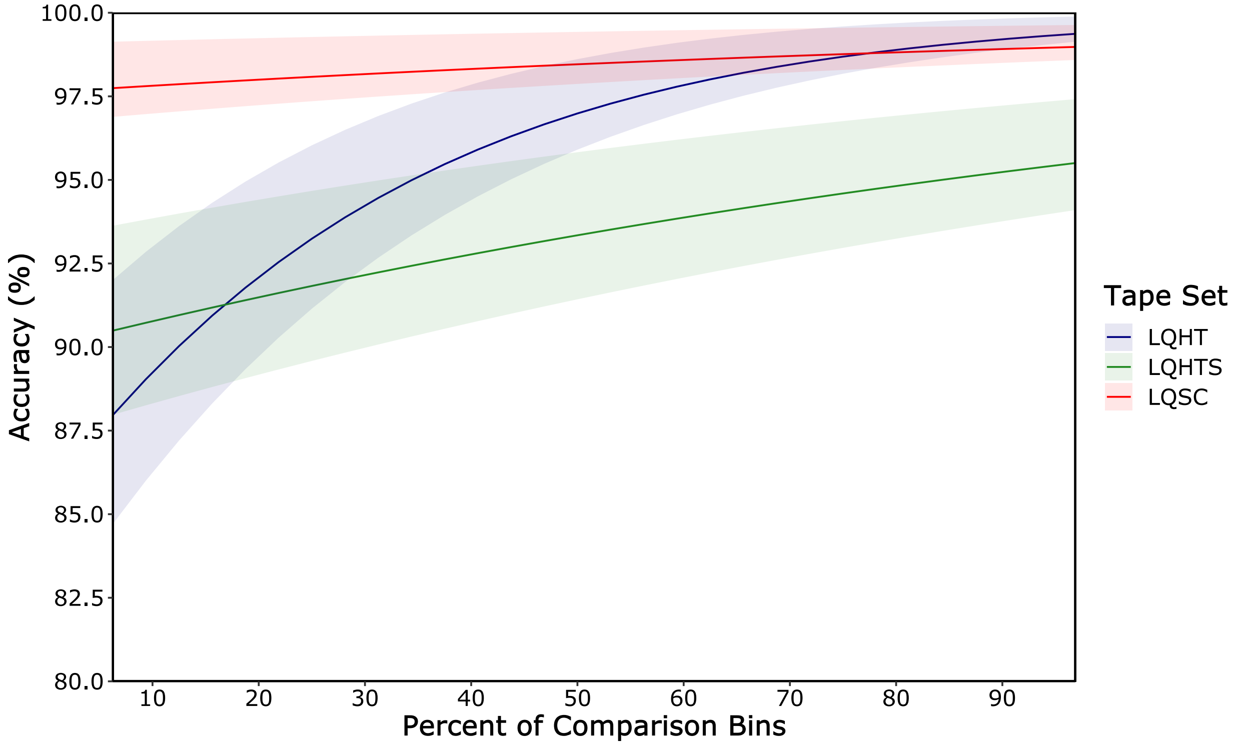


Figure 4-3. Accuracy of the ESS method as applied to partial widths of the low-quality tape subsets.

When considering the high-quality samples, the general observations made during the analysis of the full-width samples persist on the partial edges. The scissor-cut samples demonstrated high accuracies, with significantly less variability in reported ESS compared to the hand-torn samples from the same roll. This has been accredited to the reduced distortion and observation of additional features not easily observed in hand-torn samples. Conversely, the high-quality hand-torn sets demonstrated much lower accuracy rates than all other sets, which resulted from a much higher rate of distortion observed in these samples, from the tearing process and additional uncertainty introduced by the required adhesive removal process. **Figure 4-4** shows the accuracy of the ESS method when applied to portions of the high-quality tape subsets. As described for the LQ set, the solid line represents the beta-regression model of the accuracy from the five iterations performed for data extraction. The shaded portions represent the accuracy variability for each percentage of the tape width. Here, the HQSC set maintains high accuracy regardless of the missing portions of the sample, with a slight increase in variability as the edge size decreases.

On the other hand, the HQHTS set has a lower accuracy even for full-size items, caused mainly by a larger number of false negative results. The accuracy decreases from about 84% to up to 74% as the size decreases, raising a cautionary flag. Since our studies have shown that ESS values under 80% don't provide the strongest support for a fit, we recommend not using physical fit examinations as the only assessment of high-quality hand-torn duct tape when partial evidence is submitted for comparisons. In other words, a physical fit examination of partial edges on this tape category can be done to rule out a potential fit. Still, its interpretation should state potential sources of error and should be followed for additional chemical and physical examinations based on its class characteristics.

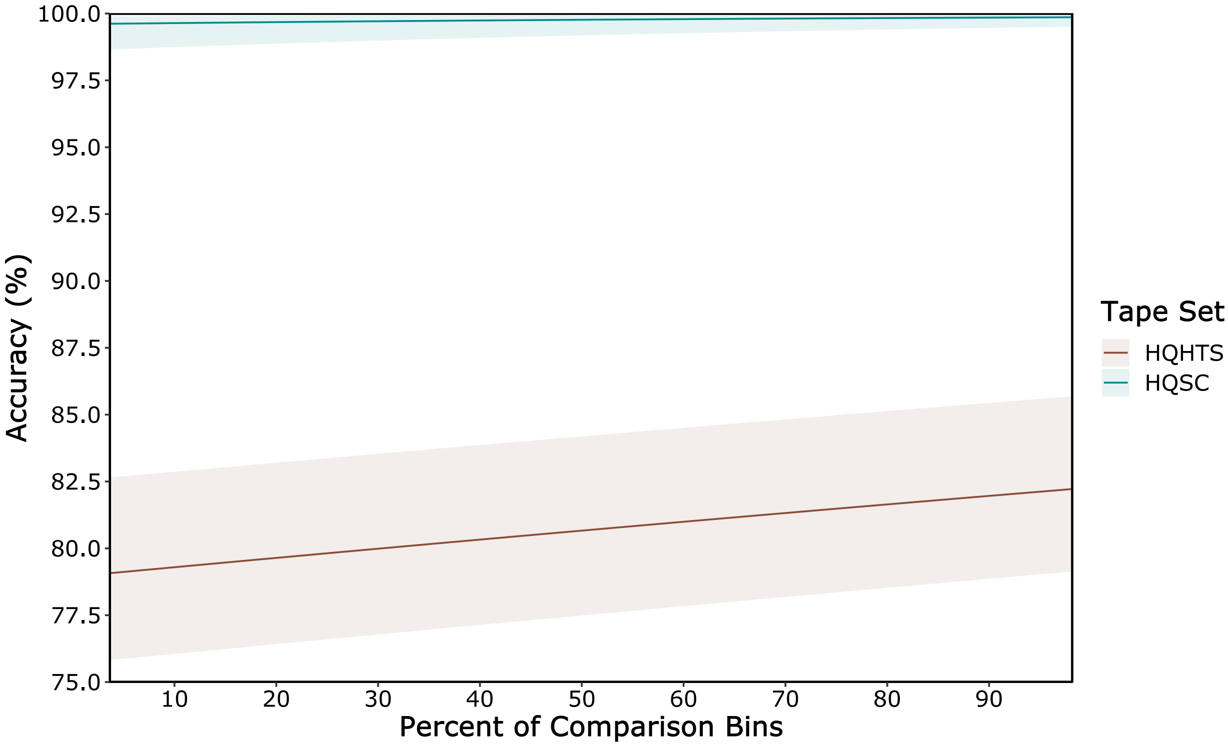


Figure 4-4. Accuracy of the ESS method as applied to partial widths of the high-quality tape subsets.

With the low-quality tapes, the scissor-cut sets demonstrate the lowest variability in accuracy, and the more tape that is present, the higher the accuracy becomes overall. In both sets, the accuracy of the stretched set is the lowest at full-size comparisons, but the accuracy of the high-quality set is more influenced by size.

The trends based on the separation method are like those observed in tapes for textiles. The hand-torn samples are generally more variable in accuracy across all partial widths, especially when the length is less than 30-40% of the width (**Figure 4-5**). On the other hand, the stabbed samples show less variability and higher overall accuracy, consistent with the findings of Andrews et al. [9]. Once again, this is likely due to the increased distortion observed in the hand-torn samples. **Figure 4-5** shows the accuracy of the ESS method when applied to portions of the textile dataset. Considering the results of these textiles, to maintain an accuracy above 90% for this dataset, the sample pair must contain at least 40% of the original width.

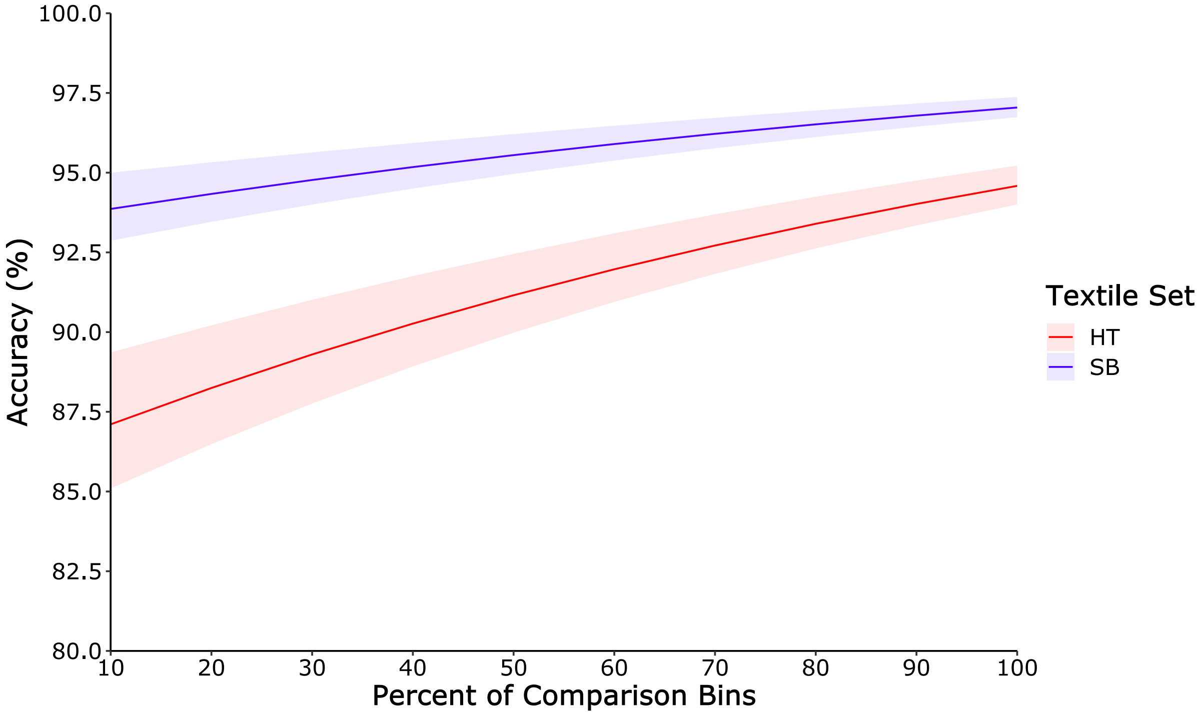


Figure 4-5. Accuracy of the ESS method as applied to partial widths of the textile samples. HT represents hand-torn samples, while SB represents stabbed samples.

There are some limitations to this approach to evaluating potential length effects. For example, in the cleaner edges seen in the stabbed textiles and the scissor-cut tapes (such as the HQSC and LQSC sets), **Figures 4-3, 4-4, and 4-5** indicate high accuracy, even at less than 10% of the comparison bins being used. This is likely due to the highly distinctive features observed in these sets. In tapes, cutting the sample preserves several features, such as the alignment, or evident misalignment, of dimples that have been cut through. In textiles, the edge is less likely to unravel and stretch during separation if cut by a blade or instrument compared to the pulling and stretching induced by tearing by hand. A critical note of the textile samples used in this study is that the sample edges are approximately two-inch wide, each cut from a larger garment such as a shirt or pants. So, in this case, 40% of an edge is only a very small proportion of the original garment. That provides an additional limitation to the interpretation of a partial textile fragment.

In actual casework, an analyst is unlikely to compare edges with less than 10% of the edge available for comparison. However, these results may help indicate when additional analysis should be performed if a physical fit is found on a partial tear. These results suggest that a partial fracture may be possible to determine at a width of at least 30-40% of the width of either tape or textiles while yielding accuracies above 90%. However, extreme caution should be taken in interpreting the value of partial comparisons. These results provide an experimental foundation on the effect of size but should not be taken as a fixed threshold.

4.4.2. Mutual information

With new attention on systematic approaches to physical fits, one of the remaining questions is identifying which features hold the most value in comparing edges. Human analysts performing examinations on sample edges may rely on specific features or criteria in their analysis that help to inform their decision. To that end, one of the main goals in developing systematic methods for physical fit comparisons is defining comparison features that can lead to standardization in the discipline.

For duct tapes, research has identified several relevant features inherent to tape manufacturing or construction that are known to remain relatively consistent along a full roll. This includes the presence or absence of calendaring striations and backing dimples, the number of warp and weft fibers, and the overall direction of the fiber weave. Other features, such as distortion, missing material, and corresponding protruding fibers and gaps, are introduced during separation [8]. In textiles, features inherent to the samples include the design and construction of the fabric, as well as any fluorescence of the fibers. Features contributed by separation include alignment of gaps, alignment of yarns across the fracture, secondary tearing, and distortion [9].

A critical aspect of feature recognition in physical fit comparisons is understanding how a feature may influence a decision. An assumption made here is that not all features share the same weight. For example, duct tapes have fiber reinforcement pressed in between the adhesive and backing layers. Due to how duct tapes are manufactured, the number of warp and weft fibers per square inch remains consistent throughout an entire roll of tape. The same is true of the orientation and location of the fibers along the edge. As such, samples in this study were observed to have true non-fitting pairs for which the microscopic alignment of the warp fibers corresponds and indicates that the samples fit together. If the warp fibers did not correspond, however, the value of that feature is substantially stronger for indicating an opinion of non-fit for that bin.

The importance of the different features is explored regarding the mutual information of the feature and the overall decision. The tape data is divided by several factors, including separation method and quality, to evaluate what features contribute most to a given subset of samples. **Figure 4-6** shows bar plots representing the mutual information for each tape comparison feature by separation method. Small value bars, such as those observed in missing material, indicate that the feature is rarely present or not particularly informative for making a decision. Some differences between hand-torn and scissor-cut samples stand out regarding the feature's importance. First, the alignment of severed dimples is substantially more informative in scissor-cut samples while not being observed in hand-torn. On the other hand, the scrim weave pattern and consistency in the weft fiber pattern are informative for both separation methods but more so for scissor-cut samples.

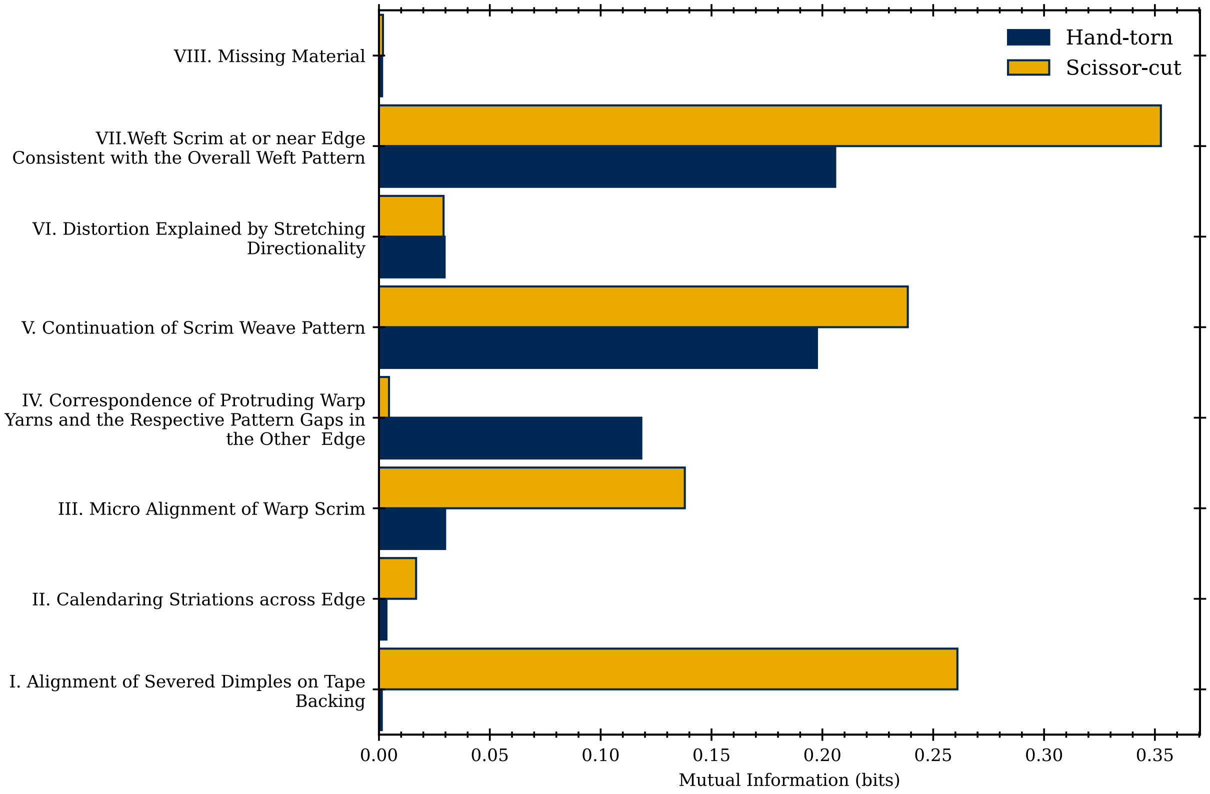


Figure 4-6. Barplots representing the mutual information of the tape features by separation method. The larger the bar, the more value the feature has for comparisons.

As seen in **Figure 4-6**, the alignment of severed dimples is critical for scissor-cut samples but not for hand-torn pieces. This results from an interesting observation made about the samples during tearing. In the tapes used for this study, the dimpling on the backing tends to align with the weft fibers on the scrim side. As a result, when torn by hand, the samples do not generally tear across a weft yarn but will tear on the side of the yarn, where the material provides less resistance to tearing. As such, the torn edges do not generally cross a dimple pattern. That is not true of the scissor-cut samples, for which the edge separation is dictated by the direction the users choose for the cut. Severed dimples and weft fibers play a critical role in comparing these samples. It is more common to see cuts that cross over the fracture, leaving multiple distinctive aligning features (or misaligning features in case of non-fits). These transversal cuts are relatively distortion-free and influence the ‘consistent weft scrim at or near the edge’ feature and the ‘continuation of scrim weave.’ These features are still present in hand-torn samples but carry more value in scissor-cut samples as they tend to deform when torn. The same applies to aligning dimples, commonly seen in the scissor-cut samples, adding substantial weight to a fit or non-fit decision for a particular bin. The opposite is true of corresponding protruding warp yarns, which are more influential for hand-torn samples. This feature is not commonly observed in the scissor-cut samples as they tend to originate when the warp yarns stretch before fracture, which does not occur when the tapes are cut. Interestingly, corresponding protruding warp yarns are only informative in the LQHTS set, despite there being another hand-torn set. This is due to the nature of the tapes. The higher-quality tape has a much higher number of scrim fibers. Specifically, it has a double-warp pattern with two warn fibers side-by-side in the alternating direction, whereas most LQ tapes only have one warp yarn. This contributed to the higher strength of the HQ roll and likely prevented the yarns from stretching as they tore. Conversely, the lower-quality tape is more prone to distortion and has fewer fibers overall. As such, protruding fibers are more common in these samples.

Figure 4-7 shows the mutual information of the features divided by set. In this case, only HQHTS, HQSC, and LQHTS are shown. Some of the same trends apply, as seen in **Figure 4-6**. The alignment of severed dimples is critical for the HQSC set, while the continuation of the scrim pattern is most critical for HQSC, followed by HQHTS and then LQHTS. Scrim weave and corresponding weft scrim are seen as influential in all sets.

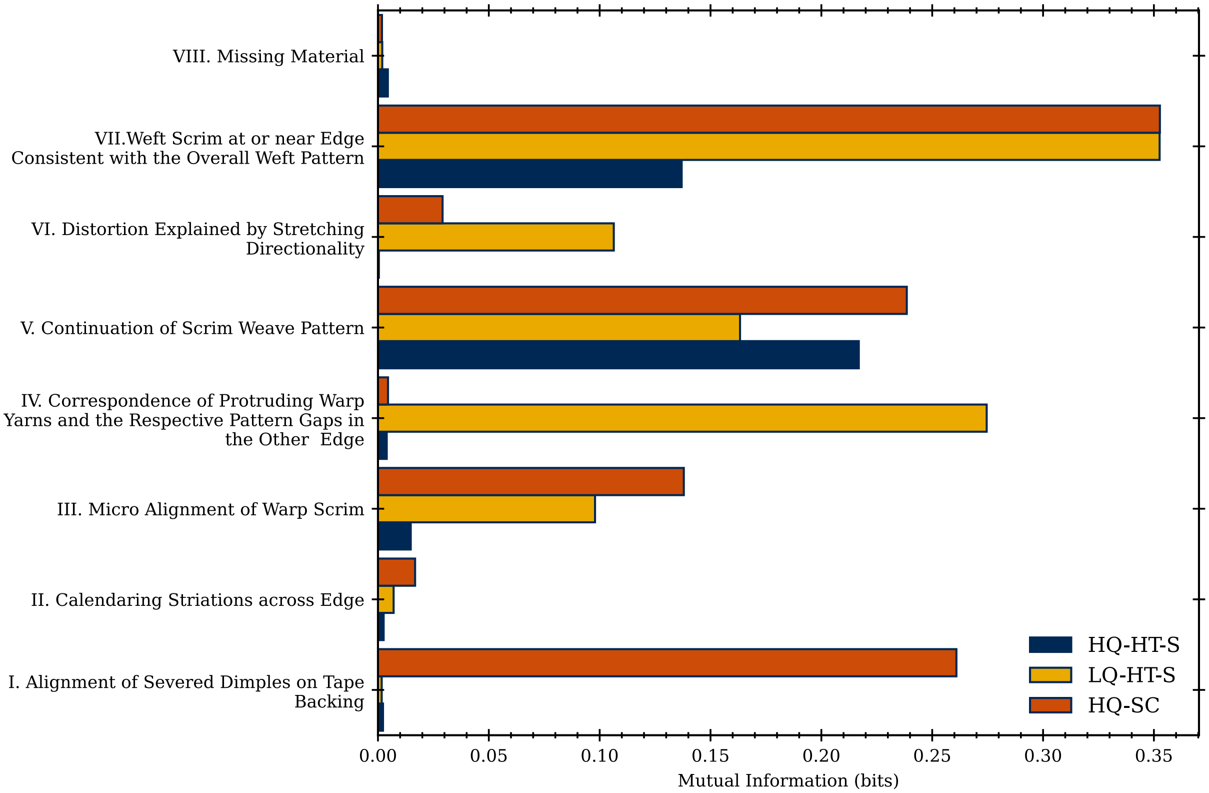


Figure 4-7. Barplots representing the mutual information of the tape features by sample subset. The larger the bar, the more value the feature has for comparisons.

Meanwhile, the mutual information of the textile dataset demonstrates similar trends. Some features hold more value in a stabbed sample or hand-torn sample. As seen in **Figure 4-8**, some features are only observed in the hand-torn samples, including secondary tearing and extreme distortion. The small bar indicated for both these features suggests that while observed in edges torn by hand only, their occurrence is still uncommon in the overall population. Hand-torn pair comparisons also rely slightly more on gap and yarn alignment than stabbed samples, because the more violent nature of hand-tearing tends to pull yarns away from the comparison edge when the fragments are pulled apart. Meanwhile, construction and print alignment are slightly more informative in stabbed sample edges. This is to be expected because these features are easier to observe on the less-frequently distorted stabbed edges. It is worth noting that the variation in the value of a feature is not as diverse between separation methods as in duct tapes, further reinforcing the need for separate features and criteria for the material in question.

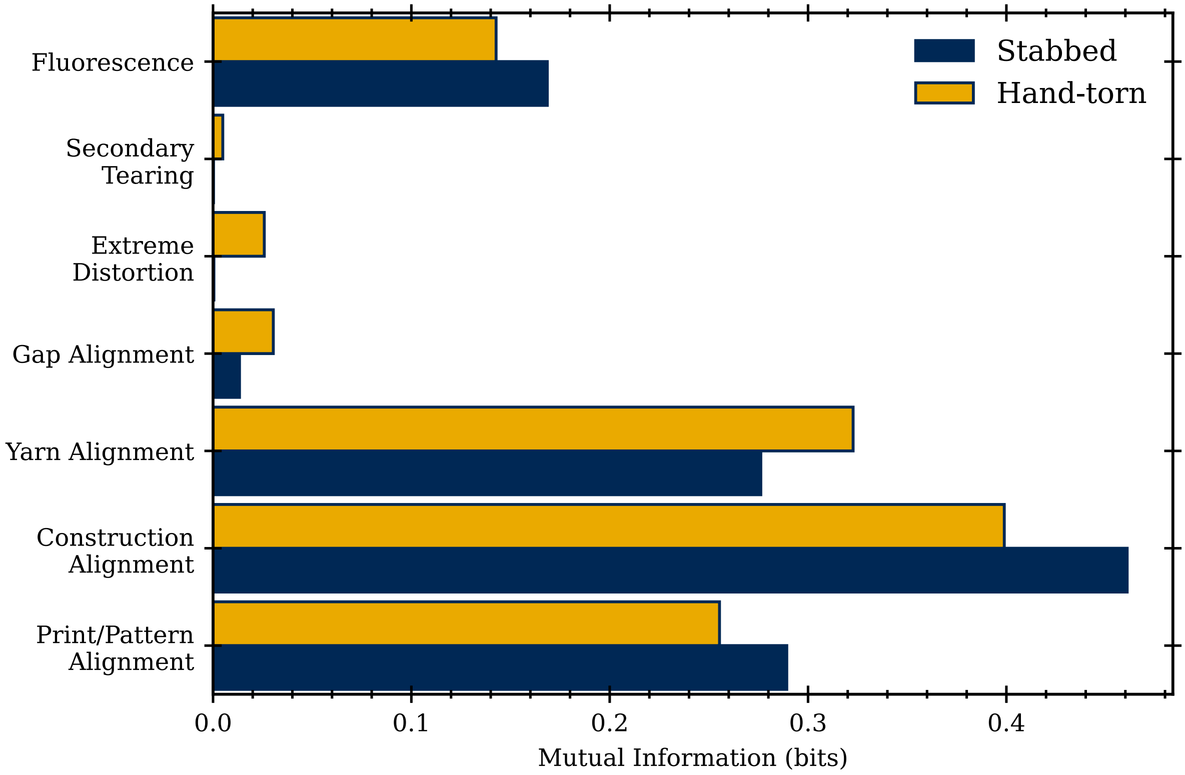
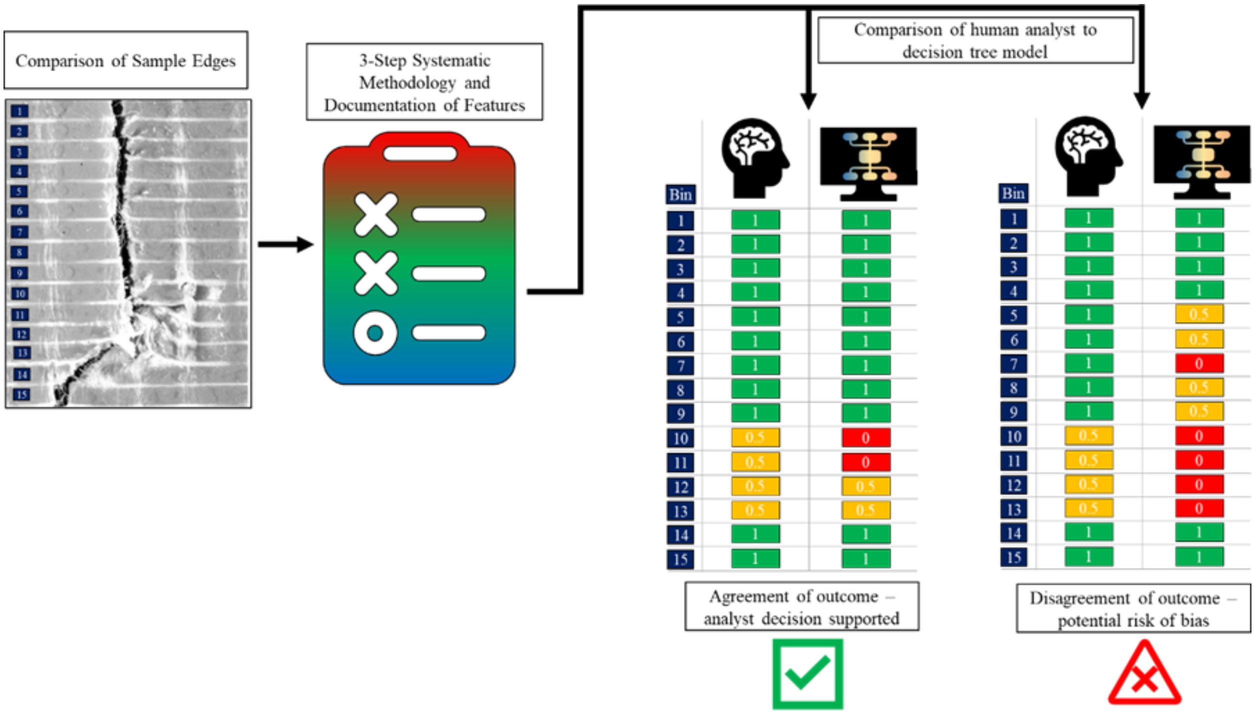


Figure 4-8. Barplots represent the mutual information of the textile features by separation method. The larger the bar, the more value the feature has for comparisons.

The feature analysis shows that a comparison feature's occurrence and informative value depend on the fracture’s circumstances, including but not limited to separation method and quality (grade). This study shows that the alignment or misalignment of a feature can hold a particular weight toward a fit or non-fit decision. This information is helpful for analysts as it provides a deeper understanding of the relative relevance that a feature may have in a fit examination. The next step towards utilizing this information more effectively is to assign quantitative values to the feature patterns to assist analysts in making decisions.

4.4.3. Decision tree

A primary goal of this study is to use the information gained from the systematic comparison of a large dataset to assist in analyst comparisons. To that end, a method using a decision tree model is presented here. The decision tree is a machine-learning model that follows a set of rules to make decisions. In this case, the decision tree uses information extracted from the comparison templates to learn from. The decision tree evaluates the pattern of features from every bin and uses the outcome of that pattern to make a decision, following a similar process to the human analysis. The purpose of this decision tree model is not to replace the human input but to supplement it with additional objective analysis to reduce the risk of bias in a comparison. For a given pair of samples, some variability in the decision reported by the analyst and the decision tree model can be expected. However, that variability should not be extensive enough to change the outcome of the decision substantially. If it does, that may indicate a risk of bias in the comparison or an opportunity to assess the rationale for those differences in the analytical process and document it with transparency. **Figure 4-9** shows a diagram of the intended comparison process.


Comparison of Sample Edges
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
3-Step Systematic Methodology and Documentation of Features
Comparison of human analyst to decision tree model
Agreement of outcome – analyst decision supported
Disagreement of outcome – potential risk of bias

Bin		
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	0.5	0
11	0.5	0
12	0.5	0.5
13	0.5	0.5
14	1	1
15	1	1

Bin		
1	1	1
2	1	1
3	1	1
4	1	1
5	1	0.5
6	1	0.5
7	1	0
8	1	0.5
9	1	0.5
10	0.5	0
11	0.5	0
12	0.5	0
13	0.5	0
14	1	1
15	1	1

Figure 4-9. Diagram demonstrating how the process of human analysis of a pair of tapes would be performed and compared to the results of the decision tree model.

The decision tree for duct tapes is trained using the tape dataset samples and the respective templates. However, the initial performance of the decision tree results in low accuracy due to contradicting decisions for feature patterns. An example of a feature pattern with the respective bin decision is shown in **Table 4-2**, where the pattern from feature edge I to VIII can be represented as (A, A, PF, A, PF, A C, NA) where A is absent, PF, present indicative of a fit, C consistent and NA not applicable. There are occurrences of that same pattern of the eight features in the human analysis that results in different decisions (i.e., fit or non-fit). This is because the analysts weighted in their minds each feature according to the observed characteristics, which is information not captured by the algorithm. In other words, the pattern sequence only provides one level of information to the algorithm. Still, it does not include a higher level of information on the relative relevance of those features. These "contradictions" of outcomes from similar patterns contribute to the poor performance of early decision tree modeling. To clarify these cases, a human analyst blindly reviews those patterns and provides a decision for a bin based solely on level-one information about the feature patterns.

An example is shown in **Table 4-3**. These revised decisions are used to retrain the decision tree model. The structure of the tape decision tree is available in **Appendix 4-B**.

Table 4-3. Example of feature pattern and bin decision assigned by an analyst.

Feature	**Response**
I. Alignment of Severed Dimples on Tape Backing	Absent
II. Calendaring Striations across Edge	Absent
III. Micro Alignment of Warp Scrim	Present - indicative of fit
IV. Correspondence of Protruding Warp Yarns and the Respective Pattern Gaps in the Other Edge	Absent
V. Continuation of Scrim Weave Pattern	Present - indicative of fit
VI. Distortion Explained by Stretching Directionality	Absent
VII. Weft Scrim at or near Edge Consistent with the Overall Weft Pattern	Consistent
VIII. Missing Material	Not applicable (no missing material)
BIN DECISION	**FIT**

Following the re-training of the model, performance improves. **Figure 4-10** shows the breakdown of performance for each tape subset. The model generally performs equally as well as the human analysis of the samples. An exception is seen in the HQHTS set, which improves true positives compared to the human analysis. but also produces a few false positive results, which are not seen in the human analysis. The HQHTS set is a particularly complex set, both for human and computational analysis. Despite being hand-torn, the samples generally separate with straight edges that do not leave distinct edge morphologies. In addition, removing the adhesive from the edge to allow observation of the edge features added some uncertainty to the decision-making process, as the fibers may have shifted from their original positions. The shifting fibers could cause alignment in non-fitting pairs or vice versa. As such, the human analysis results in higher false negative and inconclusive rates than the decision tree analysis.

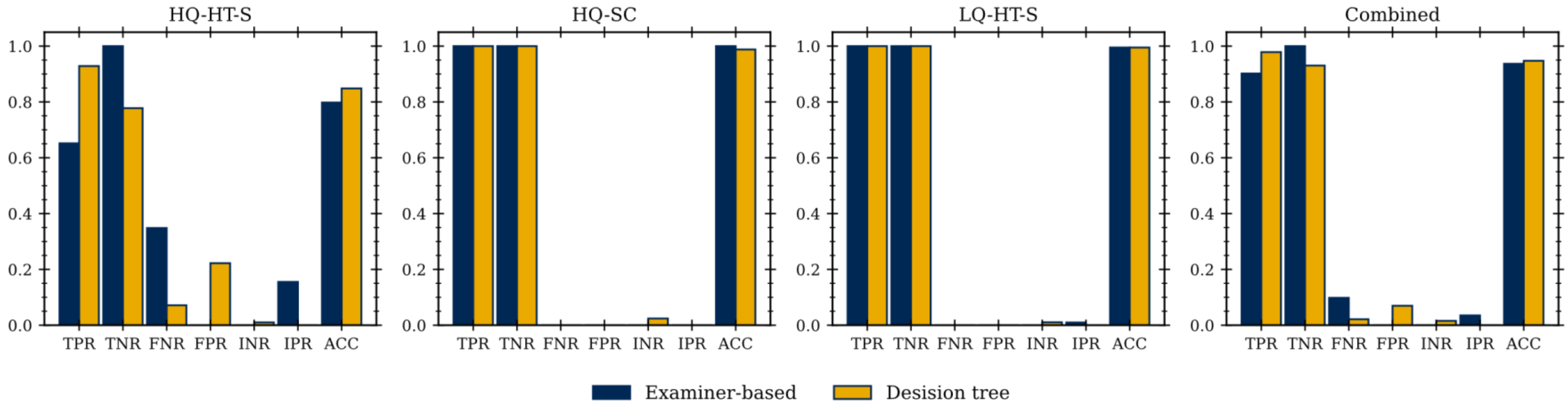


Figure 4-10. Performance of the decision tree against the human analysis for each duct tape sample set. The performance rates included are the true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR), inconclusive rates for both true non-fits (INR) and true fits (IPR), and the accuracy (ACC).

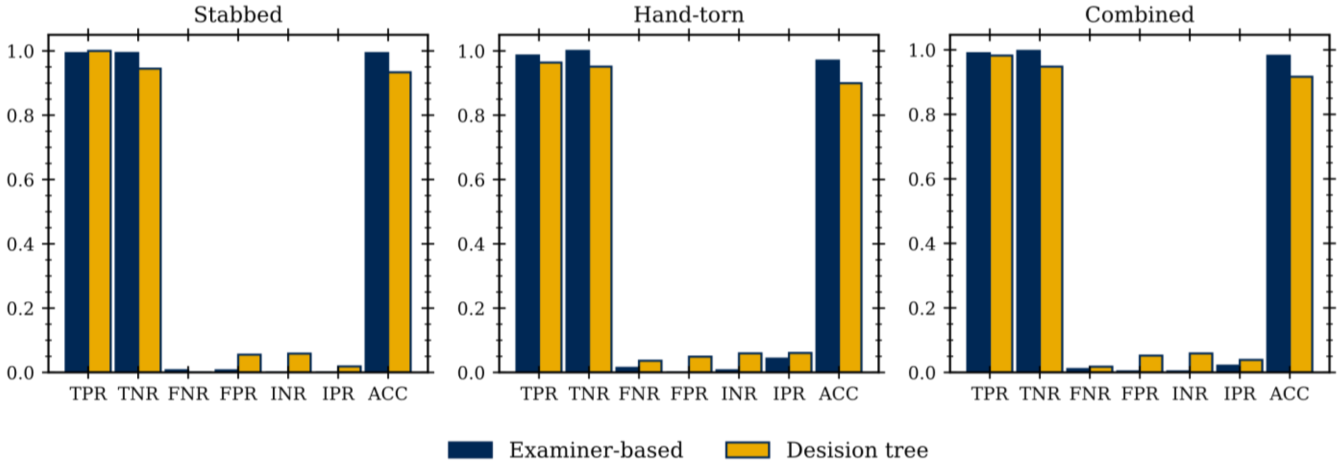


Figure 4-11. Performance of the decision tree against the human analysis for each textile sample set. The performance rates included are the true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR), inconclusive rates for both true non-fits (INR) and true fits (IPR), and the accuracy (ACC).

The decision tree for textiles is also trained on the respective data and templates. Like the initial tape model, the initial textile model suffers from reduced performance due to contradicting decisions for feature patterns. These cases are reviewed and re-assigned a decision by a human analyst based on the patterns of decisions, which improves overall performance. The model performs equally as well as the human analysis of the samples. However, the model does produce false positives and is inconclusive across both hand-torn and stabbed samples.

These results indicate that the decision tree model shows significant potential as a tool to help in the decision-making process for physical fit comparisons. While caution is needed regarding the chance of false identifications, if used in tandem with human analysis, the tool could help identify samples where a further examination is recommended if the model outcome disagrees with the analyst. It also provides additional information that adds transparency and support to the conclusion. For instance, confidence and objectivity can be demonstrated if the algorithm agrees with the analyst's decision. The algorithm removes the judgment from the decision process. It minimizes the risk of bias from the prior information from the analyst's observations of the question and known overall alignment or misalignment.

4.5.Conclusions

This study demonstrates the value of modeling data and feature extraction to provide additional information for forensic physical fit analysis. The quantitative data provided by the ESS approach can be used to model and assess the partial widths of tapes and how only having a portion of the edge for comparison could influence the method's accuracy. While, ideally, forensic practitioners would have the entire length of a fracture during comparison, the whole width of the sample may only be available in some cases. The fragments may have separated or been lost during the commission of the crime or in evidence recovery. This study addresses this important dilemma. The results demonstrate that, based on the sample sets used here, a length of at least 35% of the edge will provide acceptable accuracy in most samples. The relative size is highly dependent on the tape specimen, and samples more prone to distortion or lacking distinct features may not be suitable for comparison with only a small portion of the edge available.

In addition, the method and templates derived from analysts' examinations are used to identify specific features of the tape edges that contribute the most to a decision. In contrast, the decision tree model evaluates the value of those features quantitatively and provides a tool that analysts can use to support their decisions. The mutual information analysis of the sample templates indicates several trends in feature value for the different subsets. While Prusinowski et al. [8] and Andrews et al. [9] demonstrate that the separation method does not have much effect on the tapes and textiles ESS scores provided by the method, this study shows that the features recognized in the samples and their relative weight vary considerably. In scissor-cut tape samples, the alignment of severed dimples is much more important for making decisions, while the feature is rarely observed in hand-torn samples. In contrast, corresponding warp yarns are usually only relevant in the hand-torn samples. In addition, the continuation of the scrim weave and consistency of the weft fibers, while observed in hand-torn, hold more value in scissor-cut pairs. The mutual information approach demonstrates that the tape quality considerably impacts the features observed and their importance in comparison. The low-quality tape overall contributed to more distortion and distinct edges than observed in high-quality samples, leading to features such as distortion and exposed warp fibers being particularly important in those samples. Meanwhile, construction and design alignment in textiles contribute more to stabbed samples, while gap and yarn alignment hold more influence in hand-torn samples. Two features only observed in hand-torn samples (extreme distortion and secondary tearing) indicate those two can have substantial value for a decision.

Finally, the decision tree model trained on the sample sets demonstrates consistent accuracy with the human analysis in most datasets. While it provides more true positives in the complex HQHTS tape set, it also results in a much higher false positive rate for that dataset and both textile subsets. As a result, the outcome of this method indicates that while it should not replace human analysis entirely, it holds strong potential as a supplementary tool. Samples independently evaluated by both the human analysis and the decision tree that arrive at different decisions can be assessed in more depth, such as by additional physical examination by another independent analyst or through further chemical analysis to support the findings.

Overall, this study increases the body of knowledge and scientific foundations of physical fit examinations of duct tapes and textiles. It addresses two important gaps in this field. First, it estimates how much the sample size can be absent for comparison before yielding unreliable results. Second, for the first time, it offers experimentally based information that can assist practitioners in understanding the weight a particular feature can hold in the examination and provides a means to document the decision-making process more objectively and transparently. Future work in this area can include the addition of more sample subsets to provide additional data for mutual information and the decision tree model. In addition, we intend to introduce the decision tree model as an open-access source to allow practitioners to evaluate the tool to help support their decisions.

4.6.References

1. OSAC. Oct 27, 2022. Lexicon web. Retrieved November 28, 2022, from https://lexicon.forensicosac.org/

2. OSAC 2022-S-0029 Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials. Trace Materials Subcommittee. November 2022.

3. OSAC 2022-S-0015, Standard Guide for Forensic Physical Fit Examination. Trace Materials Subcommittee. November 2022.

4. Gross, S. Physical Fit Task Group – Trace Materials Subcommittee of NIST-OSAC. Survey of Physical Fit Protocols. 2019. Unpublished Survey, presented at NIST-OSAC Trace Subcommittee Meeting, March 2020, OK.

5. Dror IE. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Anal. Chem.* 2020. 92: 7998–8004.
6. Brooks E, Prusinowski M, Gross S, Trejos T. Forensic Physical Fits of Trace Materials: A Review. *Forensic Science International.* August 2020. Available online May 29th, 2020. https://doi.org/10.1016/j.forsciint.2020.110349
7. Prusinowski M, Brooks E, Trejos T. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. *Forensic Science International*. 2020. https://doi.org/10.1016/j.forsciint.2019.110103
8. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – physical fits of torn and cut duct tapes. *Forensic Science International*. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567.
9. Andrews Z, Prusinowski M, Neumann C, Trejos T. Assessing physical fit examinations of stabbed and torn textiles through a large dataset of casework-like items and inter-laboratory studies. *Journal of Forensic Sciences*. Under review.
10. Murphy K. Probabilistic Machine Learning: An introduction. MIT Press. 2022.
11. Ross BC. Mutual information between discrete and continuous data sets. *PLoS One*. 2014 19; 9(2): e87357. doi: 10.1371/journal.pone.0087357.
12. van Dijk CD, van Someren A, Visser R, Sjerps M. Evidential value of duct tape comparison using loopbreaking patterns. *Forensic Sci Int*. 2022 332:111178. doi: 10.1016/j.forsciint.2022.111178.
13. Spaulding J, Picconatto G. Characterization of fracture match associations with automated image processing. *Forensic Science International*. 2023. 342. 10.1016/j.forsciint.2022.111519
14. Prusinowski M, Brooks E, Neumann C, Trejos T. Forensic interlaboratory evaluations of a systematic method for examining, documenting, and interpreting duct tape physical fits. *Forensic Chemistry*. Published online March 2023. 34: 100487. https://doi.org/10.1016/j.forc.2023.100487
15. Andrews Z, Prusinowski M, Nguyen E, Neumann C, Trejos T. Assessing physical fit examinations of stabbed and torn textiles through a large dataset of casework-like items and inter-laboratory studies. *Journal of Forensic Sciences*. Under preparation.
16. Tavadze P, Lang L, Prusinowski M, Andrews Z, Trejos T, Romero A. Using convolutional neural networks to support examiners in duct tape physical-fit comparisons. *Forensic Science international*. Under review.
17. Shannon CE. A Mathematical Theory of Communication. *The Bell System Technical Journal.* 1948. 27: 379-423.
18. scikit-learn. Machine learning package in Python. https://scikit-learn.org/stable/
19. Thomas A, Cover T. Elements of information theory. United Kingdom: Wiley, 2006.
20. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Netherlands: Springer New York. 2009.

Appendix 4-A

Mutual information

Mutual information is a measure of the relationship between two random variables. Mutual information can be expressed in terms of entropy, which requires understanding conditional entropy. Conditional entropy measures the amount of uncertainty that remains in a random variable Y after observing X. It quantifies the uncertainty of Y given that X is known and is denoted as $\mathbb{H}(Y|X)$. The mathematical formula for calculating $\mathbb{H}(Y|X)$ is:

$$\mathbb{H}(Y|X) = -\sum_{x}\sum_{y} p(x,y)\log p(y|x) \tag{3}$$

Using this formula, we can rewrite equation 1 in terms of conditional entropy:

$$\mathbb{I}(X;\, Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \tag{4}$$

This equation shows that mutual information between X and Y can be interpreted as reducing uncertainty about X after observing Y, and vice versa (due to symmetry).

Data preparation

Table 4A-1. Table representing the mapping of the feature response options into the numerical format.

Feature value	Mapped to
Absent	0
Present - indicative of fit	1
Present - indicative of non-fit	2
Present and explained by stretching	0
Present and not explained by stretching	1
Consistent	0
Not Consistent	1
Not applicable (no missing material)	0
Observed Missing Material	1
Stretched	0
Not Stretched	1

Appendix 4-B

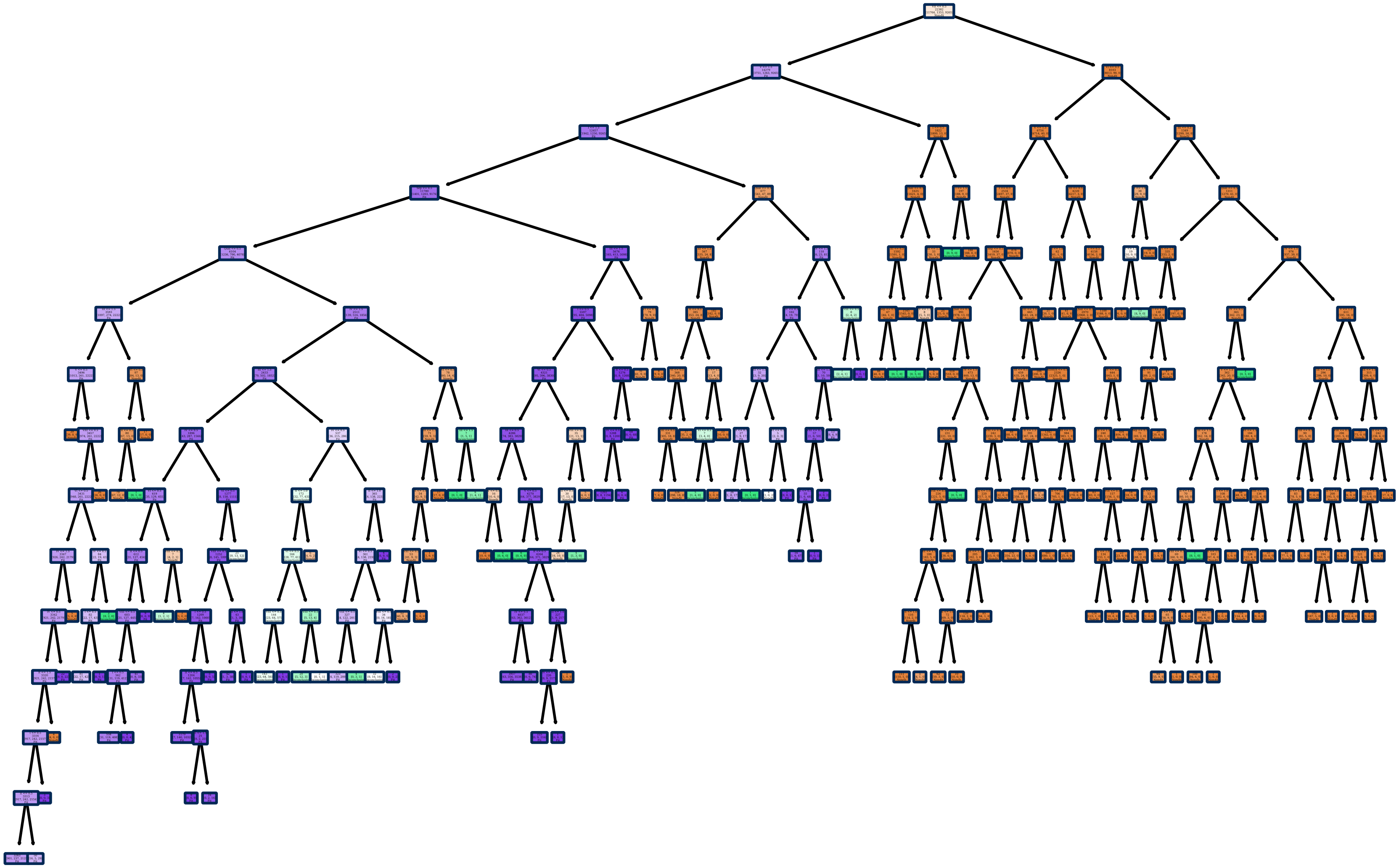


Figure 4B-1. Decision tree diagram for the physical fits of duct tapes. Nodes in blue represent those that contribute to a fit decision, nodes in green represent inconclusive, and nodes in orange represent non-fit.

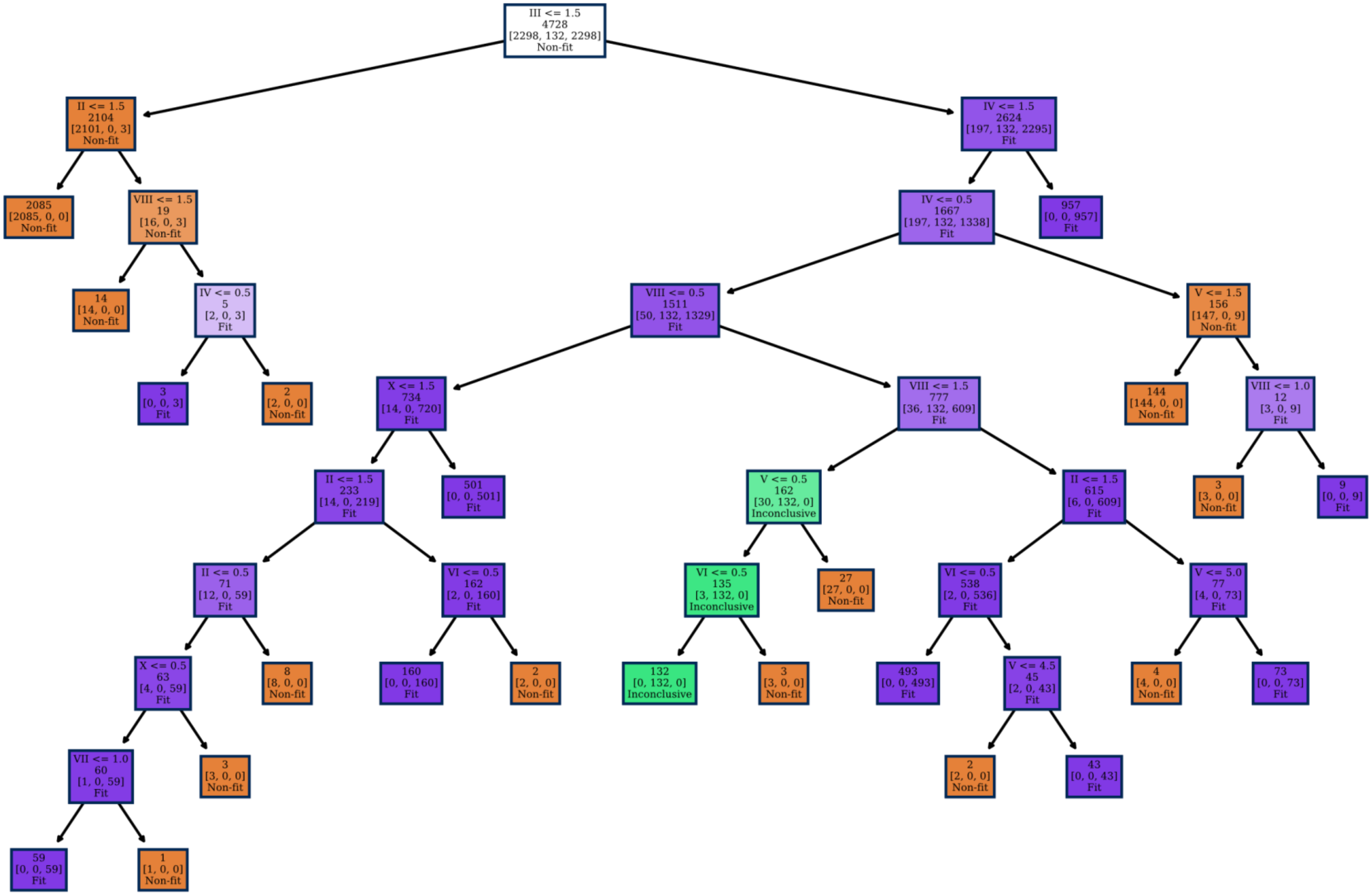


Figure 4B-2. Decision tree diagram for the physical fits of textiles. Nodes in blue represent those that contribute to a fit decision, nodes in green represent inconclusive, and nodes in orange represent non-fit.

Chapter 5: Development of a systematic comparison method for forensic physical fit analysis of automotive polymers

5.1. Overview

This study proposes a method for the physical fit comparison of brittle automotive polymers. The previously developed comparison method for duct tape edges inspired the approach, with adaptations specific to the rigid polymer's properties and the respective features imparted after a fracture. The method includes identifying relevant comparison features, deciding how the polymer edges can be divided into subunits for comparison, and developing standardized quantitative criteria. A reporting template is designed to guide analysts through the comparison and thought process documentation. The template requires documentation of qualitative comparison features. It estimates two quantitative metrics, the edge similarity score (ESS) and the feature prominence value (FPV), which describes the influence of a feature on a fit or non-fit decision.

More than 380 pairs of polymer samples are compared to evaluate the newly developed method. The pairs originate from automotive headlight and taillight assemblies that were broken in a controlled manner. The polymer sources are grouped based on the main polymer composition and morphology into three classes, translucent clear, translucent colored, and opaque colored. While analyzing the samples independently, the analysts are kept blind to the ground truth of the pairs. The report template used by the analysts documents ten comparison features for each comparison bin on the edges of the samples. Auto-populated cell options are provided to characterize the influence of the feature on a decision. These decision options are then assigned a quantitative value defined as the feature prominence value (FPV). The sum of the feature prominence values (FPS) and quantitative edge similarity scores (ESS) provides analysts with the means to assess the similarity between two given edges and estimate potential thresholds to help support an analyst's decision.

Misidentification rates of the initial comparison set are low, and no false positives are reported. The method demonstrates that most true non-fit polymers receive low ESS (0-10%) and low FPS (less than -5). True fit pairs generally receive high ESS (90-100%) and high FPS (15 or greater). Exploratory data analysis shows that polymer composition may impact the quality of a physical fit between polymer edges. Inter-analyst variation of ESS and FPS is low for samples analyzed by two independent analysts. The documentation template provides clear and transparent insight into the features that influenced the decision-making process. The proposed approach is expected to facilitate the implementation of consensus-based protocols at forensic laboratories and provide scientific foundations for data-driven opinions.

5.2. Introduction

Plastic materials are common in everyday life; as such, they are also regularly collected as potential forensic evidence. When the items are separated during criminal activity and left behind at a scene, they might be valuable evidence and assist in reconstructing events. To understand the features imparted during the fracture of plastics, one first needs to recognize the objects' main composition and physical properties. Plastic materials are composed primarily of polymers (long-chain molecule strings). Polymers can be classified based on many criteria, such as the source of the material (including if it is a natural or synthetic polymer or polymer type) or application-based (such as rigid plastics or soft plastics) [1]. Another way of classifying polymers is whether they are brittle or ductile. Brittle polymers tend to shatter once the stress passes their elastic limit and don’t deform as much. Examples of brittle polymers include plastic lenses, Styrofoam, and disposable cutlery. Ductile polymers are generally more flexible and tend to deform before fracture, and include materials such as tapes, plastic bags, and films such as cling wrap.

For forensic purposes, polymer materials are generally grouped by their application. The analytical techniques used to characterize the polymer depend on the sample type and the instrumental methods adopted in laboratories. One limitation of forensic polymer analysis is the substantial number of end-use items that can be analyzed, such as tapes, bags, automotive parts, plastic containers, fabrics, and textiles. Defining standardized protocols for every possible object type is unrealistic. So instead, the focus has been on developing specific protocols for the polymer materials most seen in forensic laboratories: fibers, tapes, and paints [2-4]. Otherwise, for other polymers, more general protocols are specified by technique rather than by material [5-6]. These techniques allow for the characterization of the main polymer types plus additional components, including colorants, fillers, and additives. As such, there is potential for evidence items to be associated based on class characteristics. However, more distinctive features are required to individualize a sample to a possible source.

To be individualized to a single source, a questioned item must align to a source with distinct and consistent edge morphology and features that indicate the two items once belonged together (e.g., a fragment from a headlight from a hit-and-run) [7-8]. The assumption has been that fracture edges are unique and not randomly reproduced. As such, many laboratories consider a physical fit conclusion sufficient not to warrant additional chemical analysis on the samples to consider class association [9]. However, in recent years, renewed calls to demonstrate the reliability of forensic pattern-matching disciplines have driven additional research in physical fit examinations [10-12]. Numerous studies have been published that estimate error rates for physical fit examinations of common forensic materials. Currently, however, only some studies propose a systematic and quantitative method for physical fit comparisons of polymer materials, and there are no consistent features or criteria that analysts can use when making decisions. Our group has developed novel methods that allow a quantitative assessment of the probative value of a fit for duct tapes and textiles. This work expands those approaches to the examination of automotive polymers.

Various studies have introduced new methods for comparing and interpreting the value of a physical fit. Most existing publications cover specific cases where physical fit examinations provide critical evidence for a criminal investigation. Still, in recent years a shift has been observed in the literature toward more extensive populational studies and statistical evaluations of the error rates and probative value of a physical fit [13]. Most studies report on the most common trace materials requested for physical fit analysis, including tapes, textiles, polymers, and paper. Ductile polymers, such as tape, have been the focus of numerous studies, and several recent studies propose quantitative comparison methods and provide error rates and statistical evaluation of physical fit [13-22].

On the other hand, a limited number of studies currently evaluate brittle polymer materials for forensic physical fit examinations. One of the few studies by Baca et al. [23] considers the fractography of glass and brittle polymers. The authors analyze polymer lenses broken using different methods and tools. The results indicate that an impact collision caused the polymer lens to crush instead of splintering, which could influence the appearance of potential fragments recovered at a crime scene [23].

Another study incorporates a computational method for comparing polymer materials. Two polymer types are evaluated – silicon-based polymer casts and Perspex (acrylic glass) plates. The two polymers represent differing polymer characteristics – the silicon polymers are elastic, while the Perspex is brittle. Two computerized systems are used to detect the contour of edges of polymer materials and compare them against a database of other samples to generate a statistical probability of a potential fit. The results find that length of the fracture is a critical feature for the algorithm. The smallest fracture lengths (0.125 cm for silicon and 1 cm for Perspex) contribute the highest error rates, at 22% and 40%, respectively. However, longer lengths improve the accuracy substantially. In the silicon polymer, increasing the fracture length from 0.125 cm to 0.250 cm drops the error rate from 22% to 8%, and when the length increases to 3 cm, the error rate further drops to 0.016%. The same is true of the Perspex, where a fracture length of 2 cm reduces the error rate to 17%, and increasing the length to over 5 cm drops the error rate to 4%. The authors note that the silicone-polymer edges produce enough information and distinct edges regarding the algorithm contour detection. In contrast, the Perspex edges have less information (less distinctive edges). Despite that, the algorithm can produce likelihood ratios that can estimate the weight of evidence for a given fit [24].

The proposed method in this study takes inspiration from the edge similarity score (ESS) method developed for duct tapes and textiles. Each sub-unit bin is independently observed, and the features and overall alignment are determined. Quantitative values are assigned to each bin depending on alignment, and the ESS is calculated as the relative percentage of alignment. True fit pairs result in high ESS, while non-fit pairs result in low ESS [21-22]. Here, we expand the approach to rigid polymers and adapt it for brittle polymer materials, specifically automotive headlight and taillight samples. The critical features needed for decision-making are material-specific. Thus, a primary goal of this study is to identify which features are most relevant for physical fit examinations of automotive polymers. In addition, brittle polymers can be observed in a three-dimensional plane, which challenges imaging the microscopic features. Overall, the developed method is intended to provide analysts with a systematic comparison method for brittle polymers, defined criteria and features, and a detailed documentation template to enhance transparency and consistency in the decision-making process.

5.3. Methods and Materials

The polymer materials are obtained from vehicle headlights and taillights salvaged from automotive body shops. The initial lamp housings are somewhat broken, originating from vehicles that replaced the damaged light. The original light assemblies are documented with images and then disassembled to the extent possible to separate the different observable polymer types (i.e., clear automotive lens cover, black housing, silver accents, and colored sections). **Figure 5-1** shows an example of one of the headlights in its original form, and then the different components are removed.

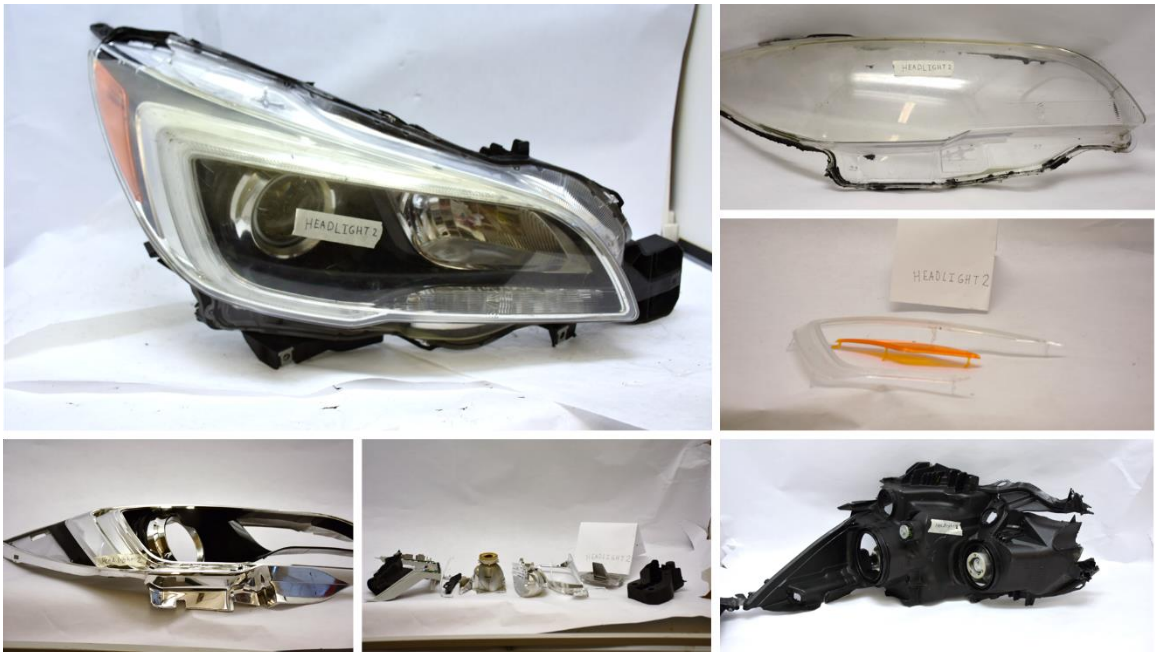
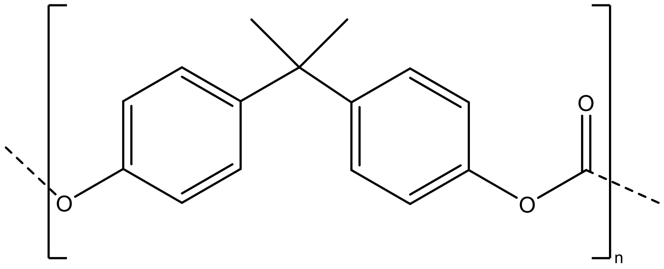
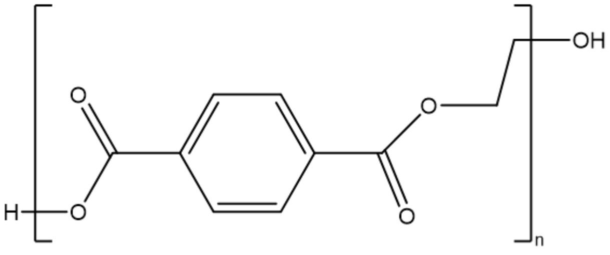


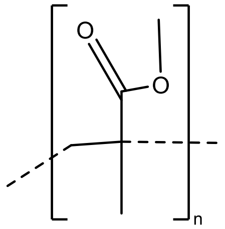
Figure 5-1. Images of an original headlight and the separated components before fracture. The initial assembly is shown in the top left image, while the deconstructed parts are in the other images. The lamp is taken apart to separate the different types of polymers and remove non-polymer parts (metal/glass).

Following the deconstruction of the automotive head/taillight assemblies, the components are sampled for analysis using Fourier Transform Infrared Spectroscopy (FTIR) to identify the primary polymer type of the different components and further understand their mechanical properties. The FTIR spectra classifications of each polymer sub-type are available in **Appendix 5A**. The parts are then grouped according to their overall appearance. The main classes of polymers and their respective compositions are described in **Table 5-1**.

Following deconstruction, the different components are broken further into smaller fragments. Images of each intact polymer are captured before fracturing using a Nikon 7200 DSLR camera with an AF-S Nikkor 18-140 mm lens. After imaging, one side of each sample is covered in painter's tape to ensure that most fragments stay together after fracturing. The polymers are fractured by placing each piece within a square concrete housing and dropping a 16 kg kettlebell directly onto the polymer sample from a consistent height of four feet. A cardboard concrete forming tube is used to guide the weight falling. The kettlebell is dropped onto the polymer and repeated to break the polymer sample into numerous fragments. After fracturing, all fragments are stored in a sealed plastic evidence bag to avoid cross-contamination between pieces from different sources.

Table 5-1. List of polymer composition and morphologies in this study. This set includes a mixture of polymer types, compositions, and morphologies from several assemblies.

Polymer Composition	**Type of Plastic**	**Study Polymer Class**
Polycarbonate 	Thermoplastic	Translucent Clear, Translucent Color
Polypropylene Terephthalate (PPT) 	Thermoplastic	Opaque Color
Polymethyl Methacrylate (PMMA)	Thermoplastic	Translucent Clear

Following the breaking, each polymer is reassembled using the original photographs as a guide. Each fragment is given a unique, partially randomized ID and packaged separately in small manila envelopes. The ID of each fragment maintains traceability of the original polymer source (headlight and taillight), the inventory number of the original polymer sample, the fragment number, and a randomized two-letter code. Non-participating analysts perform the imaging, labeling, and documentation to ensure those performing the comparisons are blind to the ground truth of which samples belong together. The analysts not responsible for examination prepare non-fit ground truth pairs by comparing pieces that were not joint but had similar edge characteristics. Finally, each pair is photographed together in the orientation intended to be analyzed; an example pair is shown in **Figure 5-2**. The individual comparison pairs all originate from the same type of polymer; as in casework analysis, samples of distinctly different polymers would be excluded before a physical fit examination. The selected pairs of samples are imaged to allow the analysts to orient the fragments correctly for comparison. While this knowledge is not available in true casework, providing the images ensures that the analysts perform comparisons only on the intended edges where the respective "truth" is known but blind to the examiner.

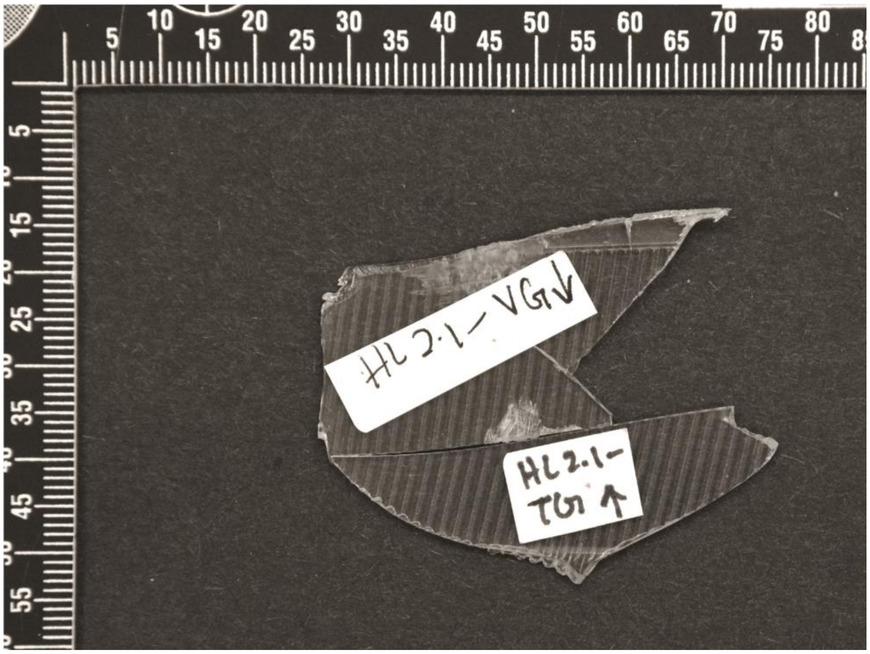


Figure 5-2. Example comparison pair of polymer samples. Each sample's label has an arrow pointing to the fragment's randomly selected "North" side. This label provides the intended orientation of the fragments and edges of interest. For example, this pair is designed to be analyzed for HL2.1-VG in the southwest orientation and HL2.1-TG in the north orientation.

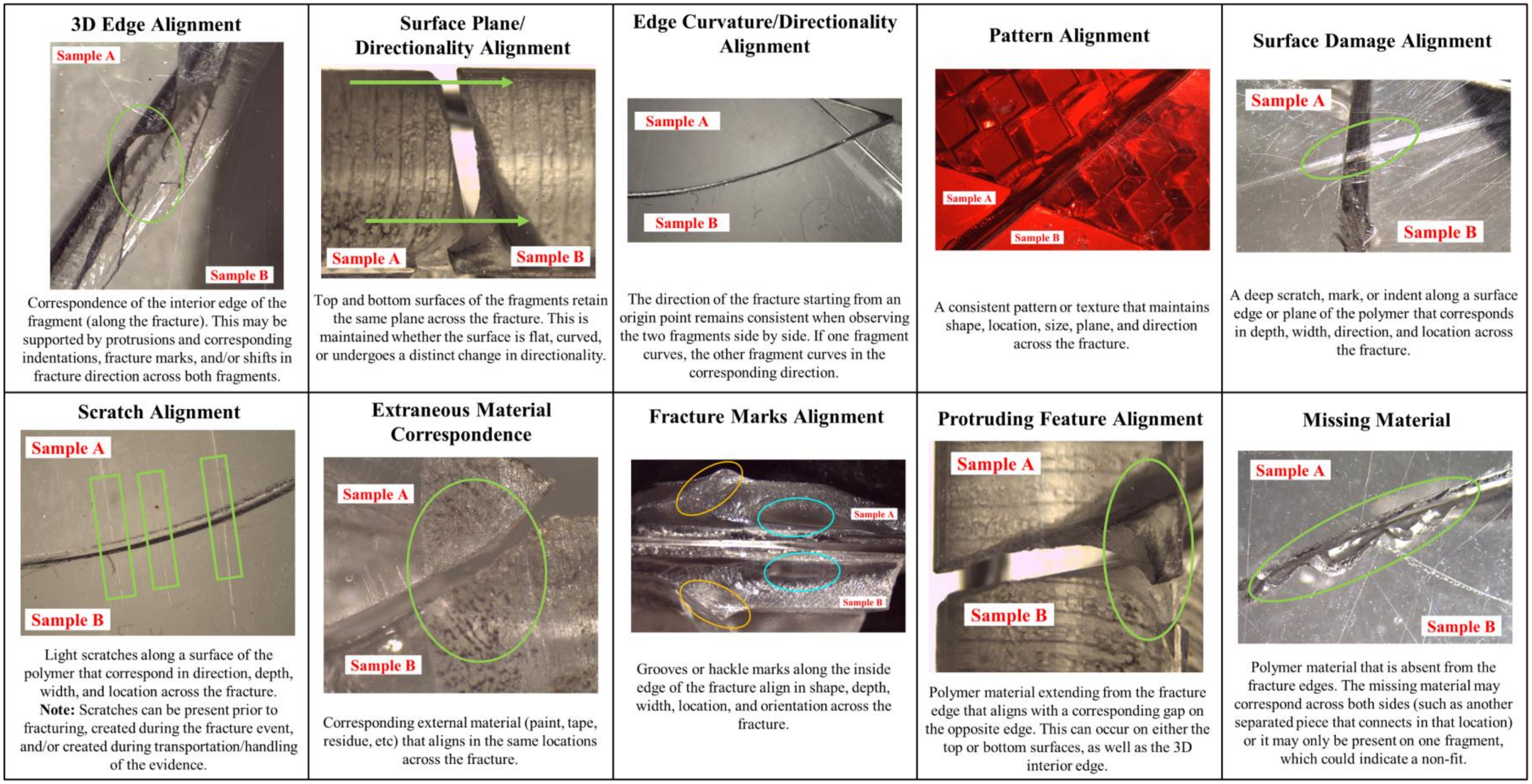
5.4. Results and Discussion

5.4.1. Development of comparison and documentation criteria

The first step of this study is to develop a comparison method and define features that analysts can use for their observations. Six true-fit pairs of polymers (clear and red) are initially used for testing and method development. Three analysts observe the six pairs independently and identify features they use during decision-making. The results are compared, and the proposed comparison characteristics are established after discussing the features that all three noted. Some of these features correspond with characteristics of other brittle materials, such as glass [25]. The terminology to describe these features and factors that can influence a fit or non-fit decision is established to ensure consistency in the reporting. In addition, an Excel documentation template is created to maintain consistent records of samples and features between analysts.

This study utilizes the previously developed edge similarity score (ESS) concept for comparing tape physical fits [21-22]. Unlike duct tapes, which have uniform areas between scrim yarns, automotive polymers have no inherent uniform sub-divisions that can be used as pre-determined comparison bins. Thus, the ESS comparison areas are based on the length of the fracture. The fractured edge is divided into five equal-length bins. To account for potential curved edges in the polymer fractures, unwaxed dental floss is used to outline the fracture edge, then divided into five sections marked on the floss with a permanent marker to help the examiner locate where each bin is. The flexibility of that marked floss makes it easy to align to the edges under the microscope during the examination and does not alter any of the pieces.

The method for this study incorporates two steps. The first step is a general observation of each sample independently and then a side-by-side comparison. Each fragment has the overall edge morphology documented. The edges are labeled as straight, curvy, serrated, or puzzle-like, and any observations about the fragment are documented. The analyst then assesses the degree of fit between them based solely on overall edge morphology. The next step incorporates the bin-by-bin assessment of the alignment of the fracture edges and features and the calculation of the ESS under a stereomicroscope (Leica EZ4, PA, USA). This step documents the presence or absence of ten defined features in each bin (**Figure 5-3**), whether that feature indicates the bin as a fit or non-fit, and describes how distinctive the feature is. The examiner has a third option: document an inconclusive bin when the observed features do not allow for a fit or non-fit decision.



3D Edge Alignment

Correspondence of the interior edge of the fragment (along the fracture). This may be supported by protrusions and corresponding indentations, fracture marks, and/or shifts in fracture direction across both fragments.

Surface Plane/ Directionality Alignment

Top and bottom surfaces of the fragments retain the same plane across the fracture. This is maintained whether the surface is flat, curved, or undergoes a distinct change in directionality.

Edge Curvature/Directionality Alignment

The direction of the fracture starting from an origin point remains consistent when observing the two fragments side by side. If one fragment curves, the other fragment curves in the corresponding direction.

Pattern Alignment

A consistent pattern or texture that maintains shape, location, size, plane, and direction across the fracture.

Surface Damage Alignment

A deep scratch, mark, or indent along a surface edge or plane of the polymer that corresponds in depth, width, direction, and location across the fracture.

Scratch Alignment

Light scratches along a surface of the polymer that correspond in direction, depth, width, and location across the fracture. **Note:** Scratches can be present prior to fracturing, created during the fracture event, and/or created during transportation/handling of the evidence.

Extraneous Material Correspondence

Corresponding external material (paint, tape, residue, etc) that aligns in the same locations across the fracture.

Fracture Marks Alignment

Grooves or hackle marks along the inside edge of the fracture align in shape, depth, width, location, and orientation across the fracture.

Protruding Feature Alignment

Polymer material extending from the fracture edge that aligns with a corresponding gap on the opposite edge. This can occur on either the top or bottom surfaces, as well as the 3D interior edge.

Missing Material

Polymer material that is absent from the fracture edges. The missing material may correspond across both sides (such as another separated piece that connects in that location) or it may only be present on one fragment, which could indicate a non-fit.

Figure 5-3. Defined comparison features for brittle polymer comparisons. The presence or absence of each feature is noted in each comparison bin, and each feature is assigned a feature prominence value corresponding with how distinctive the feature is.

For example, in each bin, the pattern alignment is documented (as needed – not all polymers have patterns or textures). In instances with a pattern, the feature is recorded as indicative of fit or non-fit. The feature is also noted as distinctive or highly distinctive. A new approach is implemented in this method to estimate the influence of each feature on a given decision. To evaluate the features quantitatively, each response option for the feature description is assigned a value, referred to as the feature prominence value (FPV). When the feature is absent or when the feature indicates an inconclusive alignment, the FPV is 0. For fit alignment, the FPV is positive, while a non-fit alignment of the feature contributes a negative FPV. The presence of more distinctive features provides higher FPV. **Table 5-2** shows the assigned values for each feature and respective descriptive options.

Table 5-2. Documented features, response options, and respective feature prominence value (FPV) for the brittle polymer comparisons. These values are summed across the ten features for each of the five comparison bins and then added to express the feature prominence sum of the comparison.

Features	**Options for a response to these features**	**Feature Prominence Value**
1. 3D Edge Alignment	Aligns and highly distinctive (indicative of fit)	2
2. Surface Plane/ Directionality Alignment	Aligns (indicative of fit)	1
	Inconclusive	0
3. Edge curvature/ Directionality	Misaligns (indicative of non-fit)	-1
	Misaligns and highly distinctive (indicative of non-fit)	-2
4. Pattern Alignment	Present and highly distinctive (indicative of fit)	2
5. Surface damage alignment	Present (indicative of fit)	1
6. Scratch alignment	Absent	0
7. Fracture marks	Present but misaligns (indicative of non-fit)	-1
8. Protruding features	Present but misaligns (highly indicative of non-fit)	-2
9. Extraneous materials	Present and highly distinctive (indicative of fit)	2
	Present (indicative of fit)	1
	Absent	0
	Present but misaligns (indicative of non-fit)	-1
	Present but misaligns (highly indicative of non-fit)	-2
10. Missing materials	Present and highly distinctive (indicative of fit)	2
	Present (indicative of fit)	1
	Absent	0
	Inconclusive	0
	Present but misaligns (indicative of non-fit)	-1
	Present but misaligns (highly indicative of non-fit)	-2

The FPV of all features is summed for each bin and then across all bins to provide the feature prominence sum (FPS). In addition, the analyst assigns a value of 0, 0.5, or 1 for each bin's overall non-fit, inconclusive, or fit alignment; the alignment values for each comparison bin are summed and expressed as a relative percentage the five total comparison areas to calculate the ESS. Finally, the final overall decision, ESS, and FPS are reported in each comparison. An example template for comparing a pair of polymer samples is shown in **Figure 5-4**.

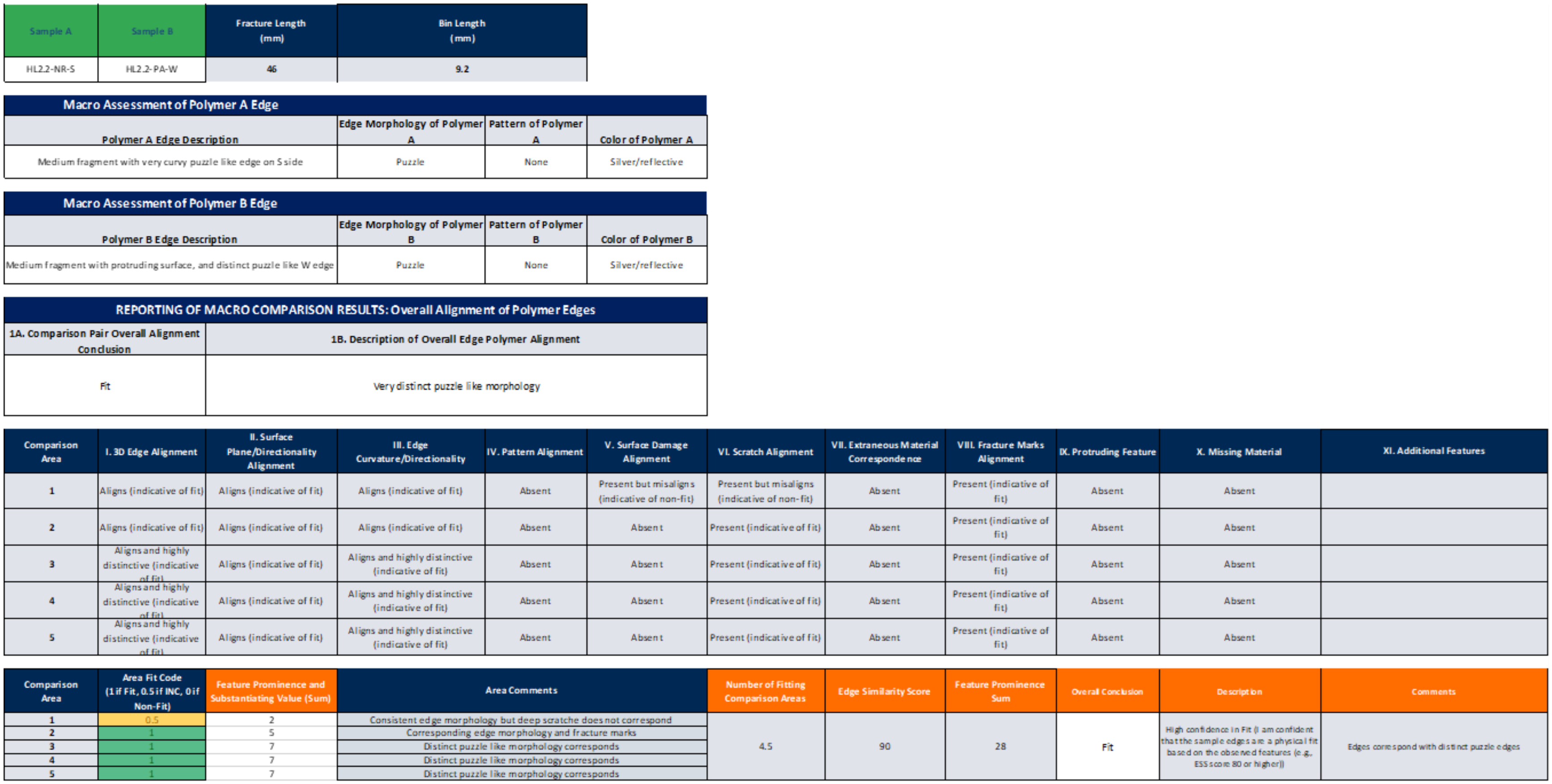
Following the development and preliminary testing of the method and documentation criteria, the analysts conduct a training session to standardize the criteria used for the examination and reporting. A set of blind comparisons is performed. The comparison set incorporates 385 individual comparison pairs completed by an analyst trained in the polymer ESS methodology. The composition of the polymers across the set is a mixture of translucent clear, translucent-colored, and opaque-colored polymers (refer to **Table 5-3**). In addition, a subset of comparisons is completed by a second analyst to evaluate inter-analyst variation in the method's performance.

5.4.2. Exploratory analysis of method performance and distributions of edge similarity scores and feature prominence on true-fit and true non-fit populations

The proposed method demonstrates good performance, with an accuracy greater than 85%. A critical outcome of the comparisons is that no false positives are reported for true non-fits. However, several false negative results are observed, along with several fit and non-fit pairs reported as inconclusive. The documentation protocols established in this study allow the evaluation of the reasons and factors that lead to those misidentifications. Some result from samples lacking distinct features on the edges, distortion caused during the fracture event, and features left post-breaking (such as scratches created during evidence packaging or comparison) that suggest the samples be non-fit.

Table 5-3. Performance rates of the analysis of the polymer pairs. This overall set includes a mixture of polymer types, compositions, and morphologies.

Performance Rates	**Overall (n = 385)**
# of True Fits/# of True Non-fits	312/73
True-Positive Rate (%)	82.7
True-Negative Rate (%)	95.9
False-Negative Rate (%)	9.9
False-Positive Rate (%)	0.0
Inconclusive Rate (True Fits) (%)	7.4
Inconclusive Rate (True Non-fits) (%)	4.1
Accuracy (%)	85.2



Sample A	Sample B	Fracture Length (mm)	Bin Length (mm)
HL2.2-NR-S	HL2.2-PA-W	46	9.2

Macro Assessment of Polymer A Edge			
Polymer A Edge Description	Edge Morphology of Polymer A	Pattern of Polymer A	Color of Polymer A
Medium fragment with very curvy puzzle like edge on S side	Puzzle	None	Silver/reflective

Macro Assessment of Polymer B Edge			
Polymer B Edge Description	Edge Morphology of Polymer B	Pattern of Polymer B	Color of Polymer B
Medium fragment with protruding surface, and distinct puzzle like W edge	Puzzle	None	Silver/reflective

REPORTING OF MACRO COMPARISON RESULTS: Overall Alignment of Polymer Edges	
1A. Comparison Pair Overall Alignment Conclusion	1B. Description of Overall Edge Polymer Alignment
Fit	Very distinct puzzle like morphology

Comparison Area	I. 3D Edge Alignment	II. Surface Plane/Directionality Alignment	III. Edge Curvature/Directionality	IV. Pattern Alignment	V. Surface Damage Alignment	VI. Scratch Alignment	VII. Extraneous Material Correspondence	VIII. Fracture Marks Alignment	IX. Protruding Feature	X. Missing Material	XI. Additional Features
1	Aligns (indicative of fit)	Aligns (indicative of fit)	Aligns (indicative of fit)	Absent	Present but misaligns (indicative of non-fit)	Present but misaligns (indicative of non-fit)	Absent	Present (indicative of fit)	Absent	Absent	
2	Aligns (indicative of fit)	Aligns (indicative of fit)	Aligns (indicative of fit)	Absent	Absent	Present (indicative of fit)	Absent	Present (indicative of fit)	Absent	Absent	
3	Aligns and highly distinctive (indicative of fit)	Aligns (indicative of fit)	Aligns and highly distinctive (indicative of fit)	Absent	Absent	Present (indicative of fit)	Absent	Present (indicative of fit)	Absent	Absent	
4	Aligns and highly distinctive (indicative of fit)	Aligns (indicative of fit)	Aligns and highly distinctive (indicative of fit)	Absent	Absent	Present (indicative of fit)	Absent	Present (indicative of fit)	Absent	Absent	
5	Aligns and highly distinctive (indicative of fit)	Aligns (indicative of fit)	Aligns and highly distinctive (indicative of fit)	Absent	Absent	Present (indicative of fit)	Absent	Present (indicative of fit)	Absent	Absent	

Comparison Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)	Feature Prominence and Substantiating Value (Sum)	Area Comments	Number of Fitting Comparison Areas	Edge Similarity Score	Feature Prominence Sum	Overall Conclusion	Description	Comments
1	0.5	2	Consistent edge morphology but deep scratche does not correspond	4.5	90	28	Fit	High confidence in Fit (I am confident that the sample edges are a physical fit based on the observed features (e.g., ESS score 80 or higher))	Edges correspond with distinct puzzle edges
2	1	5	Corresponding edge morphology and fracture marks						
3	1	7	Distinct puzzle like morphology corresponds						
4	1	7	Distinct puzzle like morphology corresponds						
5	1	7	Distinct puzzle like morphology corresponds						

Figure 5-4. Documentation template for comparison of brittle polymer samples. The first step covers the overall assessment of each edge independently and then a side-by-side comparison, where the analyst documents whether they observe the pair as a fit or non-fit. The second step covers the bin-by-bin assessment of the compared edges, which are separated into five macroscopic bins based on the overall length of the fracture. Ten major features are documented for their absence or presence. If present, the analyst documents whether the feature indicates a fit or non-fit, and the degree of distinctiveness of the feature is assigned a prominence value. The analyst reports each bin as fit, non-fit, or inconclusive by coding it as 1, 0, or 0.5, which is then automatically calculated as the ESS. The conclusion, ESS, and the sum of prominence values (FPS) are reported. The complete Excel file clean template can be accessed through this link.

The distribution of ESS for the polymer set is shown in **Figure 5-5** and documents the reported ESS for the true non-fit and fit pairs based on the ground truth. **Figure 5-5A** shows that the ESS distribution for the polymer comparison set for true non-fit pairs generally falls below 30, with the majority at an ESS of 0. There are two inconclusive true non-fit pairs at an ESS of 40. This indicates that even with some observed fit features, the overall ESS was still low for those samples. Conversely, the ESS distribution for the true fit pairs is more variable. While most fall at 90 to 100, more inconclusive and false negative pairs are recorded with low ESS values. Despite the wider variation, there is a clear separation between the ESS distribution of fits and non-fits, indicating that the score can be used as a metric of fit quality. **Figure 5-5B** shows the distribution of ESS as a histogram. The true non-fit pairs are centered on an ESS of 0, but several true fit pairs are reported with an ESS of 0.

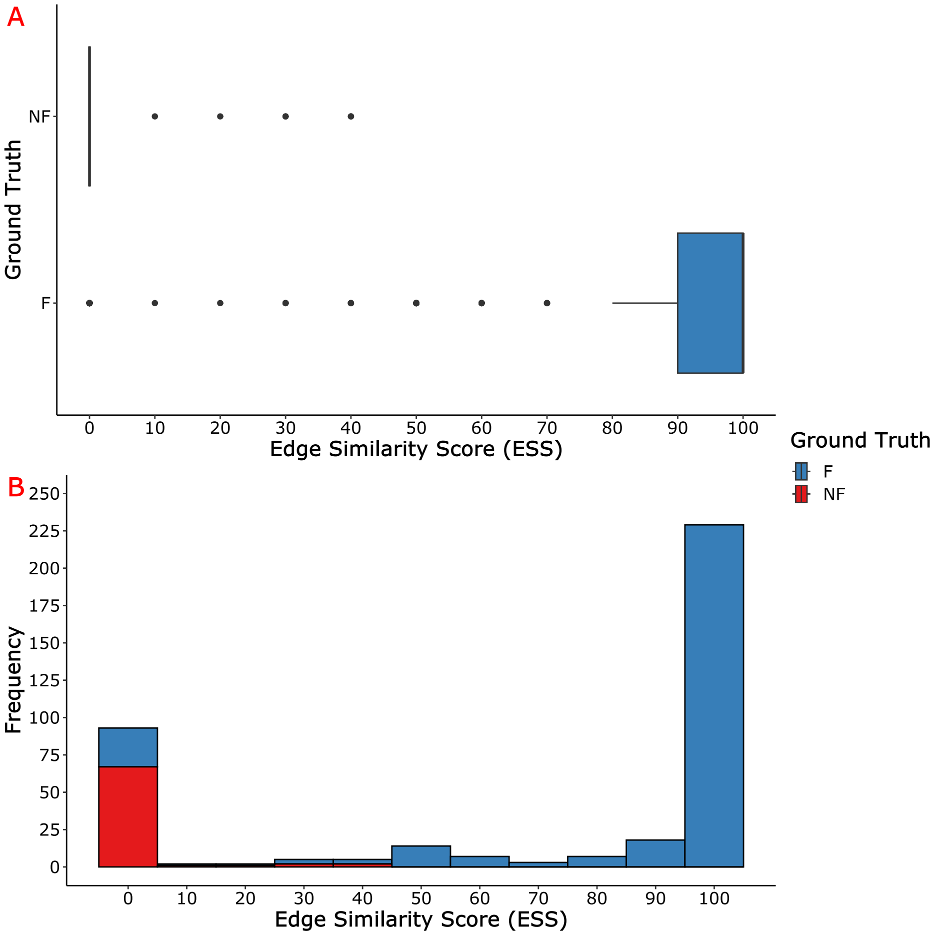


Figure 5-5. Distribution of ESS for true fit (TF) and true non-fit (TNF) pairs from the polymer analysis (n=385 pairs). No false positives are reported, and only a few TNF pairs are reported as inconclusive. **Figure 5-5A** shows boxplot distributions of the ESS values of the polymer samples. The distribution of ESS for TF pairs is much broader. Those correctly reported as fits have an overall median ESS of 100; several pairs receive lower ESS and are misidentified as inconclusive or non-fits. **Figure 5-5B** shows the histogram of ESS values, where the number of misidentified true fit pairs is visible, and generally minimal overlap is observed between ESS on the two groups.

An additional exploratory analysis of these polymer comparison pairs is through the distributions of the FPS. **Figure 5-6** shows the frequency distribution of the FPS for the sample set. All the true positive pairs receive FPS values of at least 15 or above, while true non-fit pairs have FP sums of -5 or lower. Most samples reported as inconclusive range in FPS from -5 to 10, but all the inconclusive pairs that are true non-fits receive an FPS of 0 or less. Meanwhile, all the inconclusive pairs that are true fits have FPS between 0 to 20. As such, the preliminary analysis would suggest that an FPS of 15 or higher supports a fit conclusion, while an FPS of less than -5 supports a non-fit decision. Meanwhile, an FPS between -5 to 10 could indicate a pair that features that demonstrate both a fit and non-fit.

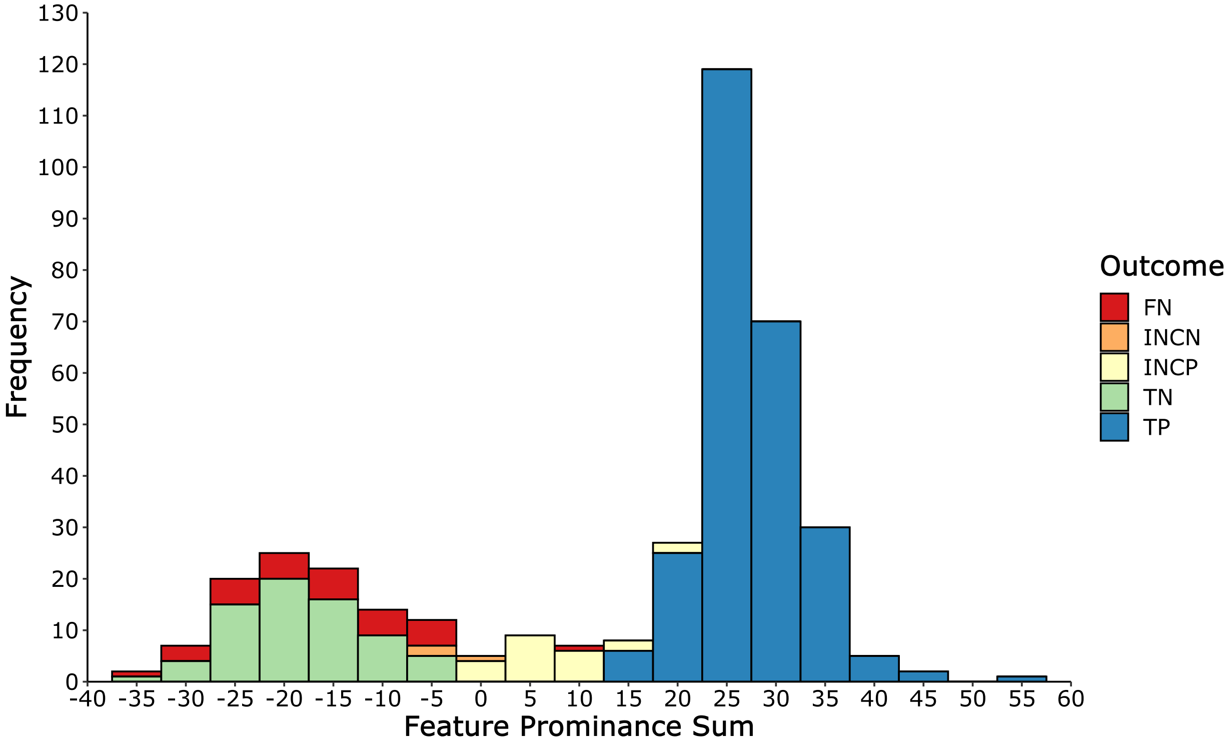


Figure 5-6. Frequency distribution of FPS for the initial polymer set. The histogram is color-coded by outcome: True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), Inconclusive-True-Non-fit (INCN), and Inconclusive-True-Fit (INCP). No FPs are reported, and a few TNF and TF are reported inconclusive. The distributions of TP sums generally are at 15 or higher, while the TN pairs have sums of -5 or lower.

The additional analysis assesses the polymer type's influence on the method's performance. **Table 5-4** shows the performance rate of the polymer set, divided by polymer class. Translucent clear polymers are the most common in this set, followed by the opaque colored and then the translucent colored. It is important to note that there is an imbalance between the number of true fits and true non-fits generated for this study. Generating non-fits for the brittle polymers is more complex than in other materials, as there is no easy way to reproduce edges with consistent features and morphologies. As such, one of the limitations of this initial study is the imbalance and the limited number of overall pairs; however, it is worth noting that the analysts performing the comparisons were blind to the ground truth in addition to the number of expected fit or non-fit pairs for the set.

Table 5-4. Performance rates of the initial analysis of the polymer pairs, grouped by polymer class of translucent clear, translucent color, or opaque color.

Performance Rates	**Opaque Colored (n=40)**	**Translucent Clear (n=314)**	**Translucent Colored (n=32)**
# of True Fits/ # of True Non-fits	31/9	253/62	28/2
True Positive Rate (%)	90.3	83.8	64.3
True Negative Rate (%)	88.9	96.8	100.0
False Negative Rate (%)	3.2	7.5	28.6
False Positive Rate (%)	0.0	0.0	0.0
Inconclusive Rate (True Fits) (%)	6.5	8.7	7.1
Inconclusive Rate (True Non-fits) (%)	11.1	3.2	0.0
Accuracy (%)	90.0	86.3	66.7

Of the three sets, the translucent-colored set generated a higher false negative rate and a high inconclusive rate for true fits. The misidentifications in this set reduce the accuracy for these polymers to 66.7%. During the comparisons, the analysts note that these polymer fragments, particularly the orange fragments originating from a headlight, tend to distort more substantially than some other polymers. This trend can be observed in **Figure 5-7**, which shows the boxplot distributions of pair ESS based on polymer class and composition. While there are two main polymer types in the transparent clear class, the distribution of the ESS for true fits and true non-fits is similar. The translucent colored set, however, has a broader distribution of ESS for both true fits and non-fits.

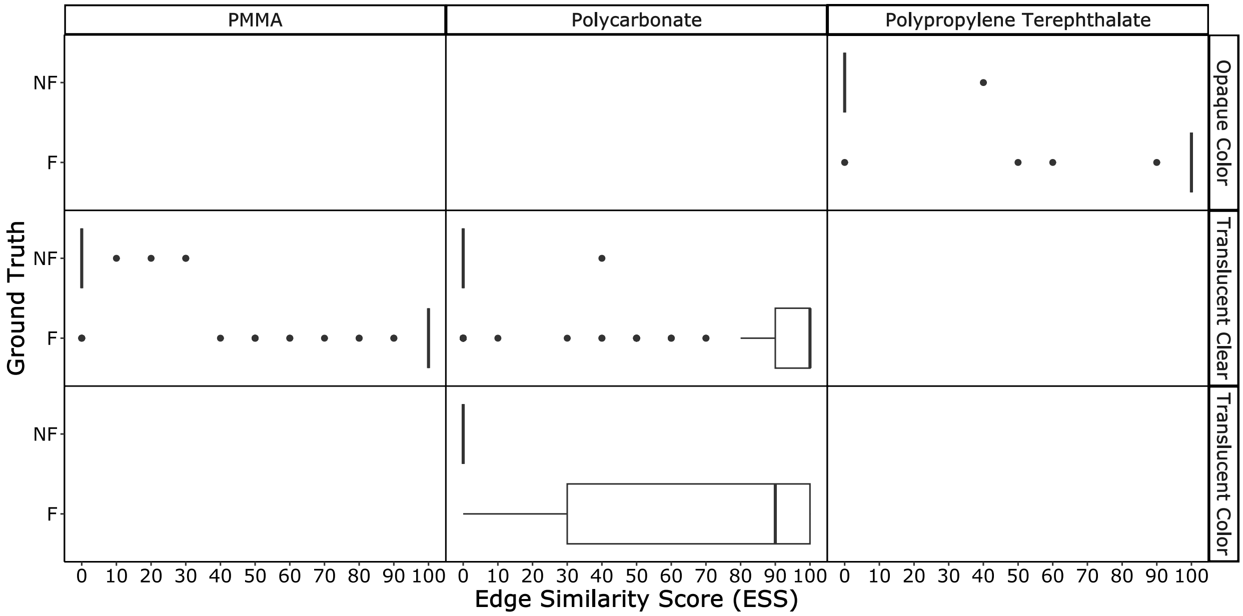


Figure 5-7. Boxplot distributions of reported ESS grouped by polymer class (translucent clear, translucent colored, and opaque colored) and polymer composition (PMMA, polypropylene terephthalate, or polycarbonate). The separation between the ESS of true fits and true non-fits for the translucent clear and opaque color sets is relatively strong, and most pairs have ESS of 100 or 0 for fits and non-fits, respectively. Conversely, the distribution of ESS for the translucent color fragments is more variable than the other classes.

While both the translucent color set and some of the translucent clear sets are made of polycarbonate, the features of the samples are much different. Polycarbonate is generally very resistant to impact [1]. However, the translucent color samples in this study demonstrate more distortion than observed in other samples. An example of the orange translucent polymer is shown in **Figure 5-8**. The translucent colored samples have more texture and patterning than the translucent clear ones. It is observed that the sample tends to fracture between the patterns rather than through them. The material is thinner in these samples than in the clear plastic samples, which could contribute to more distortion as the sample fractures.

Even though the translucent colored polymers tend to have more distinct patterns than the other classes, the more extensive distortion interfered with observing multiple features. This includes surface, pattern, and 3D edge alignment, three main characteristics to inform the overall edge assessment. An example of a misidentified orange translucent polymer pair is shown in **Figure 5-8**. These samples have more visible surface damage, likely contributing to the misidentification. This pair has a noticeable gap between the edges between the MN-E (east direction) and AZ-W (west direction), and AZ has visible surface damage on the W and E sides. This causes the sample to be misclassified as non-fit with an ESS of 0 and an FPS of -32.

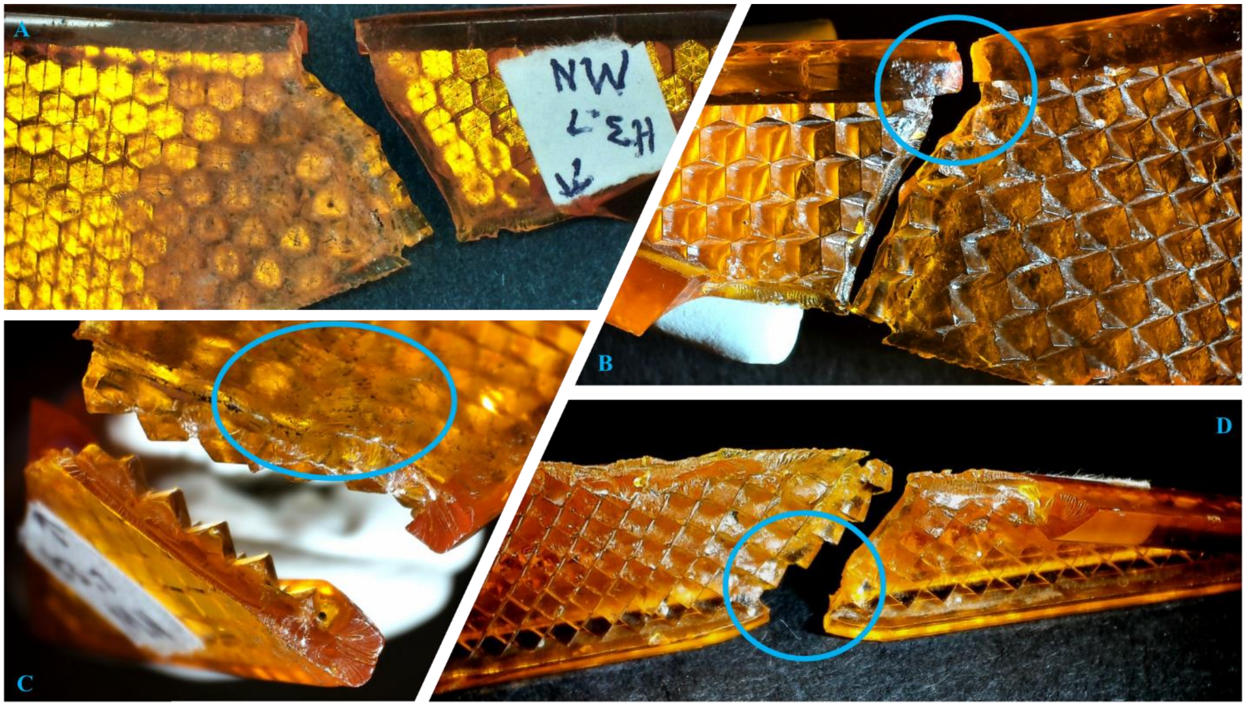


Figure 5-8. Example pair from translucent color polymers. The two fragments have visible surface damage, and the respective comparison edge has a noticeable gap between the material of the two edges in a manner that indicates they do not fit together. **Figures 5-8A, 5-8B,** and **5-8D** show the gap between the edges caused by missing material and distortion of the edges. In contrast, **5-8C** shows an indent in the surface of AZ top sample that causes misalignment of the pattern (texture) and the edge of the samples.

The overall trend remains consistent between classes when evaluating the FPS values for the polymer classes (**Figure 5-9**). True positive pairs receive high FPS, while true negative pairs receive low FPS, and the inconclusive pairs sit between approximately -5 to 10. The documentation templates provided insight into the reason for this trend; some true non-fit pairs have features that indicate similarity between edges, particularly surface plane alignment, which contributes to positive FPV. As seen in the performance, both the translucent clear and translucent colored classes demonstrate more false negatives. The range of assigned FPS corresponds with the true negative distributions, meaning there are fragment pairs that indicate highly distinctive features that suggest non-fit conclusions.

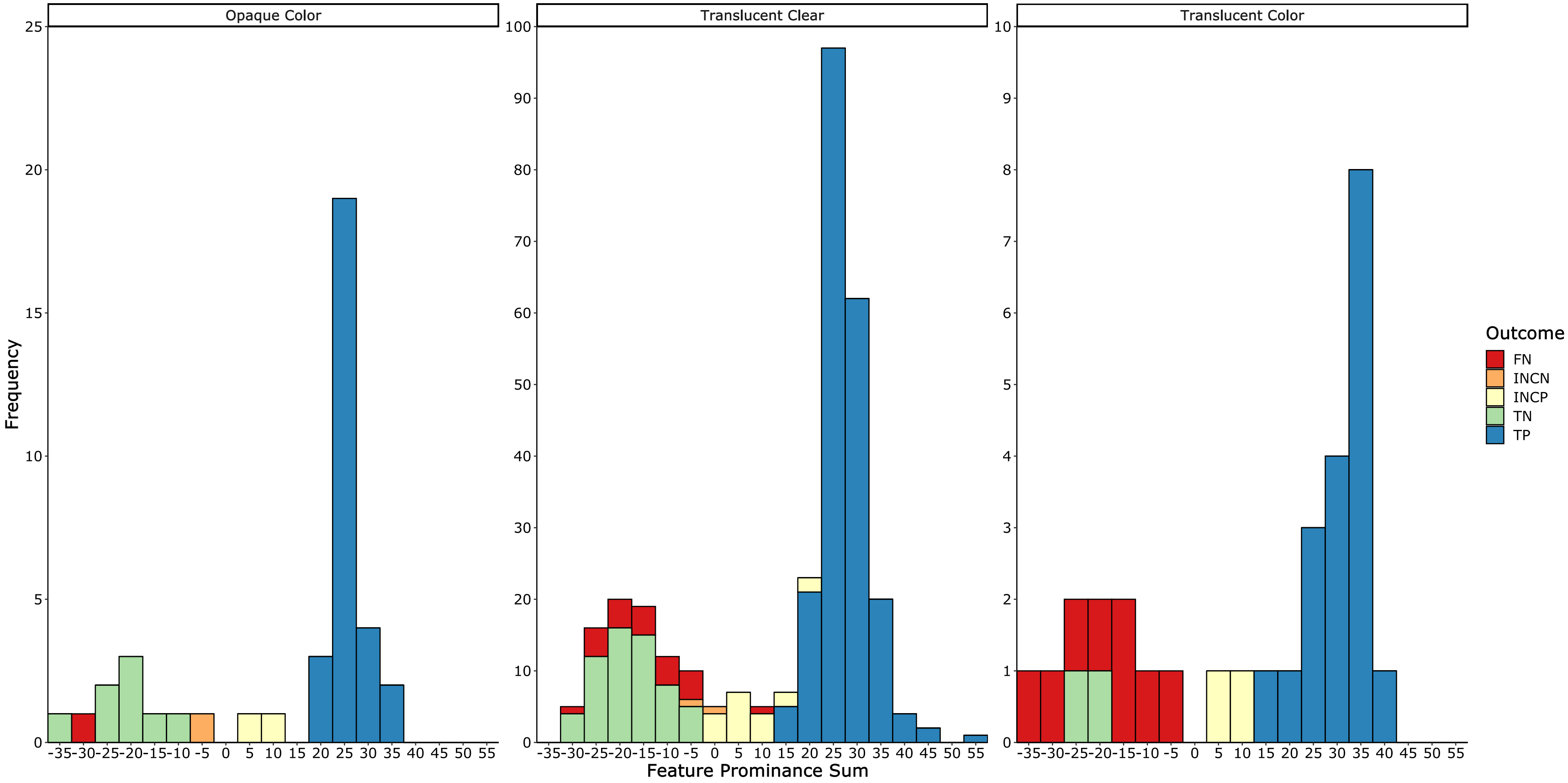


Figure 5-9. The outcome color-coded the frequency distribution of FPS for the three polymer classes. True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), Inconclusive-True-Non-fit (INCN), and Inconclusive-True-Fit (INCP) are shown for each class. The distributions of TP sums generally start at 15 or higher, while the TN pairs have sums of -5 or lower for each of the three sets. The translucent clear and translucent color classes demonstrate several false negatives with very low FPS.

Overall, while the initial examination demonstrates that the method performs relatively well, analysts must consider the suitability of the samples for comparison. Fragments prone to distortion may become misshapen, limiting the utility of the edge for comparison. Despite this, the polymer samples demonstrate many valuable features for decision-making on physical fit examinations. The proposed method identifies which features are more informative to the examiner's opinion and is a first step for assisting standardization in analyzing and reporting these types of automotive polymers.

The proposed methodology also demonstrates that performing both the macroscopic and microscopic comparison of the samples often proves beneficial for comparison. **Figure 5-10** shows the outcomes of the samples based on the decision made at the macroscopic or microscopic conclusion. As shown, the observations made at the microscopic level improve the overall rate of correct identifications. From the macroscopic to microscopic comparison, true positives and true negatives increase, and inconclusive decisions of both true fits and true non-fits decrease. There is a slight increase in false negatives, but critically, false positives are observed in the macroscopic decisions that are not present in the microscopic decisions. As such, this demonstrates that making observations at both the macroscopic and microscopic levels generally improve the accuracy rates and minimizes the risks of false positives. Nonetheless, it is worth noting that if there are substantial differences at the macroscopic level, an analyst can be justified in reporting a non-fit without performing a microscopic comparison.

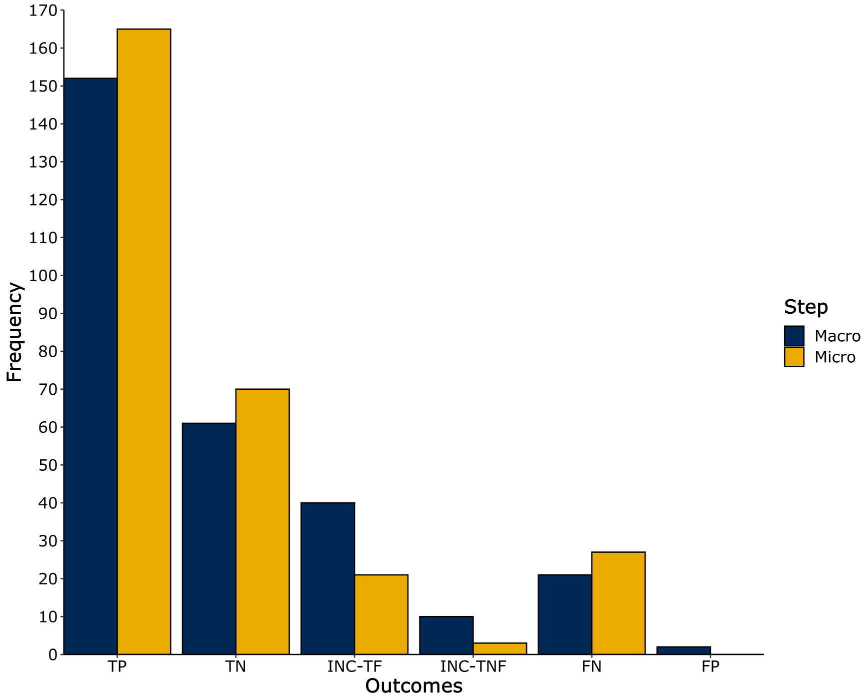


Figure 5-10. Bar graph of outcomes from overall polymer comparison set. Based on the general conclusions, there is an improvement in accuracy from the macroscopic to microscopic decisions, indicating it is generally vital to complete the microscopic comparison unless there are distinct visual differences at the macroscopic level.

Similar trends are observed by polymer class when exploring the difference between the macroscopic and microscopic decisions. **Figure 5-11** shows the breakdown of macroscopic versus microscopic decisions by class, and the trend is consistent in all three – improvements in true positives and negatives and decreased inconclusive decisions.

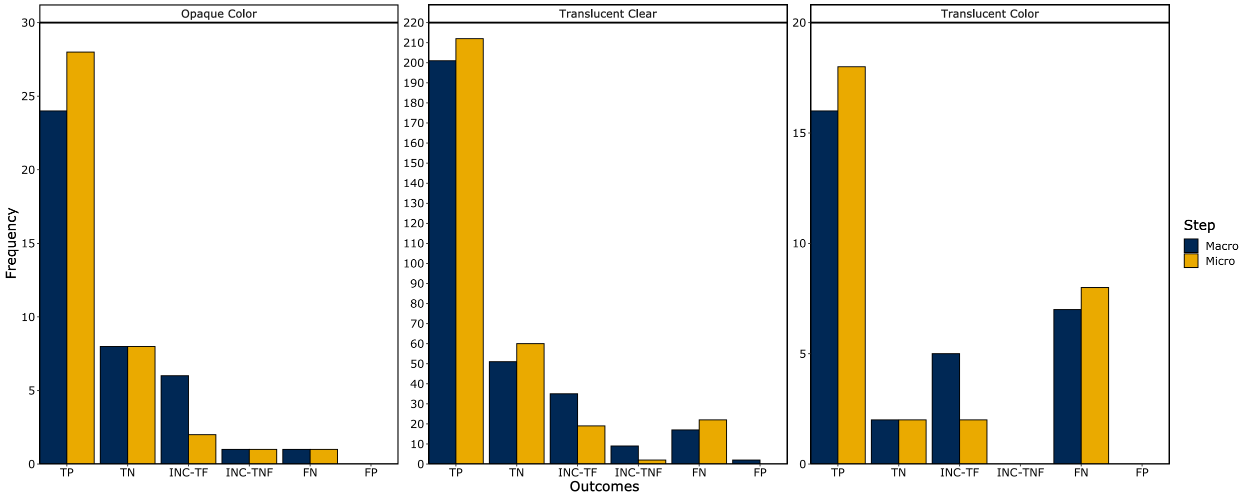


Figure 5-11. Bar graph of outcomes from overall polymer comparison set split by polymer category. Generally, there is an improvement in accuracy from the macroscopic to microscopic decisions across all three classes.

5.4.3. Inter-examiner variation

To further evaluate the method's performance, a subset of samples from all three polymer classes is independently assessed by a second analyst. This subset contains 187 pairs purposely selected to include the most challenging samples, including those pairs where the first analyst observes more misclassifications. The performance of the method for these samples by both analysts is shown in **Table 5-5.**

Overall, the accuracy is comparable between the two analysts, however, Analyst A has a slightly higher false negative and inconclusive rate for true fit pairs. Most of the pairs misidentified by one analyst are also misidentified by the other, and neither reports a false positive. Some variation in performance is not unexpected. Differences in the interpretation of features and the degree of distinction of the features can contribute to variation in performance, along with each analyst's tolerance for risk. The evaluation of the ESS template, its color-coded bin-by-bin comparison, and the documentation of the features that lead the examiners' decisions help identify the sources of variability between analysts. One conclusion that is derived from this assessment is that because we are using only five bins, a small discrepancy in a single bin can lead to a difference of 10% to 20% in the ESS (i.e., assigning a 0.5 versus 0 or 1 in one of the bins). As a result, it is recommended to increment the number of comparison bins to at least ten to minimize the variability in reported scores.

Table 5-5. Performance rates of the inter-analyst examination. This set is a subset of the initial set and contains a mixture of the polymer classes.

Performance Rates	Analyst A	Analyst B
# of True Fits/# of True Non-fits	114/73	114/73
True Positive Rate (%)	72.8	81.6
True Negative Rate (%)	95.9	98.6
False Negative Rate (%)	14.9	11.4
False Positive Rate (%)	0.0	0.0
Inconclusive Rate (True Fits) (%)	12.3	7.0
Inconclusive Rate (True Non-fits) (%)	4.1	1.4
Accuracy (%)	81.8	88.2

The ESS between the two analysts is also comparable. Analyst A has a slightly wider distribution of ESS for true fits, but both analysts have the median ESS for fits at 100 and non-fits at 0. There is limited overlap between the ground truth ESS distributions (**Figure 5-12**).

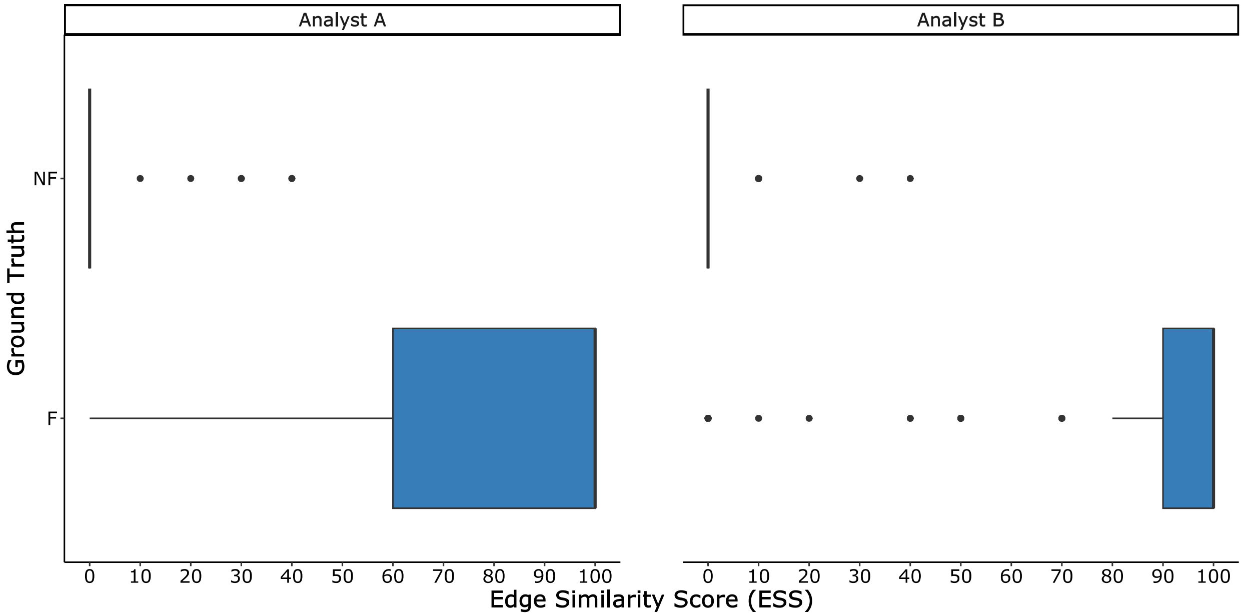


Figure 5-12. Boxplot distribution of ESS for true fit (TF) and true non-fit (TNF) pairs from the inter-analyst polymer analysis. No false positives are reported, and only a few TNF pairs are reported as inconclusive for both analysts.

Moreover, the FPV and FPS metrics reveal a similar weight given by the analysts to the features that lead to a particular bin decision, this provides evidence that the metrics are a promising step towards the standardization of the polymer fit examinations. **Figure 5-13** shows the comparison of FPS for the two analysts on the inter-analyst polymer set. The distributions are similar, sharing the general trends in samples correctly identified as fits and non-fits. For both analysts, inconclusive samples tended to have FPS between 0 to 10.

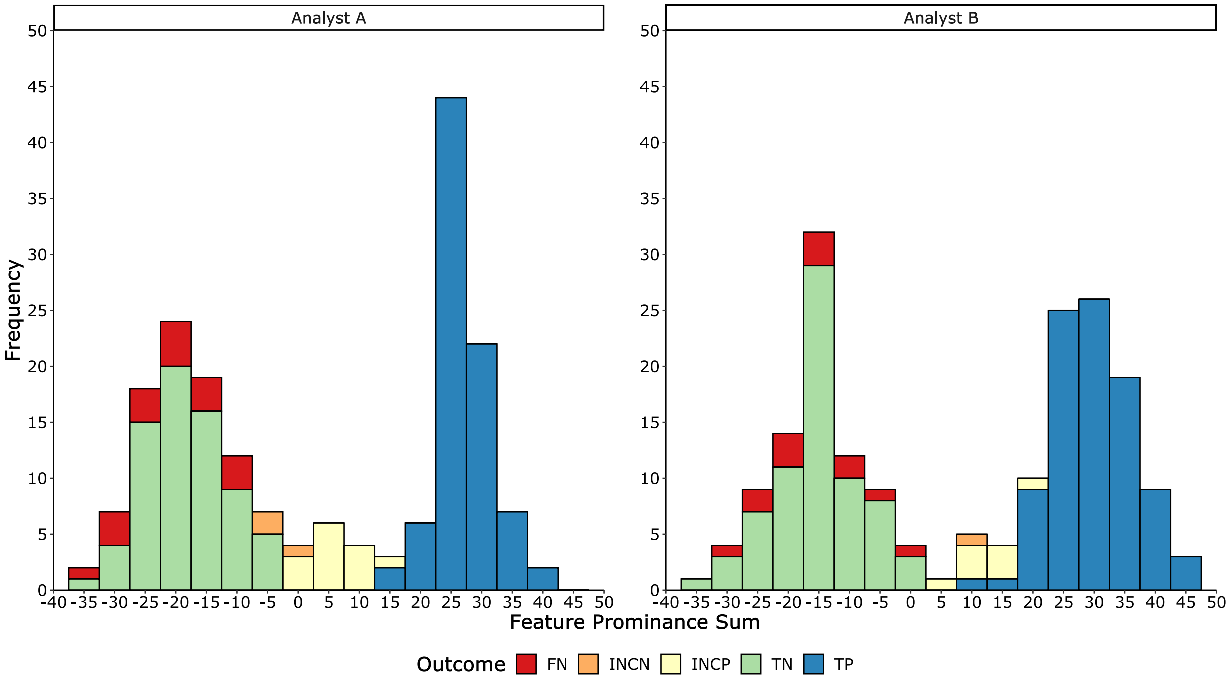


Figure 5-13. The frequency distribution of FPS for the 187 inter-analyst pairs. True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), Inconclusive-True-Non-fit (INCN), and Inconclusive-True-Fit (INCP) are shown for each analyst. The distributions of FPS are similar for both analysts, with only minor variations in the distributions of the inconclusive pairs.

Interestingly, when comparing both analysts' macroscopic and microscopic decisions, the same trends apply, where performing the microscopic observations improves the overall accuracy. **Figure 5-14** shows the outcomes of the samples based on the decision made at the macroscopic or microscopic decision for both Analyst A and B. For both analysts, the number of true positives and negatives increases, and inconclusive decisions of true fits and true non-fits decrease. Both analysts demonstrate false positives in the macroscopic comparisons that are not present in the microscopic decisions. As such, regardless of the analyst, performing both macroscopic and microscopic observations improves overall performance.

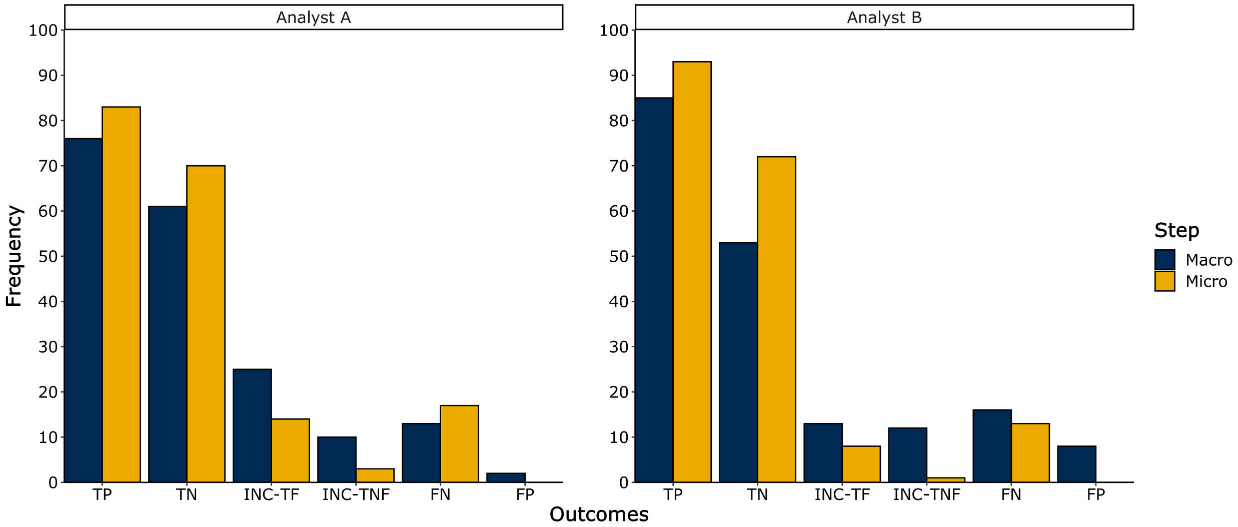


Figure 5-14. Bar graph of outcomes from the two analysts on the inter-examiner polymer set. Both analysts demonstrate improved performance from the macroscopic to microscopic decisions.

The reporting approach makes comparing results between the two analysts straightforward and easy to review. The documentation template ensures that analysts evaluate the same bin areas using standardized features and criteria. However, the interpretation of the features is still decided by the analyst, which allows analysts to express and document their own opinions and allow for mutual discussion of the samples even long after the samples are compared initially.

5.5. Conclusions

The edge similarity score method demonstrates good overall performance for physical fit examinations of brittle automotive polymers — the initial comparison set of 385 comparisons results in an overall accuracy of 85.2%, with most of the false negative misidentifications driven by distortions during the breaking process. The absence of false positives indicates that no true non-fit pairs demonstrated enough features to show they belonged together. It is observed that even some true non-fit pairs have characteristics that show similarity between edges, particularly surface plane alignment. However, there are still sufficient distinct features to assist in correctly identifying when observed under the microscope (distinctive fracture marks, surface damage, and curve directionality). The novel addition of the feature prominence sum provided an additional quantitative metric to assess the similarity between edges and evaluate which features hold more support for the analyst decisions. This study demonstrates preliminary ranges that can be used to support an analyst's decisions: true fits with FPS greater than 15 and true non-fits with FPS less than -5.

In addition, inter-analyst performance shows consistency, with both analysts demonstrating similar overall accuracies, ESS, and FPS distributions when using the proposed examination and reporting method. The three-dimensional nature of the polymers examines fractured edges more challenging, as its more complex to position the side of the fragment by the side to observe the features in all the dimensions. Also, unlike other materials of simpler and predictive construction patterns (like duct tape), it is not straightforward to identify the correct top and bottom edge of the compared items, including additional subjectivity in this assessment. Based on evaluating the reporting templates, we hypothesize that these two factors contribute to some variability between examiners. However, the approach proposed here is anticipated to provide a first step toward more systematic comparison criteria and documentation. It is also anticipated that the future evaluation of this tool by practitioners can lead to improvements in reproducibility.

The initial mixed polymer set indicates that polymer composition and mechanical properties may influence the fit's quality. According to the analyst's observations, the translucent colored fragments derived from headlights tend to deform more. However, with the limited number of comparisons performed in this class, it is currently unknown whether this is due to the polymer construction alone. More data is needed to evaluate polymer composition and construction's significance in edge morphology post-breaking.

This study establishes for the first time a set of standard features and terminology for comparing brittle polymer edges and provides a documentation template incorporating the features for physical fit examinations. The initial comparison set provides insight into potential factors that could influence the quality of a fit and the possible error rates associated with polymer physical fit analysis. The documentation of the ESS and FPS of these pairs provides a means to quantitatively assess the similarity between edges and how the features correspond to an analyst's decision. It is important to note that there are some misidentified samples, meaning that there is a reasonable chance that some polymer evidence samples may not be suitable for physical fit examinations, so analysts must be aware of potential sources of error in comparisons and assess if there is a chance that the samples could have been further damaged after breaking. Future work in this area will continue to refine the comparison method and documentation template. With additional comparisons, there is an opportunity to utilize computational models to estimate the weight of the comparison features and provide more representative feature prominence values. In addition, with an increased dataset, additional statistical analysis on the reported ESS and FPS for polymer comparisons to estimate the probative value of a potential fit will be reported in a follow-up work.

5.6. References

1. Causin V. *Polymers on the Crime Scene: Forensic Analysis of Polymeric Trace Evidence.* Springer. 2015. ISBN: 978-3-319-15494-7. DOI:10.1007/978-3-319-15494-7
2. ASTM International. ASTM E2225 Standard Guide for Forensic Examination of Fabrics and Cordage. 2021. DOI: 10.1520/E2225-18.
3. ASTM International. ASTM E1610-18 Standard Guide for Forensic Paint Analysis and Comparison. 2018. DOI: 10.1520/E1610-18
4. ASTM International. ASTM E3260-21 Standard Guide for Forensic Examination and Comparison of Pressure Sensitive Tapes. 2022. DOI: 10.1520/E3260-21

5. ASTM International. ASTM E3296-22, Standard Guide for Using Pyrolysis Gas Chromatography and Pyrolysis Gas Chromatography-Mass Spectrometry in Forensic Polymer Examinations. 2022. DOI: 10.1520/ E3296-22
6. ASTM International. ASTM 2809-22 Standard Guide for Using Scanning Electron Microscopy/Energy Dispersive X-Ray Spectroscopy (SEM/EDS) in Forensic Polymer Examinations. 2022. DOI: 10.1520/2809-22
7. OSAC Trace Materials Subcommittee. OSAC 2022-S-0029 Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials. https://www.nist.gov/system/files/documents/2022/02/28/OSAC%202022-S-0029%20Standard%20Guide%20for%20Interpretation%20and%20Reporting%20in%20Forensic%20Comparisons%20of%20Trace%20OPEN%20COMMENT.pdf
8. OSAC Trace Materials Subcommittee. OSAC 2022-S-0015. *Standard Guide for Forensic Physical Fit Examination.* 2022. https://www.nist.gov/system/files/documents/2021/12/06/OSAC_2022-S-0015_Standard_Guide_for_Forensic_Physical_Fit_Examination_DRAFT_OSAC_PROPOSED.pdf
9. Gross, S. Physical Fit Task Group – Trace Materials Subcommittee of NIST-OSAC. Survey of Physical Fit Protocols. 2019. Unpublished Survey, presented at NIST-OSAC Trace Subcommittee Meeting, March 2020, OK.
10. National Academy of Sciences (NAS). Strengthening Forensic Science in the United States: A Path Forward. 2009. doi: 0.17226/12589
11. President's Council of Advisors on Science and Technology (PCAST). Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/otsp/PCAST/pcast_forensic_science_report_final.pdf
12. OSAC Research Needs Trace Materials Subcommittee. Development of Quantitative Assessment and Evaluation of Error Rates in Physical Fit Determinations of Trace Materials. 2021. https://www.nist.gov/system/files/documents/2021/06/23/Trace%20Materials%20SC_R%26D%20Need_Development_of_Quantitative_Assessment_and_Evaluation_of_Error_Rates_in_Physical_Fit_Determinations_of_Trace_Materials.pdf
13. Brooks E, Prusinowski M, Gross S, Trejos T. Forensic Physical Fits of Trace Materials: A Review. *Forensic Science International.* August 2020. Available online May 29th, 2020. https://doi.org/10.1016/j.forsciint.2020.110349
14. Bradley MJ, Gauntt JM, Mehltretter AM, Lowe PC, Wright DM. A Validation Study for Vinyl Electrical Tape End Matches. *J For Sci.* 2011;56(3):606-611.
15. M.J. Bradley, R.L. Keagy, P.C. Lowe, M.P. Rickenbach, D.M. Wright, M.A. LeBeau, A validation study for duct tape end matches, J. Forensic Sci. 51 (2006) 504– 508, doi:http://dx.doi.org/10.1111/j.1556-4029.2006.00106.x.

16. K.R. McCabe, F.A. Tulleners, J.V. Braun, G. Currie, E.N. Gorecho, A quantitative analysis of torn and cut duct tape physical end matching, J. Forensic Sci. 58 (2013) S34–S42.
17. F.A. Tulleners, J. Braun, The Statistical Evaluation of Torn and Cut Duct Tape Physical End Matching, Natl. Inst. Justice 2011, July. Rep. No. 235287. (n.d.).
18. W. Ristenpart, F.A. Tulleners, A. Alfter, Quantitative algorithm for the digital comparison of torn duct tape, Final Rep. to Natl. Inst. Justice Grant 2013-R2- CX-K009, Univ. Calif. Davis Davis, CA, 2017.
19. van Dijk CD, van Someren A, Visser R, Sjerps M. Evidential value of duct tape comparison using loopbreaking patterns. *Forensic Sci Int*. 2022 332:111178. doi: 10.1016/j.forsciint.2022.111178.
20. Spaulding J, Picconatto G. Characterization of fracture match associations with automated image processing. *Forensic Science International*. 2023. 342. 10.1016/j.forsciint.2022.111519
21. Prusinowski M, Brooks E, Trejos T. Development and validation of a systematic approach for quantitatively assessing the quality of duct tape physical fits. *Forensic Science International*. 2020. https://doi.org/10.1016/j.forsciint.2019.110103
22. Prusinowski M, Andrews Z, Neumann C, Trejos T. Assessing significant factors that can influence physical fit examinations – physical fits of torn and cut duct tapes. *Forensic Science International*. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567.
23. A.C. Baca, J.I. Thornton, F.A. Tulleners, Determination of fracture patterns in glass and glassy polymers, J. Forensic Sci. 61 (2016) 92–101, doi: http://dx.doi.org/10.1111/1556-4029.12968.
24. Y. Yekutieli, Y. Shor, S. Wiesner, T. Tsach, Physical matching verification, Final Rep. to United States Dep. Justice Grant 2005-IJ-R-051, Natl. Crim. Justice Ref. Serv., Rockville, MD, 2012.
25. Hayes, Michael, et al. Fractography in Failure Analysis of Polymers, Elsevier Science & Technology Books, 2015. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/wvu/detail.action?docID=2049242.

Appendix 5-A

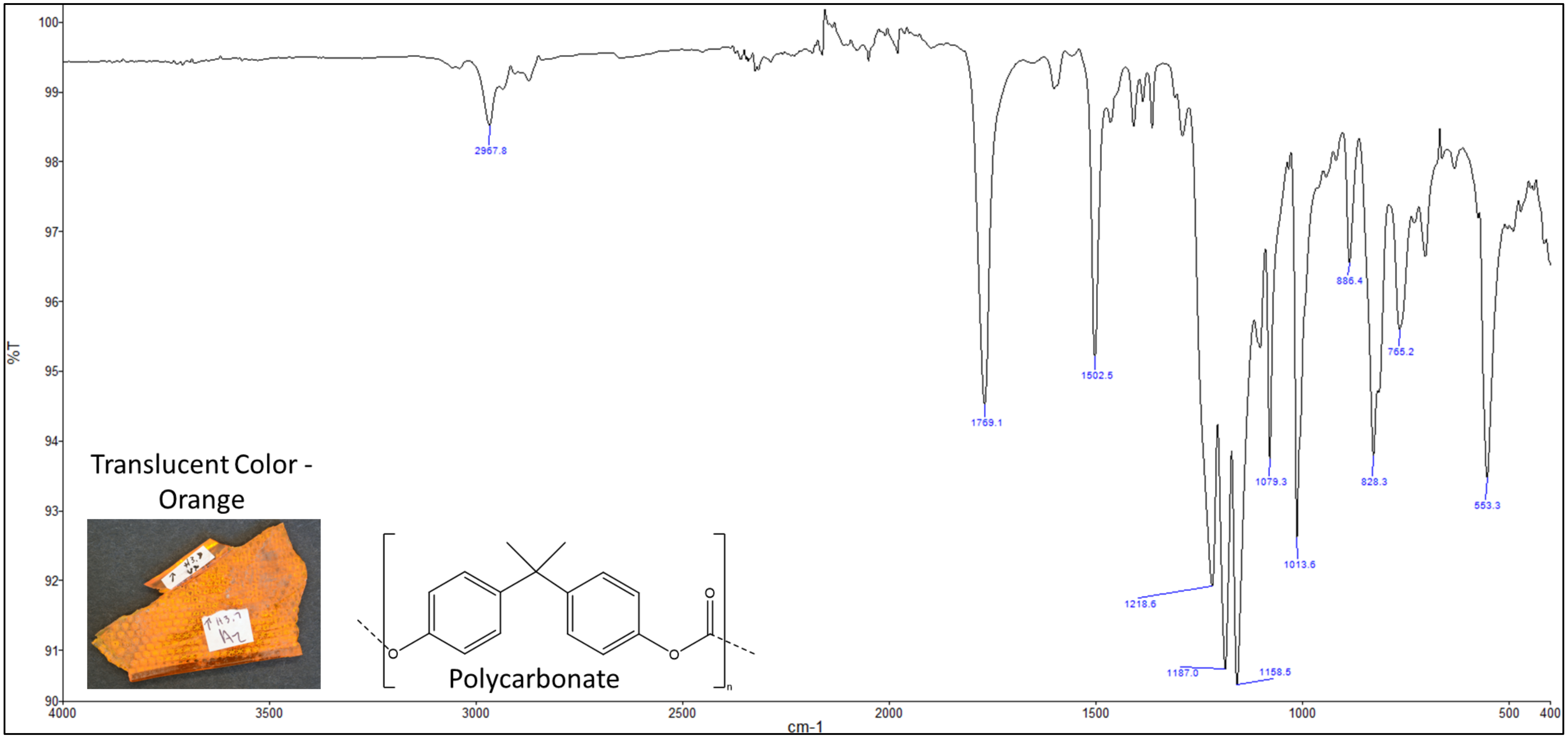


Figure 5A-1. FTIR spectra of the translucent color (orange) samples. The main polymer type is polycarbonate.

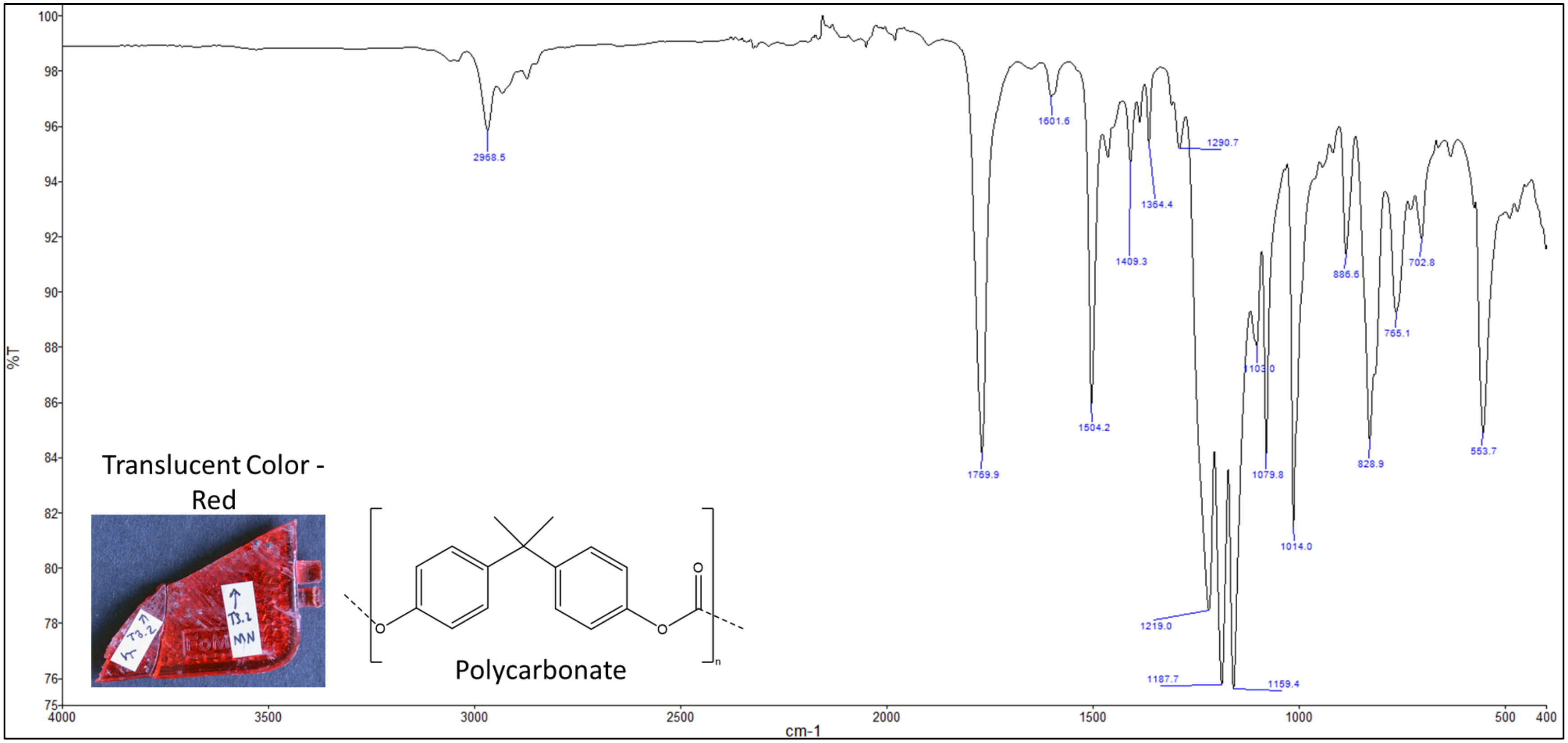


Figure 5A-2. FTIR spectra of the translucent color (red) samples. The main polymer type is polycarbonate.

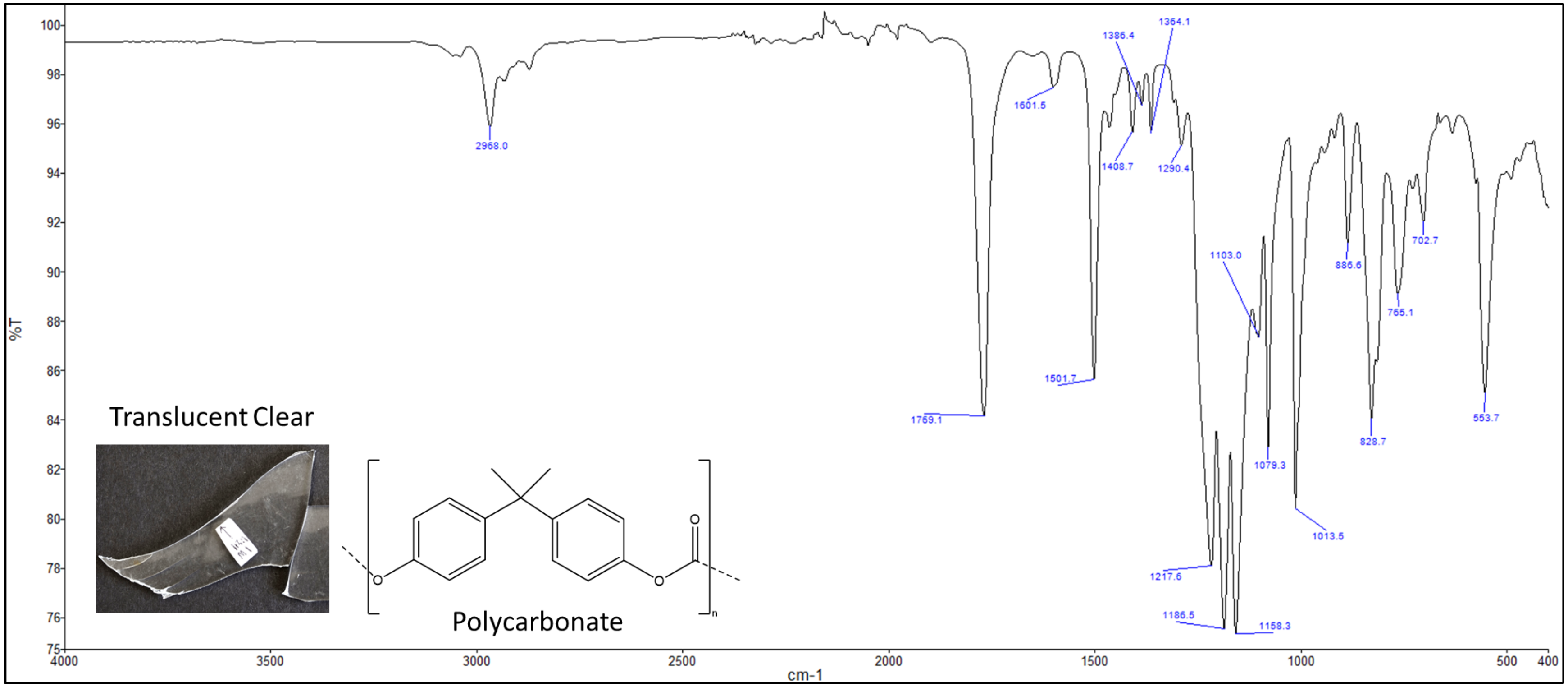


Figure 5A-3. FTIR spectra of one of the translucent clear sources. The main polymer type is polycarbonate.

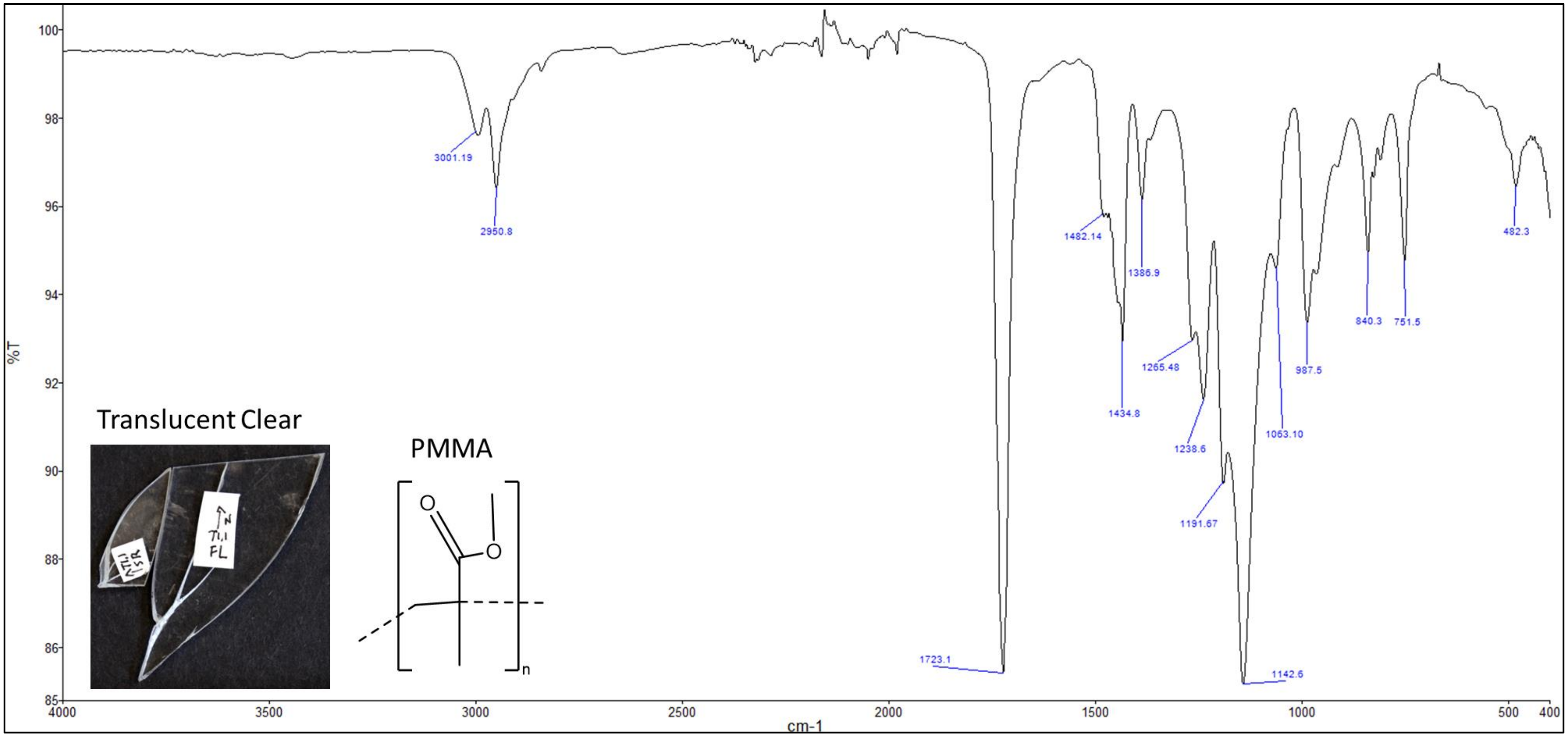


Figure 5A-4. FTIR spectra of the other translucent clear source. The main polymer type is poly(methyl methacrylate) or PMMA.

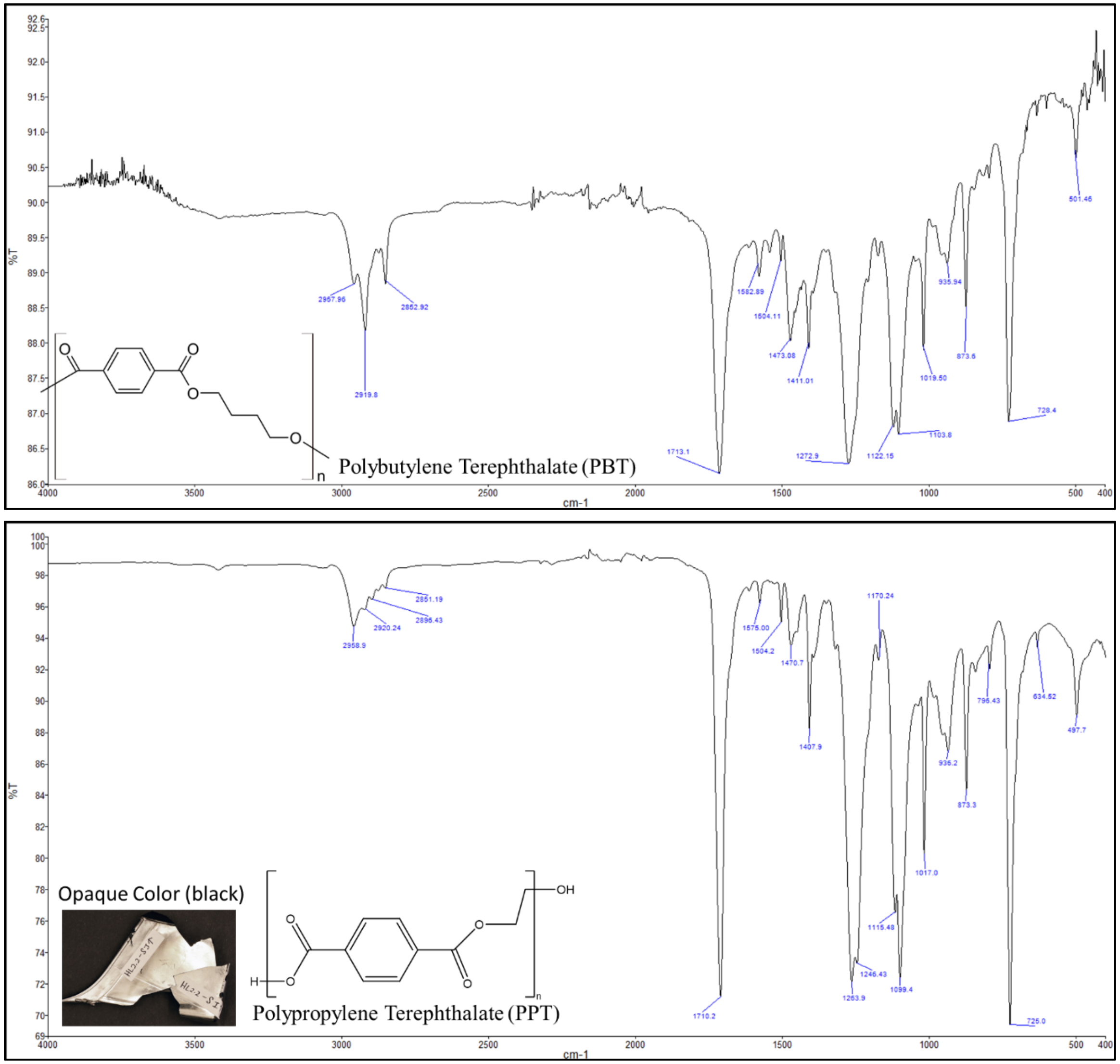


Figure 5A-5. FTIR spectra of the opaque color polymer. The silver coating on the top surface is polybutylene terephthalate (PBT), while the main body (the black portion) is polypropylene terephthalate (PPT).

Chapter 6: Development of methods for objective comparison of XRF spectra

6.1. Overview

Providing statistical support for X-ray Fluorescence (XRF) spectral comparisons using quantitative similarity measures is addressed in this study. A set of electrical tapes, most originating from different sources (94 rolls) and an additional subset originating from the same source (20 samples from the same roll), are characterized via XRF. Noise in spectra is filtered using Fast Fourier Transform (FFT), and baselines are corrected using a second derivative-constrained weighted regression. Then, spectral contrast angle ratios (SCAR) are calculated for each pairwise comparison (n=4,561). This criterion optimizes a threshold to assess associations or exclusions to minimize error rates. In addition, SCAR is used to classify samples using cluster analysis. An automated approach to sample comparison utilizing a random forest algorithm assists in identifying the basis for similarities or differences between compared spectra. This study presents an objective approach to reporting opinions and probabilistic determinations of spectral data that can be used as a model for other fields and materials. The SCAR metric can capture information on the variability between the compared samples and the variability within same-source samples. Also, the SCAR can support the forensic examiner's decision-making process and add transparency in various ways. It offers an automated and quantitative basis for associations or exclusions of spectral comparisons. It also performs well for classifying samples based on the class characteristics of elemental compositions and making inferences about the sample's origin.

6.2. Introduction

Pattern recognition tasks relying on spectral comparison between samples are common in the analytical and forensic chemistry fields [1]. Identifying unknown samples using reference libraries, comparing questioned samples to control samples to infer source identity, and classifying or discriminating samples are some examples of activities requiring the comparison of two or more spectra [1]. Various heuristic means and quantitative metrics exist to compare spectra. However, interpretation of the level of similarity between spectra in the context of the chosen pattern recognition task usually relies on the scientist's judgment, and decisions become more complex when facing high-dimensional spectra with subtle differences or when comparing sets of spectra to each other and having to account for within-source variability.

Heuristic comparison methods that rely primarily on human intervention, such as spectral overlay, are commonly used in Fourier Transformed Infrared Spectroscopy (FTIR) and Raman spectroscopy, X-ray fluorescence (XRF), Scanning Electron Microscopy-Energy Dispersive Spectroscopy (SEM-EDS), Laser-induced Breakdown Spectroscopy (LIBS), and microspectrophotometry (MSP) data [2-5]. A spectral overlay is a process that involves superimposing spectra from multiple samples to visually determine if any exclusionary differences are present that would indicate that the samples originate from different sources. This may include observable differences in peak heights, areas, or an element's presence or absence. A spectral overlay is valuable for distinguishing between samples with noticeable differences quickly. Still, it becomes more complex when the spectra are more similar or when the comparisons are performed using sets of spectra for accounting for within-source variability. In these cases, decisions based on spectral overlay may be inconsistent across examiners. The problem of subjectivity in human-based spectral comparisons is found across forensic disciplines and spectrochemical methods (e.g., comparison of IR, XRF, EDS, Raman, and mass spectra). In a recent review paper, Lavine and co-authors discuss the need for statistical support in comparing IR spectra to assess the quality of a spectral comparison and minimize the appearance of personal bias [1].

Quantitative metrics for the comparison of spectra have also been proposed. For instance, Lavine et al. and other research groups broadly report the statistical comparison of IR spectra of automotive paints for library searching and pattern recognition purposes [1, 13-25]. Other methods reported for spectral comparison are based on distance estimations (Euclidean, Mahalanobis, or Manhattan city block distance), correlation coefficients, or similarity indices [26-29]. Note that while using quantitative metrics may remove some of the bias involved with spectral overlay, using metrics in an inference scheme for comparing sets of spectra has yet to be satisfactorily addressed.

The Euclidean distance is one of the most common comparison metrics that estimate the distance between two vectors in two or three dimensions [30]. Euclidean distance has been used in various applications, such as paints, cotton, inks, and the profiling of drugs [1, 26-35]. The Mahalanobis distance is another metric helpful during the classification of items since it enables the normalization of multivariate measurements on samples [30, 36]. Another reported comparison criterion is the correlation coefficient (Pearson product-moment correlation coefficient). Correlation coefficients have been reported in the library, searching for inks and comparing cotton contaminants and spray paints [29, 34-38]. The correlation coefficient measures the strength and direction of the relationship between two variables. The value of the correlation coefficient ranges from -1 to +1. The closer the absolute value of the coefficient is to 1, the stronger the relationship between variables. In addition, the sign of the correlation coefficient indicates whether the relationship between variables is direct or inverse [30, 39, 40].

The probative value of the result of comparing sets of spectra is a central question of interest to the trier of fact. Using quantitative metrics is a step towards developing transparent and reproducible models to infer the identity of the source of samples characterized by analytical spectra. However, the field is still lacking these models. Thus, the need to develop more objective and statistical methods to support examiner decisions on multi-dimensional spectral data has been acknowledged as a critical research need by the Trace Materials Subcommittee of the Organization of Scientific Area Committees (OSAC) [6].

Tape materials such as duct tape and electrical tape have been extensively analyzed by various analytical techniques, including FTIR, SEM-EDS, µ-XRF, LIBS, laser-ablation inductively coupled plasma-mass spectrometry (LA-ICP-MS), and pyrolysis gas chromatography-mass spectrometry (Py/GC-MS) [3-5, 7-10]. Discrimination capabilities have been reported for each method independently and when combined in a complete analytical scheme. These studies found that elemental analysis is the most discriminating and informative technique for electrical tape samples [5, 7, 8].

Some statistical methods, such as analysis of variance (ANOVA), principal component analysis (PCA), and quadratic discriminant analysis (QDA), have been used to evaluate the elemental data collected from electrical tapes using XRF [5, 7, 8]. These methods are advantageous as they complement spectral overlay and lead to identifying groups of samples with similar formulations. Nonetheless, the spectral overlay is the predominant practice for analyzing x-ray data collected from forensic polymer, tape, and glass items [1, 11, 12]. Furthermore, XRF analyses of tape material show high within-sample variability when performed on different apparatuses. The lack of appropriate metrics and the high within-sample variability result in XRF data being restricted to comparing single spectra to each other and in limited applicability of XRF spectra for database searching.

This study proposes spectral contrast angle ratios (SCAR) as a complement to human-based comparisons. Using SCARs allows for the objective and transparent comparison of sets of spectra and for reducing the within-sample variability between XRF measurements made on different apparatuses (thus enabling library searches). Furthermore, SCAR is compatible with a two-stage inference model for assessing the probative value of material characterized by XRF. SCARs are based on the spectral contrast angle (SCA, also known as cosine dot product), a similarity measure often used in a mass spectral library [41, 44]. In this study, SCAs are calculated for pairs of spectra known to originate from a given source and for pairs of spectra potentially originating from different sources. The SCAR represents the ratio of the between-source and within-source SCAs. A previous publication by our group is used as the basis of this study, where SCAR is utilized to evaluate comparisons of X-ray fluorescence spectra of black electrical tape samples [8]. Our past results show that using SCARs dramatically improves the consistency of the discrimination of samples from different sources, even when the samples are analyzed on several configurations of XRF instruments [5,8]. This study expands our preliminary work by 1) improving the algorithms used to pre-process XRF spectra, 2) estimating the distribution of SCARs across populations known to originate from the same tape roll and different rolls, 3) identifying SCAR-based criteria for similarities and exclusions, and their respective error rates, and 4) utilizing SCAR for classification purposes. The classification approach adds value, as the elemental composition of electrical tape backing has shown potential for providing investigative information but is currently underutilized. The classification of sources using SCAR is validated using random forest models [45, 46]. These models are used to understand sample grouping based on spectral characteristics.

6.3. Methods and Materials

The data is collected from 94 black electrical tape samples previously characterized by our research group and other researchers [3-5, 7, 8]. The electrical tape samples are analyzed on a Thermo Scientific ARL QUANT'X energy-dispersive XRF spectrometer, utilizing filters to improve the signal-to-noise at specific energy levels. An aluminum filter enhances the signal-to-noise ratio of lower energy elements (low Zc), such as calcium, titanium, and iron. The mid-range energies (mid-Zc) are amplified using a thick palladium filter, while the high-range energies (high Zb) are amplified using a thick copper filter during collection. Three replicates are collected from each sample, and an additional subset of 20 samples collected from the same roll is used for an intra-roll assessment. The samples are pre-treated using hexane to remove the adhesive before analysis.

A spectral overlay is used to evaluate initial groupings based on similarities of elements and observed differences in peak heights. Pre-processing spectra remove noise and areas with minimal information, correct the baseline, and normalize the intensity. The spectra for the low filter are cleaned by removing the flat tail after 12 KeV, filtering noise using Fast Fourier transform (FFT), and then correcting the baseline using second derivative constrained weighted regression. The mid-filter data is cleaned by removing the tail after 33 KeV and filtering the noise using FFT. Finally, the high filter data is processed by removing the tails before 1.5 KeV and after 33 KeV, followed by reducing noise using FFT. Each spectrum is then normalized by dividing it by the intensity of the maximum peak.

Spectral contrast angles between two spectra x_i and x_j are calculated as described in [39], where $\langle x_i, x_j \rangle$ is the dot product and $\|x_i\|$ is the L2 norm:

$$SCA = acos\left(\frac{\langle x_i, x_j \rangle}{\|x_i\|\|x_j\|}\right) \qquad \text{(Equation 1)}$$

Angles between samples of the same source are calculated as *within-source angles*. For example, source A with three samples (A1, A2, A3) has within-sample angle comparisons of A1 vs. A2, A1 vs. A3, and A2 vs. A3. For each pairwise comparison, the angles between the samples from (potentially) different sources are calculated as the *between-source angles*. For two sources, A and B, this refers to comparisons between A1 vs. B1, A1 vs. B2, A1 vs. B3, A2 vs. B1, A2 vs. B2, A2 vs. B3, A3 vs. B1, A3 vs. B2, and A3 vs. B3.

SCARs are calculated by taking the ratio between the circular mean of the between-source angles and the circular mean of the within-source angles [8, 41].

$$Spectral\ Contrast\ Angle\ Ratio\ (SCAR) = \frac{Mean\ \theta\ (between\ source\ angle)}{Mean\ \theta\ (within\ source\ angles)} \qquad \text{(Equation 2)}$$

Hierarchical clustering is performed on the proximity matrix representing all pairwise SCARs between the 94 tapes. The resulting dendrogram is compared to the initial manual comparison based on the spectral overlay. Since SCARs are univariate measurements summarizing all similarities and differences between two sets of spectra, it is complicated to directly and easily interpret the clusters observed in the dendrogram. We interpret and validate the dendrogram clustering using random forest classifiers and compare the results with the analyst-based feedback determined by spectral overlay classification. The classifiers are designed to automatically detect common and differing peaks (based on intensity and slopes) between sets of XRF spectra. Commonalities and differences are assessed based on the importance of each wavelength (measured by its Gini index) to reach a correct classification [47].

All algorithms to process the spectra, calculate SCA and SCAR, and perform hierarchical clustering and random forest classification are implemented in R (version 4.1.3).

6.4. Results and Discussion

6.4.1. Sample comparison using spectral overlay

The electrical tape backings are first evaluated by spectral overlay to determine which chemical elements are present in the samples. This allows samples with noticeably different compositions to be quickly differentiated. All three spectra of one tape are overlaid over the spectra of another to perform the overlay. Present elements are recorded using a criterion of a signal-to-noise (SNR) ratio greater than 3, and areas of differences are observed and documented. Spectral overlay allows for a visual depiction of the variation within- and between sources. The spectra overlays are performed with a blind approach, where the ground truth is hidden from the analyst to avoid bias regarding the brand information of the tapes or how previous analysis methods have grouped them. While most tapes can be differentiated, numerous pairs cannot be distinguished or demonstrate only minor similarities and differences.

From these comparisons, 41 distinct groups are formed, with additional differentiation between sources through observing more minor differences in peak shape or height, leading to 61 total sub-groups (**Figure 6-1**). Comparisons are made blindly and independently corroborated by two analysts. Any reported discrepancy is discussed and resolved with a third reviewer. Based on the spectral overlay of XRF data, a discrimination power of 97.0% is obtained (132 indistinguishable pairs out of a total of 4371 comparisons) [8]. The manual process of overlaying spectra is tedious due to the extensive pairwise comparisons. Still, it proved helpful for the analyst in gaining knowledge of the elemental compositions of the tapes and setting the basis of the observed differences or similarities.

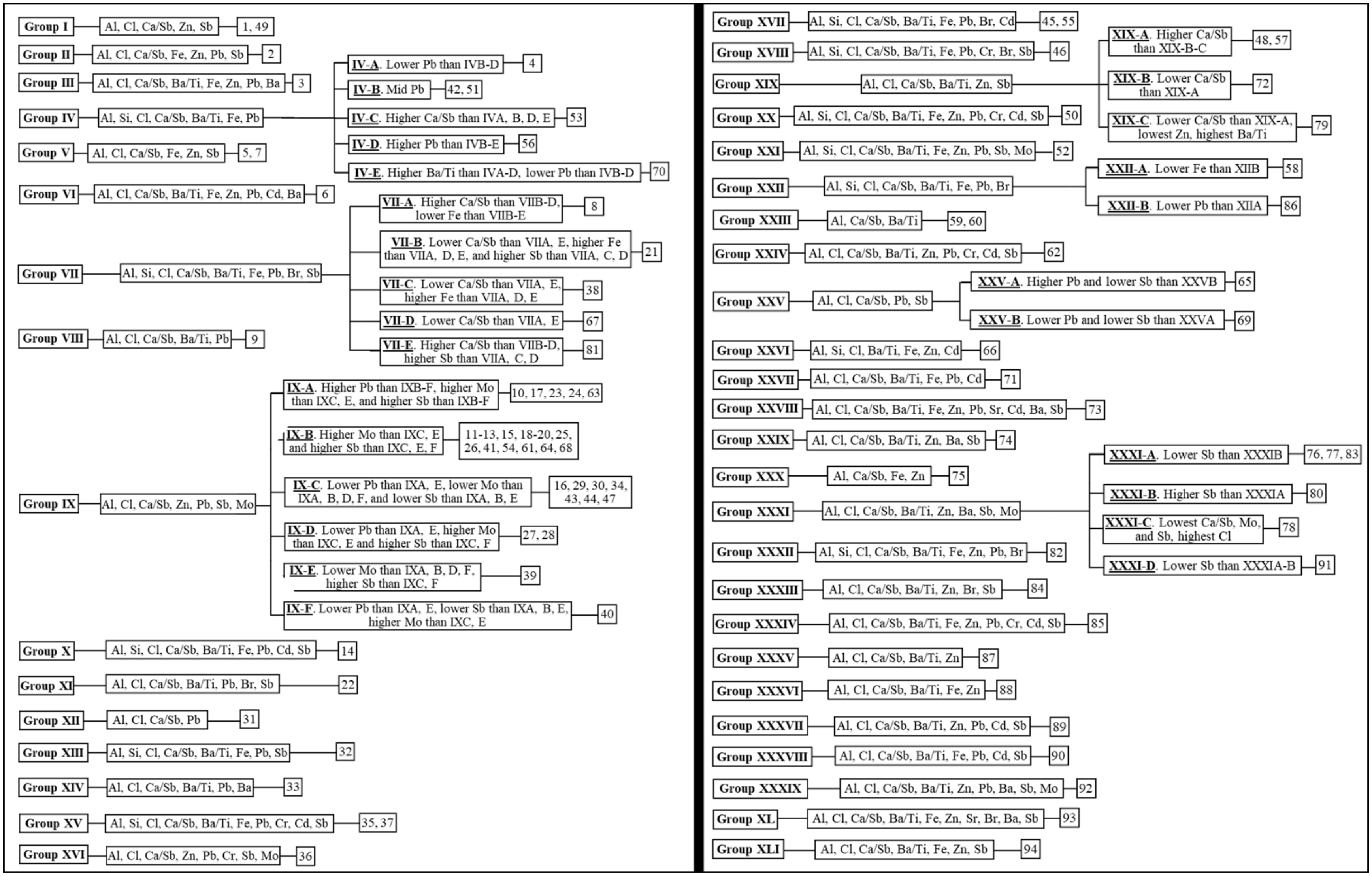

Group I
Al, Cl, Ca/Sb, Zn, Sb
1, 49
Group II
Al, Cl, Ca/Sb, Fe, Zn, Pb, Sb
2
Group III
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Ba
3
Group IV
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb
IV-A. Lower Pb than IVB-D
4
IV-B. Mid Pb
42, 51
IV-C. Higher Ca/Sb than IVA, B, D, E
53
IV-D. Higher Pb than IVB-E
56
IV-E. Higher Ba/Ti than IVA-D, lower Pb than IVB-D
70
Group V
Al, Cl, Ca/Sb, Fe, Zn, Sb
5, 7
Group VI
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Cd, Ba
6
Group VII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Br, Sb
VII-A. Higher Ca/Sb than VIIB-D, lower Fe than VIIB-E
8
VII-B. Lower Ca/Sb than VIIA, E, higher Fe than VIIA, D, E, and higher Sb than VIIA, C, D
21
VII-C. Lower Ca/Sb than VIIA, E, higher Fe than VIIA, D, E
38
VII-D. Lower Ca/Sb than VIIA, E
67
VII-E. Higher Ca/Sb than VIIB-D, higher Sb than VIIA, C, D
81
Group VIII
Al, Cl, Ca/Sb, Ba/Ti, Pb
9
Group IX
Al, Cl, Ca/Sb, Zn, Pb, Sb, Mo
IX-A. Higher Pb than IXB-F, higher Mo than IXC, E, and higher Sb than IXB-F
10, 17, 23, 24, 63
IX-B. Higher Mo than IXC, E and higher Sb than IXC, E, F
11-13, 15, 18-20, 25, 26, 41, 54, 61, 64, 68
IX-C. Lower Pb than IXA, E, lower Mo than IXA, B, D, F, and lower Sb than IXA, B, E
16, 29, 30, 34, 43, 44, 47
IX-D. Lower Pb than IXA, E, higher Mo than IXC, E and higher Sb than IXC, F
27, 28
IX-E. Lower Mo than IXA, B, D, F, higher Sb than IXC, F
39
IX-F. Lower Pb than IXA, E, lower Sb than IXA, B, E, higher Mo than IXC, E
40
Group X
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Cd, Sb
14
Group XI
Al, Cl, Ca/Sb, Ba/Ti, Pb, Br, Sb
22
Group XII
Al, Cl, Ca/Sb, Pb
31
Group XIII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Sb
32
Group XIV
Al, Cl, Ca/Sb, Ba/Ti, Pb, Ba
33
Group XV
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Cr, Cd, Sb
35, 37
Group XVI
Al, Cl, Ca/Sb, Zn, Pb, Cr, Sb, Mo
36
Group XVII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Br, Cd
45, 55
Group XVIII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Cr, Br, Sb
46
Group XIX
Al, Cl, Ca/Sb, Ba/Ti, Zn, Sb
XIX-A. Higher Ca/Sb than XIX-B-C
48, 57
XIX-B. Lower Ca/Sb than XIX-A
72
XIX-C. Lower Ca/Sb than XIX-A, lowest Zn, highest Ba/Ti
79
Group XX
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Cr, Cd, Sb
50
Group XXI
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Sb, Mo
52
Group XXII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Br
XXII-A. Lower Fe than XIIB
58
XXII-B. Lower Pb than XIIA
86
Group XXIII
Al, Ca/Sb, Ba/Ti
59, 60
Group XXIV
Al, Cl, Ca/Sb, Ba/Ti, Zn, Pb, Cr, Cd, Sb
62
Group XXV
Al, Cl, Ca/Sb, Pb, Sb
XXV-A. Higher Pb and lower Sb than XXVB
65
XXV-B. Lower Pb and lower Sb than XXVA
69
Group XXVI
Al, Si, Cl, Ba/Ti, Fe, Zn, Cd
66
Group XXVII
Al, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Cd
71
Group XXVIII
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Sr, Cd, Ba, Sb
73
Group XXIX
Al, Cl, Ca/Sb, Ba/Ti, Zn, Ba, Sb
74
Group XXX
Al, Ca/Sb, Fe, Zn
75
Group XXXI
Al, Cl, Ca/Sb, Ba/Ti, Zn, Ba, Sb, Mo
XXXI-A. Lower Sb than XXXIB
76, 77, 83
XXXI-B. Higher Sb than XXXIA
80
XXXI-C. Lowest Ca/Sb, Mo, and Sb, highest Cl
78
XXXI-D. Lower Sb than XXXIA-B
91
Group XXXII
Al, Si, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Br
82
Group XXXIII
Al, Cl, Ca/Sb, Ba/Ti, Zn, Br, Sb
84
Group XXXIV
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Pb, Cr, Cd, Sb
85
Group XXXV
Al, Cl, Ca/Sb, Ba/Ti, Zn
87
Group XXXVI
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn
88
Group XXXVII
Al, Cl, Ca/Sb, Ba/Ti, Zn, Pb, Cd, Sb
89
Group XXXVIII
Al, Cl, Ca/Sb, Ba/Ti, Fe, Pb, Cd, Sb
90
Group XXXIX
Al, Cl, Ca/Sb, Ba/Ti, Zn, Pb, Ba, Sb, Mo
92
Group XL
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Sr, Br, Ba, Sb
93
Group XLI
Al, Cl, Ca/Sb, Ba/Ti, Fe, Zn, Sb
94

Figure 6-1. The 94 different source electrical tapes are classified into 41 distinct groups based on composition and discrimination using spectral overlay. Sources with the same elemental composition are grouped but can be further divided into 61 subgroups based on observed differences in elemental peak heights. Sources and groups listed here are not arranged in any hierarchical order.

Of note for these groups are the tapes that could not be distinguished. These sources are often not distinguishable by other analytical methods or a combination of methods used, supporting our findings [3-5, 7, 8]. For example, using the spectral overlay of the XRF data, Group IX contains a total of 30 sources, differentiated into six major subgroups (A to F) based on the relative concentration of the elements present. These tapes originate from the manufacturer 3M, all made in the United States (see **Appendix A**). As such, it is unsurprising that most tapes have very similar physical and chemical compositions.

Interestingly, Group XXXI also contains tapes manufactured by 3M in the United States but do not have detectable lead, as seen in the tapes from Group IX. This same grouping is seen using both XRF and LA-ICP-MS and is further demonstrated with the clusters observed in PCA analysis of the samples conducted by Martinez-Lopez et al. [7]. Remarkably, similar classifications are observed when using only elemental analysis (LA-ICP-MS or XRF) or a comprehensive analytical protocol (microscopic examination, FTIR, SEM-EDS, and Py-GC-MS), demonstrating the highly informative power of sensitive elemental techniques for these tapes and its potential to use it as a rapid screening tool. These findings can lead to more cost-effective changes in current casework workflows by introducing elemental analysis earlier in the analytical sequence, reducing backlogs when exclusionary differences are detected.

In addition, tapes originating from different countries could be identified as different, even if made by the same manufacturer. For example, 3M tapes manufactured in Taiwan can be distinguished from those made by the same company in the United States [7, 8]. The consistency of the results from these methods demonstrates the value of elemental analysis of electrical tape backings. It shows that XRF can be a valuable technique for analyzing tapes with high specificity and sensitivity. This type of information can become helpful in providing leads during an investigation. In a word of caution, it should be noted that a single manufacturer can have the same product produced at different sites, meaning sources labeled under the same product name could potentially have different chemical compositions. On the other hand, tapes made by the same manufacturer may be labeled under different product names by the distributors, leading to differently labeled sources having a similar chemical composition. These factors may affect the interpretation of potential manufacturing sources.

6.4.2. Application of spectral contrast angle ratio

Despite the value of spectral overlay for determining the composition and identification of different samples, there are instances where it is difficult for the analyst to judge whether two samples have true differences beyond expected intra-sample variation. For example, tapes 10 and 26 had the same elemental compositions. While there are some minor observable variances in height at the calcium and antimony peaks (3.4 KeV) and zinc (8.6 KeV) (**Figure 6-2**), those differences are insufficient for some examiners to report the samples as distinguishable.

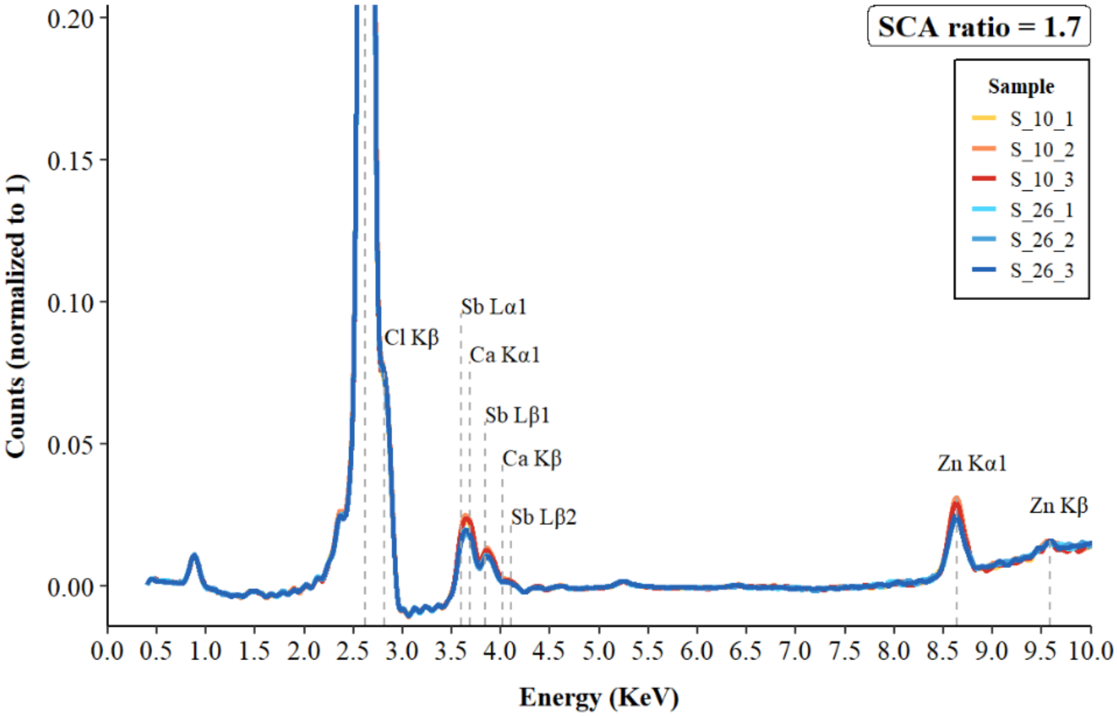


Figure 6-2. Spectral overlay of samples from tapes 10 and 26. There is a minor height difference in the calcium/antimony peaks at 3.4 KeV and the zinc peak at 8.6 KeV, where sample 10 has a slightly higher concentration of those elements. The corresponding SCAR for these samples is low (1.7).

Spectral contrast angle ratios are calculated to provide a means to compare the spectra quantitatively and objectively. At the same time, spectral contrast angles can represent the difference between two individual spectra. Utilizing SCA ratios allows for the variability within and between samples to be accounted for [8, 39]. SCARs are calculated for all pairwise comparisons of the 94-tape set and the 20 intra-roll samples for the data collected under all three filters.

Figure 6-3 shows the SCAR distribution of all pairwise comparisons for 94 tapes (n=4371) and 20 intra-roll samples (n=190). Samples originating from the same roll (intra-roll) result in SCARs between 0.94 and 2.04. Meanwhile, samples from sources identified during spectral overlay as being classified into different groups (inter-group) have ratios ranging from 0.98 to 112. Samples from sources belonging to a common subgroup (e.g., Group VII-A) result in ratio values between 0.96 and 4.34. In contrast, inter-subgroup pairs (i.e., tapes belonging to different subgroups with subtle differences, such as Group VII-A and Group VII-B) result in ratios ranging between 0.98 and 14.97.

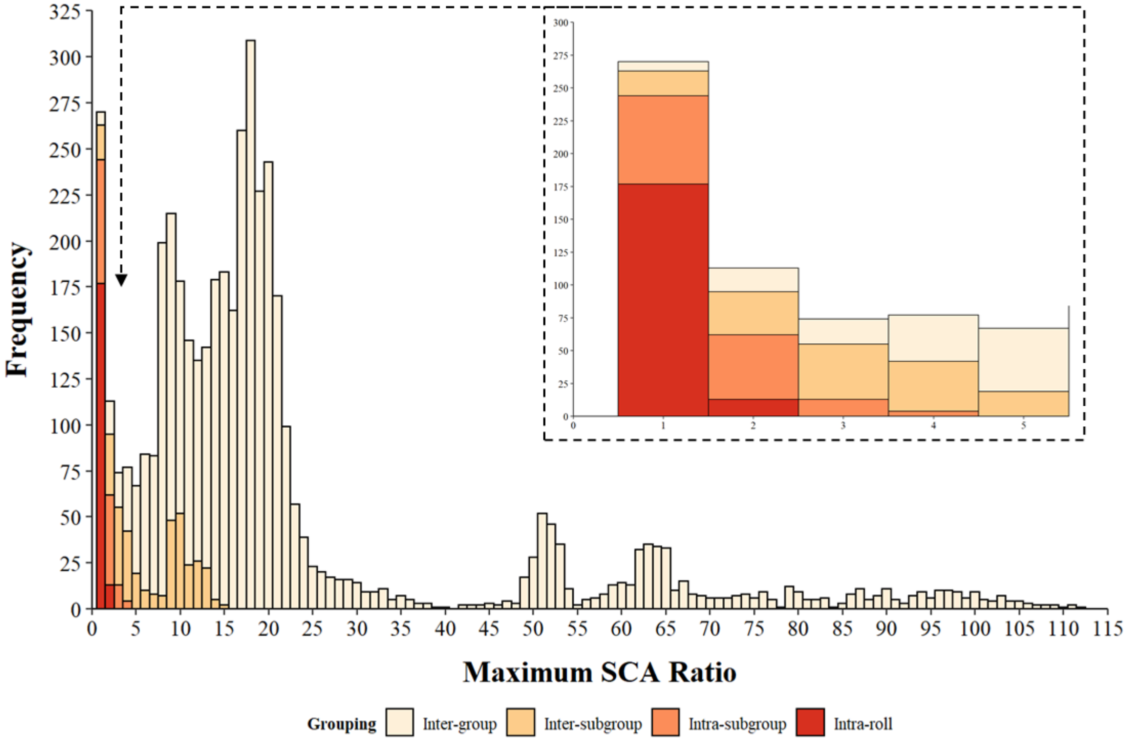


Figure 6-3. Histogram representing the frequency distribution of SCAR for all pairwise comparisons of electrical tape samples from different-source and same-source. The zoomed insert shows the range of 0 to 5 in more detail, indicating that most intra-roll pairs had SCA ratios less than 2.

In theory, tapes with the same composition could be expected to have an SCA ratio of 1.0, as the between and within angles should be the same. In practice, however, minor variability in the analytical process or the tape composition produces some deviations from a SCAR of 1.0. Still, in these same-source cases, the SCAR values are much lower than those from different-source items. For example, intra-roll samples 2 and 5 (**Figure 6-4**) have an SCA ratio of 1.4, with no observable differences in the spectral overlay in the low-energy region.

The tapes with SCA ratios close to 1.0 are more likely to be reported as indistinguishable by spectral overlay, even if they originated from different rolls. For example, samples 1 and 49 (**Figure 6-5A**) demonstrate only subtle variations in peak height for calcium and antimony and have an SCA ratio of 2.7 within the low filter data. These samples are assigned to the same group by spectral overlay observations. In comparison, tapes classified into different subgroups during spectral overlay showed slightly higher SCA ratios, such as tapes 8 and 81 (**Figure 6-5B**), with an SCA ratio of 3.2 in the low-energy region. These tapes are classified as Groups VII-A and VII-E, respectively, due to a slightly higher concentration of calcium/antimony in tape 8, observable at approximately 3.6 KeV.

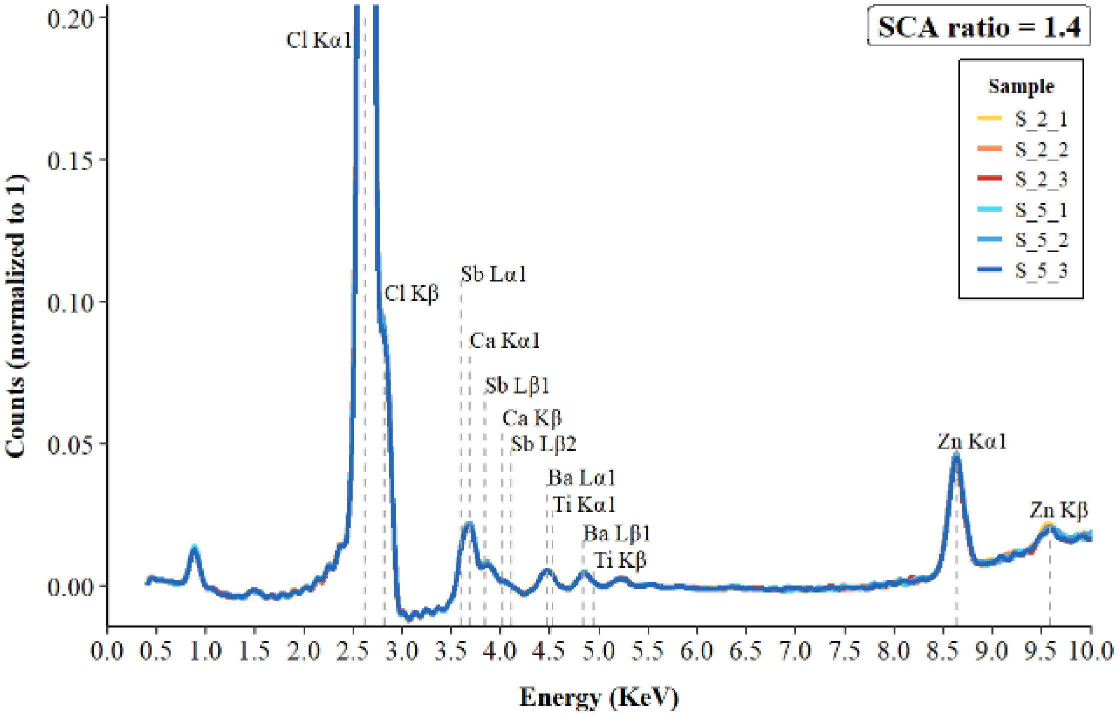


Figure 6-4. Spectral overlay of intra-roll samples 2 and 5. In the low-energy region, the samples contain chlorine, antimony, calcium, barium, titanium, and zinc. In the low-energy region, the emission lines of antimony/calcium and barium/titanium cannot be resolved, so they are labeled together. The SCAR for these samples from the same roll is 1.4.

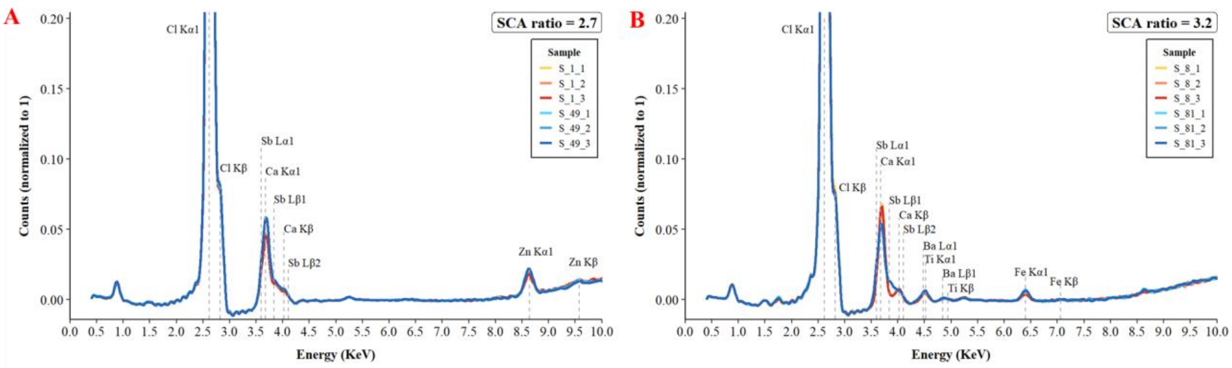


Figure 6-5A (left). Spectral overlay of samples from tapes 1 and 49, which are classified into the same group (I, **Figure 6-1**). While minor height variations exist, the spectral overlay could not distinguish the tapes. **Figure 6-5B** (right) shows the overlay of tapes 8 and 81, which belong to the same major group VII, but are separated into subgroups VII-A and VII-E due to higher concentration of calcium/antimony (3.6 KeV) in sample 8.

In contrast, tapes that showed more considerable visual differences during spectral overlay displayed large SCA ratios. An example is shown in **Figure 6-6**, with tapes 39 and 66, 4 and 50, and 5 and 85, indicating SCA ratios of 8.7, 15.9, and 60.8, respectively.

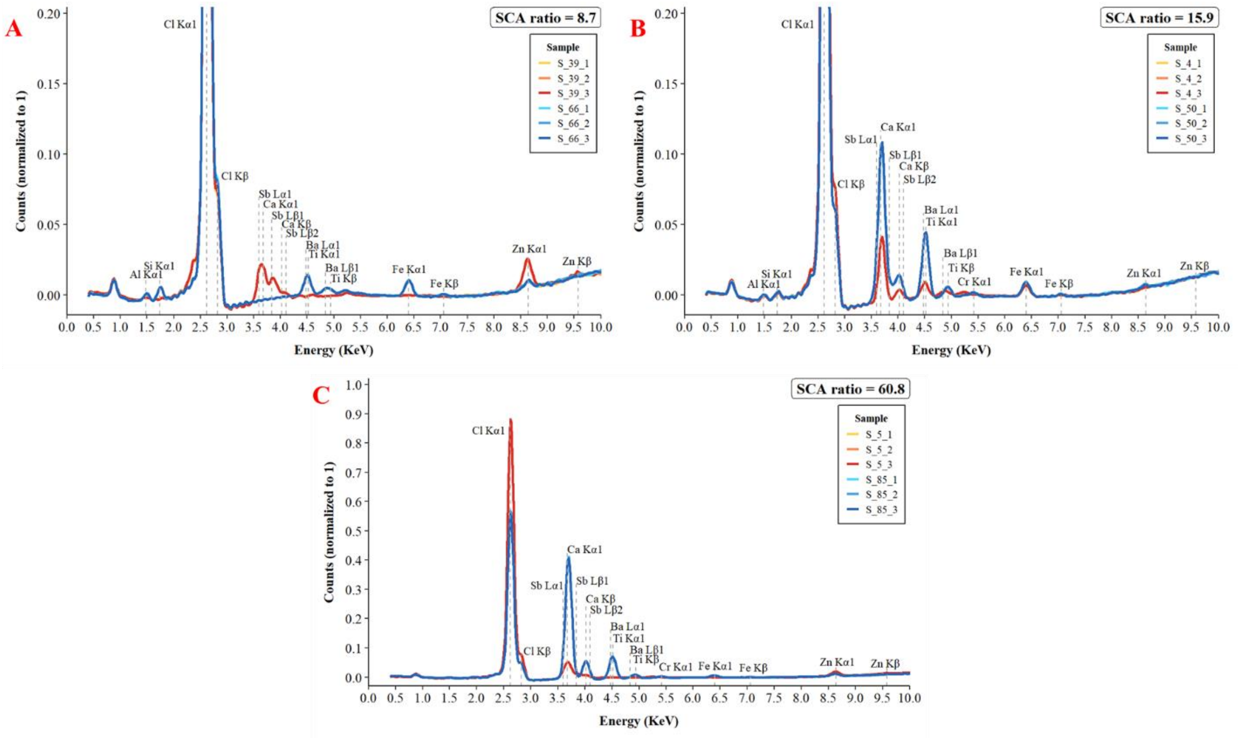


Figure 6-6. Spectral overlays of tapes are found in different groups by spectra overlay, all with high SCA ratios in the low-energy region. **Figure 6-6A** demonstrates tape 39 with an elevated antimony/calcium and zinc concentration, while tape 66 has higher barium/titanium and iron. **Figure 6-6B** shows tape 50 with a higher concentration of antimony/calcium and barium/titanium than tape 4. **Figure 6-6C** shows tape 5 with a substantially higher chlorine concentration, while 85 has a high Ca/Sb and Ba/Ti concentration.

Same-source and different-source SCAR are well separated, showing the potential to assist analysts when making determinations of distinguishability. Therefore, a threshold is calculated to determine the SCA ratio value that minimizes the false inclusion and false exclusion rates (**Figure 6-7**). Using the SCAR for each pairwise comparison, a threshold of 1.95 resulted in a 3.4% false inclusion rate (148 of 4371 different-source comparison pairs) and a 2.1% false exclusion rate (4 of 190 same-source comparison pairs). As a result, the discrimination power of the set is 96.4% when using this SCAR criterion.

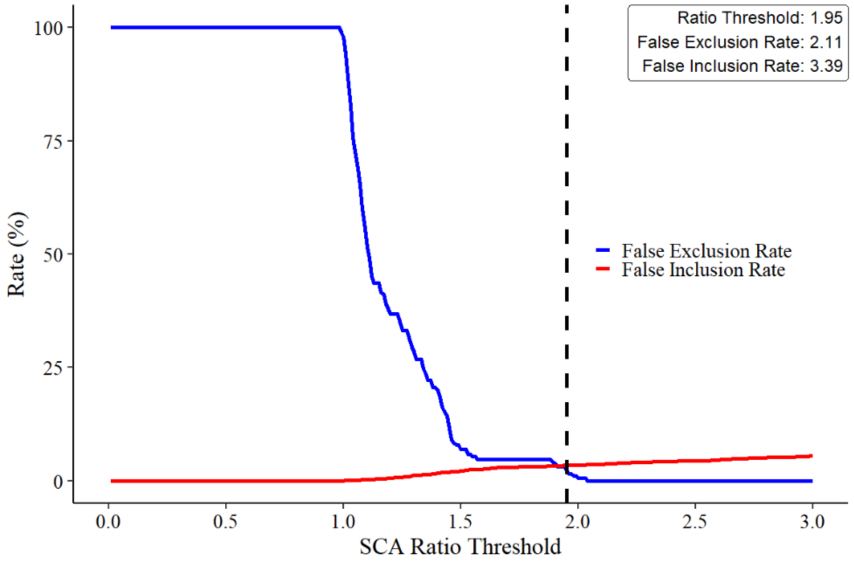


Figure 6-7. Rates of false inclusion and false exclusion of all comparison pairs at different SCA ratios. The ratio value that minimized the false exclusion and false inclusion rate is 1.95.

However, while some additional pairs of tapes are differentiated with the inclusion of the SCAR threshold of 1.95, many intra-subgroup pairs identified by spectral overlay remain indistinguishable. This is primarily observed in tapes from Group IX. The lack of complete discrimination aligns with previous findings, indicating that some samples in this tape collection are similar enough in their formulation to be indistinguishable by XRF and other techniques (e.g., LA-ICP-MS, SEM-EDS, FTIR, Py-GC-MS) [3-5, 7, 8]. Moreover, the agreement of spectral overlay and SCA ratios shows that SCA ratios can be estimated as a complementary means to further support the examiner's observations as an independent metric. More importantly, the SCAR provides a metric that complements the binary decision of association or exclusion with a continuous quantitative value that assesses the level of similarity and, thus, can assist in evaluating the value of the evidence.

6.4.3. Automated clustering and sample comparison

The development of methods to assign the probative value of comparison between tapes, such as the two-stage approach, requires estimating some measure of the rarity of the tapes' characteristics. This involves studying many tapes and defining a measure to assign probabilities. Section 3.1 shows that tapes can be based on their elemental spectra and that some clusters have more members than others. This observation creates a method to assign match probabilities to tape elemental composition based on clustering many randomly selected tapes.

Unfortunately, the manual grouping reported in Section 3.1 is a tedious process, particularly as the number of tapes increases, and may need more reproducibility if performed by the spectral overlay. In this section, we propose a method relying on the hierarchical clustering of tapes based on their pairwise SCARs. The matrix of pairwise SCARs used to build the dendrogram is a weighted sum of the matrices of the pairwise SCARs for each filter used in the XRF analysis (low, medium, and high), as described above. The weights are determined by optimization as 28.98 for the low, 21.63 for the mid, and 1.15 for the high filter. Hence, the low filter has the most significant influence on the distance between samples. This is consistent with the observations made during spectral overlay, as elements accentuated in the low filter often provided the most discrimination between samples. The mid filter also substantially influences the resulting distance between samples, but the high filter only has a minor influence. However, the high filter is where the resolved peaks for antimony can be observed, providing additional discrimination between samples. We validated the hierarchical clustering method by comparing the clusters resulting from the hierarchical clustering with the grouping described in **Figure 6-1**.

Since SCARs are univariate measures summarizing all similarities and differences between two sets of spectra into a single value, we propose to use a random forest classifier to interpret the grouping done by hierarchical clustering. Note that hierarchical clustering is unsupervised and can be used without prior knowledge of the number of clusters or the assignment of tapes to given classes. On the contrary, random forest classifiers require the definition of several classes and to development of a training set of tapes with known target classes. In this project, we rely on hierarchical clustering to group the tapes. Once grouped, we train the random forest classifier to study the similarities and differences between the tapes in pairs of clusters.

Figure 6-8 shows the dendrogram with 61 classes labeled and color-coded. Most of these clusters correspond with groups and subgroups identified by the manual comparison using spectral overlay. We note that the hierarchical clustering complements spectral overlay by providing a visual representation of those tapes that display similarities and differences, with a relative dimension of the level of correspondences or distinctions among groups.

Most tapes are consistent with the groups observed by spectral overlay, with a few exceptions (4 out of 61 groupings). Those exceptions, seen in clusters 29, 38, 40, and 51 (**Figure 6-8**, highlighted in yellow), tend to be tapes identified as having slightly different compositions during spectral overlay. Still, the elemental differences are generally minimal, or the elements are very close to the signal-to-noise ratio of 3. Moreover, because spectral overlay cannot assign hierarchical grouping, some clusters are called by observed differences through pairwise comparisons and not grouped by the closeness of similar profiles between subgroups. The complexity of capturing similarity information by spectral overlay increases with the number of comparisons, thus missing that valuable information in the comparison process.

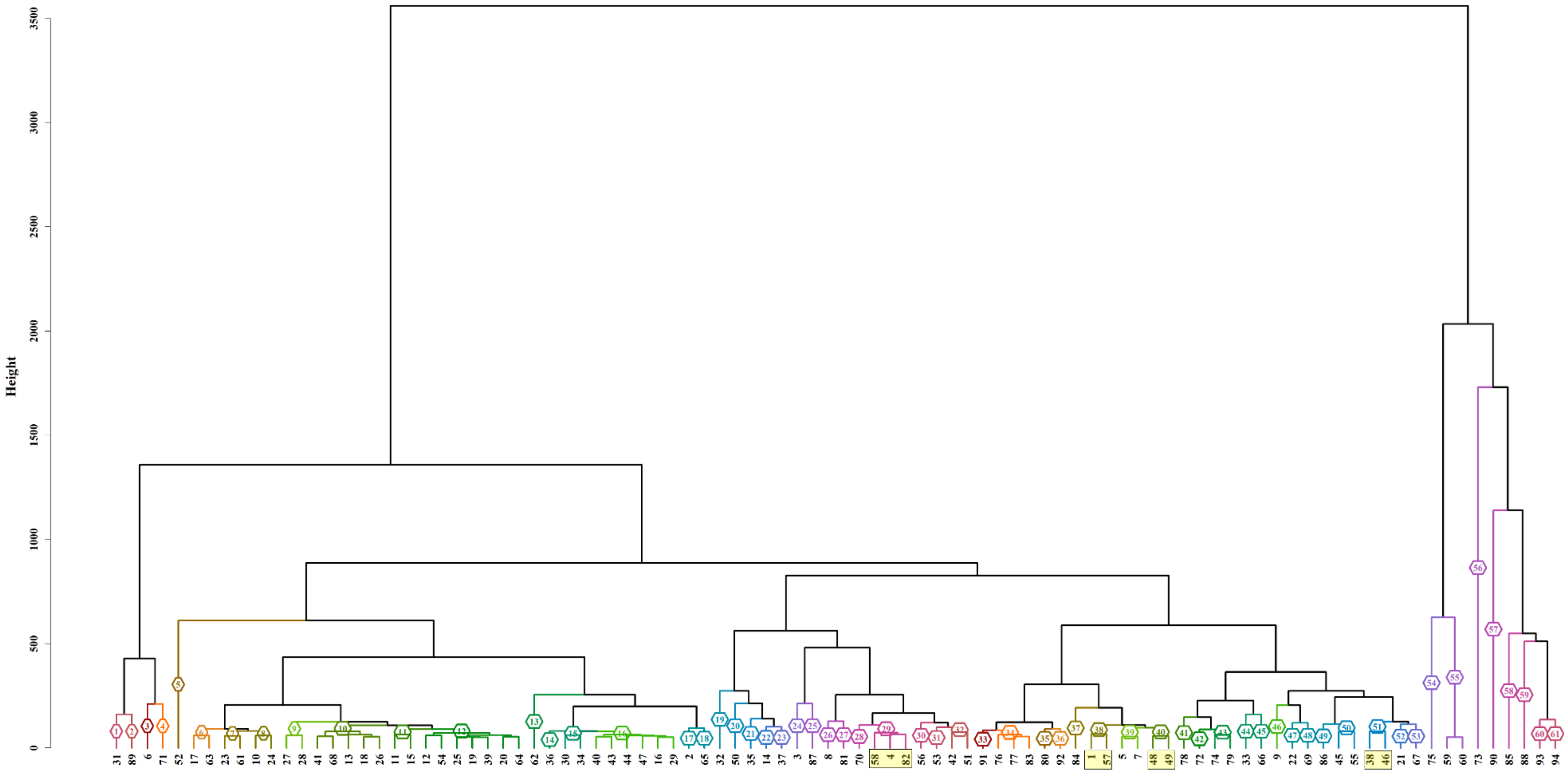


Figure 6-8. Dendrogram representing the clustered tapes from the dataset of 94 different sources based on the weighted average of SCA ratios from each filter. The data is divided into 61 distinct clusters to compare the clustering analysis to the original 61 groups identified during spectral overlay. Only 4 clusters in this dendrogram are not generally consistent with the spectral overlay groups, highlighted in yellow (clusters 29, 38, 40, and 51).

A discrepancy is observed between spectral overlay and SCAR-clustering in cluster 29 (tapes 4, 58, 82), where each member is separated into their main groups by spectral overlay (**Figure 6-1**). However, the elemental composition of those tapes all had aluminum, chlorine, calcium/antimony, barium/titanium, iron, and lead in common. The major difference is that tape 4 does not show bromine, while tape 82 is noted to have zinc (**Figure 6-9**).

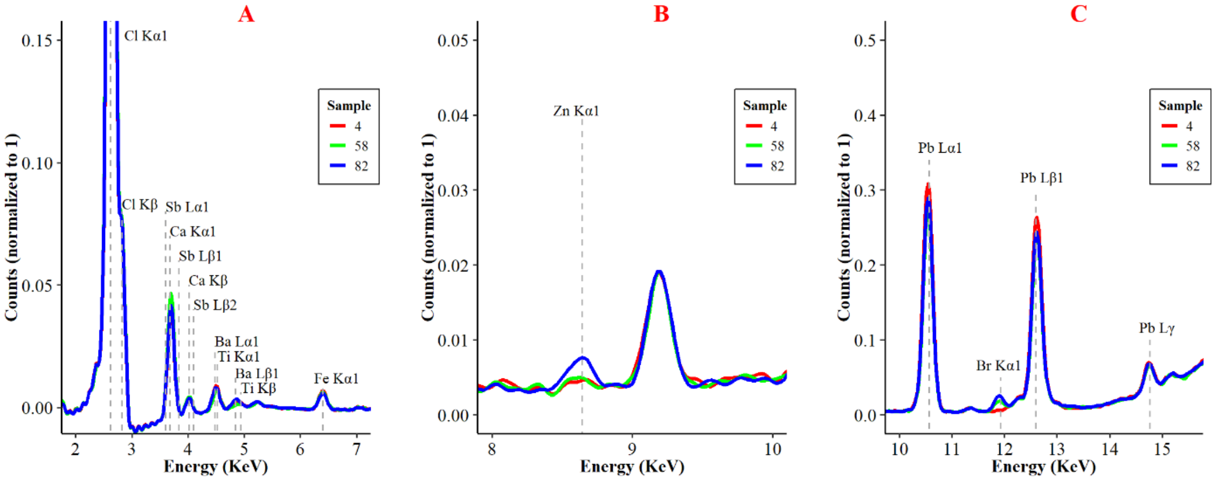


Figure 6-9. Spectral overlay comparison of tapes 4, 58, and 82. Despite being separated into different main groups by spectral overlay alone, these tapes have a comparable content of major elements such as chlorine, calcium/antimony, barium/titanium, and iron (seen in the low filter region – **Figure 6-9A**). They are also consistent in lead content, as seen in the mid-filter region (**Figure 6-9C**). There is a slight height difference in the zinc peak at 8.6 KeV (**Figure 6-9B**) and the bromine peak at 11.8 KeV (**Figure 6-9C**).

The bromine content in tapes 58 and 82 is visible when observing the overlaid spectra. Still, it is relatively small compared to the content of other elements, such as calcium/antimony, barium/titanium, and lead, consistent with all three tapes (**Figure 6-9**). The same is true of the zinc content, indicating those two elements are not influential on the final clustering of these tapes. Moreover, the calculated SCAR values for these three tapes are consistent with the cluster observation. Comparing samples 4 to 58, 4 to 82, and 58 to 82 resulted in SCAR values less than 1.5 for low, mid, and high filters, which puts all SCAR values below the calculated threshold of 1.95 to distinguish different-source samples.

The discrepancy between the manual and automated grouping resulting in cluster 51 (tapes 38 and 46) had a similar explanation. Tape 36 is initially separated into Group VII by spectral overlay, while tape 46 is in Group XVIII (**Figure 6-1**). These groups had various elements in common, including aluminum, silicon, chlorine, calcium/antimony, barium/titanium, iron, lead, bromine, and antimony (resolved in the high filter region). However, when looking at the tapes by spectral overlay, as shown in **Figure 6-10**, there is an observable difference in the antimony content in the high region (between 26 and 30 KeV).

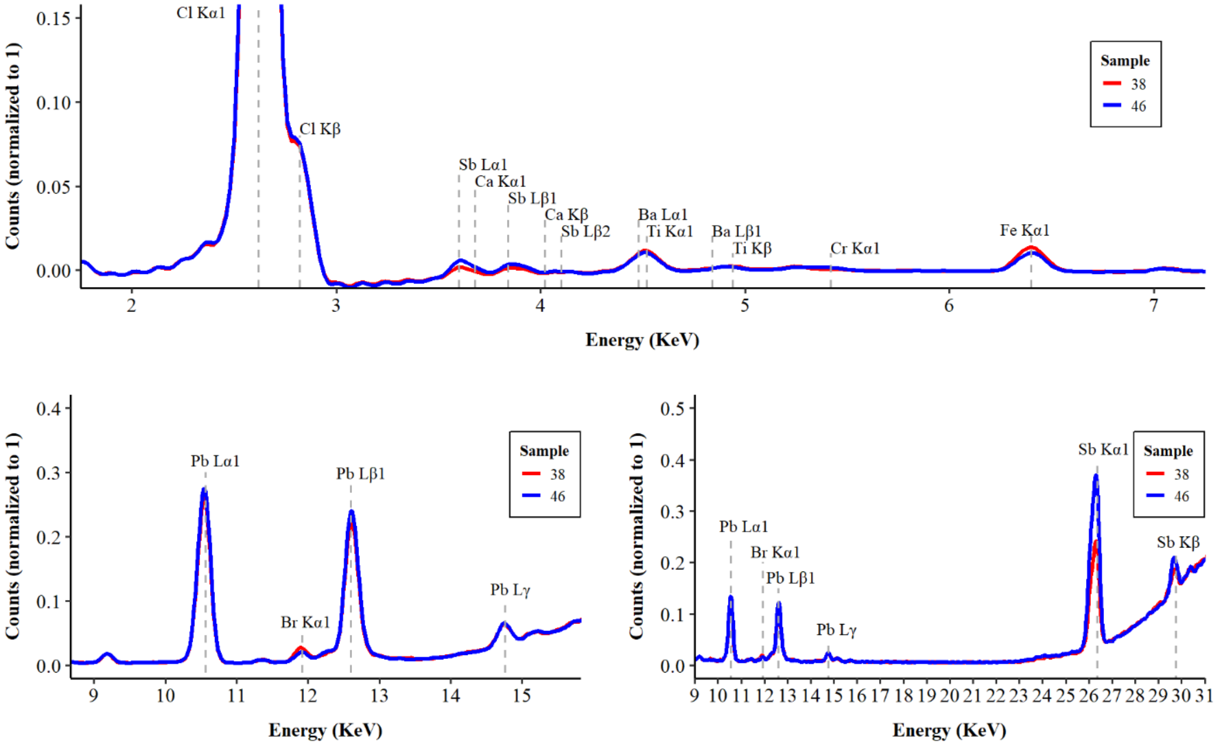


Figure 6-10. Spectral overlay comparison of tapes 38 and 46. The tapes have comparable compositions in the low region (top figure) with similar amounts of chlorine, calcium/antimony, barium/titanium, and iron. They are also consistent in lead and bromine content, as seen in the mid-filter region (bottom left figure). An observable height difference in the antimony peak in the high filter region (bottom right figure) at 26.5 KeV.

The high filter SCAR value for these two tapes is also slightly higher, with a value of 3.6. However, as described previously, the high filter is weighted less than the other two regions for the cluster analysis. As such, the SCAR values of 1.7 and 1.2 for the low and mid filters for these tapes significantly influenced the relatively small distance between tapes 38 and 46.

Finally, the last case of discrepancy between the dendrogram and manual spectral overlay analysis included tapes 1, 57 (cluster 38), and tapes 49 and 48 (cluster 40). Tapes 1 and 49 are grouped by spectral overlay into Group I, while tapes 57 and 48 are sorted into Group XIX-A. However, with cluster analysis, tape 1 is closest to 57 than to 48 and 49. This is likely due to the calcium\antimony content observed in the low region and differences in the antimony content seen in the high region (**Figure 6-11**). These samples resulted in relatively low SCAR values, all less than 3. Pairwise comparisons of 1 to 57 and 48 to 49 are the smallest, at SCARs of 1.4 and 1.2, respectively.

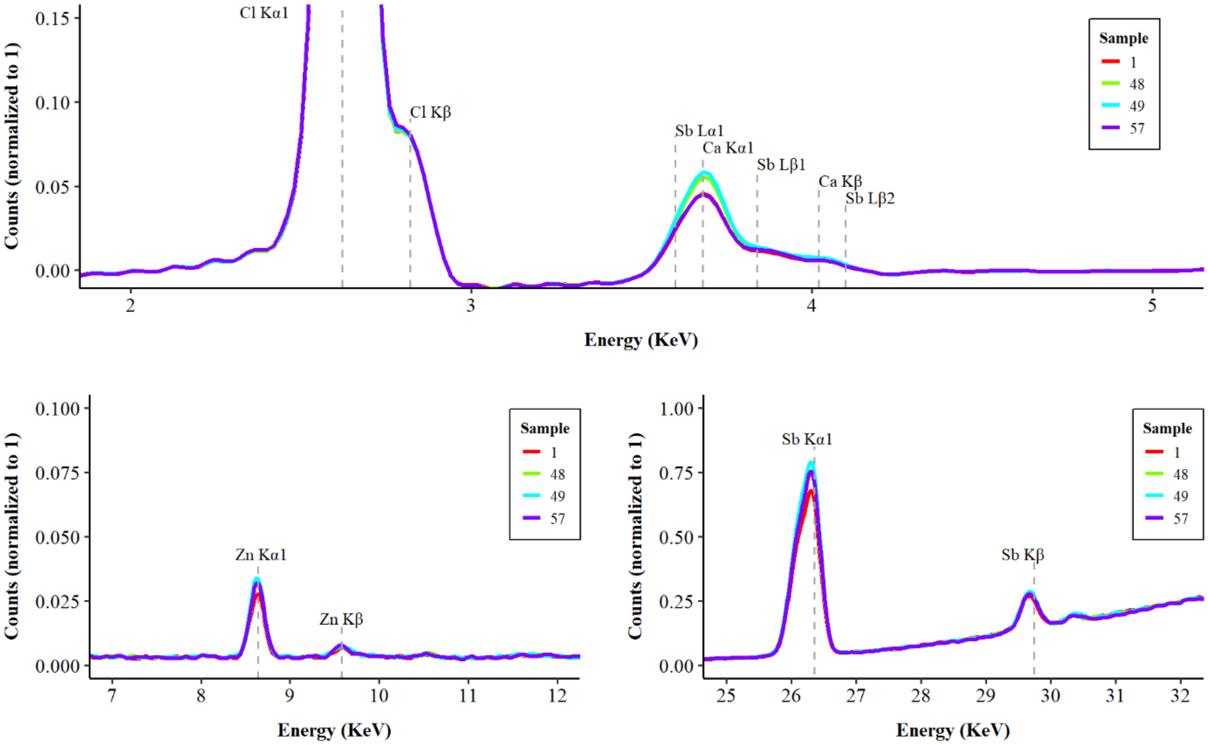


Figure 6-11. Spectral overlay comparison of tapes 1, 48, 49, and 57. In the low region (top figure), the tapes have comparable compositions with similar amounts of chlorine and calcium/antimony, though tapes 48 and 49 have a slightly higher concentration. They are also consistent in zinc content, as seen in the mid-filter region (bottom left figure), although the zinc concentration is relatively low overall. An observable height difference in the antimony peak in the high filter region (bottom right figure) at 26.5 KeV.

Overall, hierarchical clustering demonstrates the effectiveness of using the SCAR as a metric to assess the similarity of spectra and as an informative input for cluster analysis for large populations of samples. The spectral overlay provides a quick means of understanding the elemental composition, the intra-variability of samples, and the similarities and differences between individual samples. However, the spectral overlay of many sources is unrealistic. It fails to represent any hierarchy between multiple samples or a quantitative representation of how comparable the composition of groups of tapes is.

The random forest classifiers and spectral overlay often detect the same similarities and differences. Below, we discuss examples from the most challenging group, classified by spectral overlay as group IX (subgroups IX-A to IX-F) and sharing the same tree node for clusters 7 to 12. All tapes on this major node correspond to 3M tapes manufactured in the United States with a high degree of similarity and some minor differences described below.

For example, clusters 8 (tapes 10, 24) and 10 (tapes 41, 68, 13, 18, 26) (**Figures 6-1** and **6-8**) consist of samples that originate from the same manufacturer, 3M, and generally consist of products from one of two premium brands, Super 88 and Super 33+ (**Appendix 6A, Table 6A-1**). As such, their elemental composition is expected to be somewhat similar. However, as seen in **Figure 6-12**, there are some areas where the automated classification denotes some peaks that are useful for differentiating samples between these clusters. These areas primarily correspond to the chlorine peak at around 2.3 KeV, the calcium/antimony peak at approximately 3.6 KeV, the lead peaks between 10-16 KeV, and the antimony peaks between 26-30 KeV, all highlighted in red and orange.

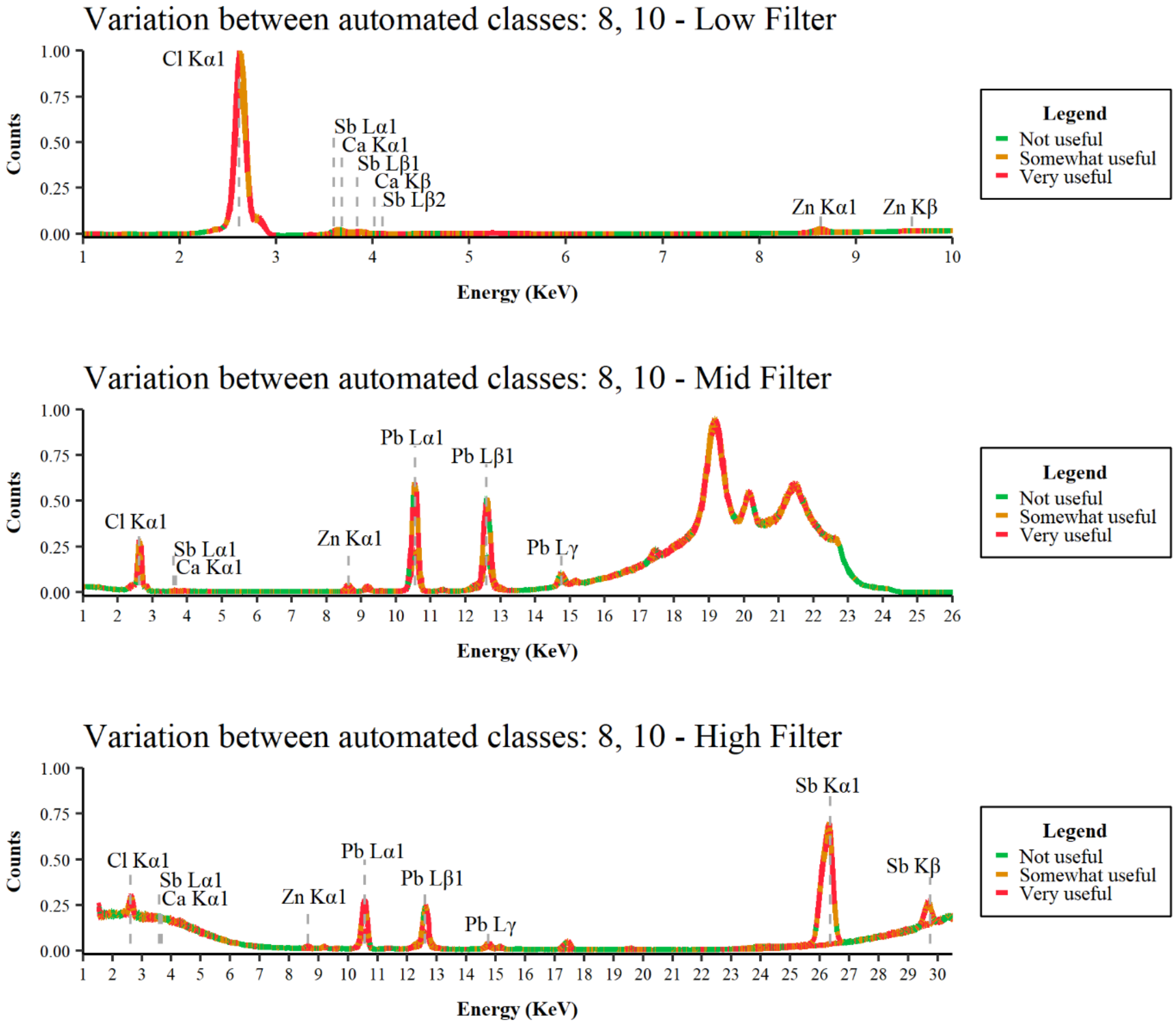


Figure 6-12. The results of the random forest classification of tapes from clusters 8 and 10 from the dendrogram (**Figure 6-8**) and the determination of the importance of each peak intensity. Areas in red are denoted as highly important, areas in orange are somewhat important, and areas in green are not particularly important for classification.

These trends are consistent with what is observed in a spectral overlay of tapes from those particular clusters. For example, tapes 10 (cluster 8) and 13 (cluster 10) have observable differences in the calcium/antimony content between 3.5 and 4.2 KeV (**Figure 6-13A**), as well as differences in lead content (**Figure 6-13B** and **6-13C**) and antimony (**Figure 6-13D**). Correspondingly, these tapes 10 and 13 belong to subgroups IX-A and IX-B when using spectral overlay.

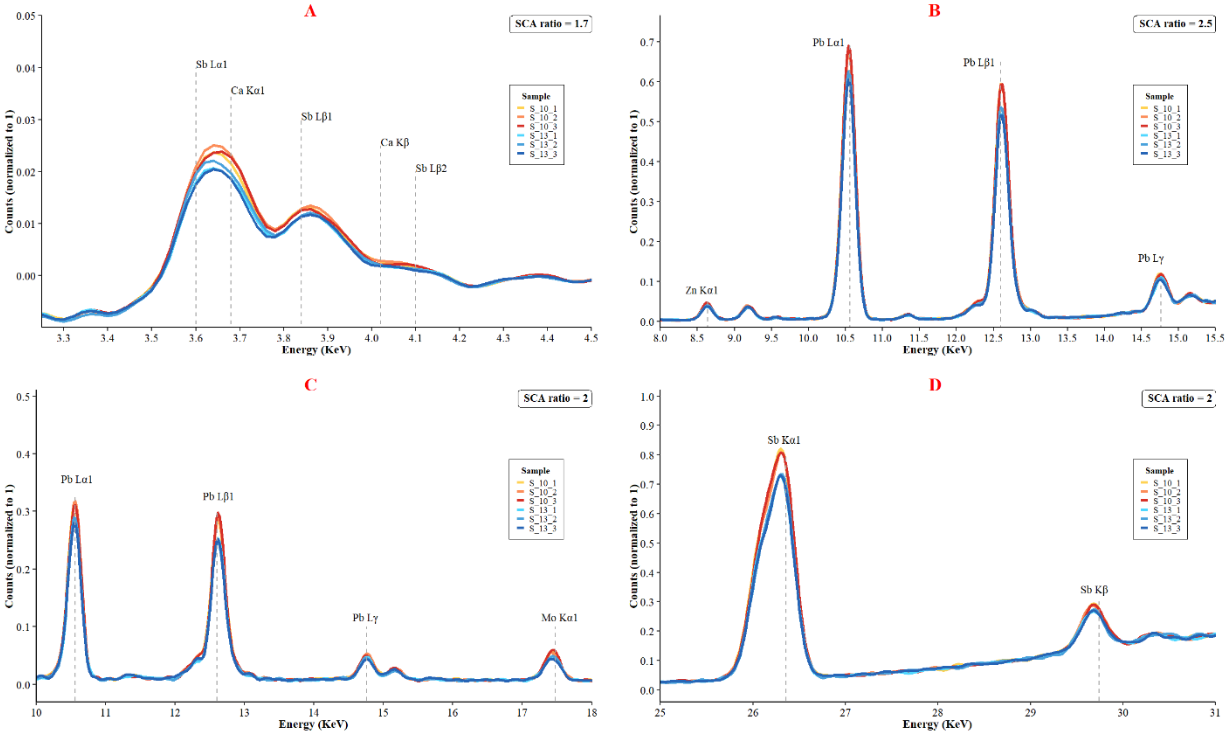


Figure 6-13. Spectral overlay of tapes 10 and 13, with all samples. There are subtle differences in the content of calcium and antimony in the low region (**Figure 6-13A**), with more distinguishable differences in the lead content in both the mid-filter region (**Figure 6-13B**) and the high-filter region (**Figure 6-13C**). In addition, tape 10 has a greater antimony concentration in the high filter **(Figure 6-13D**) region.

The random forest classifiers and dendrogram demonstrate additional discrimination of some tapes that are difficult to identify by spectral overlay, including tapes from the same manufacturer and labeled under the same brand. For example, cluster 10 (samples 41, 68, 13, 18, 26), cluster 11 (samples 11, 15), and cluster 12 (samples 12, 19, 20, 25, 39, 54, 64) (**Figure 6-8**) consist of 3M-made tapes from the brand Super 33+ (**Appendix 6A, Table 6A-1**). These tapes show considerable similarities and are all classified under sub-group IX-B by spectral overlay, except for tape 39 (subgroup IX-E) (**Figure 6-1**). When using the cluster analysis, it is notable that the distance between the clusters is relatively small, reflecting predominant similarities. When compared by random forest, the slight differences between the clusters become notable. Lead content in the mid-region is critical for differentiating between tapes in these clusters and some additional contributing elements (zinc, chlorine, and antimony).

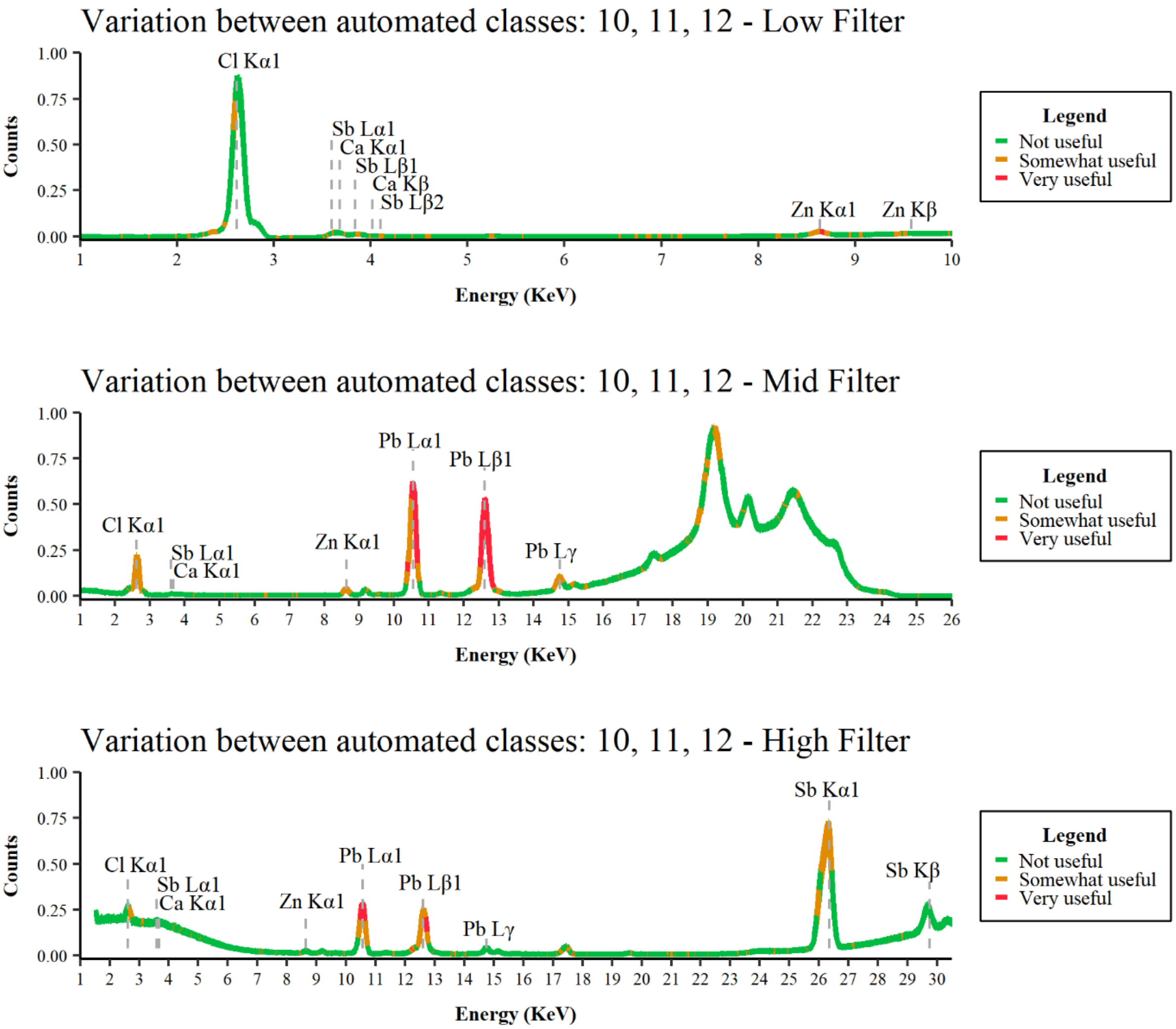


Figure 6-14. The results of the random forest classification of tapes from classes 10, 11, and 12 from the dendrogram (**Figure 6-8**) and the determination of the importance of each peak intensity. Areas in red are denoted as highly important, areas in orange are somewhat important, and areas in green are not particularly important for classification. For these classes, lead is the most crucial element for classification, followed by zinc (low region) and antimony (high region).

An example of the spectral comparison of tapes in these clusters is shown in **Figure 6-15**. Sample 13 (cluster 10), sample 15 (cluster 11), and sample 20 (cluster 12) show minor differences in zinc, lead, and antimony. However, these differences are minimal when compared by spectral overlay, so an examiner with a conservative approach to classification could be expected to report these tapes as indistinguishable. The SCAR values resulting from comparing these tapes are generally low (all but one ratio remained under the threshold of 1.95), supporting the classification of these tapes as indistinguishable and the proximity of these clusters.

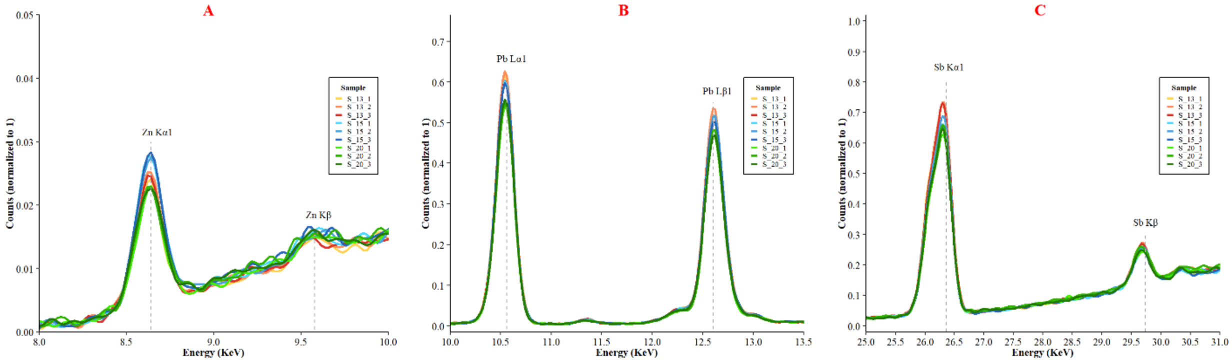


Figure 6-15. Spectral overlay spectra of tapes 13, 15, and 20, with all samples. There are differences in zinc content in the low region (**Figure 6-15A**) with tape 15 having the highest concentration. Tape 20 has the lowest lead content in the mid-region (**Figure 6-15B**), while tape 13 has slightly higher antimony content in the high region (**Figure 6-15C**).

Overall, the combined approach of SCA ratio calculations and the random forest model classifications provides an effective means of comparing electrical tape backing samples on an individual basis as well as to the larger population of tapes to provide a quantitative metric of similarity or difference, as well as the relevant characteristics of those differences to support the examiner's decisions.

6.5. Conclusions

Human intervention is essential in data interpretation in the forensic comparison of XRF spectra. The training and experience of the practitioner can provide valuable input regarding the chemical composition of the items evaluated, along with the case context and the samples' physical characteristics. Nonetheless, the forensic community recognizes the need for statistical support in the comparison process of spectral data, mainly because visual comparisons such as spectral overlay can introduce bias by unintentionally focusing on similarities, not detecting small difference thresholds, or having different perceptions or attention to detail among practitioners. Also, when conclusions rely considerably on human-based abilities for pattern recognition, the transparency in demonstrating the decision-making process can be challenged.

This study demonstrates that spectral contrast angle ratios provide a means to evaluate data collected from XRF through a metric that accounts for the similarities and differences between and within samples, making them an effective quantitative tool to express the significance of an association or exclusion. When combined with spectral overlay, spectral contrast angle ratios can provide a quantitative criterion to help examiners decide on spectra, even for very similar ones. Application of the spectral contrast angle ratios can be relatively straightforward, with the calculation of the ratios made practical with a standard Excel macro or open-source application (R code).

For the tapes evaluated in this study, a clear distinction of SCAR values is observed for items originating from the same roll compared to items originating from different sources. Moreover, samples with low SCA ratios tend to belong to subclasses of similar brands or quality grades of tape rolls. Thus, SCAR can offer a more objective and reproducible approach to assist forensic examiners in assessing the value of their evidence beyond a binary association/exclusion conclusion. In addition, the random forest classification model allows for the automated and rapid identification of peaks of interest for differentiation between spectra, which can generally only be determined through a manual spectral overlay. These tools become increasingly helpful when the number of measurements or compared items increases. The introduction of random forest algorithms adds value to the SCA ratio calculation by comparing a large population of samples (> 4,000 pairwise comparisons).

Overall, SCAR has shown to be a valuable input for classifying samples, providing insights as to potential lead information such as manufacturers or tape classes when unknown samples are compared to a larger population. The approaches reported here allow for a visual representation of the data that can be easily explained to the trier of fact. This approach can be expanded to various analytical chemistry disciplines, spectrochemical data, and other forensic materials and has the potential to provide additional probabilistic interpretation to support decisions made on analytical spectra.

6.6. References

1. Lavine BK, Almirall J, Muehlethaler C, Neumann C, Workman J. Criteria for Comparing Infrared Spectra – A Review of the Forensic and Analytical Chemistry Literature. *Forensic Chem.* 2020. 18, 100224. https://doi.org/10.1016/j.forc.2020.100224
2. ASTM International. E2809-13 Standard Guide for Using Scanning Electron Microscopy/X-Ray Spectrometry in Forensic Paint Examinations. West Conshohocken, PA; ASTM International, 2013. https://doi.org/10.1520/E2809-13
3. Mehltretter AH, Bradley MJ, Wright DM. Analysis and Discrimination of Electrical Tapes: Part I Adhesives. *J. Forensic Sci*. 2011. 56 (1), 82–94. https://doi.org/10.1111/j.1556-4029.2010.01560.x.
4. Mehltretter AH, Bradley MJ, Wright DM. Analysis and discrimination of electrical tapes: Part II Backings. *J. Forensic Sci*. 2011. 56 (6), 1493–1504. https://doi.org/10.1111/j.1556-4029.2011.01873.x.
5. Prusinowski M, Mehltretter AH, Martinez-Lopez C, Almirall J, Trejos T. Assessment of the utility of X-ray Fluorescence for the chemical characterization and comparison of black electrical tape backings. *Forensic Chem*. 2019. 13, 100146. https://doi.org/10.1016/j.forc.2019.100146.
6. Objective comparisons of highly-dimensional spectral data. OSCA Research Needs Assessment Form; Trace Materials Subcommittee. Feb 2021.

https://www.nist.gov/system/files/documents/2021/06/23/Trace%20Materials%20SC_R%26D%20Need_Objective_Comparisons_of_Highly-Dimensional_Spectral_Data.pdf

7. Martinez-Lopez C, Trejos T, Mehltretter AH, Almirall J. Elemental analysis and characterization of electrical tape backings by LA-ICP-MS. *Forensic Chem*. 2017. 4, 96–107. https://doi.org/10.1016/j.forc.2017.03.003
8. Brooks E, Mehltretter AH, Prusinowski M, Trejos T. Optimization and Validation for Qualitative Elemental Analysis of Electrical Tape Backings by X-ray Fluorescence. *Forensic Chemistry*. 2020. 21, 100291. https://doi.org/10.1016/j.forc.2020.100291
9. Sun Z, Quan Y, Sun Y. Elemental analysis of white electrical tapes by wavelength dispersive X-ray fluorescence spectrometry. *Forensic Sci International*. 2013. 232 (1), 169-172. http://dx.doi.org/10.1016/j.forsciint.2013.07.009
10. Mehltretter AH, Wright DM, Dettman JR, and Smith MA. Intra-roll and intra-jumbo roll variation of duct tapes. *JASTEE*. 2015. 6 (1), 21-41.
11. ASTM Standard E2926-17, Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry. ASTM International, West Conshohocken, PA. 2017. https://doi.org/10.1520/E2926-17.
12. ASTM Standard E3260-21, Standard Guide for Forensic Examination and Comparison of Pressure Sensitive Tapes. ASTM International, West Conshohocken, PA. 2021. https://doi.org/10.1520/E3260-21.
13. Lavine BK, White CG**,** Allen MD, Weakley A. Pattern recognition assisted infrared library searching of the paint data query database to improve investigative lead information from automotive paint trace evidence. *Appl. Spec*. 2017. 71 (3), 480-495. DOI: 10.1177/0003702816666287
14. Lavine BK, White CG**,** Allen MD, Fasasi A, Weakley A. Evidential significance of automotive paint trace evidence using a pattern recognition based infrared library search engine for the paint data query forensic database. *Talanta*. 2016. 159, 317-329. https://doi.org/10.1016/j.talanta.2016.06.035
15. Lavine BK, White CG**,** Allen MD. Forensic analysis of automotive paints using a pattern recognition assisted infrared library searching system: Ford (2000-2006). *Microchem. J*. 2016. 129, 173-183. https://doi.org/10.1016/j.microc.2016.06.024
16. Fasasi A, Mirjankar N, Stoian RI, White C, Allen M, Sandercock MP, Lavine BK. Pattern recognition assisted infrared library searching of automotive clear coats. *Appl. Spec*. 2015. 69, 84-94. DOI: 10.1366/14-07578
17. Lavine BK, Fasasi A, Mirjankar N, White C, Sandercock M. Search prefilters to assist in library searching of infrared spectra of automotive clear coats. *Talanta*. 2015. 132, 182-190. http://dx.doi.org/10.1016/j.talanta.2014.08.061
18. Lavine BK, Fasasi A, Mirjankar N, Sandercock M, Brown SD. Search prefilters for mid-IR spectra of clear coat automotive paint smears using stacked and linear classifiers. *J. Chemometrics*. 2014. 28, 385-394. DOI: 10.1002/cem.2598

19. Lavine BK, Fasasi A, Mirjankar N, White C. Search prefilters for library matching of infrared spectra in the pdq database using the autocorrelation transformation. *Microchem. J***.** 2014. 113, 30–35. https://doi.org/10.1016/j.microc.2013.11.001

20. Lavine BK, Fasasi A, Mirjankar N, Sandercock M. Development of search prefilters for infrared library searching of clear coat paint smears. *Talanta*. 2014. 119, 331–340. http://dx.doi.org/10.1016/j.talanta.2013.10.066

21. Lavine BK, Mirjankar N, Ryland S, Sandercock M. Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics. *Talanta*. 2011. 87, 46-52. https://doi.org/10.1016/j.talanta.2011.09.039

22. Lambert D, Muehlethaler C, Esseiva P, Massonnet G. Combining spectroscopic data in the forensic analysis of paint: application of a multiblock technique as a chemometric tool. *Forensic Sci. International*. 2016. 263, 39-47. DOI: 10.1016/j.forsciint.2016.03.049

23. Maric M, van Bronswick W, Lewis SW, Pitts K. Synchrotron FTIR characterization of automotive primer surfacer paint coatings for forensic purposes. *Talanta*. 2013. 118, 156-161. https://doi.org/10.1016/j.talanta.2013.10.016

24. Muehlethaler C, Massonnet G, Esseiva P. Discrimination and classification of FTIR spectra of red, blue, and green spray paints using a multivariate statistical approach. *Forensic Sci. International*. 2014. 244, 170-178. https://doi.org/10.1016/j.forsciint.2014.08.038

25. Martyna A, Michalska A, Zadora G. Interpretation of FTIR spectra of polymers and Raman spectra of car paints by means ofutilizing likelihood ratio approach supported by wavelet transform for reducing data dimensionality. *Anal. Bioanal. Chem.* 2015. 407, 3357-3376. https://doi.org/10.1007/s00216-015-8558-9

26. Varmuza K, Penchev PN, Scsibrany H. Maximum common substructures of organic compounds exhibiting similar infrared spectra. *J. Chem. Inform. Comput Sci*. 1998. 38, 420-427. https://doi.org/10.1021/ci9700889

27. Boruta M. FT-IR search algorithm – assessing the quality of a match. *Spec*. 2012. 27 (8), 1-6. https://www.spectroscopyonline.com/view/ft-ir-search-algorithm-assessing-quality-match

28. Szafarska M, Wozniakiewicz M, Pilch M, Zieba–Palus J, Koscielniak P. Computer analysis of ATR-FTIR spectra of paint samples for forensic purposes. *J. Molecular Structure*. 2009. 924-926, 504-513. https://doi.org/10.1016/j.molstruc.2008.11.048

29. Loudermilk JB, Himmelsbach DS, Barton III FE, de Haseth JA. Novel search algorithms for mid-infrared spectral library of cotton contaminants. *Appl. Spec*. 2008. 62, 661-670. https://doi.org/10.1366/000370208784657968

30. Bodis L, Ross A, Pretsch E. A novel spectra similarity measure. *Chemom. Intell. Lab. Syst*. 2007. 85, 1–8. https://doi.org/10.1016/j.chemolab.2005.10.

31. Li J, Hibbert DB. Comparison of spectra using a Bayesian approach. An argument using oil spills as an example. *Anal. Chem*. 2005. 77, 639–644. https://doi.org/10.1021/ac048894j

32. Esseiva P, Gaste L, Alvarez D, Anglada F. Illicit drug profiling, reflection on statistical comparisons. *Forensic Sci International*. 2011. 207, 27-34. https://doi.org/10.1016/j.forsciint.2010.08.015
33. Li T, Hua Z, Meng X, Liu C. A Simple and Effective Physical Characteristic Profiling Method for Methamphetamine Tablet Seized in China. *J. Forensic Sci*. 2018. 63 (2), 541-547. https://doi.org/10.1111/1556-4029.13554
34. Neumann C, Margot P. New perspectives in the use of ink evidence in forensic science Part II. Development and testing of mathematical algorithms for the automatic comparison of ink samples analysed by HPTLC. Forensic Sci. International. 2009. 185, 38–50.
35. Neumann C, Ramotowski R, Genessay T. Forensic examination of ink by high-performance thin layer chromatography—The United States Secret Service Digital Ink Library. Journal of Chromatography A. 2011. 1218(19), 2793-2811.
36. Bonetti J, Quarino L. Comparative Forensic Soil Analysis of New Jersey State Parks Using a Combination of Simple Techniques with Multivariate Statistics. *J. Forensic Sci*. 2014. 59 (3), 627-636. https://doi.org/10.1111/1556-4029.12375
37. Dirwono W, Park JS, Augustin-Camacho MR, Kim J, Park HM. Application of micro-attenuated total reflectance FTIR spectroscopy in the forensic study of questioned documents involving red seal inks. *Forensic Sci. International*. 2010. 199, 6-8. https://doi.org/10.1016/j.forsciint.2010.02.009
38. Muehlethaler C, Massonnet G, Hicks T. Evaluation of infrared spectra analyses using a likelihood ratio: a practical example of spray paint examination. *Sci. Just.* 2016. 56, 61-72. https://doi.org/10.1016/j.scijus.2015.12.001
39. Krooshof P, Tran T, Postma G, Melssen W, Buydens L. Effects of including spatial information in clustering multivariate image data. *Trends in Analytical Chemistry*. 2006. 25 (11), 1067-1080.
40. Schober P, Boer C, Schwarte L. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*. 2018. 126 (5), 1763-1768.
41. Wan K, Vidavsky I, Gross M. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J Am Soc Mass Spectrom*. 2002. 13, 85-88. https://doi.org/10.1016/S1044-0305(01)00327-0
42. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectrom*. 1994. 5 (9), 859-866.
43. Swartz ME, Brown PR. Use of mathematically enhanced spectral analysis and spectral contrast techniques for the liquid chromatographic and capillary electrophoretic detection and identification of pharmaceutical compounds. *Chirality*. 1996. 8 (1), 67-76.
44. Zhvansky ES, Pekov SI, Sorokin AA, Shurkhay VA, Eliferov VA, Potapov AA, Nikolaev EN, Popov IA. Metrics for evaluating the stability and reproducibility of mass spectra. 2019. *Sci Rep*. 9, 914. https://doi.org/10.1038/s41598-018-37560-0
45. Breiman L, Random forests. Mach. Learn. 2001. 45, 5–32. https://doi.org/10.1023/A:1010933404324

46. Abreu Fontes J, Anzanello MJ, Brito J, Bucco GB, Fogliatto FS, Puglia F. Combining wavelength importance ranking to the random forest classifier to analyze multiclass spectral data. *Forensic Sci International*. 2021. 328, 110998. https://doi.org/10.1016/j.forsciint.2021.110998
47. Li L, Yu , Bai S, Cheng J, Chen X. Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO. *Journal of Sensors*. 2018. 2018, 1-9. https://doi.org/10.1155/2018/1578314

Appendix 6-A

Table 6A-1. Brand information for all samples used in this study [8].

Sample	Brand	Product	Country
1	Marcy Enterprises, Inc.	MA 750	Taiwan
2	Advance®	AT7, BS3924, 31/90Tp	England
3	Work Saver™ (Royal Tools)	Stock no. 55, 5 color PVC Tape Assortment	China
4	tesa tape, Inc.	40201, No. 111 E52811A	Taiwan
5	Tape It, Inc.	E-60	Taiwan
6	Qualpack®	1346, 6-Color	China
7	Marcy Enterprises, Inc.	MA 750	Taiwan
8	Manco®	200 MPH, AE-66	Taiwan
9	Archer® (Radio Shack)	64-2349	Taiwan
10	3M Scotch™	Super 88, 054007-06143	USA
11	3M Scotch™	Super 33+, 10414 NA	USA
12	3M Scotch™	Super 33+, 10455 NA	USA
13	3M Scotch™	Super 33+	USA
14	Frost King®	ET-60	Taiwan
15	3M Scotch™	Super 33+, 10455 NA	USA
16	3M	Tartan™ 1710, part no. 054007 49656	USA
17	3M Scotch™	Super 88, 054007-06143	USA
18	3M Scotch™	Super 33+, Cat. 195NA	USA
19	3M Scotch™	Super 33+, Cat. 195NA	USA
20	3M Scotch™	Super 33+, 10414 NA	USA
21	Manco®	P-66	Taiwan
22	Manco®	667 Pro Series™	Taiwan
23	3M Scotch™	Super 88, 054007-06143	USA
24	3M Scotch™	Super 88, 054007-06143	USA
25	3M Scotch™	Super 33+ 054007-06143	USA
26	3M Scotch™	Super 33+ 054007-06143	USA
27	3M	Tartan™ 1710, part no. 054007 49656	USA
28	3M	Tartan™ 1710, part no. 054007 49656	USA
29	3M	Temflex™, 1700, 54007-69764	USA
30	3M	Temflex™, 1700, 54007-69764	USA
31	Regal®	Model ET-6	Taiwan
32	GE	GE2472-3DD	Taiwan
33	3M Scotch™	Cat. 190	USA
34	3M	Tartan™ 1710, part no. 054007 49656	USA
35	Frost King®	ET60	Taiwan
36	3M	Tartan™ 1710, part no. 49656	USA
37	National	All-Purpose Grade	Taiwan

38	Manco®	P-660	Taiwan
39	3M Scotch™	Super 33+, 3744NA	USA
40	3M	Tartan™ 1710, part no. 054007 49656	USA
41	3M Scotch™	Super 33+, 200NA	USA
42	National	All-Purpose Grade	Taiwan
43	3M	Tartan™ 1710, part no. 054007 49656	USA
44	3M	Tartan™ 1710, part no. 054007 49656	USA
45	Calterm®	49605	Taiwan
46	Manco®	P-20	Taiwan
47	3M	Tartan™ 1710, part no. 054007 49656	USA
48	Tape It, Inc.	36-T	USA
49	Tape It, Inc.	36-T	USA
50	GE	GE2472-31D	Taiwan
51	National	No. 101, E52811A	Taiwan
52	Frost King®	ET60FR	USA
53	National	No. 101, E52811A	Taiwan
54	3M Scotch™	Super 33+, 03404NA	USA
55	Manco®	1219-60	Taiwan
56	Victor Automotive Products (Thermoflex)	33-UL60, No. 101, E52811A	Taiwan
57	United Tape Company	UT-602	Taiwan
58	Frost King®	ET60	Taiwan
59	Tuff™ Hand Tools		China
60	Tuff™ Hand Tools		China
61	3M Scotch™	88T	USA
62	Nitto Denko	No. 228	Taiwan
63	3M Scotch™	Super 88, 054007-06143	USA
64	3M Scotch™	Super 33+, 10455 NA	USA
65	3M Scotch™	700 Commercial Grade, 054007-04218	USA
66	L.G. Sourcing, Inc.	19453	Taiwan
67	Manco®	P-66	Taiwan
68	3M Scotch™	Super 33+	USA
69	3M	Tartan™ 1710, part no. 054007 49656	Taiwan
70	Tyco Adhesives (National)	No. 101, E52811A	Taiwan
71	Qualpack®	1346, 6-Color	China
72	Nitto Denko	Nitto® No. 228	Taiwan
73	Frost King®	ET60FR	China
74	3M Scotch®	700 Commercial Grade, 054007-04218	USA
75	3M Scotch™	Linerless Electrical Rubber Splicing Tape, 2242, 06165	USA
76	3M Scotch®	Super 33+, Cold Weather Electrical Tape, 16736NA	USA
77	3M Scotch®	Super 33+, 054007-06132	USA

78	3M	Tartan™ 1710 General Use, 054007-49656	Taiwan
79	3M Scotch®	700 Commerical Grade, 054007-04218	USA
80	3M Scotch®	Super 88, 054007-06143	USA
81	Ace (Henkel)	All Weather	Taiwan
82	Ace (Henkel)	Weather Resistant	Taiwan
83	3M Scotch®	Super 33+, 10414 NA	USA
84	3M	Tartan™ 1710 General Use, 054007-49656	Taiwan
85	Frost King®	ET60FR	China
86	Duck (Henkel)	Vinyl Electrical Tape	Taiwan
87	Nitto Denko	No. 21E	China
88	Frost King®	ET60FR	China
89	Power Pro Craft	ETF	China
90	Duck (Henkel)	Extra wide electrical tape	China
91	3M Scotch®	Super 33+	USA
92	3M Scotch®	Super 88, 054007-06143	USA
93	Commercial Electric (Home Depot)	EE-100	China
94	3M	3M Economy 1400	Taiwan

Chapter 7: Conclusions and Future Work

7.1. Summary of the Study

This research addressed the overall goal of improving and enhancing how forensic trace evidence is analyzed and compared. This was accomplished through two primary goals: developing and validating systematic comparison methods for performing physical fit examinations and applying a quantitative metric for spectral data to enhance the comparison of visually similar spectra. We anticipate that the methodologies, readily available comparison templates, and developed statistical measures will aid and support forensic practitioners in comparing forensic trace evidence and improve transparency in how decisions are made. The main findings and achievements of the research studies are discussed in the following sections.

7.2. Outcomes of Study Objectives

7.2.1. Objective 1. Expand the validation of a human-based systematic comparison method for duct tape physical fits with a more extensive dataset of duct tapes and interlaboratory studies.

This objective aimed to expand and enhance the previously developed systematic comparison method for physical fits of duct tapes. To do this, we expanded the current dataset of duct tape comparison pairs and explored the performance of the edge similarity score (ESS) method through two interlaboratory exercises. Expanding the duct tape dataset to incorporate more than 3000 comparisons provided sufficient data to explore the relationship between the various experimental factors controlled during sample preparation – quality (grade) of tape, separation method, and stretching, and their influence on reported ESS values. The interlaboratory exercises allowed for refinement of the training and communication of the method and enhancements made to the documentation templates through feedback provided by current practitioners.

The final tape dataset consisted of more than 3000 comparisons, divided into subsets of samples based on tape grade, separation method, and stretching. The ESS method demonstrated good overall performance, generally over 90% accuracy, and low false negative rates. None of the subsets reported any false positive results, indicating that it is not common for false positive results to be reported when the whole edge is present for comparison. For the subsets that demonstrated lower accuracy (around 85%), the leading cause was the tape's grade and the resulting features and characteristics of the edges post-separation. The higher-quality grade contributed to higher uncertainty in the examination due to generally straighter edges and the possibility of shifted fibers, leading to more misclassification of the samples as non-fits. Statistically, the data demonstrated a relationship between grade and separation method and showed that both factors impact the ESS values reported for true-fitting pairs. Stretching contributed minor variability in reported ESS but did not significantly affect the overall conclusion. This study established criteria for ESS thresholds: pairs with ESS greater than 80 strongly support a fit decision, while pairs with scores less than 20 strongly support a non-fit conclusion. Critically, regardless of grade or separation method, all the tapes used in this study demonstrated features valuable for comparison that make physical fit examinations feasible.

The two interlaboratory studies provided results from thirty-eight different participants for a total of 266 sample comparisons. Each exercise demonstrated high overall performance, between 90-100%. Some discrepancies were noted in the first study regarding the participant's reported ESS values and overall decisions. These discrepancies were resolved in the second study as a product of improvements to training and documentation of the examination process. The participants generally agreed on the ESS method's potential for facilitating systematic and transparent peer review and training. The improved documentation templates defined comparison features and criteria, and methods of communication that were demonstrated to be practical tools in the standardization of forensic examination of physical fits.

Overall, the outcome of these studies demonstrates that not every physical fit conclusion holds the same confidence. In addition, while there may be consistency in overall decisions between analysts and generally low error rates, the specific features observed on a given pair of samples and interpretation thereof may vary between individual analysts. As such, standardizing observed features, their value in comparison, and how similarity and differences between sample edges should be reported is a critical objective for forensic practitioners. The value of the proposed ESS method is that it provides a means of accompanying a reported physical fit with an assessment of its probative value, enhancing a physical fit comparison from a single binary decision to a more nuanced expression of the edge alignment and the significance of that comparison in context to a larger population of samples.

7.2.2. Objective 2. Compare the performance of a computational algorithm for comparisons of duct tape physical fits against the human-based results.

This objective aimed to explore the performance of a developed computational model for physical fit comparisons of duct tapes. While other algorithms have been developed to assess the edges of duct tapes and report on large populations of samples, this approach is novel in its use of both sides of the samples, which our research has demonstrated to provide critical features for comparison. Trained on a population of over 4000 images, the proposed computational model performs well with comparable high accuracies (71-88%) and low error rates on samples previously analyzed by human analysts.

Indeed, layer-wise relevance propagation can demonstrate which pixels and tape characteristics the computational model relies on to decide for a given pair. Furthermore, the model recognized several features the human analysts reported for the same sample pairs, reinforcing the value of specific edge characteristics for comparison. In addition, the model provided objective, statistical support for physical fit examinations by reporting membership probabilities.

However, it is worth noting that the model does produce false positive results not seen in human analysis and has room for improvement. As the initial dataset used is relatively small, adding more datasets in the future can assist in the learning process and performance of the model. Nevertheless, the model presented here demonstrated the value of computational models for the forensic community to use to support decision-making in physical fit examinations. In time, we anticipate the development of additional databases and models for duct tapes and other materials of forensic interest.

7.2.3. Objective 3. Extract and interpret edge feature data of duct tape and textile physical fit examinations using computational models

This objective addressed several enduring questions in physical fit analysis: the minimum sample length necessary to perform reliable physical fit examinations and identify the most critical comparison features to substantiate an examiner’s opinion. The quantitative data provided by the ESS approach of more than 1000 tapes and 600 textiles were extracted to accomplish this objective.

The partial width of the sample edges demonstrated that lengths of at least 35% of the tape or textile edges would provide acceptable accuracy in most samples. However, the relative size is highly dependent on the specimen, and samples more prone to distortion or lacking distinct features may not be suitable for comparison with only a small portion of the edge available. The mutual information analysis of the sample templates indicates several trends in feature value for the different subsets. Duct tape quality and separation method considerably impacted the features observed and their importance in comparison. For textiles, different features held more value based on the separation method of the samples.

Finally, the decision tree model trained on the sample sets demonstrated consistent accuracy with the human analysis in most datasets. However, it produced higher false positive rates than observed in human analysis. As a result, the outcome of this method indicated it has the potential as a valuable supplementary tool for analysts in making decisions for specific pairs of tape or textile samples. Samples independently evaluated by the decision tree to have a different conclusion from the human analyst can be assessed further through another independent analyst's verification or chemical analysis.

7.2.4. Objective 4. Develop and validate a systematic method to quantify the quality of fracture fits of hard plastics.

This objective included the development of a method for the physical fit comparison of brittle automotive polymers. Specifically designed to address the distinctive features and properties of the rigid polymer material, the new method provides not only the quantitative edge similarity score (ESS) but also a feature prominence value (FPV), which describes the influence of a feature on a fit or non-fit decision.

The three-dimensional nature of the polymers provided specific challenges for comparisons. For example, identifying the correct top and bottom surfaces of the compared item and the correct orientation of the fragments contributed to some of the discrepancies in performance and reported similarity metrics. In addition, the preliminary analysis indicated that polymer composition and mechanical properties might influence the fit's quality, such as thinner samples that deformed more during separation.

The proposed ESS method applied to a population of 385 comparisons resulted in 85% accuracy, with no false positives reported. Most false negative misidentifications resulted from features influenced by distortions left during the breaking process. The analysis of decisions made at the macroscopic and microscopic levels demonstrated that performing the microscopic examination of the samples is valuable in most cases unless the samples have clear and distinct differences at the macroscopic level. In addition, the feature prominence sum (FPS) criteria demonstrated another potential metric that supports an analyst's decisions. They provided information on the most valuable features for a detailed comparison. Inter-analyst performance on a subset of the polymer sets demonstrated consistency between two analysts regarding accuracies (82% and 88%) and reported ESS and FPS distributions.

7.2.5. Objective 5. Develop and validate a systematic method to compare XRF data of electrical tapes.

This objective aimed to demonstrate a novel approach for quantitative and objective comparison of complex X-ray Fluorescence (XRF) spectral data. A set of electrical tapes, with 94 different-source samples and 20 same-source samples, were characterized, and the data was processed to normalize and reduce noise. The pairwise comparison of all different source samples (n=4371) was completed using spectral overlay to classify the samples. While the spectral overlay is valuable for the rapid comparison and identification of visually distinct compositions, there is a need for statistical support in the comparison process of spectral data due to risks of bias, a lack of consensus-based thresholds for detecting and interpreting minor differences, and a lack of transparency in demonstrating the decision-making process.

Spectral contrast angle ratios (SCAR) were calculated for each pairwise comparison of different-source and same-source samples (n=4,561). The SCAR ratios were evaluated to determine trends in the sample analysis, and a random forest model was designed to identify critical peak differences in the XRF spectra between groups. This study demonstrated that SCAR served as an effective tool to evaluate and classify samples by accounting for similarities and differences between and within samples; this allowed them to be used to express the significance of an association or exclusion. Furthermore, applying the spectral contrast angle ratios was straightforward, with the calculation of the ratios made practical with Excel macros and open-source applications. As such, when used with spectral overlay, spectral contrast angle ratios served as a valuable quantitative criterion to help analysts decide on chemically similar samples.

In addition, the random forest classification model provided automated and rapid identification of peaks of interest for differentiation between spectra. These differences would generally only be determined through a manual spectral overlay, which could be time-consuming when performed on large populations in casework. Cluster analysis of the sample populations using SCAR also helped identify consistent trends in manufacturing and source information, which could prove critical in an investigation. Furthermore, SCAR holds potential value as a means to estimate a probabilistic assessment of the characterization and comparison of specific samples.

7.3. Future Work

The research discussed here forms a promising foundation for future research into comparison criteria and methods for forensic trace evidence. The systematic methods developed for physical fits of duct tape and automotive polymers demonstrate the value of defined comparison criteria and quantitative similarity metrics. The large population of duct tape samples allowed for statistical testing of the experimental factors. It revealed a relationship between grade and separation method on ESS value. The one for duct tapes inspired the method developed for automotive polymers. By analyzing additional samples, we plan to evaluate the fit quality for that material statistically. As with such a database, additional samples originating from different sources will provide more support and insight into the value of the features and experimental factors we are exploring. The next phase of this research will incorporate additional materials and larger populations of samples. A critical next stage for physical fit analysis of all the materials discussed in this research is the implementation of the probabilistic interpretation. This study demonstrates the potential use of score-based likelihood ratios for duct tape physical fits to provide a proxy for the probative value of the evidence. This approach can be tested on other materials, and additional data will enhance the recommended ESS thresholds and their interpretation.

Furthermore, with the feedback provided by the participants of the interlaboratory studies, we improved upon the method training and documentation templates used for the duct tapes. These changes are anticipated to help with the future implementation of such systematic comparison methods in laboratories. This method and the corresponding documentation template hold substantial promise for training and verification processes in forensic laboratories. Several laboratories, including some of those that participated in the interlaboratory exercises, have expressed interest in additional collaboration on our method for tapes and other materials. Further work in this area will continue to address the concerns and feedback provided by practitioners in physical fit examinations, aiming to improve and expand the application, communication, and interpretation of the ESS method on duct tapes, textiles, automotive polymers, and other materials.

The development of computational models, such as the convolutional neural networks and decision trees developed for the analysis of physical fits in this research, have demonstrated their potential to assist analysts in informing and supporting their decisions. Furthermore, the results of both models were comparable to human analysis for the same samples. However, additional data and training for both approaches will enhance their performance. In addition, the application of these computational models can serve as a foundation for developing additional models for application to duct tape edges and other forensic-relevant materials to provide objective support for physical fit examinations.

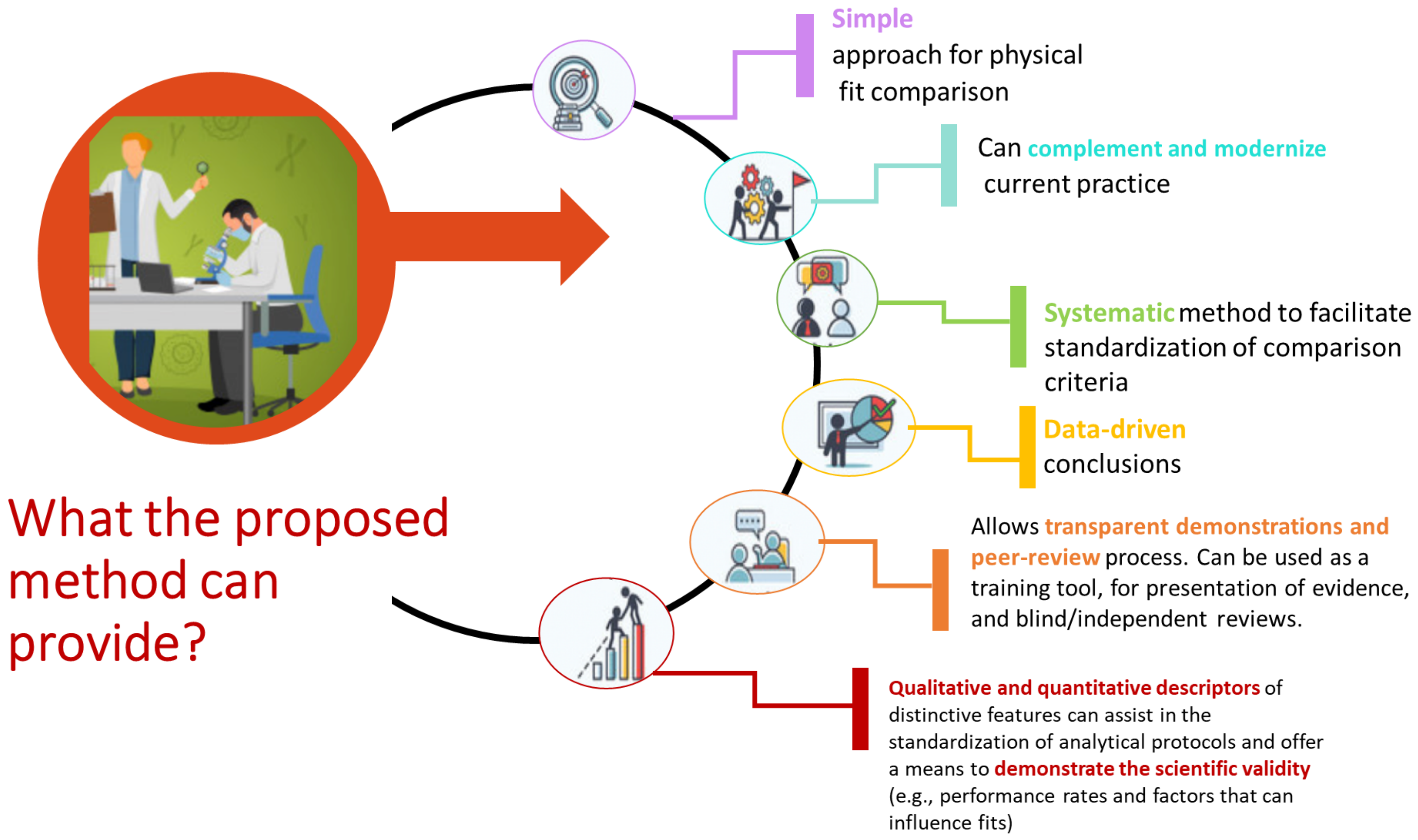
Finally, developing a quantitative approach to comparing X-ray fluorescence data for electrical tapes has demonstrated the value of an objective and computational comparison metric for semi-quantitative data. Forensic analysts have relied on spectral overlay for comparisons, which may contribute to discrepancies between analysts in an overall decision in cases of complex or very similar data. Determining a potential threshold for the spectral contrast angle ratio provides a quantitative value that analysts can use to support their decisions on pairs will similar compositions. However, this approach can be advanced by analyzing new samples on modern instruments to determine if updated manufacturing and compositions for electrical tapes and new and more sensitive detectors on XRF instruments will impact the recommended threshold values for SCARs.

7.4. Impact on the Criminal Justice System

Over the course of this project, results have been disseminated and presented to the forensic and scientific community and received very positive feedback. The main highlights of the value of the physical fit approach are described in **Figure 7-1**. The ESS methods developed in this research offer a simple approach for fit comparison that can complement and modernize fit examinations. One of the main benefits of the methods is that they are systematic and facilitate the standardization of comparison criteria using data-driven conclusions. An added value is the capability to allow transparent demonstration of the thought process in peer review and training. The documentation of qualitative and quantitative descriptors helps not only in the standardization of protocols but also in the demonstration of scientific validity that can serve as a foundation for supporting the value of physical fits and a better understanding of the factors that can influence accuracy. We anticipate further interest in this subject by publishing new methods and models to support physical fit examinations of additional materials.

Furthermore, thanks to the feedback gathered during the interlaboratory exercises, the input of the current practitioners is incorporated into method development. More than twenty laboratories participated in the duct tape interlaboratory studies, and dissemination of the results of that study has attracted the attention of laboratories across the United States and internationally. Interpreting physical fits has become much more collaborative, involving practitioners from various organizations, statisticians, psychologists, attorneys, legal scholars, and others. This collaboration has led to the development of a standard guide for physical fit examination and, hopefully, in the future, the adoption of more standardized procedures for physical fit examination and additional studies that can help us further establish the reliability of physical fit examinations in casework and testimony.

Another impact of the project is the results of the quantitative method for the analysis and interpretation of spectral data. One current research need for forensic trace evidence has been quantitative and objective approaches for interpreting complex spectrochemical data. This research addresses this gap in the field. The method demonstrated in this study would provide an open-source and straightforward process for comparing spectral data that would allow analysts to compare their samples and have the resulting quantitative metric fit within a relevant population to describe the respective weight of the comparison.



What the proposed method can provide?

Simple approach for physical fit comparison

Can **complement and modernize** current practice

Systematic method to facilitate standardization of comparison criteria

Data-driven conclusions

Allows **transparent demonstrations and peer-review** process. Can be used as a training tool, for presentation of evidence, and blind/independent reviews.

Qualitative and quantitative descriptors of distinctive features can assist in the standardization of analytical protocols and offer a means to **demonstrate the scientific validity** (e.g., performance rates and factors that can influence fits)

Figure 7-1. Diagram denoting the main benefits of the physical fit methodologies developed in this research.

Appendix: Curriculum Vitae

Meghan Prusinowski Morgantown WV 26505
mnp0006@mix.wvu.edu 585.208.8436

EDUCATION

Doctor of Philosophy – *Forensic Science* with Certificate in *University Teaching* Aug. 2019 – Present
Department of Forensic and Investigative Science
West Virginia University, Morgantown, WV
Expected Graduation Spring 2023

Master of Science - *Forensic and Investigative* Science Aug. 2019
Department of Forensic and Investigative Science
West Virginia University, Morgantown, WV
(27 credit hours in forensic classes, including forensic chemistry and trace evidence examination, 3 credit hours of statistical analysis, and 6 credit hours of research specializing in forensic analytical chemistry)

Bachelor of Arts - *Chemistry* with Minor in *Biology* May 2017
State University of New York at Geneseo, Geneseo, NY
(39 credit hours in chemistry, 6 credit hours of physics, equivalent credit for differential and integral calculus)

WORK EXPERIENCE

Graduate Research Assistant Jan. 2021 – Present
WVU Forensic and Investigative Science Department, Morgantown

Researching and collecting data for NIJ award- Assessing the Strength of Trace Evidence Fracture Fits through a Comprehensive, Systematic and Quantifiable Approach. Award # 2020-DQ-BX-0012, (Jan. 2021 – Dec. 2022) Tatiana Trejos (PI), Aldo Romero (co-PI)

- Method development of systematic and quantitative comparisons of duct tape and polymer materials
- Sample preparation and imaging of sample populations of duct tapes, textiles, paper, stamps, and automotive polymers
- Development and coordination of an interlaboratory study of duct tape physical fits, incorporating responses from more than 20 participants at state and federal laboratories across the US and internationally
- Coding in Python to assist in the development of an automated comparison algorithm to compare edges of separated materials

Graduate Teaching Assistant Aug. 2017 – May 2021
WVU Forensic and Investigative Science Department, Morgantown

Trace Evidence Analysis (FIS 414, 416, and 614) Jan. 2021 – May 2021

- Prepared laboratory experiments and supervised instrument operation of Laser-induced breakdown spectroscopy (LIBS), Fourier Transform Infrared Spectroscopy (FTIR), Glass Refractive Index Measurement (GRIM), Scanning electron microscopy-energy dispersive spectroscopy (SEM-EDS)
- Assisted and guided students during laboratory exercises, mock crime scenes, and mock courtroom testimony
- Graded laboratory reports and assignments and assisted with online administration of course material
- Assisted with the creation and editing of laboratory manual

Forensic Chemical Analysis (FIS 340 – Lecture and FIS 341 – Laboratory) Aug. 2020 – Dec. 2020

- Prepared laboratory experiments and supervised instrument operation of Laser-induced breakdown spectroscopy (LIBS), Fourier Transform Infrared Spectroscopy (FTIR), Glass Refractive Index Measurement (GRIM), Scanning electron microscopy-energy dispersive spectroscopy (SEM-EDS), and Ultraviolet–visible spectroscopy (UV-Vis)
- Assisted and guided students during laboratory exercises and mock courtroom testimony
- Graded laboratory reports and assignments
- Developed, filmed, and edited demonstration videos to complement laboratory experience and assisted with online administration of course material.
- Assisted with the creation and editing of laboratory manual.

Analytical Forensic Toxicology (FIS 470 – Lecture and FIS 471 – Laboratory) Aug. 2020 – Dec. 2020

- Prepared laboratory experiments and supervised instrument operation
- Assisted and guided students during laboratory exercises

First Year Seminar (FIS 191) Aug. 2020 – Dec. 2020

- Assisted with online administration of course material
- Graded assignments
- Assisted with advising first-year students on scheduling for upcoming semesters

Analysis of Seized Drugs (FIS 460 – Lecture and FIS 461 – Laboratory) Jan. 2020 – May 2020

- Prepared laboratory experiments and supervised instrument operation of Ultraviolet–visible spectroscopy (UV-Vis), Fourier Transform Infrared Spectroscopy (FTIR), and Gas chromatography-mass spectrometry (GCMS)
- Assisted and guided students during laboratory exercises
- Graded laboratory reports and assignments

Arson and Explosives Analysis (FIS 451 – Lecture and FIS 452 – Laboratory) Jan. 2020 – May 2020

- Prepared laboratory experiments and supervised instrument operation of Gas chromatography-mass spectrometry (GCMS)
- Assisted and guided students during laboratory exercises

Crime Scene Investigation Overview (FIS 202) Jan. 2018 – May 2018

- Assisted with online administration of course material and graded assignments

Expert Testimony Perspectives (FIS 306) Jan. 2018 – May 2018

- Assisted with online administration of course material and graded assignments

Social Relations-Forensic/Law (FIS 380) Jan. 2018 – May 2018

- Assisted with online administration of course material and graded assignments

Introduction to Forensic Science (FIS 201) Aug. 2017 – Dec. 2017

- Assisted with administration of course material and graded assignments

First Year Seminar (FIS 191) Aug. 2017 – Dec. 2017

- Graded assignments
- Assisted with advising first-year students on scheduling for upcoming semesters
- Planned events for FIS Living-Learning Community members

- Offered tours and informational visits for prospective students
- Attended recruiting events for prospective students

Graduate Teaching Assistant Aug. 2019 – Dec. 2019
WVU Biology Department, Morgantown

- Taught as primary instructor for two sections of general biology laboratory
- Assisted and guided students during laboratory exercises
- Prepared laboratory experiments and graded assignments

Graduate Assistant, Residential Learning and Assessment Aug. 2018 - May 2019
WVU Housing Department, Morgantown

- Designed and implemented specialized programs for the Forensic and Global Living Learning Communities
- Recruited for 12 Living-Learning Communities at WVU
- Evaluated best practices for living-learning communities
- Assessed programming and residential interactions

RESEARCH

Doctoral Research. *Enhancing the forensic comparison process of common trace materials through the development of practical and standardized methods.* [In progress]

- Developing systematic methods for comparison of duct tape and polymeric materials, and assessment of reliability using interlaboratory studies and computer algorithms.

  Research activities include:

  - Method development of systematic and quantitative comparisons of 3000 duct tape and 1000 polymer samples
  - Sample preparation of populations of duct tapes, textiles, paper, stamps, and automotive polymers
  - Development of templates for documentation of features, edge morphology, and final comparison decisions
  - Use of microscopes (stereomicroscope and digital microscope) to view samples under various magnifications and lighting conditions to observe and document features
  - Use of imaging (high-resolution scanner, touchless/3D scanner, and photography) to document samples for computer algorithm use
  - Development and coordination of an interlaboratory study of duct tape physical fits, incorporating responses from more than 20 participants at state and federal laboratories across the US and internationally
  - Coding in Python and R to assist in the development of an automated comparison algorithm to compare edges of separated materials
  - Data and statistical analysis using Microsoft Excel, Python, R, and JMP
  - Supervision of master's and undergraduate research students in sample preparation, training on systematic methods, instrument usage (microscope, Video Spectral Comparator (VSC), micro-X-ray Fluorescence (µ-XRF), Fourier Transform Infrared Spectroscopy (FTIR) and Glass Refractive Index Measurement (GRIM)), programming and coding in R and Python, and statistical analysis
  - Dissemination of results through presentations at various forensic conferences and published peer-reviewed manuscripts
  - Assisting with reports to funding agency

Master's Thesis. *Assessing the reliability of physical end matching and chemical comparison of pressure sensitive tapes*.

- Developed and assessed a systematic quantitative method for physical end comparisons of hand-torn and scissor cut duct tapes. More than 2000 edge comparisons completed since May 2017. Statistical analysis of error rates and score likelihood ratios calculated to offer support for examiner conclusions on duct tape physical fit determinations.
- Evaluated the use of micro-X-ray Fluorescence as a technique to compare the elemental composition of 90 electrical tape backings. Compared three XRF instrumental configurations to scanning-electron microscopy (SEM-EDS) and laser-ablation inductively coupled plasma mass-spectrometry (LA-ICP-MS).
- Evaluated semi-quantitative method of performing spectral comparisons as a more objective alternative to conventional spectral overlay comparisons.

CONFERENCE PRESENTATIONS

1. Development of systematic and practical documentation templates for tape and textile physical fit comparisons. **Meghan Prusinowski**, Zachary Andrews, Tatiana Trejos. Presented at MAFS/ASTEE annual conference. September 2022. [**Oral**]

2. Assessing significant factors that can influence physical fit examinations of tape and textiles. **Meghan Prusinowski**, Zachary Andrews, Cedric Neumann, Tatiana Trejos. Presented at European Academy of Forensic Sciences (EAFS) conference. June 2022. [**Poster**]

3. Validation of a Systematic Method for Duct Tape Physical Fits Through Inter-Laboratory Studies. **Meghan Prusinowski**, Evie Nguyen, Tatiana Trejos. Presented at AAFS Conference 2022. February 2022. [**E-Presentation]**

4. Development of Systematic Methods for the Physical Edge Comparison of Trace Materials. **Meghan Prusinowski**, Zachary Andrews, Tatiana Trejos. Presented at Escola de Ciências Forenses (ESCIF). October 2021. **[Oral/E-Presentation]**.

5. Microfiber Alignment in Stamp Edges for Physical Fit Comparisons. Colton Diges, Zachary Andrews, **Meghan Prusinowski**, Tatiana Trejos. Presented at WVU Summer 2021 Undergraduate Research Symposium. July 2021. **[Oral/E-Presentation]**.

6. Assessing the Value of Microfiber Alignment Between Stamp Edges for Physical Fit Comparisons. Zachary Andrews, Colton Diges, **Meghan Prusinowski**, Tatiana Trejos. Presented at 2021 Online Trace Evidence Symposium. July 2021. **[Oral/E-Presentation]**.

7. Lessons Learned from an Interlaboratory Study on Physical Fits of Duct Tape. **Meghan Prusinowski**, Tatiana Trejos. Presented to members of the NIST-OSAC Material/Trace Subcommittee. June 2021. **[Oral/E-Presentation]**.

8. OSAC Paint Interlaboratory Study on Interpretation. Andria Mehltretter, **Meghan Prusinowski**, Tatiana Trejos. Presented to members of the NIST-OSAC Material/Trace Subcommittee, Chemistry Scientific Area Committee, Forensic Science Standards Board and statisticians. March 2021. **[Oral/E-Presentation]**.

9. Development of Systematic Approaches for Physical Fit Comparisons of Trace Materials. **Meghan Prusinowski**, Zachary Andrews, Evie Nguyen, Tatiana Trejos. Presented at AAFS Conference (virtual). February 2021. **[E-Poster]**.

10. Development and Validation of Systematic Methods for Physical Fit Examinations. **Meghan Prusinowski**, Zachary Andrews, Evie Brooks, Tatiana Trejos. NEAFS/ASTEE combined Annual Meeting (virtual). October 2020. **[Oral/E-Presentation]**.

11. Development of a Systematic Method for Textile Physical Fit Comparisons. Evie Brooks, Zachary Andrews, **Meghan Prusinowski**, Tatiana Trejos. Presented at 2020 Online Forensic Symposium. July 2020. **[E-Poster]**.

12. Feasibility study of Spectral Comparison Contrast Methods for Interpretation of XRF Spectra of Electrical Tape Bindings. Evie Brooks, **Meghan Prusinowski**, Tatiana Trejos. Presented at 2020 Online Forensic Symposium. July 2020. **[E-Poster]**.

13. Validation of a Quantitative Method for Assessing Physical Fits of Trace Materials through Inter-Laboratory Studies and Computational Algorithms. **Meghan Prusinowski**, Evie Brooks, Pedram Tavadze, Aldo Romero, Tatiana Trejos. Presented at Online Forensic Graduate Symposium, West Virginia University, Morgantown, WV. May 2020. **[E-Poster].**

14. Evaluation of Spectral Comparison and Contrast Methods for the Interpretation of X-ray Fluorescence Spectra of Electrical Tape Backings. Evie Brooks, **Meghan Prusinowski**, Andria Mehltretter, Tatiana Trejos. Presented at Online Forensic Graduate Symposium, West Virginia University, Morgantown, WV. May 2020. **[E-Poster]**.

15. Duct Tape Fracture Fit Interlaboratory Study: Lessons Learned. **Meghan Prusinowski**, Evie Brooks, Tatiana Trejos. Presented to members of the NIST-OSAC Material/Trace Subcommittee, Chemistry Scientific Area Committee, Forensic Science Standards Board and statisticians. March 2020. **[Oral].**

16. Assessing the Value of a Physical End Match in Trace Evidence: Comparison of Human-based and Computational-based Approaches. **Meghan Prusinowski**, Evie Brooks, Pedram Tavadze, Aldo Romero, Tatiana Trejos. Presented at AAFS Conference, Anaheim CA, February 2020. Nominated for Forensic Science Foundation Emerging Forensic Science Award. **[Oral]**.

17. Developing a Systematic Method for the Comparison of Duct Tape Edges. **Meghan Prusinowski**, Evie Brooks, Tatiana Trejos. Presented to members of the NIST-OSAC Material/Trace Subcommittee, Chemistry Scientific Area Committee, Forensic Science Standards Board and statisticians. July 2019. **[Oral].**

18. Putting the Pieces Together: Developing and Applying a Systematic Method for the Comparison of Duct Tape Edges. **Meghan Prusinowski**, Evie Brooks, Tatiana Trejos. Presented at MAAFS conference, Morgantown, WV. May 2019. **[Oral].**

19. The Forensic Evaluation of Cut and Torn Duct Tape Fractures: A Validation Study to Quantify the Quality of a Physical Match. **Meghan Prusinowski**, Tatiana Trejos. Presented at AAFS conference, Baltimore, MD. February 2019. **[Poster].**

20. Characterization and Comparison of Electrical Tape Backings by X-Ray Florescence (XRF). **Meghan Prusinowski**, Andria Mehltretter, Claudia Martinez-Lopez, Jose Almirall, Tatiana Trejos. Presented at AAFS conference, Baltimore, MD. February 2019. Nominated for Forensic Science Foundation Emerging Forensic Science Award. **[Poster]**.

21. Application of Score Likelihood Ratios for the Assessment of Duct Tape Fracture Comparisons. Mollie Fisher, **Meghan Prusinowski**, Tatiana Trejos. Presented at CBDIAI 2017 conference, Morgantown, WV. April 2018. **[Poster]**.

22. Evaluation of Performance Rates in the Determination of Duct Tape Fracture Matches. Aaron Brake, **Meghan Prusinowski**, Tatiana Trejos. Presented at joint MAFS/SAFS/ASTEE 2017 conference, Cincinnati, OH. September 2017. **[Poster and Oral]**.

23. Assessment of the Validity of Duct Tape End Matches in Forensic Comparisons. Aaron Brake, **Meghan Prusinowski**, Tatiana Trejos. Presented at WVU Summer Undergraduate Research Symposium, Morgantown, WV. July 2017. **[Poster]**.

PUBLICATIONS

1. **Meghan Prusinowski**, Evie Brooks, Cedric Neumann, Tatiana Trejos. Forensic interlaboratory evaluations of a systematic method for examining, documenting, and interpreting duct tape physical fits. Forensic Chemistry. 2023. 34: 100487. https://doi.org/10.1016/j.forc.2023.100487

2. Zachary Andrews, **Meghan Prusinowski**, Cedric Neumann, Tatiana Trejos. Assessing physical fit examinations of stabbed and torn textiles through a large dataset of casework-like items and inter-laboratory studies. *Journal of Forensic Sciences*. Under review.

3. Pedram Tavadze, Logan Lang, **Meghan Prusinowski**, Zachary Andrews, Tatiana Trejos, Aldo Romero. Using convolutional neural networks to support examiners in duct tape physical fit comparisons. *Forensic Science International*. Under review.

4. **Meghan Prusinowski**, Zachary Andrews, Cedric Neumann, Tatiana Trejos. Assessing significant factors that can influence physical fit examinations – physical fits of torn and cut duct tapes. *Forensic Science International*. Published online January 2023. 343. https://doi.org/10.1016/j.forsciint.2023.111567

5. Evie Brooks, Andria Mehltretter, **Meghan Prusinowski**, Tatiana Trejos. Optimization and Validation for Qualitative Elemental Analysis of Electrical Tape Backings by X-ray Fluorescence. *Forensic Chemistry*. October 2020. https://doi.org/10.1016/j.forc.2020.100291

6. Evie Brooks, **Meghan Prusinowski**, Susan Gross, Tatiana Trejos. Forensic Physical Fits of Trace Materials: A Review. *Forensic Science International*. August 2020. Available online May 29th 2020. https://doi.org/10.1016/j.forsciint.2020.110349

7. **Meghan Prusinowski**, Evie Brooks, Tatiana Trejos. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. *Forensic Science International*. Feb. 2020. Available online Dec. 19th 2019. https://doi.org/10.1016/j.forsciint.2019.110103

8. **Meghan Prusinowski**. Assessing the reliability of physical end matching and chemical comparison of pressure sensitive tapes. *Graduate Theses, Dissertations, and Problem Reports*. 2019. 4026. https://doi.org/10.33915/etd.4026

9. **Meghan Prusinowski**, Andria Mehltretter, Claudia Martinez-Lopez, Jose Almirall, Tatiana Trejos. Assessment of the utility of X-ray Fluorescence for the chemical characterization and comparison of black electrical tape backings. *Forensic Chemistry*. May 2019. Available online Feb 4st 2019. https://doi.org/10.1016/j.forc.2019.100146

PROFESSIONAL ASSOCIATIONS AND ORGANIZATIONS

1. National Institute of Standards and Technology – Organization of Scientific Area Committees (NIST-OSAC) Trace Subcommittee – Full member (since October 2022), Affiliate Member (July 2019 to October 2022)
2. American Academy of Forensic Science (AAFS) – Student Affiliate
3. American Society of Trace Evidence Examiners (ASTEE) – Student Affiliate

HONORS, AWARDS, AND RELEVANT COLLABORATIONS

1. Selected as member of NIST-OSAC Materials/Trace Subcommittee – A prestigious nomination for a student. NIST-OSAC is an organization of more than 550 experts in forensic science disciplines, as well as research, measurement science, statistics, law, and policy. The mission of OSAC is to develop and evaluate scientific standards for use in the forensic science disciplines.

   Main contributions:

   - Started as an affiliate member in July 2019, was nominated and accepted for full membership starting October 1st, 2022.
   - Member of the physical fit task group, assisted with design of physical fit survey, collaborated on an interlaboratory study, presented research results, and provided introductory training on edge similarity score approach. Assisted task group in writing the first standard on examination of physical fits.
   - Currently participating in manuscript preparation on a Paint Interpretation Interlaboratory Study conducted by Materials (Trace) Subcommittee, and working on the development of a Glass Interpretation Interlaboratory Study
   - Co-author of review paper which will serve as a foundation for development of future standard methods.
   - Member of the Polymers Task group, assisted with writing of polymer XRF standard and currently assisting with physical characteristics of tape standard
   - Attended bi-annual subcommittee meetings since July 2019 to present data on interlaboratory study results and XRF analysis of electrical tape data.
2. Interdisciplinary research collaborations with West Virginia University Physics Department, South Dakota State University Statistics Department, Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF), and the Federal Bureau of Investigation (FBI) Chemistry Unit.
3. Participated as assistant instructor of Forensic Examinations of Physical Fits—Past, Present, and Future Workshop. Hosted at MAFS/ASTEE annual meeting, September 14th, 2022.
4. Received American Society of Trace Evidence Examiners (ASTEE) Travel Award. Sept. 2022.
5. Received American Society of Trace Evidence Examiners (ASTEE) Scholarship Award. Feb. 2020.
6. Received Forensic Sciences Foundation (FSF) Student Scholarship Award. Nov. 2019.
7. Dean's List at SUNY Geneseo Fall 2014, Spring 2015, Spring 2016.

SKILLS AND MAIN QUALIFICATIONS

1. Competence in chemistry

   a. Bachelor’s degree in Chemistry, Master’s degree in Forensic Science with an emphasis in Forensic and Analytical Chemistry, Doctoral Degree in Forensics with emphasis in Forensic and Analytical Chemistry (anticipated graduation Fall 2022).

2. Robust foundation on fundamentals of manufacturing, sample preparation, recovery, forensic examination, and interpretation of trace materials following ASTM test methods and practices.
   a. **Glass** (analysis of refractive index following ASTM E1967-19, elemental analysis of glass on XRF following ASTM E2926-17, elemental analysis of glass using ICP-MS following ASTM E2330 and LA-ICP-MS following ASTM E2927-16)
   b. **Paint** (analysis following ASTM E1610-18, FTIR analysis of paint following ASTM E2937-18, and elemental analysis following ASTM E2809)
   c. **Tape** (analysis following ASTM E3260-21, FTIR analysis following ASTM E3085-17, and relevant SWGMAT guidelines for physical examinations and characteristics)
   d. **Gunshot Residue (GSR)** (SEM-EDS analysis following E1588-20)
   e. **Hairs and fibers** (microscopic examination and analytical techniques following relevant SWGMAT guidelines)

NOTE: Experience acquired through:
   a. Graduate research
   b. Graduate coursework (Trace Evidence Examination, Advanced Forensic Chemistry, Introduction to Microscopy, Forensic Statistics, Foundations of Criminalistics, Court Testimony, Forensic Casework Practicum, Research Design in Forensic Science, Forensic Informatics, Professional Ethics in Forensic Science)
   c. Teaching assistance (Trace Evidence Examination, Forensic Chemical Analysis, Introduction to Forensic Science)
   d. Participation in workshops and conference presentations (e.g., Workshop on LIBS, ICP-MS and LAICP-MS for Glass and Gunshot Residue, hosted by WVU Forensic and Investigative Science Department (14-June-2022 to 17-June-2022).

3. Experience in fundamentals of sample preparation, recovery, forensic examination and interpretation of forensic chemical and biological evidence (illicit drugs, biological materials, fire debris)

NOTE: Experience acquired through:
   a. Graduate coursework (Advanced Forensic Chemistry, Introduction to Microscopy, Forensic Statistics, Foundations of Criminalistics, Court Testimony, Forensic Casework Practicum, Forensic Informatics, Professional Ethics in Forensic Science, Forensic Laboratory Management, Advanced Forensic Biology)
   b. Teaching assistance (Forensic Chemical Analysis, Analysis of Seized Drugs, Arson and Explosives Analysis, Forensic Toxicology, Introduction to Forensic Science)

4. Proficient in various analytical instrumentation and interpretation of results:
   a. Scanning Electron Microscope -Energy Dispersive Spectroscopy (SEM-EDS)
   b. micro-X-ray Fluorescence (µ-XRF)
   c. Gas Chromatography-Mass Spectrometry (GC-MS)
   d. Laser-Induced Breakdown Spectroscopy (LIBS)
   e. Laser-Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS)
   f. Video Spectral Comparator (VSC)

   g. Fourier Transform Infrared Spectroscopy (FTIR)
   h. Ultraviolet–visible Spectroscopy (UV-Vis)
   i. Glass Refractive Index Measurement (GRIM)
   j. Microscopy (Stereomicroscope, Polarizing-light microscope, and Comparison microscope)

5. Experience coding in R, Python and Linux

6. Experience with statistical packages (R, JMP, Excel)

7. Ability for problem-solving and initiative acquired through:
   a. Development and revision of methods for systematic comparison and feature documentation
   b. Analysis of data for research in various formats (R, Excel, etc) and interpretation of statistical models for data analysis
   c. Assisting colleagues and external collaborators in data analysis and displaying data in graphical format
   d. Troubleshooting of instrumentation (FTIR, GRIM, µ-XRF, digital microscope, high-resolution scanner)

8. Adaptability and organizational skills developed through:
   a. Planning and coordination of research activities for undergraduate and graduate research students
   b. Participation in organization of a department-hosted research symposium
   c. Organization of interlaboratory studies and data collection
   d. Collection, organization, and documentation of physical and digital sample databases

9. Strong oral and written communication and collaboration skills:
   a. Training undergraduate and graduate students and researchers on instrumentation usage and fundamentals, standard operating procedures, and data analysis
   b. Collaboration with colleagues to complete research activities and data analysis
   c. Participation in interlaboratory exercises with external collaborators (including Florida International University, Federal Bureau of Investigation, Organization of Scientific Area Committee, etc)
   d. Publication of peer-reviewed manuscripts and reports to funding agencies (see **Publications**)
   e. Numerous oral and poster presentations at scientific conferences and meetings (see **Conference Presentations**)