Document Title:	Assessing the Strength of Trace Evidence Fracture Fits through a Comprehensive, Systematic and Quantifiable Approach
Author(s):	Tatiana Trejos, Aldo Romero
Document Number:	309918
Date Received:	December 2024
Award Number:	2020-DQ-BX-0012

FINAL RESEARCH REPORT

Agency: National Institute of Justice

Award number: 2020-DQ-BX-0012

Project Title: Assessing the Strength of Trace Evidence Fracture Fits through a Comprehensive, Systematic and Quantifiable Approach

PI: Tatiana Trejos, Associate Professor
Tatiana.trejos@mail.wvu.edu
304.293.6978

Co-PI: Aldo Romero, Professor
Aldo.Romero@mail.wvu.edu
304.293.6317

Submitting official: Tatiana Trejos, Associate Professor, Department of Forensic and Investigative Science, WVU

Submission date: 06/30/23
DUNS: 191510239
EIN: 550665758

Recipient Organization: West Virginia University Research Corporation
886 Chestnut Ridge Road P.O. Box 6845, Morgantown, WV 26506-6845

Award Period: 01/01/2021 to 06/30/23
Award Amount: $466,543.00

Reporting Period End Date: 06/30/23

FINAL RESEARCH REPORT

Signature of Submitting Official:

Tatiana Trejos, Assistant Professor

Katie Schneller
Director, Office of Sponsored Programs

FINAL RESEARCH REPORT

Assessing the Strength of Trace Evidence Fracture Fits through a Comprehensive, Systematic and Quantifiable Approach

Tatiana Trejos[1], Aldo Romero[2], Meghan Prusinowski[1], Zachary Andrews[1], Pedram Tavadze[1,2], Cedric Neumann[3]

[1]West Virginia University
Department of Forensic and Investigative Science

[2]West Virginia University
Department of Physics and Astronomy

[3]Battelle Memorial Institute

TABLE OF CONTENTS

Assessing the Strength of Trace Evidence Fracture Fits through a Comprehensive, Systematic and Quantifiable Approach

I SUMMARY OF THE PROJECT

1.1. Abstract

Criminal activities such as sexual assaults, kidnappings, hit and runs, and homicides often lead to the fracturing of materials. The realignment between fragments left at the scene and those items recovered from an individual, or object of interest, could become crucial evidence during the investigation. These fracture fits are often regarded as the highest degree of association of trace materials due to the common belief that fracture edges often produce individualizing patterns. Nonetheless, there is a need to demonstrate the scientific validity of this assumption. Currently, the examination of fracture edges involves the subjective judgment of the examiner without quantifiable uncertainty. There are no consensus-based standard methodologies for the identification of distinctive features on a fractured edge, a systematic criterion for informing a fit/non-fit decision, or methods for assessing the weight of the evidence. Thus, there is a critical need to develop, validate, and standardize fracture fit examinations and their respective interpretation protocols. In the absence of such foundations, the assessment of the value of the evidence and the reliability of the expert's testimony would remain challenging.

The overall goal of this research was to develop an effective and practical approach that provides an empirically demonstrable basis to assess the significance of trace evidence fracture fits. Our specific goals were first, to develop a systematic method for the comparison of fracture fits of common trace materials such as duct tapes, textiles, and automotive plastics, using both human-based protocols and automated computational algorithms. Second, to develop a relevant extensive database of nearly 9,000 samples to evaluate performance rates in this field, and assess the weight of a fracture fit using similarity metrics, probabilistic estimates, and score likelihood ratios. Third, to evaluate the utility and reliability of the proposed approach through inter-laboratory studies that can establish consistency base rates. The partnership of experienced forensic researchers, computational material science physicists, statisticians, and practitioners was crucial for developing strategies to facilitate the future adoption of the developed approaches within crime laboratories.

This research specifically addressed several research needs identified by NIST-OSAC[1] and the NIJ-TWG[2] (quantitative assessment of error rates, scientific foundations, standardization, validation, interpretation, casework review, and proficiency assessment). As a result, this study is anticipated to transform current trace evidence practice by providing—for the first time—harmonized examination protocols and decision thresholds, effective mechanisms to ensure transparent and systematic peer-review process and interlaboratory testing, and quantitative basis that substantiate the evidential value of fracture fit conclusions.

1.2. Major Goals and Objectives

This study aims to contribute to the advancement of the trace evidence discipline by developing a practical approach with an empirically demonstrable basis to assess the probative value of fracture fits. The overall goal of this proposal is to answer the question: "how significant is a given fracture fit between two objects?" Specifically, the study is designed to answer this question within a context relevant to U.S. criminal justice by providing a quantifiable basis for assessing the weight of the evidence and reliable scientific grounds. As a result, it is anticipated that this study will provide a necessary foundation to help trace evidence examiners to move away from subjective conclusions of fracture matches and give the trier of fact tangible and understandable resources to assess the relevance of the evidence. The specific objectives of this research are to:

1) **Objective 1:** Develop a systematic method for comparing fracture fits of common trace materials such as duct tapes, textiles, and plastics, using more objective human-based protocols and automated computational algorithms.
2) **Objective 2:** Develop an extensive collection of trace physical fractures to validate methods for a quantitative assessment of the evidence and test the methods proposed under objective 1. This collection will be encapsulated in a digital database management system.
3) **Objective 3:** Evaluate the utility and reliability of the proposed approach in the casework context through inter-laboratory studies.

These goals are accomplished through the following specific tasks:

1) **Task 1 (Objective 1)**—Develop systematic methods to compare fracture fits of common trace materials such as duct tapes, textiles, and plastics.
2) **Task 2 (Objective 1)** —Develop and validate automated computational algorithms to compare fracture fits.
3) **Task 3 (Objective 2**) —Develop an extensive database on trace physical fractures of duct tape, textiles, and plastics, and test the methods proposed under Objective 1.
4) **Task 4 (Objective 2)** —Validate quantitative methods for assessing the probative value of fracture fits.
5) **Task 5 (Objective 3)** —Design interlaboratory studies for the evaluation of error rates of the proposed comparison approach among practitioners

DISCLAIMER: This report summarizes the main findings of this research project. Some of these findings have been published in scientific journals[3-6], thesis, dissertations [7-8], or have been submitted for publication and are under review.[9-12] Therefore, some content, tables, and figures are an adaptation of published articles. As per journal copyright policies, the authors are entitled to re-use portions, excerpts, and their own figures or tables in other works that are not published commercially, without permission or payment (with full acknowledgment of the original article). More detailed information about the methods, data analysis, and results can be found in the published manuscripts listed in Section 3.1 and cited in the respective sections of this document.

1.3. Research Questions

The separation of materials such as tape, plastics, and textiles from their original source frequently occurs during violent activities, leaving distinct patterns along the fractured edges. These features assist examiners when attempting to determine the source of the sample. During the examination of these materials, the analyst compares a known and questioned item to determine if they fit back together like a puzzle. The 3D realignment along the edges of the objects is known as a physical match and is often regarded as conclusive evidence. However, there are still several challenges in demonstrating the reliability and "individuality" of physical fits due to a lack of knowledge of error rates and uncertainty associated with such conclusions. Crime laboratories have no consensus-based standard methods describing which characteristics are most discriminating, or how to evaluate and document significant fracture features. The process is, therefore, extremely subjective. Moreover, when deciding if the edges are similar enough to be considered a match, the examiner does not have scientifically established criteria to inform their opinion.

A decade ago, the National Academy of Sciences (NAS) raised awareness of the need for reporting error rates and uncertainties associated with subjective analysis in pattern evidence. These concerns resurfaced with the President's Council of Advisors on Science and Technology (PCAST) report in 2016 and statements from the American Statistical Association in 2019.[13-15] Also, the Organization of Scientific Area Committees (NIST-OSAC) recently identified a major gap in research on trace evidence's physical fits[1], which coincides with the 2019 Forensic Science Research and Development Technology Working Group (TWG)[2] identification of top priorities in this field.

As a result, this study was designed to answer fundamental questions to build stronger scientific foundations of fracture fit examinations and provide the criminal justice system with reliable resources to assess the relevance of the evidence. These research questions are:

1. Do all physical fits hold the same probative value?
2. Which individual or class characteristics can be evaluated in fractured edges to assist the forensic examiner during a physical fit examination?
3. Are these features dependent on the fractured or separated material?
4. Which factors influence the occurrence of these features and the quality of a physical fit?
5. What are the performance rates of physical fit examinations? Are the performance rates dependent on the type of object?
6. Can quantitative metrics demonstrate the quality of a fit and be used for the probabilistic interpretation of the evidence?
7. Can computational and mathematical models be used to complement human-based examinations?
8. What strategies can be developed to minimize potential bias and subjectivity during the forensic examination of physical fits?
9. Can standardized protocols be developed for the examination, documentation, and interpretation of physical fits through the assessment of the method via large datasets and interlaboratory studies?

1.4. Research Design, Methods, Data Analysis

1.4.1. Research design and methods of analysis

Project tasks and methodology

The methodology and experimental design are described below within five major tasks to address the major research questions and objectives of this study (See **figure 1).**

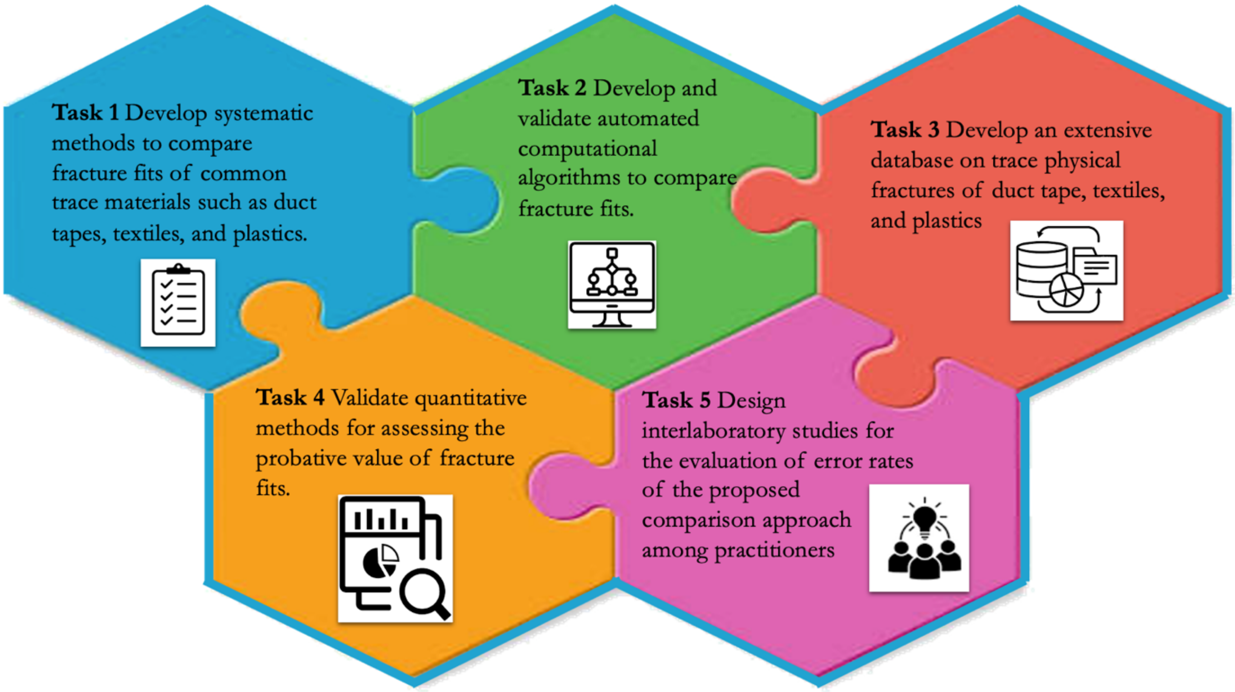

Task 1 Develop systematic methods to compare fracture fits of common trace materials such as duct tapes, textiles, and plastics.
Task 2 Develop and validate automated computational algorithms to compare fracture fits.
Task 3 Develop an extensive database on trace physical fractures of duct tape, textiles, and plastics
Task 4 Validate quantitative methods for assessing the probative value of fracture fits.
Task 5 Design interlaboratory studies for the evaluation of error rates of the proposed comparison approach among practitioners

Figure 1. *Summary of the main five experimental research tasks.*

Task 1 (Objective 1)— Develop a systematic method for the comparison of fracture fits of common trace materials such as duct tapes, textiles, and plastics

In this task, we have developed material-specific systematic methods for identifying and comparing relevant features along fracture edges. To optimize the impact of this study on criminal justice, we utilized a survey conducted by the newly formed NIST-OSAC Physical Fit Task Group[16] to select the most prevalent materials submitted to forensic agencies. Thus, the three materials evaluated in this research were duct tapes, textiles, and plastics that are typically fractured or separated in a variety of crimes such as sexual assaults, homicides, suicides, and vehicular offenses, to mention some.

Trace materials can be recovered from crime scenes as microscopic units often invisible to the naked eye or as larger pieces left behind during the contact between objects and individuals. The pieces should be relatively large (~ cm long) rather than small micro-traces to conduct a fracture alignment. For instance, sizeable pieces of duct tape are often received in cases where victims have been gagged or bound and in the construction of improvised explosive devices. Textile damage is observed in stabbing and tearing during forceful contact between individuals or sharp objects. Plastics are more commonly observed as a product of the bending of vehicle bumpers or the breaking of headlights during car collisions. As a result, our empirical datasets have been designed to simulate samples generated under these types of scenarios closely.

Each of the materials of interest has different chemical and physical properties and manufacturing construction that inevitably imparts various features during the separation of objects (**Figure 2**).
As a result, the first step in this task was to develop means to standardize the examination protocols. We have defined for each material: 1) what features are relevant for the comparison, 2) what is the smallest subunit "bin" of comparison across the tear or fracture, 3) how to document the examiner observations and decisions, 4) how to convert the qualitative observations to a quantitative measure, and 5) how to apply the qualitative and quantitative criteria in the assessment of the evidential value.

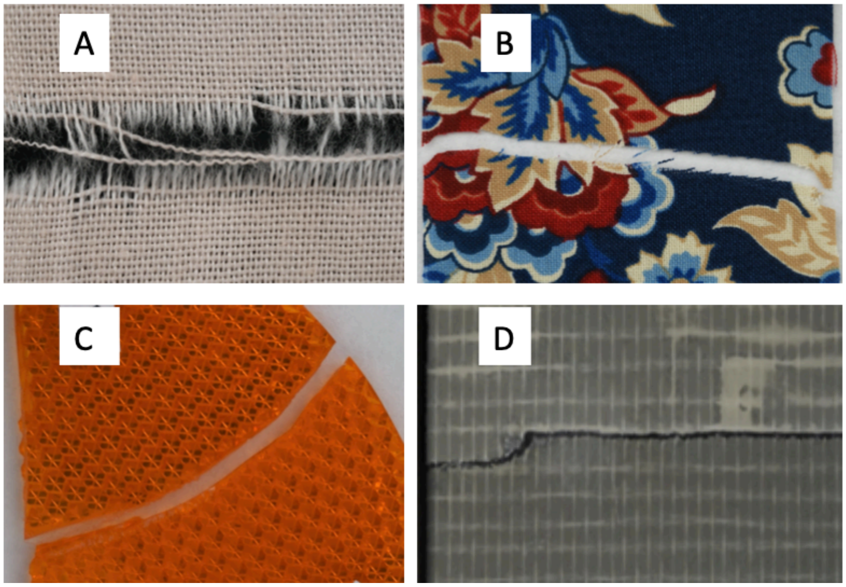


Fig. 2. *Example of physical fits of textiles (A, B), plastic (C), and duct tape (D).*

Duct tapes

For duct tapes, we developed a systematic approach to compare edges. A typical duct tape has a backing layer, a reinforcement scrim fiber, and adhesive. All three components are considered during the examination. Duct tapes of different quality grade (high quality, HQ, medium quality MQ, M, and low quality LQ, L) were separated by hand-torn, HT, or scissor-cut, SC). A subset was also stretched to simulate complex case circumstances (S). We defined terms to describe overall fractured patterns and relevant features. Microscopic features include alignment of the edges, scrim in the warp direction across the tapes, weft fibers across the tear, stretching or distortion in the direction of tearing, backing dimples or markings across the edge, and protrusions and indentations or other loss of material across the edges). The examiners put the joint edges together and observed the tear patterns under a microscope at 10-40x. We used a stereo microscope with an automated stage and reflected and transmitted light and alternate UV and IR light to boost important features such as fillers or fiber fluorescence.

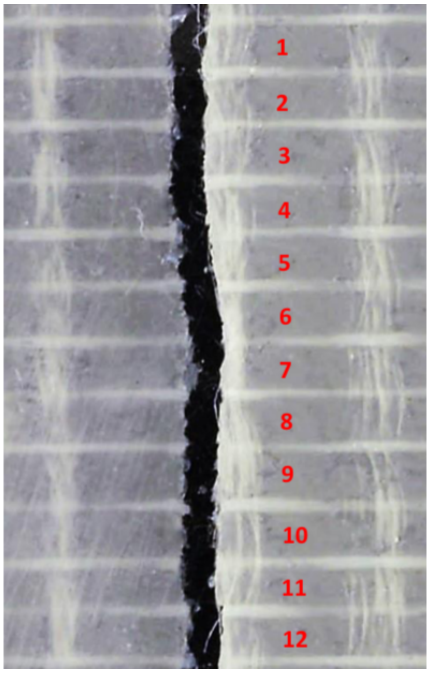


Fig. 3. Example of alignment of scrim bins to estimate ESS

Also, we have defined the smallest comparison bin as the torn area between a pair of warp scrims. The rationale for this selection is that the number of scrims is a constant feature across a tape roll and therefore provides a systematic means to evaluate the similarities and dissimilarities between tapes. The examiners then report an edge similarity score (ESS) for each tape comparison. An example edge is shown in **Figure 3**, where 12 scrim bins are visible. The examiner would divide the number of matching areas by the total number of scrim bins across the entire tape width to calculate the ESS.

The comparison is then performed by flipping the tape and inspecting the alignment at the backing side. This has been shown to improve other methods that used relative alignment lengths. The estimation of the size of a fracture is somehow arbitrary because a tape fracture is rarely straight. In contrast, our method of estimating a score by scrim bins assures that the examiners will be looking at the same areas and the same number of regions when performing an examination. Moreover, the method applies to a large variety of tapes, regardless of their number of scrims, which is often associated with the tape grade quality. The match score is then utilized to make quantitative assessments of the weight of the evidence and provide a way to use a consistent reporting criterion across examiners. Also, an overall edge alignment and macro evaluation of the distinctive surface and edge features were incorporated, considering the feedback received by examiners during interlaboratory studies.

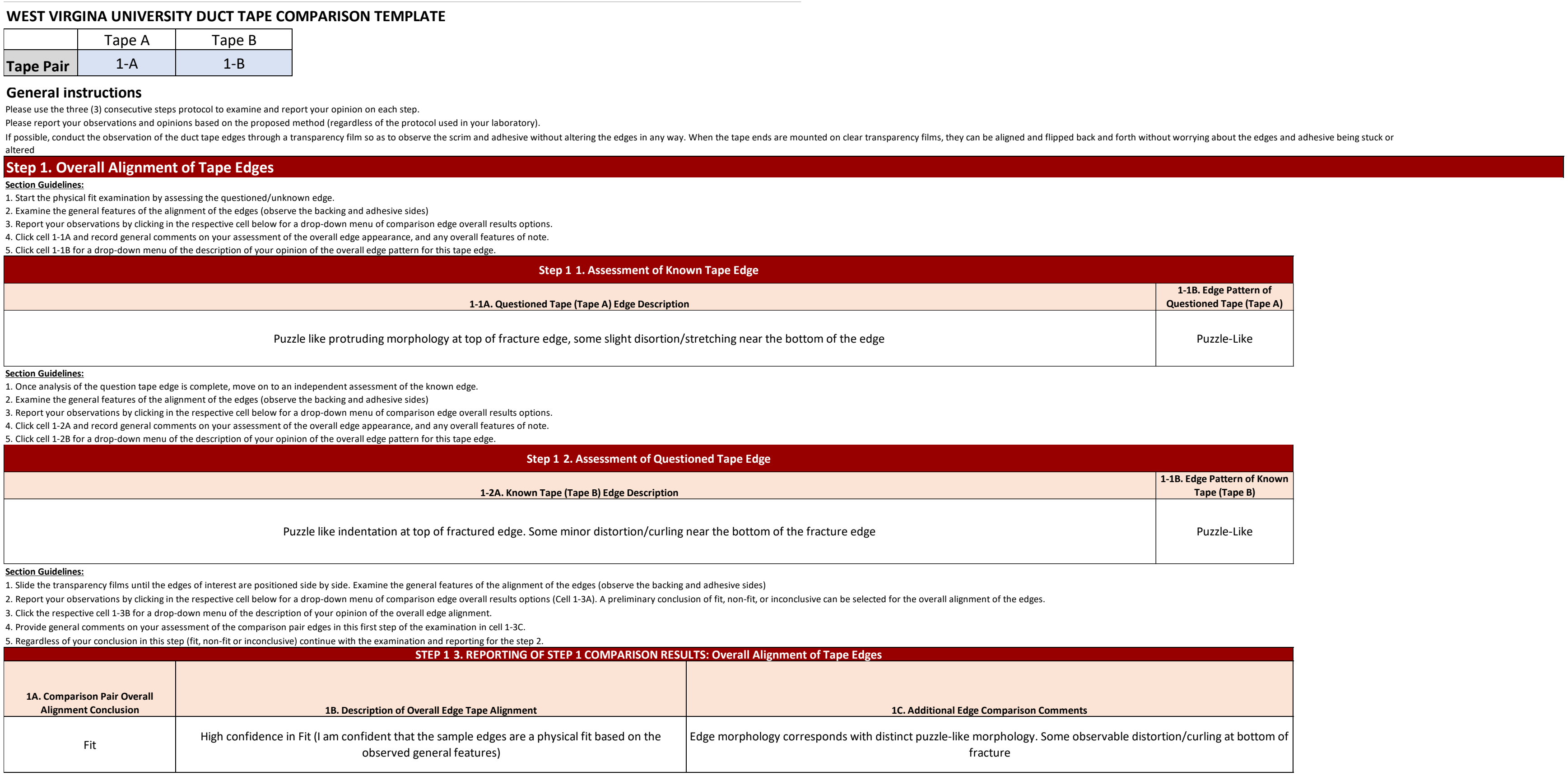
In terms of documentation, the examiner notes any significant factors holding weight in their decision. Through this study, we found that incorporating detailed documentation of the comparison edge features by comparison bin has proven effective during the peer-review process, adds transparency to the decision-making process, and facilitates standardization of procedures. The use of bin-to-bin annotations on digital images has been a beneficial approach, as examiners can independently compare the same regions that served as the basis for their decisions and evaluate any potential discrepancies.

Simple and practical Excel templates were designed with step-by-step instructions to document the observations, including auto-populated cells for each feature of interest and an embedded macro that color-codes each bin decision to aid the examiner in the quick assessment of the fit and non-fit areas. For example, the bin score cells have a built-in 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). The templates also calculate the ESS metric as the data is entered by the examiner and guide the examiner to document final decisions, opinions, and observations in a systematic and reproducible manner. A template example is shown in **Figures 4-1, 4-2,** and **4-3**, where each step of the comparison is broken down into three sections.

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 (see **Figure 5**). [7] 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.

In the second step, more detailed examination and documentation are conducted to identify major features in relatively large regions of the tape using macroscopic and microscopic observations. To aid with inter-examiner consistency, the template includes eight major features that we have found to hold the most weight in the decision process of a fit or non-fit in this study. Examples of these features are shown in **Figure 6**. 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 provide important systematic qualitative criteria for the examination of duct tape fits.

Finally, the last step consists of observation and documentation of the same eight defined features in step two, but here the observations are made on each bin-by-bin comparison, while the samples are compared under the microscope. An advantage of this step is that it adds objectivity since examiners must use standardized criteria to make a data-driven assessment of each bin independently to the results of previous bins. It is in this step that ESS is automatically calculated and displayed in the template to assist the examiner in forming their opinion.



WEST VIRGINA 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.

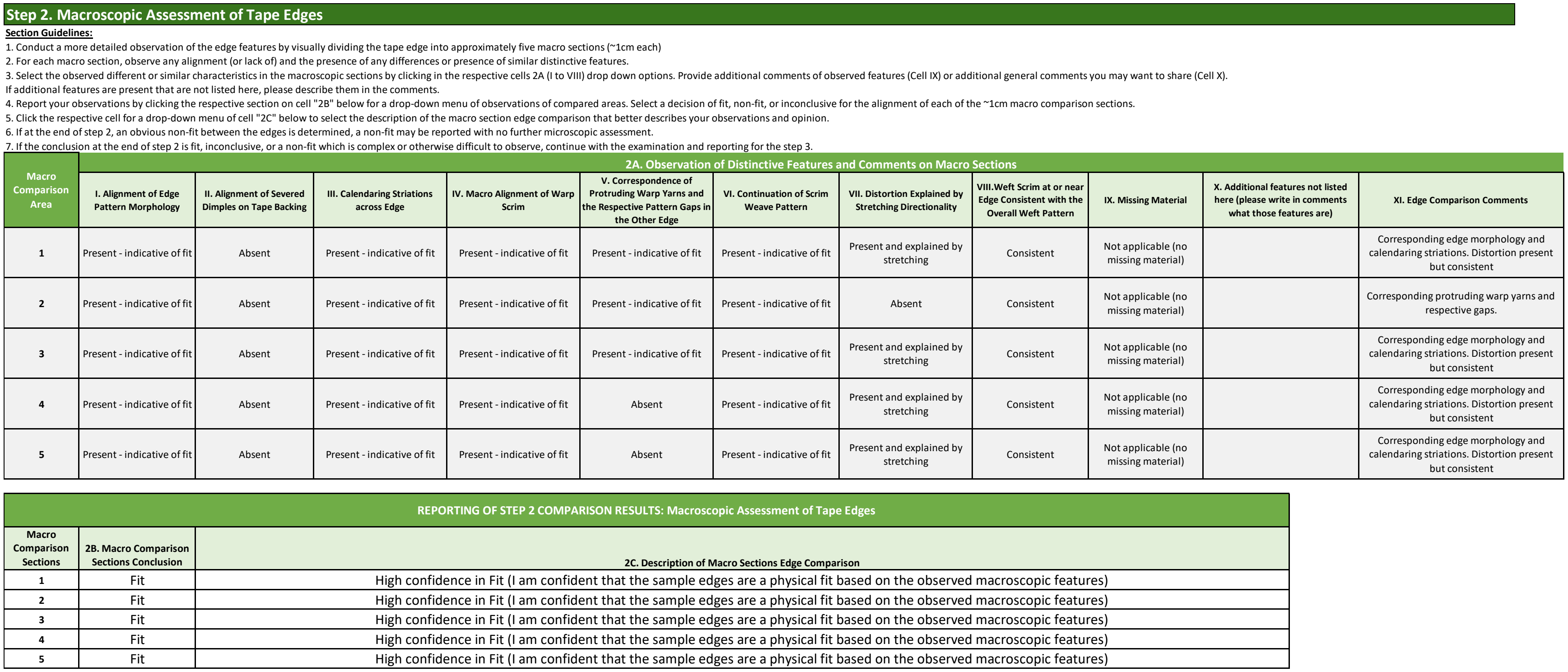
Step 1 2. Assessment of Questioned Tape Edge	
1-2A. Known Tape (Tape B) Edge Description	1-1B. Edge Pattern of Known Tape (Tape B)
Puzzle like indentation at top of fractured edge. Some minor distortion/curling near the bottom of the fracture edge	Puzzle-Like

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 4-1 *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.*[3]



Step 2. Macroscopic Assessment of Tape Edges

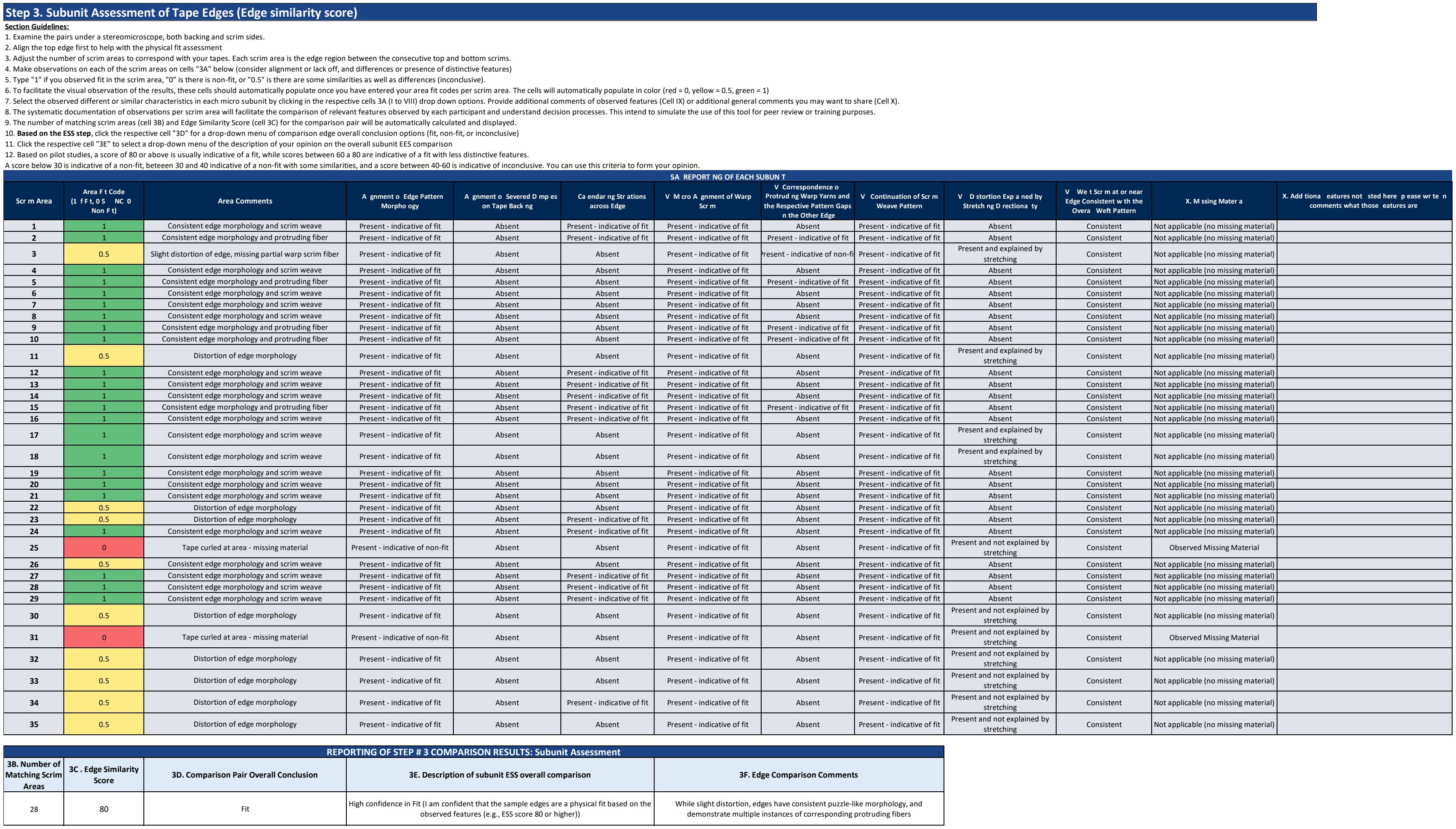
Section Guidelines:

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 4-2. *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.* [3]



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 REPORT NG OF EACH SUBUN T									
Scr m Area	**Area F t Code (1 f F t, 0 5 NC 0 Non F t)**	**Area Comments**	**A gnment o Edge Pattern Morpho ogy**	**A gnment o Severed D mp es on Tape Back ng**	**Ca endar ng Str ations across Edge**	**V M cro A gnment of Warp Scr m**	**V Correspondence o Protrud ng Warp Yarns and the Respective Pattern Gaps n the Other Edge**	**V Continuation of Scr m Weave Pattern**	**V D stortion Exp a ned by Stretch ng D rectiona ty**	**V We t Scr m at or near Edge Consistent w th the Overa Weft Pattern**	**X. M ssing Mater a**	**X. Add tiona eatures not sted here p ease wr te n comments what those eatures 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 4-3. *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.*[3]

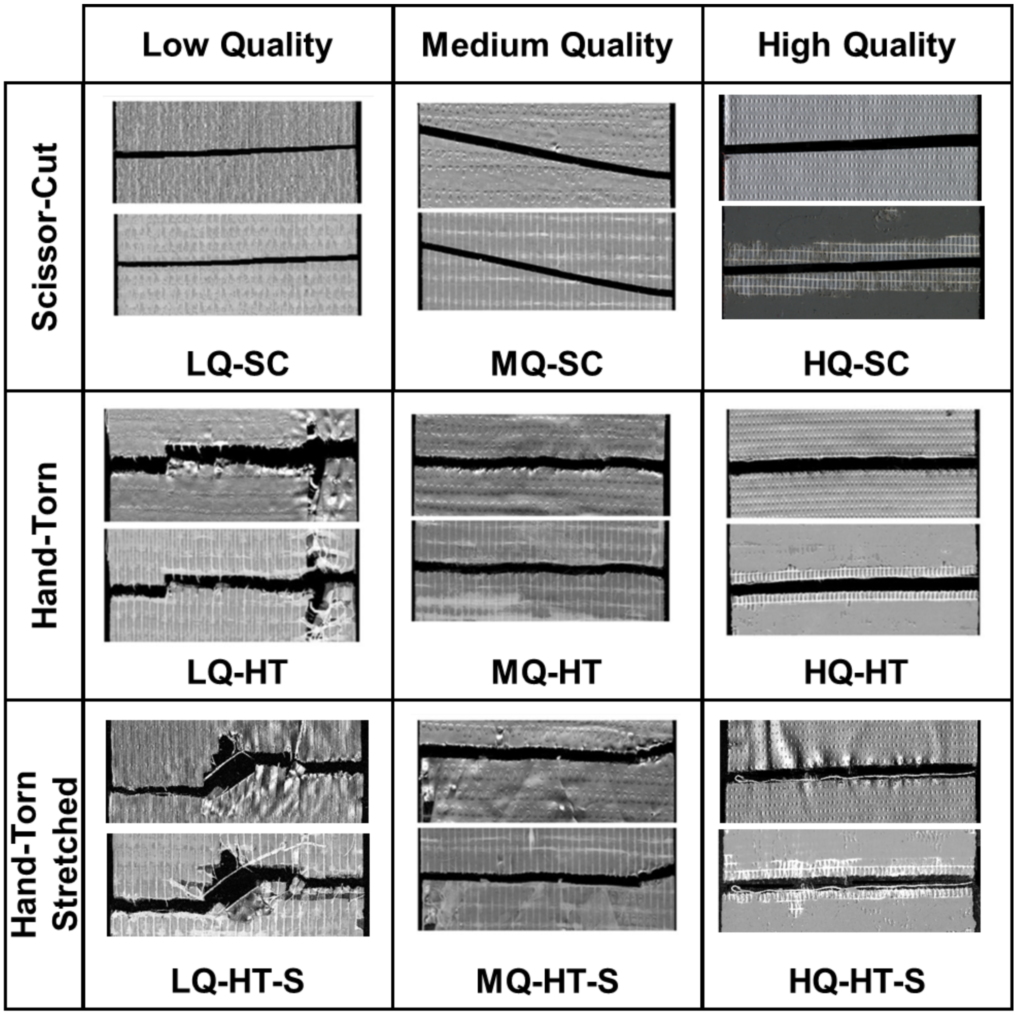
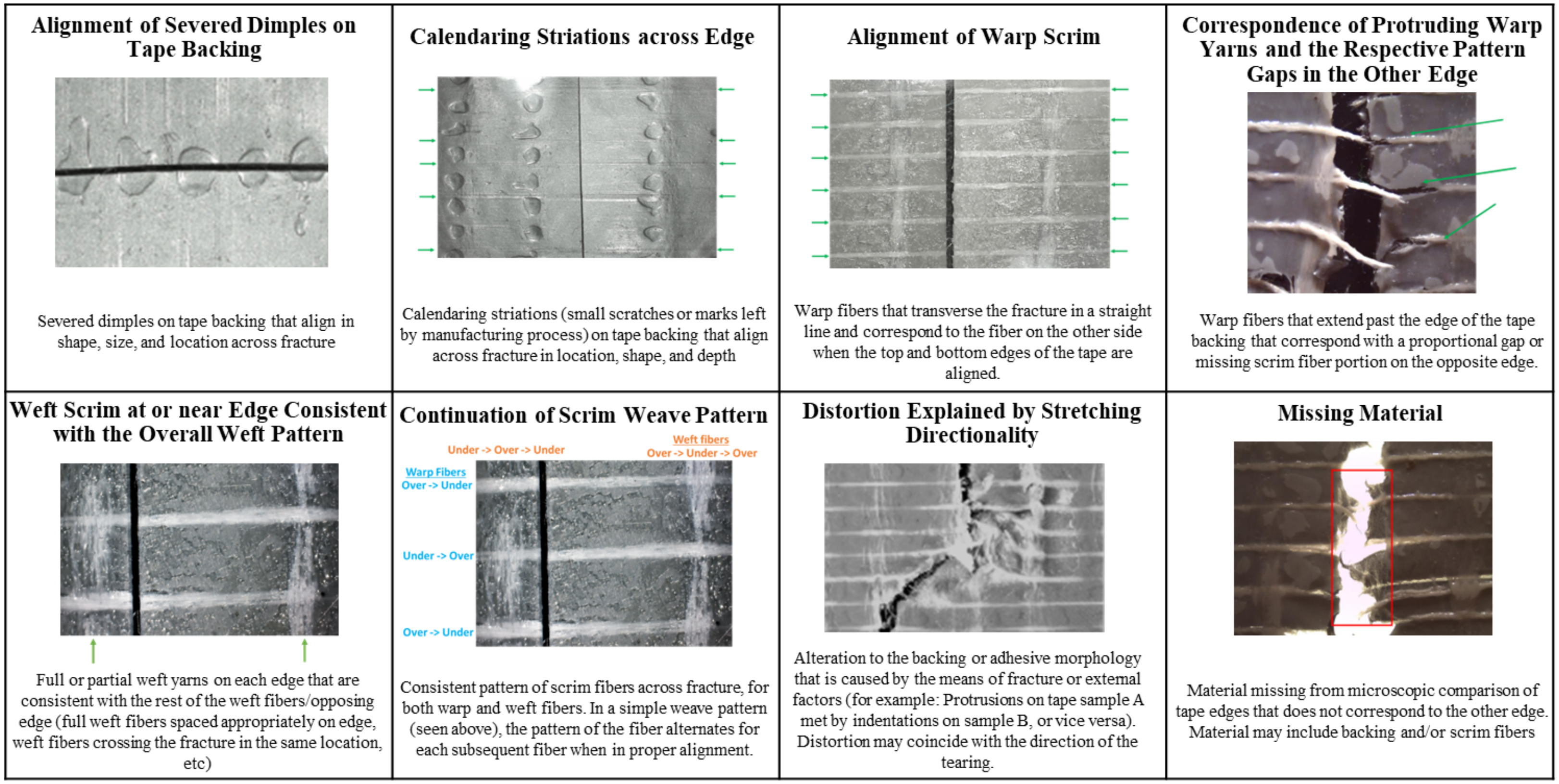


Figure 5. *Examples of true fit pairs from the tape sets. The images show the distorted morphology observed in the MQ-HT-S, LQ-HT, and LQ-HT-S edges. Despite also being hand-torn, the edges observed in the HQ-HT set are very straight and less distinctive, even when stretched.*[4]



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

Fig 6. *Examples of the eight descriptive features documented for duct tape edge comparisons. These features are some of the most observed on duct tapes regardless of grade. Oftentimes, they help establish and document standardized criteria during sample comparison. Alignment of severed dimples: these are severed dimples on tape backings that align from one edge to the other in shape, size and location across the fracture. This feature is only applicable on the backing side. When it is present and aligns across the separated edges, it can provide strong support to a fit decision because these manufactured-imprinted marks have some inherent variability across a single roll and when split through the fracture, those patterns are very unlikely to align by chance. Likewise, when there is major misalignment of the severed dimples, the feature provides strong support to the non-fit decision.* [4,7]

Textiles

In the case of textiles, yarns are used to create fabric by either weaving or knitting. Yarns are then interlaced by various constructions, depending on the desired properties of the end-product. Modern weave and knit machines provide very consistent yarn constructions. After the fracturing process, samples are compared visually, including examining the general size and shape, weave/knit type, fiber type, and twist.

In the proposed method, examiners first analyze a comparison pair's overall edge morphology and general fracture alignment using a stereomicroscope at 10-40x magnification with reflected and transmitted light. Here, features such as weave alignment and pattern/print alignment are observed. Weave alignment occurs when the direction of the weave is consistent between both samples being compared. Pattern or design alignment occurs when designs printed on the fabric, such as stripes or flowers, align across the fractured edge. Both features generally increase an examiner's confidence in the presence of a physical fit between two samples. The examiner may also observe distortion in the form of curling or stretching of the fracture edge, which may limit their ability to determine the presence of a physical fit. During the development and optimization stage for textiles, we defined the main relevant features and terminology to standardize the observations, reporting, and criteria used during the decision-making process. (See **Figure 7**).

Following this, the examiner subdivides the length of the fracture edge into ten (10) comparison bins or areas of equal size. The examiner conducts an independent comparison within each respective bin and identifies it as a fit, non-fit, or inconclusive, assigning a quantitative value of 1, 0, or 0.5, respectively for the bin. The examiner also uses UV light to identify any fluorescence exhibited by the fibers on either sample. The presence or absence of fluorescence on both samples being compared may increase an examiner's confidence in the presence of a physical fit between the two samples.

As described in the duct tape section above, the examiner also documents the overall morphological features and individual features observed in each comparison bin in a digital template throughout this process. This digital template follows similar approaches as the ones described for tapes, but it is customized for evaluating the textile's features. The examiner documents the qualitative features per bin location, allowing for a straightforward peer review process, as previously discussed. Finally, the template macro will automatically generate an Edge Similarity Score (ESS). An example of the template is provided in **Figure 8**.

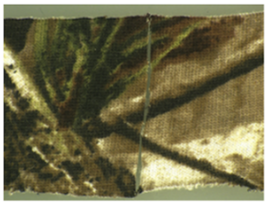
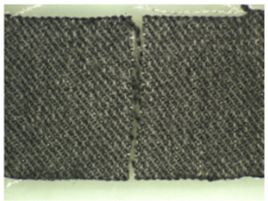
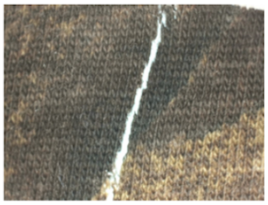
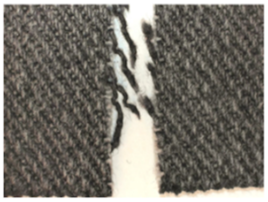
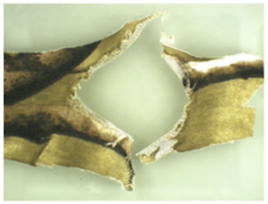
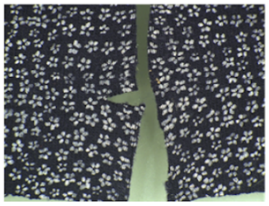
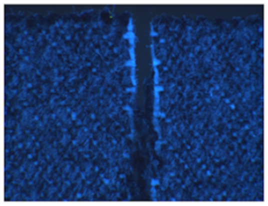
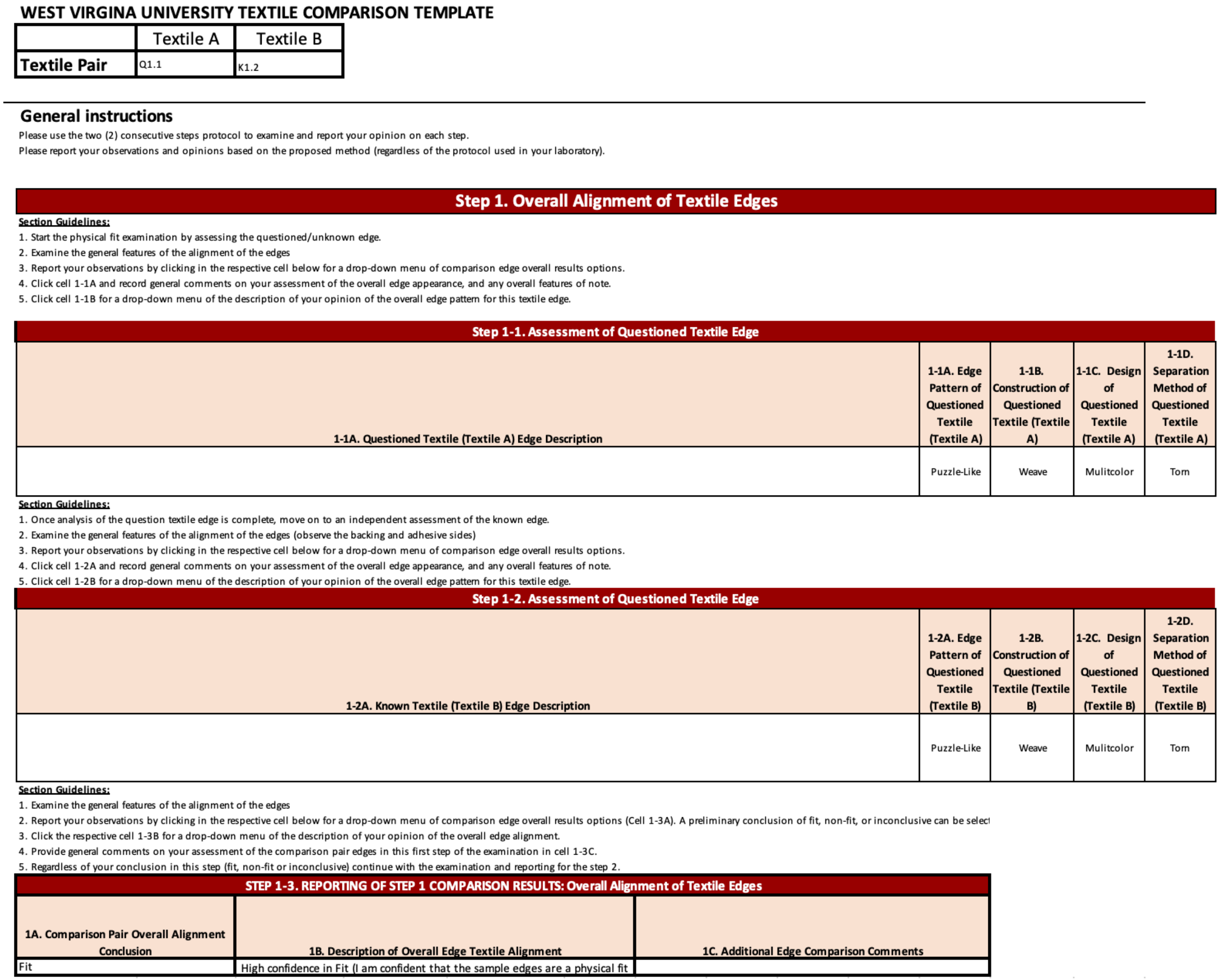
<table>
<tr><th>Design Alignment</th><th colspan="2">Construction Alignment</th><th>Edge Alignment</th></tr>
<tr><td>Consistency and alignment of yarn color and pattern between two textile fragments</td><td colspan="2">Consistency and alignment of construction, including type (weave/knit) and direction, between two textile fragments. Consistency in the thread or yarn count between the two fragments is also considered</td><td>Overall edge shape alignment. Identified edge shapes include straight, wavy, and puzzle-like</td></tr>
<tr><td></td><td colspan="2"></td><td></td></tr>
<tr><th>Yarn Alignment</th><th>Extreme Distortion</th><th>Extreme Distortion</th><th>Fluorescence</th></tr>
<tr><td>Alignment of yarns that have been pulled out of the fracture edge between two textile fragments</td><td>Force applied during the fracture event causes distortion that can mask other features</td><td>A secondary, perpendicular tear that is not the primary fracture that is being compared</td><td>Fluorescence of individual yarns can aid in the identification of a physical fit</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>

Figure 7. *Prominent textile-relevant features and terminology defined in this study to assist analysts in making determinations of physical fits.*[8]



WEST VIRGINIA UNIVERSITY TEXTILE COMPARISON TEMPLATE

	Textile A	Textile B
Textile Pair	Q1.1	K1.2

General instructions

Please use the two (2) 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).

Step 1. Overall Alignment of Textile 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
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 textile edge.

Step 1-1. Assessment of Questioned Textile Edge

1-1A. Questioned Textile (Textile A) Edge Description	1-1A. Edge Pattern of Questioned Textile (Textile A)	1-1B. Construction of Questioned Textile (Textile A)	1-1C. Design of Questioned Textile (Textile A)	1-1D. Separation Method of Questioned Textile (Textile A)
	Puzzle-Like	Weave	Mulitcolor	Torn

Section Guidelines:

1. Once analysis of the question textile 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 textile edge.

Step 1-2. Assessment of Questioned Textile Edge

1-2A. Known Textile (Textile B) Edge Description	1-2A. Edge Pattern of Questioned Textile (Textile B)	1-2B. Construction of Questioned Textile (Textile B)	1-2C. Design of Questioned Textile (Textile B)	1-2D. Separation Method of Questioned Textile (Textile B)
	Puzzle-Like	Weave	Mulitcolor	Torn

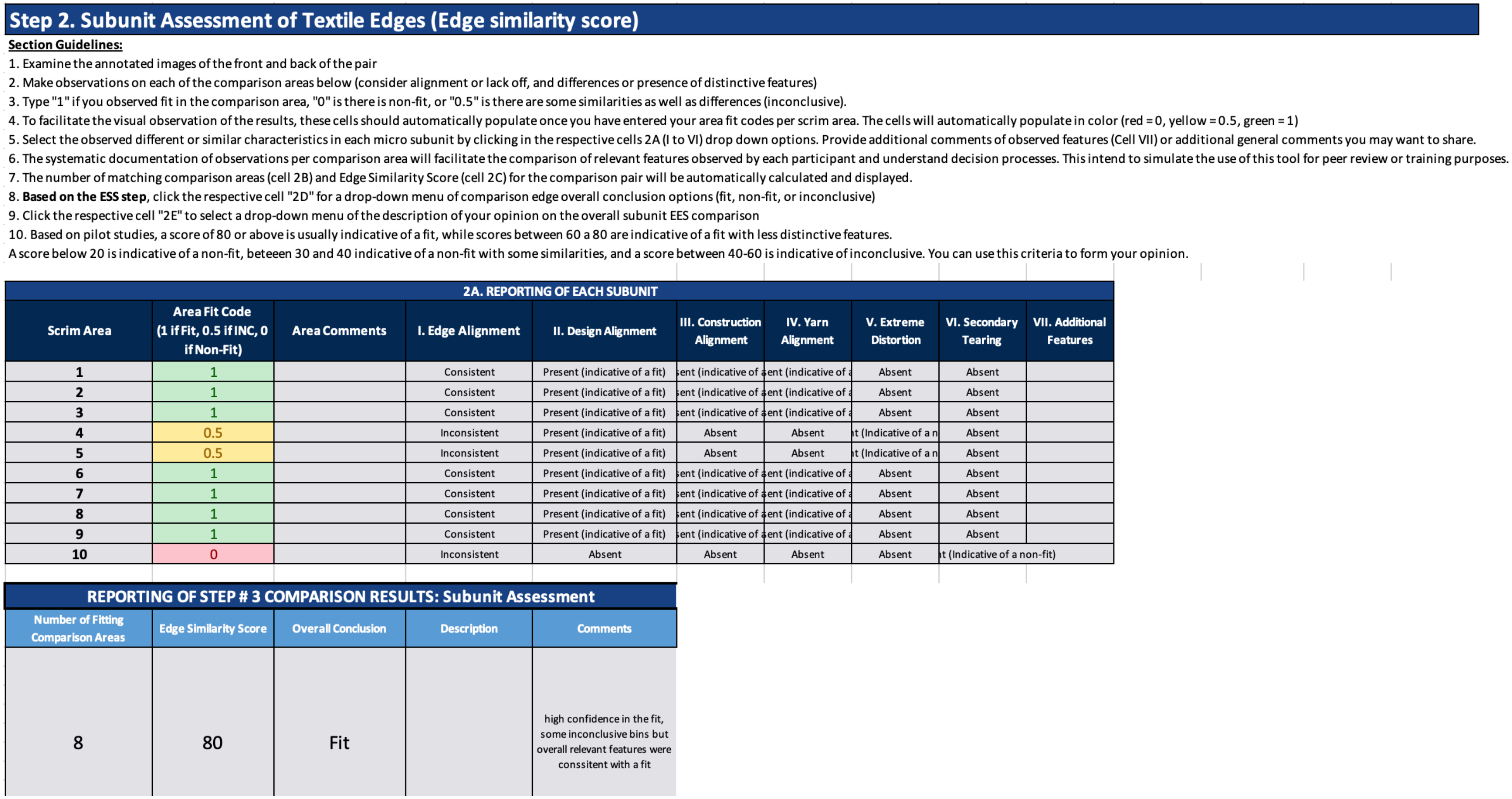
Section Guidelines:

1. Examine the general features of the alignment of the edges
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 select
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 Textile Edges

1A. Comparison Pair Overall Alignment Conclusion	1B. Description of Overall Edge Textile Alignment	1C. Additional Edge Comparison Comments
Fit	High confidence in Fit (I am confident that the sample edges are a physical fit	

Fig. 8-1. Example of filled-out cells for step 1 for textile comparisons; this step covers the macroscopic assessment of the compared pairs.[8]



Step 2. Subunit Assessment of Textile Edges (Edge similarity score)

Section Guidelines:

1. Examine the annotated images of the front and back of the pair
2. Make observations on each of the comparison areas below (consider alignment or lack off, and differences or presence of distinctive features)
3. Type "1" if you observed fit in the comparison area, "0" is there is non-fit, or "0.5" is there are some similarities as well as differences (inconclusive).
4. 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)
5. Select the observed different or similar characteristics in each micro subunit by clicking in the respective cells 2A (I to VI) drop down options. Provide additional comments of observed features (Cell VII) or additional general comments you may want to share.
6. The systematic documentation of observations per comparison 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.
7. The number of matching comparison areas (cell 2B) and Edge Similarity Score (cell 2C) for the comparison pair will be automatically calculated and displayed.
8. **Based on the ESS step**, click the respective cell "2D" for a drop-down menu of comparison edge overall conclusion options (fit, non-fit, or inconclusive)
9. Click the respective cell "2E" to select a drop-down menu of the description of your opinion on the overall subunit EES comparison
10. 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 20 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.

2A. REPORTING OF EACH SUBUNIT

Scrim Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)	Area Comments	I. Edge Alignment	II. Design Alignment	III. Construction Alignment	IV. Yarn Alignment	V. Extreme Distortion	VI. Secondary Tearing	VII. Additional Features
1	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
2	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
3	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
4	0.5		Inconsistent	Present (indicative of a fit)	Absent	Absent	t (Indicative of a n	Absent	
5	0.5		Inconsistent	Present (indicative of a fit)	Absent	Absent	t (Indicative of a n	Absent	
6	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
7	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
8	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
9	1		Consistent	Present (indicative of a fit)	ent (indicative of a	ent (indicative of a	Absent	Absent	
10	0		Inconsistent	Absent	Absent	Absent	Absent	t (Indicative of a non-fit)	

REPORTING OF STEP # 3 COMPARISON RESULTS: Subunit Assessment

Number of Fitting Comparison Areas	Edge Similarity Score	Overall Conclusion	Description	Comments
8	80	Fit		high confidence in the fit, some inconclusive bins but overall relevant features were consistent with a fit

Fig. 8-2. Example of filled-out cells for step 2 for textile examinations; this step covers the microscopic assessment of the compared pairs. The edges are visually separated into ten sections, and seven 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. At the bottom, the number of fitting areas and the edge similarity score is displayed along with the final opinion of the examiner. [8]

Plastics

Plastics have a significantly different composition and construction than duct tape and fibers. Unlike the soft polymers used in duct tape, the composition of rigid plastics is designed to keep the material firm. Therefore, hard plastics are brittle, breaking by tension when enough force is applied. Automotive hard plastic objects commonly submitted to crime laboratories include the vehicle's headlights, taillights, and fragments from the bumpers found in larger-sized fragments at scenes such as hit-and-runs and car accidents.

To select the samples used in this study, we investigated first some of their mechanical properties and chemical compositions. The main polymers used in manufacturing these parts are polypropylene, polyurethane, and polycarbonate, but they can also include materials such as nylon or polyvinyl chloride. These polymers are common because they are durable and resistant to many environments or substances that would otherwise damage or weaken the polymer material. Automotive plastic objects can sometimes be marked with a serial number that can be used to establish the composition of the polymer material as reported by the manufacturer.

The intended purpose of the material influences macroscopic features on the surface of the hard plastics. Polymers from taillights and headlights are generally transparent, may have some degree of curvature or patterning, and contain striations and markings across the surface. Plastics from bumpers may have several layers of paint on top of the polymer core, as well as striations or marks left from wear and tear on the bumpers. 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 main fractured edge, overlapping material, along with distinctive fracture patterns and directionality, can inform the examiner's opinion. Prior to fracturing, each intact polymer sample was analyzed using an ATR-FTIR spectrometer to identify the primary chemical makeup of the polymer. A summary of these findings can be seen below in **Table 1,** which also includes the color and opacity of each polymer analyzed.

Table 1. *Description of polymer color, opacity, and chemical composition of the hard plastics collection set.*

Item	Color	Opacity	Chemical Compounds
H1.2	Clear	Translucent	Polycarbonate
H1.5	Clear	Translucent	Polycarbonate
H2.14	Red	Translucent	Polycarbonate, solvent red 111
H3.1	Silver/Black	Opaque	Polypropylene Terephthalate (PPT)
H3.2	Clear	Translucent	Polycarbonate
H3.7	Orange	Translucent	PMMA, solvent orange 60
H3.9	Clear	Translucent	Polycarbonate

Item	Color	Opacity	Chemical Compounds
H4.6	Silver/Black	Opaque	PolybutyleneTerephthalate (PBT)
H4.19	Black	Opaque	Polycarbonate, PMMA Acrylic
H4.25	Clear	Translucent	Polycarbonate
H4.36	Clear	Translucent	Polycarbonate
H4.37	Clear	Translucent	Polycarbonate
H4.40	Silver/Black	Opaque	Polypropylene Terephthalate (PPT)
H5.4	Orange	Translucent	Polycarbonate, orange 60
H5.6	Silver/Black	Opaque	Polypropylene
T1.1	Clear	Translucent	Polybutylene Terephthalate (PBT)
T3.2	Red	Translucent	PMMA Acrylic, solvent red 111
T3.3	Clear	Translucent	Polycarbonate

Images were captured of each intact polymer prior to fracturing using a DSLR camera. After imaging, the samples were taped with blue painter's tape and fractured. Following this, each polymer was reassembled, using the original photographs as a guide, by individuals who would not be conducting any of the comparisons and then renamed with a unique identifier using random number generator. After the fragments were relabeled, they were packaged separately in labeled manila envelopes.

Like tapes and textiles, a systematic method was developed and evaluated for the analysis and documentation of automotive plastics (see **Table 2** and **Figure 9**). First, the examiner provides a brief description of each sample and documents the edge shape, color, and pattern (straight, curvy, puzzle-like, or serrated). They then align the two fragments and offer a preliminary conclusion based on macroscopic observations. Following this, regardless of their preliminary conclusion, the examiner proceeds to the microscopic examination of the samples. The comparison edge of the polymer fragments was divided into five bins of equal length. Because the edges of fractured polymers are not often straight lines, strands of dental floss were aligned along the fracture edge and cut to size. Then, the dental floss was measured using a scale and divided into 5 equal lengths, which are marked on the floss. The floss is aligned to the fracture edge during physical fit comparison to determine where each bin starts and ends. During the comparison, examiners will look for ten features in the polymer, 3D edge alignment, surface plane/directionality alignment, edge curvature/directionality, pattern alignment, surface damage alignment, scratch alignment, fracture marks alignment, protruding feature, missing material, and extraneous material. Images and descriptions of each feature can be seen in **Table 2** below.

Table 2-1. *Polymer features observed during physical fit comparison.*

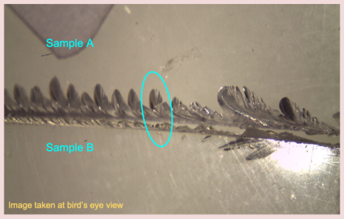
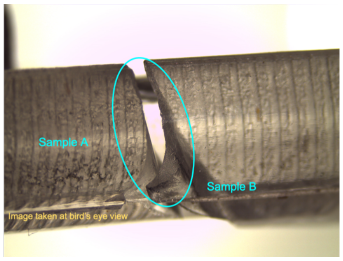
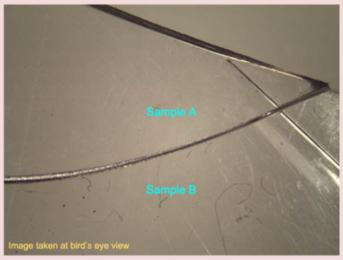
Feature	Description	Image
3D Edge Alignment	The interior edge of the fragment (along the fracture) corresponds. This may be supported by protrusions and corresponding indentations, fracture marks, and/or shifts in fracture direction that should be consistent across both fragments	
Surface Plane/Directionality Alignment	The 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	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	

Table 2-2. *Polymer features observed during physical fit comparison.*

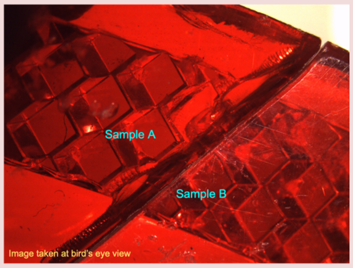
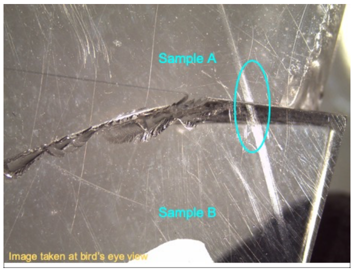
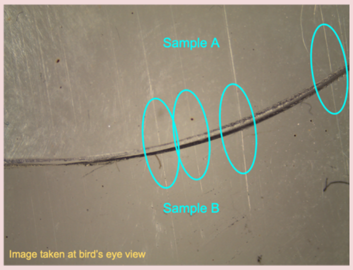
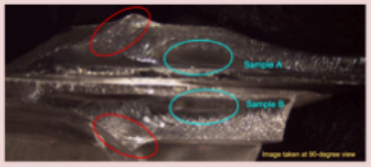
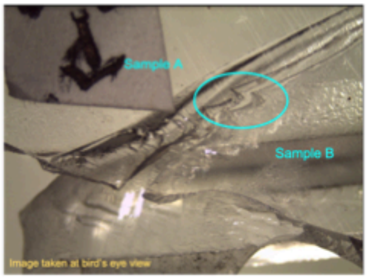
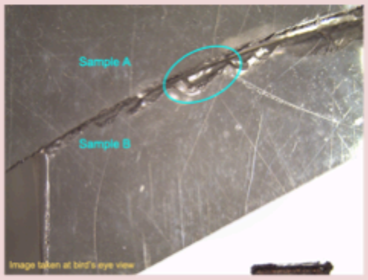
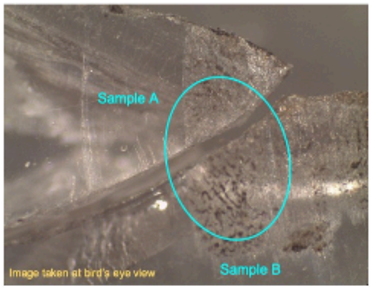
Feature	Description	Image
Pattern Alignment	A consistent pattern or texture that maintains shape, location, size, plane, and direction across the fracture.	
Surface Damage Alignment	A deep scratch, mark, or indent along a surface edge of the polymer that corresponds in depth, width, direction, and location across the fracture.	
Scratch Alignment	Light scratches along the 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 impact and breaking of the polymer, or created during handling of the broken evidence.	

Table 2-3. Polymer features observed during physical fit comparison.

Feature	Description	Image
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 both surfaces, as well as the 3D interior edge.	
Missing Material Misalignment	Polymer material that is absent from the fracture edges in a manner that **does not** correspond to the other edge.	
Extraneous Material Alignment	Corresponding external material (paint, tape, residue, etc.) that aligns in the exact same locations across the fracture.	

The features observed and the conclusions for each pair by the examiner are documented on a custom-made template as an Excel sheet. In addition to the ESS metric, the examiner includes the effect of each feature on their decision-making process, by assigning each feature a prominence value (FPV). The prominence values for each feature in each comparison area are summed together to generate an overall Feature Prominence Sum (FPS). In general, positive FP number indicate the presence of a physical fit or an inconclusive decision, whereas negative FPVs indicate non-fits or inconclusive decisions. The examiner classifies each of the five individual comparison areas as fit, non-fit, or inconclusive. They then report the pair overall conclusion using the qualitative and quantitative criteria.

The main analysts in this task are students who have followed at least 2-month material specific introductory training (i.e., duct tape, textiles, and/or plastics) established by the PI that includes:

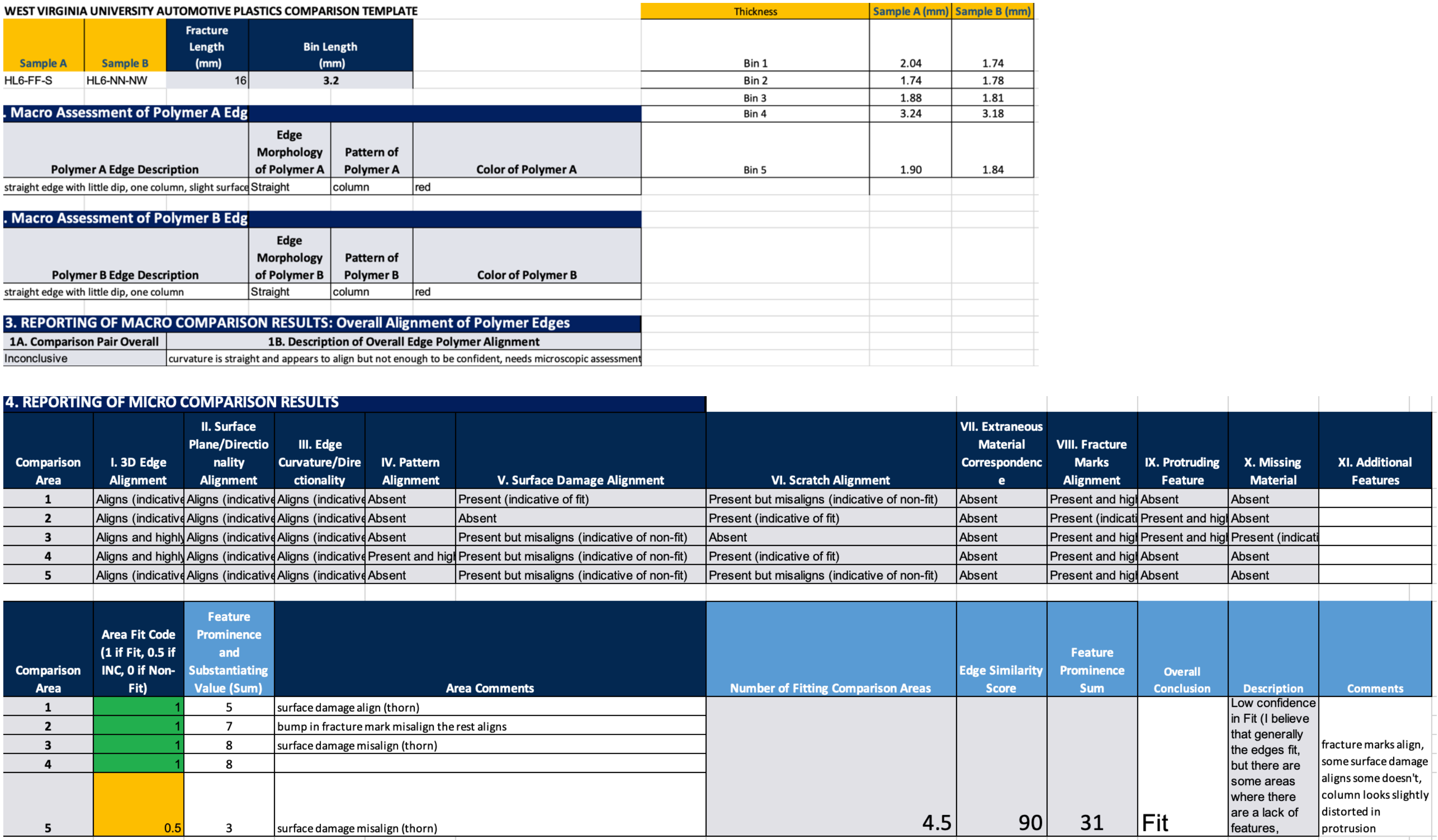
a) reading and discussion of relevant literature,

b) testing on three modules in which the students will build knowledge of each material of interest (composition, manufacture, distribution, common cases, typical examination protocol, fracturing mechanisms, data analysis, and interpretation),

c) hands-on training sets for the examination of fracture fits, for which the "ground" truth is known but maintained blind to the student,

d) blind hands-on training tests in which the student's ESS or FPS scores are compared to those reported by a consensus panel and the overall conclusions are evaluated based on the ground truth (i.e., known fit, or known-non fit).

Our training quality control has set the accordance and concordance of +/- 10%. Accordance is associated with the probability that two identical samples tested by the same individual under the same conditions but at different times will be assigned the same similarity score and associating features. Concordance is related to the probability that two identical samples tested by separate individuals under the same conditions will be given the same similarity score. The analysis of accordance and concordance is focused on descriptive statistics.



WEST VIRGINIA UNIVERSITY AUTOMOTIVE PLASTICS COMPARISON TEMPLATE

Sample A	Sample B	Fracture Length (mm)	Bin Length (mm)
HL6-FF-S	HL6-NN-NW	16	3.2

Thickness	Sample A (mm)	Sample B (mm)
Bin 1	2.04	1.74
Bin 2	1.74	1.78
Bin 3	1.88	1.81
Bin 4	3.24	3.18
Bin 5	1.90	1.84

. Macro Assessment of Polymer A Edg

Polymer A Edge Description	Edge Morphology of Polymer A	Pattern of Polymer A	Color of Polymer A
straight edge with little dip, one column, slight surface	Straight	column	red

. Macro Assessment of Polymer B Edg

Polymer B Edge Description	Edge Morphology of Polymer B	Pattern of Polymer B	Color of Polymer B
straight edge with little dip, one column	Straight	column	red

3. REPORTING OF MACRO COMPARISON RESULTS: Overall Alignment of Polymer Edges

1A. Comparison Pair Overall	1B. Description of Overall Edge Polymer Alignment
Inconclusive	curvature is straight and appears to align but not enough to be confident, needs microscopic assessment

4. REPORTING OF MICRO COMPARISON RESULTS

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	Aligns (indicative	Aligns (indicative	Absent	Present (indicative of fit)	Present but misaligns (indicative of non-fit)	Absent	Present and higl	Absent	Absent	
2	Aligns (indicative	Aligns (indicative	Aligns (indicative	Absent	Absent	Present (indicative of fit)	Absent	Present (indicati	Present and higl	Absent	
3	Aligns and highly	Aligns (indicative	Aligns (indicative	Absent	Present but misaligns (indicative of non-fit)	Absent	Absent	Present and higl	Present and higl	Present (indicati	
4	Aligns and highly	Aligns (indicative	Aligns (indicative	Present and higl	Present but misaligns (indicative of non-fit)	Present (indicative of fit)	Absent	Present and higl	Absent	Absent	
5	Aligns (indicative	Aligns (indicative	Aligns (indicative	Absent	Present but misaligns (indicative of non-fit)	Present but misaligns (indicative of non-fit)	Absent	Present and higl	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	1	5	surface damage align (thorn)	4.5	90	31	Fit	Low confidence in Fit (I believe that generally the edges fit, but there are some areas where there are a lack of features,	fracture marks align, some surface damage aligns some doesn't, column looks slightly distorted in protrusion
2	1	7	bump in fracture mark misalign the rest aligns						
3	1	8	surface damage misalign (thorn)						
4	1	8							
5	0.5	3	surface damage misalign (thorn)						

Fig. 9. Sections 1-3: Example of filled-out cells for step 1 for plastic polymer comparisons; this step covers the macroscopic overall assessment of the compared pairs. Section 4: Example of filled-out cells for step 2 for plastics examinations; this step covers the microscopic assessment of the compared pairs. The edges are visually separated into five sections, and ten major features are documented for their absence or presence, and the weight these features hold in the bind decision, which is automatically estimated into a numeric feature prominence value then, the template automatically estimates the quality metrics of edge similarity score and feature prominence sum.

Task 2 (Objective 1) — Development and validation of automated computational algorithms for the comparison of fracture fits.

The primary aim of this task is to supplement the ESS method by introducing a computational comparison model for fractured edges to provide additional objective support for a physical fit examination. Robust and practical computational algorithms are developed for fracture fit analysis for three main purposes: 1) establish a platform to create a database of fractured edges, 2) predict if a compared pair of samples present a physical fit or not using image-recognition algorithms, and 2) gain further understanding of the importance of evaluated features on the decision of a fit or non-fit.[17-20]

To achieve this, we utilize a machine-learning model to process images of the fractured items.[21-31] We first developed an open-source Python package designed for image analysis tasks, including edge detection, background noise-reduction, and image filtering for materials of interest in the field of forensics science.[32] Additionally, the package contains a database handler to manage the flow of data to and from machine-learning models. We then construct a convolutional neural network (CNN) model that classifies the tape images into fits and non-fits, providing a fit membership probability output. We apply the CNN approach to the scrim and the backing of the scan images separately and combine the results using a decision tree classifier.

There are different situations in forensic science where matching is performed, and the dimensionality of the objects control the method that can be used. In this project, we focus only on cases where images are captured from the objects, and edge analysis is performed to define the contour image. The approach evaluates the performance of machine learning algorithms based on neural networks (fully connected, FNN and convolutional, CNN) to predict if images of two pieces of evidence coincide. The matched objects are then classified and used as a discrete analysis of the matching process (singular prediction) or a more continuous statistical analysis, after a large number of those are performed, where trends on the information within the images can be extracted.

ForensicFit: Fractured edges database

Image database set preparation

In this project, we center our investigation on duct tapes, textiles, and vehicle plastics and respective images from the fractured edges. A more detailed description of the samples is provided in the following task 3 section. The database has been created by producing images of high resolution using an EPSON 12000XL scanner with images scanned using SilverFast 8, version 8.8.0r14, an interface at a resolution of a minimum 600 dots-per-inch. Hard plastics, on the other hand, present several challenges to capturing the edge features due to the presence of various features in three dimensions, diverse angles and planes of surfaces of the materials. Also, the refractive and the reflective nature of some of the polymers produce some artifacts and difficulties in focusing at different depths of field. Regardless of efforts made with enhanced photography and microscopic imaging devices, the images did not meet optimal details to feed the CNN. In this dataset, images of 2D planes are stored to demonstrate some of the features. However, more advanced 3D imaging technology or 3D molds are recommended in the future to build image databases of hard polymers. Instead, we utilize information from the human-based examination and documentation to utilize computational and mathematical models to further the information derived from plastic fit examinations.

The dataset of tape images includes tapes generated from three different qualities (low, medium, and high grade) and two separation methods (hand-torn or scissor-cut). As a result, there are six total subsets of tape samples. The database consists of images scanned from 900, 200, and 898 low-, medium-, and high-quality tape samples, respectively, for a total of 1998 individual tapes and 3996 images from the backing and adhesive/scrim sides. To improve contrast and consistency across images, the samples were placed on top of black cardboard. Each tape was scanned twice, once to capture the top surface of the tape (backing layer) and the second to capture the underside (adhesive/scrim layer).

The textiles image dataset consists of 793 textiles samples of various constructions (knit, woven, or mixed), modes of separation (stabbed or hand-torn), compositions (cotton, polyester, rayon, or mixed), and color design (unicolor or multicolor). Each fractured textile was scanned with a white or black cardboard backing to enhance contrast, depending on the color if the fabric.

Minor corrections to the images were made during scanning to enhance the contrast and visibility of the edges and features, such as setting the black point of the image to the posterboard to ensure the background was the darkest part of the image (or white for textiles of dark color). Additional corrections are performed using Adobe Photoshop on some images to address specific issues to remove artifacts. Each tape image is stored in a 2-dimensional matrix where each element represents a pixel intensity value between 0 and 255, corresponding to black and white, respectively.

After the images are taken, data preprocessing consists of many steps such as data cleaning, transformation, feature extraction, and reformatting. Before setting up the architecture of the network the image dimensions are reduced by; 1) using the smallest image resolution where the edge surface details are still visible; 2) focusing only on the important part of the image— the comparison edge. Here, a python package (ForensicFit [32]) has been developed to bridge the gap from raw images to data suitable for a machine-learning model. ForensicFit was developed to analyze images collected from materials of interest in forensic science. Additionally, it can receive different image formats and store them efficiently on a general and flexible database. This database is accessible from other parts of the code for image processing, statistical analysis, and training a machine-learning model. The essentials of the package are explained in Supplementary Information. The source code is hosted on GitHub. [32] For this study, ForensicFit provides the means to automatically crop the image to only include the comparison edge of the tape.

The dots-per-inch (dpi) resolution was set to the minimum scanned dpi value (600 dpi). A window of 410×2400 px^2 ($pixels^2$) was selected around the comparison edge. The x-dimension (length of the tape) was achieved with relative cropping from the comparison edge (see Supplementary Information for more details). For the y-dimension (width of the tape), because tapes originating from different rolls may have different widths, they do not have the same size in the y-dimension. The width of the tapes used in this study range from 2200 to 2600 *px*. Because the CNN requires consistent inputs, the images were cropped on the borders of the tape and resized to 2400 *px*. Resizing can cause small alteration of the image; it is important to note that this type of alteration is different from the physical distortion due to the stretching of the tape. Physical stretching follows shearing and straining constraints that can cause the tape's edge to warp in a wavy pattern, whereas the resized scanned image remains unchanged in its overall appearance. Nevertheless, because all tapes undergo the same distortion, it does not influence the outcome.

The output comparison edge image was then further resized to be as small as possible and still retain the fine details of the tape. This resizing was done for computational efficiency and to accommodate GPU memory limitations. In this case, the edge images were reduced by half, leading to an edge image with a size of 205×1200 px^2 and a resolution of 300 dpi. **Figure 10** shows an example of the output of the reduction and the concatenated input resulting in two images (scrim and backing) of size 410×1200 px^2 ready to be passed on to the CNN.

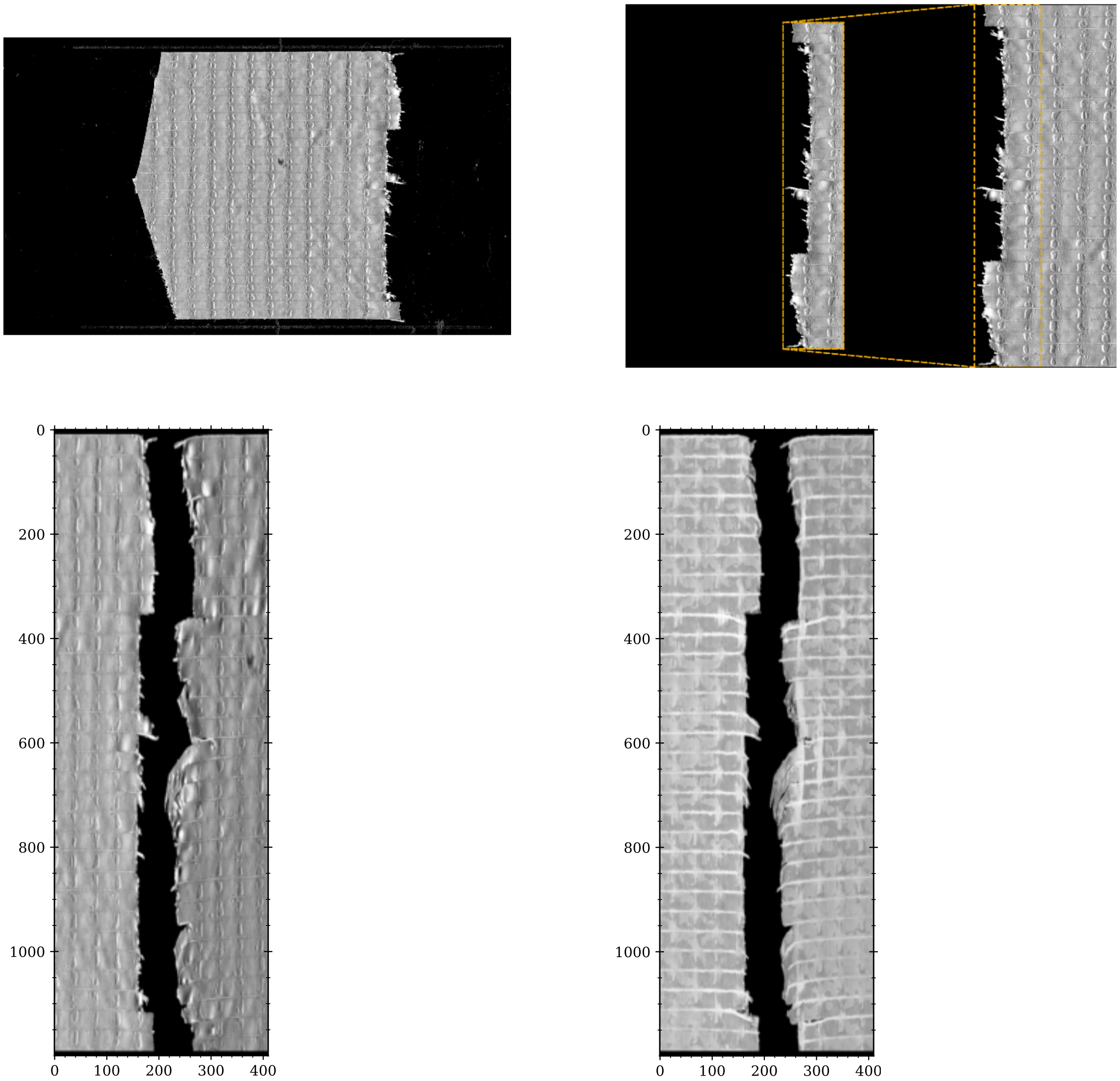


***Figure 10**. Top Left: Scanned image of a low-quality grade tape. Image shows the backing side of the tape. One of the edges has been cut into an arrow shape, representing a non-comparison edge. For this publication this image was manually cropped. Top Right: Preprocessing of tape image by ForensicFit. The image is automatically split in the middle of the tape, its background cleaned, rotated to be horizontal, and cropped to its boundaries in the y direction by ForensicFit. The dashed golden box shows area selected from the tape that is passed on as the input for the convolutional neural. network. Bottom: examples of the convolutional neural network image inputs. Left: Concatenated image of two tape edges on the backing side. Right: Concatenated image of two tape edges on the scrim side.*

ForensicFit database

ForensicFit is a well-controlled and efficient database where the user can store, query, analyze, and use the data created for a particular application. ForensicFit uses state-of-the-art image processing methods to analyze and store the generated data. The data is compatible with well-known machine-learning packages such as TensorFlow[21], PyTorch, and SciKit-learn[22]. It utilizes NumPy[23], SciPy[24], matplotlib[25], OpenCV[26-27], scikit-image[28], PyMongo, and GridFS. Also, the package follows PEP-257[29] and PEP-484[30] for documentation and type hints, respectively.

The package is organized into three main sub-packages, *core*, *database*, and *utils*. A brief description is provided here, but more detailed information is provided in the Supplementary section and within the package instructions as well. The subpackage *core,* as the name suggests, contains the most important functionalities within the package. It contains python classes that manage the read/write, analysis, and metadata storage. These classes provide a skeleton for the data structure used in the package. Moreover, they define the standards for future implementations used for different types of materials. The *database*, provides an efficient and flexible method for storing and retrieving the raw and preprocessed data. The functionality of the rest of the package does not depend on this sub-package. It was added merely to simplify the storage and query process of the data. One can still store and access the raw or analyzed data using the traditional image storage approaches. Lastly, *utils*, contains all the image manipulation, plotting, and memory access tools that are used in different sections of the package. In addition to package documentation, three stand-alone python scripts accompany the package for batch processes, *create_metadata.py*, *preprocess_bin_based.py*, *store_on_db.py*.

Convolutional neural network configuration

This model uses a convolutional neural network (CNN) followed by a fully connected neural network as implemented in TensorFlow [21] to train on the prepared images. The CNNs contain a series of convolutional layers followed by a fully connected network. The convolutional layers carry out the tasks of pattern recognition (feature extraction) and dimensionality reduction, while the fully connected layers make the decision on whether the items' pairs are a fit or non-fit.

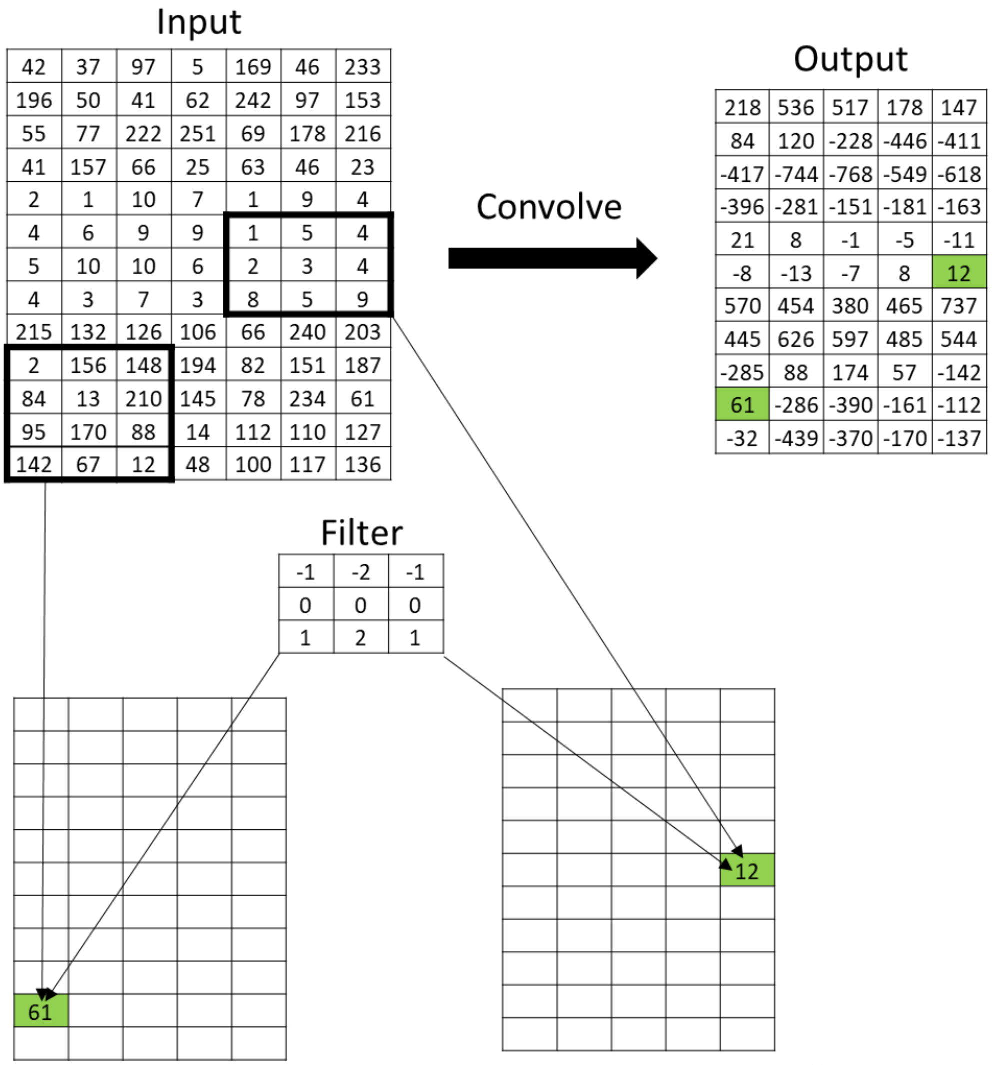
The network was built from a series of convolution layers, where filters with small kernel of 3×3 px^2 window (smallest size capable of capturing the notion of left/right, up/down, and center and strides of 1×1 was used. The convolutional layers used Rectified Linear Unit (ReLU) (31) activation functions and were followed by 2×2 pixel window Max-pooling layers to handle the dimension reductions.

The CNN architecture was inspired by the popular VGG-16 [33], which with a simple architecture achieves remarkable results. The number of convolution layers was selected by considering the size of the reduced dimensions of the image and available GPU memory for training. At the end of the convolutional layer, the image is flattened to a 1-dimensional vector of size 136,192 elements. Compared to the raw flattened input (1200×405=492,000), this significantly reduces the number of parameters the network needs to learn. Finally, three fully connected dense layers of size 500, 100, and 1 are added. The 500,100 layers use the ReLU activation function [34], whereas the final layer has a sigmoid activation function to map results between 0 and 1 used in a binary classification. A 0.5 weighted dropout layers is used to fight the overfitting [35]. **Figure 11** and **Table 3** show an overview of the architecture of the CNN.

The dataset was divided into training and validation with a ratio of 80:20. A five-fold cross-validation scheme was used to maximize the model's familiarity with the data without risking overfitting. Model hyperparameters dictate how the learning is performed. These hyperparameters determine the learning process and must be carefully tuned to ensure a robust convolutional neural network. The batch size, which is the number of images loaded into the memory and processed simultaneously, was set to 5. This choice considered the size of the images, network's dimensions, and the available GPU memory. The substantial size of both the network and the images justified the use of smaller batch sizes.

Table 3. Convolutional neural network architecture. The network consists of a series of consecutive convolutional filters followed by a fully connected neural network.[12]

Network type	Layer name	Activation function	Kernel/Pool size	Strides	Number of filters/units	Tensor shape
Convolutional	Input	-	-	-	-	1200×410×1
	Convolution	ReLU	3×3	1×1	32	1200×410×32
	Max-pooling	-	2×2	1×1	-	600×205×32
	Convolution	ReLU	3×3	1×1	64	600×205×64
	Max-pooling	-	2×2	1×1	-	300×103×64
	Convolution	ReLU	3×3	1×1	128	300×103×128
	Max-pooling	-	2×2	1×1	-	150×52×128
	Convolution	ReLU	3×3	1×1	256	150×52×256
	Max-pooling	-	2×2	1×1	-	75×26×256
	Convolution	ReLU	3×3	1×1	512	75×26×512
	Max-pooling	-	2×2	1×1	-	38×13×512
	Convolution	ReLU	3×3	1×1	1024	38×13×1024
	Max-pooling	-	2×2	1×1	-	19×7×1024
Fully connected	Flatten	-	-	-	-	136192
	Dropout	-	-	-	-	136192
	Dense	ReLU	-	-	500	500
	Dropout	-	-	-	-	500
	Dense	ReLU	-	-	100	100
	Dense	Sigmoid	-	-	1	1



Input

42	37	97	5	169	46	233
196	50	41	62	242	97	153
55	77	222	251	69	178	216
41	157	66	25	63	46	23
2	1	10	7	1	9	4
4	6	9	9	1	5	4
5	10	10	6	2	3	4
4	3	7	3	8	5	9
215	132	126	106	66	240	203
2	156	148	194	82	151	187
84	13	210	145	78	234	61
95	170	88	14	112	110	127
142	67	12	48	100	117	136

Output

218	536	517	178	147
84	120	-228	-446	-411
-417	-744	-768	-549	-618
-396	-281	-151	-181	-163
21	8	-1	-5	-11
-8	-13	-7	8	12
570	454	380	465	737
445	626	597	485	544
-285	88	174	57	-142
61	-286	-390	-161	-112
-32	-439	-370	-170	-137

Filter

-1	-2	-1
0	0	0
1	2	1

Figure 11. *An example of convolution applied with a 3×3 filter and a stride of 1×1.* [12]

The loss function, which measures the model's accuracy in predicting the training data, was selected as binary cross-entropy. The optimizer, responsible for guiding the model towards minimizing the loss function, was set to the Adaptive Moment Estimation (Adam) algorithm.[36] The learning rate, which defines the optimization step-size during the model training, was set to an initial value 10^{-4} and gradually decreased to 10^{-5} over 25 training epochs using a second-degree polynomial function. The learning rate and the number of epochs were determined through trial and error. It was observed that using a constant learning rate resulted in a highly variable validation loss, which may be attributed to oscillation around the problem's global minimum.

Finally, in the case of tapes, the combination of scrim and backing CNNs is necessary to capture the most features on each side. For this, two identical CNN models were independently trained on the scrim and backing sides of the image tapes, resulting in two separate predictions for each of tape pairs. To combine the outcomes from both CNNs, a range of supervised learning techniques was assessed, including Gradient Boosting Classifier, K-nearest Neighbors, Decision Tree, Support Vector Machine, Logistic Regression, Random Forest, and AdaBoost. The decision tree algorithm was ultimately chosen, considering the separation of distribution of fit membership probabilities assigned to true fits and true non-fits, as well as its performance on various statistical metrics.[12]

Textile images preprocessing and convolutional neural network

The image preprocessing and CNN of textiles followed similar strategies described above for tapes with some modifications. The first step is determining a way to represent the images in a 1-dimensional vector while reducing the “ the curse of dimensionality” and preventing to destroy the spatial context of local features inside the image during the flattening of the array. To address this, the images undergo preprocessing to reduce the size and focus on the edge where the fit features are. The ForensicFit package, developed by us, provides the means to crop the image to only include the edge of the textiles. The dpi resolution is set to the minimum dpi (600 dpi,600 dpi) of all the raw images, this keeps pixel to inches ratio consistent through all images.

Now, the size of the output edge image will be (610 pixels, 2600 pixels). The *x*-dimension (length of the textile) is achieved with relative cropping from the edge. The *y*-dimension (width of the textile) is more complicated as textiles originating from different sources will have massive differences in sizes, and therefore may not have the same size in the y-dimension. The width values of the textiles used in this study range from 2200-2600 pixels. At this same point in the tape preprocessing, the tape images were resized to the y dimension. This was done as the tape images were very similar in morphology, therefore the distortions introduced by the resizing did not cause a significant problem. However, textiles' morphology can be quite different from image to image, therefore there is a need to maintain the aspect ratio. The algorithm needs to have a consistent input size; therefore, the images are cropped near the edge in the y-dimension, then a padding of zeros was added equally on both sides in the y-dimension if the size of the crop is less than the maximum preset size 2600 pixels. The output edge image is then resized to be as small as possible and still retain the fine details of the textile. In this case, the edge images are reduced by half leading to an edge image with a size of (305 pixels, 1300 pixels) and a resolution of 600dpi. See **Figure 12** for the output of the reduction.

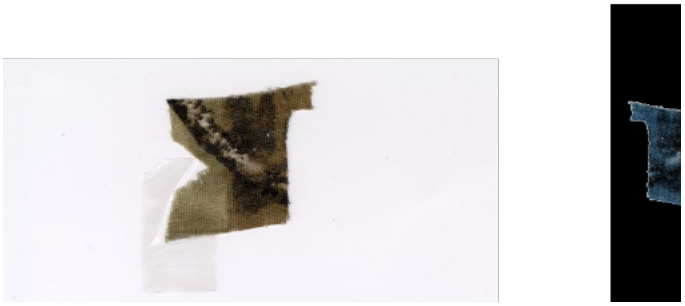


Figure 12. *Example input of a textile image. This image shows one side of the textile. One of the edges has been cut in a triangle shape, representing a non-comparison edge. The preprocessing reduced the original image to the two edges, and the non-comparison edge will not be used.*

The textiles model uses a convolutional neural network (CNN) followed by a fully connected neural network as implemented in TensorFlow to process the database of images. The CNN has two main purposes. Convolutional layers act as a feature extractor and dimensionality reduction technique for the textile pairs, and the fully connected layers make the decision whether the textile pairs are a fit or non-fit. A concatenated input of the textiles along the *x*-direction was selected to give the input pair a size of (1300, 610). An example of the input following preprocessing is shown in **Figure 13**. In the figure, it is of note one of the textiles images is flipped from the original stated and concatenated to the left of the other. This is done so the network can learn on the cross-correlations between the edges of the textiles image. The network should also be able to recognize the symmetry of the inputs. It should not matter which textiles are flipped and concatenated to the other as they are the same fit. This symmetry is taken into account by randomly selecting the textiles to be flipped.

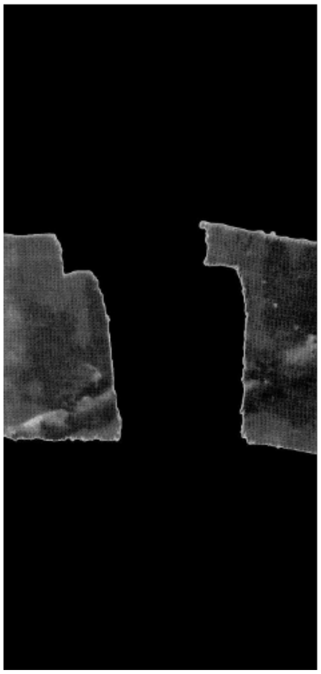


Figure 13. *Example image inputs. Two textile edges. Each image has a size of 1600 × 610 pixels.*

The textile CNN network is built from a series of convolution layers, where filters with small receptive field of 3×3 pixel window (smallest size capable of capturing the notion of left/right, up/down, and center) and a strides of 1 was used. The convolutional layers used ReLU activation functions and were followed by 2×2 pixel window Max-pooling layers to handle the dimension reductions. At the end of the convolutional layer, the image is flattened to a 1-dimensional vector of size 215,040 elements. Compared to the raw flattened input (1300×610=793,000), this reduces the number of parameters the network needs to learn. Finally, three fully connected dense layers of size 500, 100, and 1 are added. The 500,100 layers use the ReLU activation function , whereas the final layer has a sigmoid activation function to map results between 0 and 1 to be used in binary classification. A 0.5 weighted dropout layer is used to fight the overfitting. Training and validation is performed as described for tapes above.

Algorithms for extracting and interpreting edge feature data for physical fits

A data analysis algorithm using mutual information and a decision tree has been developed to do a physical fit evaluation based on data received from the physical fit examinations performed by examiners with the reasoning of their decisions. For each material, the pairs are examined by a trained analyst using a standardized documentation spreadsheet to record the occurrence 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.

Reliability of partial comparison items: establishing criteria for minimal sample size for fit comparisons using mutual information theory

To evaluate the value and performance rates of partial comparisons when a full edge sample is not available for comparison, the bin documentation 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 bin areas. Next, 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. The criteria for performance rates include a decision of non-fit if the ESS of the partial width is less than 40, inconclusive if the ESS is 40–60, or a fit if the ESS is 60 or higher. Finally, the recorded outcome of the partial width comparison is evaluated versus the known ground truth (*i.e.*, known true fit, known true non-fit).[9]

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.

Mutual information (MI) is used to analyze the data from the analyst's reporting templates. The numerical code of the template indicates the analyst's decision for each bin (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, as illustrated in **Figures 4, 8, and 9**, for tapes, textiles, and plastics templates, respectively. 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. The mutual information of each feature is then calculated. **Figure 14** 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.[37]

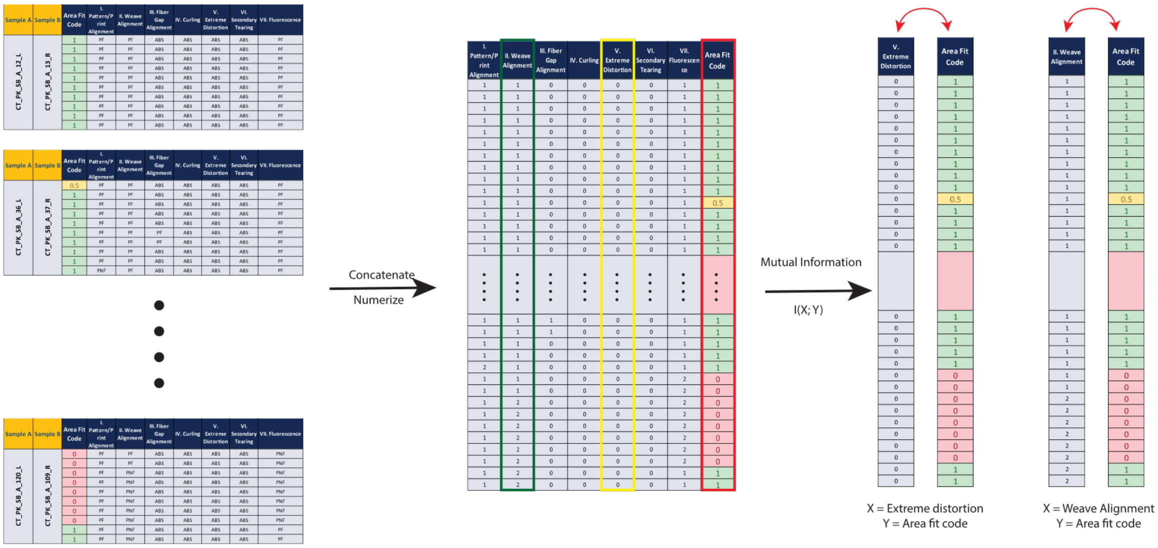


Figure 14. Process of data manipulation for the calculation of mutual information.

Computational decisions on fit comparisons using decision tree algorithms

Decision trees are used here as supervised machine-learning algorithms to break 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. A single tree can fully describe the feature space partitioning. Creating the decision tree is equivalent to finding the optimum partitioning for an n-dimensional feature space.

Finding the optimal partitioning of the feature space is shown to be a nondeterministic polynomial-time complete (NP-complete), a type of computational problem that no efficient solution algorithm has been found to solve[38] , 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. This study uses the DecisionTreeClassifier function implemented in the Scikit learn package using the analyst's reporting templates as input. Entropy is used as the criterion for growing 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.

Task 3 (Objective 2) — Develop an extensive database on trace physical fractures of duct tape, textiles, and plastics, and test the method proposed under Objective 1.

The data generated from this study will serve as an essential body of knowledge for interpreting fracture fit evidence. We have created the most extensive available collection set on trace physical fractures to serve as the basis for the validation of decision criteria and statistical methods for quantitative assessment of the evidence. We collected nearly 9,000 items to generate 4,733 independent physical comparison pairs (**Table 4**). Since the "ground truth" of the source of each sample is known, the datasets are used to generate training and testing sets (known true fits and known true non-fits). The proposed dataset reflects the three most common types of materials commonly fractured or separated from their original source. Finally, the proposed dataset is structured such that the most common factors believed to influence fracture appearance (for each type of material) can be studied (**Figure 15**). The different sample sizes for the dataset have been defined such that performance studies can detect the effects of the different factors (and their two-way interactions) on the ESS scores and respective performance rates.

Table 4. *Sample information for the development and validation of the database of fracture fits.*

Material	**Dataset size**	**Composition**	**Sources**
Duct tapes	3321 comparison pairs, each composed of 2 fractured objects (obtained from >6000 samples)	3 Tape grades (high, medium, and low quality); 2 separation methods (hand-torn and cut); post-fracture stretching.	Duct tape rolls were purchased at retailer stores and online.
Textiles	967 comparison pairs (obtained from ~1200 samples)	2 patterns (unicolor and multicolored); 2 separation methods (torn and stabbed); 2 fabric constructions (knit and weave), 3 fiber compositions (100% cotton, polyester, and mixed)	Fabrics were collected from donated clothing items.
Plastics	445 comparison pairs (obtained from 1337 samples)	Several automotive plastic types (lights, mirror housing, and bumper).	Automotive parts collected at junk yards.

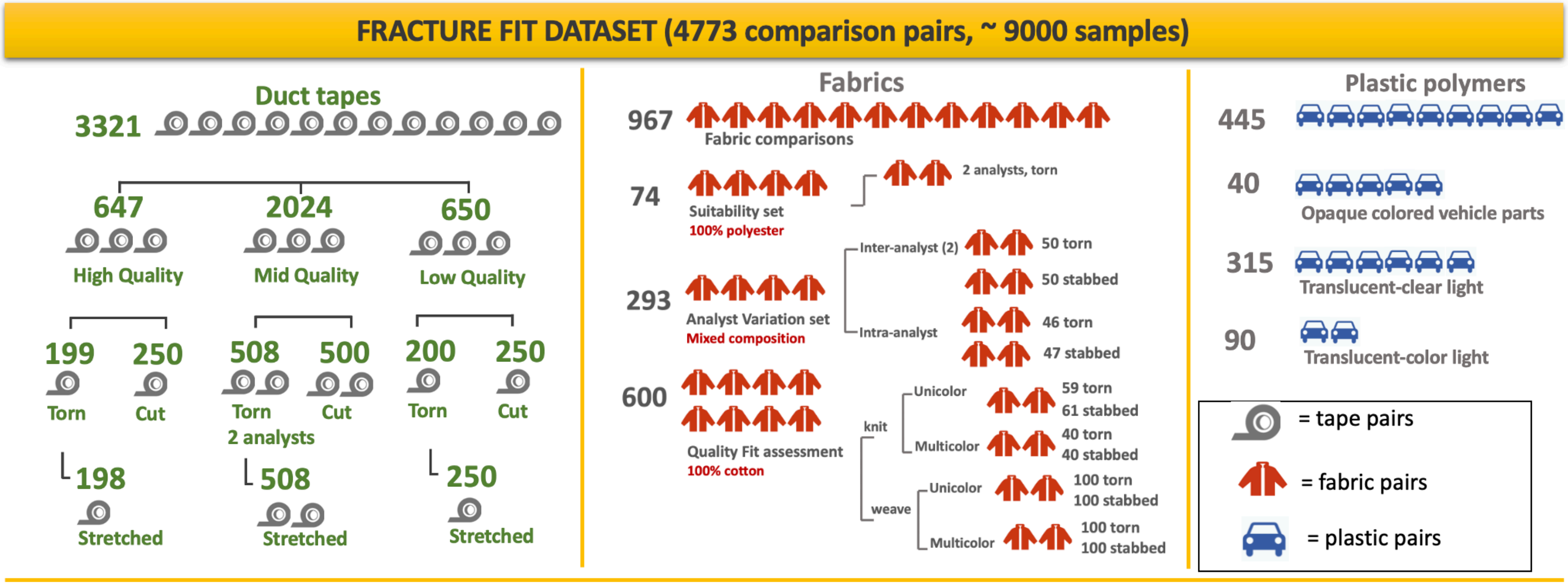


Figure 15. Left: Breakdown of subsets for the physical fits dataset. The tape set consists of subsets of samples originating from each of three grades of tape, low quality (LQ), medium quality (MQ), or high quality (HQ). Edges are scissor-cut (SC), hand-torn (HT), or hand-torn with additional stretching (HT-S). Mid: The textiles set distribution by composition (polyester, cotton, mixed), construction (knit, weave), design (unicolor, multicolor), and separation method (hand-torn ror stabbed). In the textile study, the intra-analyst set uses textile samples from the same set as the inter-analyst study. In the suitability set, two analysts independently analyze 37 comparisons (74 comparisons by the two analysts). Right: distribution of vehicle plastics by color and material type.

Two main types of data are generated in the study, 1) metadata in the standardized reporting templates containing qualitative descriptions and numeral data, and 2) images of the samples scanned and curated as explained in the previous task 2 section. The samples were prepared by non-participating students, maintaining the examiners blind to the origin of the samples during the examination and data analysis. A random number generator was then used to create new sample ID numbers to minimize bias and the ground truth of training and analysis dataset was maintained under the custody of the PI. Specific naming convention codes were generated to label data and samples with unique and traceable identifiers.

Duct tape dataset sample preparation

The tape rolls used to create the subsets for this study were of different grades (**Figure 15**). All the within-set pairwise comparisons were prepared using pieces from the same roll by either tearing the tape by hand or cutting it with a pair of scissors. To simulate complex samples, a subset was also stretched by removing three times the tape from the acetate and stretching it in the width and length directions. The fractured pieces are approximately four inches long and placed onto transparent acetate film to facilitate manipulation of the sample under the microscope without altering the edges.

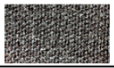
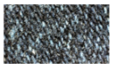
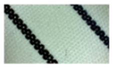
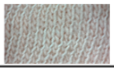
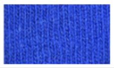
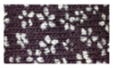
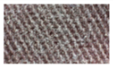
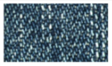
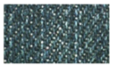
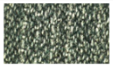
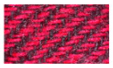
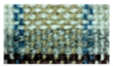
The participating examiners were given a standard reporting template to fill out with the comparison pairs pre-listed (Microsoft Excel spreadsheet) and asked to examine the assigned pairs, first documenting observation for the questioned (set arbitrarily as left side sample in the list), followed by the known sample, and then placed them side by side. The low and mid-quality tape sets had semi-transparent adhesive, allowing the scrim to be seen through the acetate and the adhesive. Observations were made using transmitted and/or reflected light sources. The thickness of the high-quality tape's adhesive prevented the observation of scrim features. Thus, during the comparison of high-quality tape, the adhesive was removed in a thin strip from each edge using liquid nitrogen and hexane to prevent distortion. Observations and annotations were made before and after removing the adhesive.

Textiles dataset sample preparation

The textile physical fit study originally consisted of sampling and analysis of 600 total 100% cotton comparison pairs split evenly between plain weave, pattern weave, and plain knit design and construction types. However, due to some stretching and suitability issues observed in some fabric configurations, the set was increased to 967 comparison pairs to account for other factors (see **Figure 15**). The textiles used for this study were collected via anonymous donations, including clothing with some normal wear. More than 100 articles of clothing were donated for this study. Donated garments were separated by composition, construction, and pattern.

Comparison pairs were then generated by dressing selected garments onto a foam mannequin. Hand-torn samples were collected by first creating a small 0.25-inch incision to facilitate the tearing process. The garment was then torn by hand to produce an approximate 3 inches fracture. Stabbed pairs were collected using a brand-new 8 inches chef's knife. A guard was placed on the knife's blade at 2.5 inches from the tip of the knife to help control penetration depth, and thus, fracture length. The garments were stabbed at a height of 18 inches from the point of contact using a controlled motion using the elbow only. A summary of the main textiles used in the study is presented in **Table 5.**

Table 5. *Table of fabrics used in this study, including their composition and construction, separated by set. The number in parenthesis in the description column represents the respective textile ID number.*[8]

	Description	Composition	Construction	Image
Suitability	(1) Tan bolt fabric	100% polyester	Knit	
Inter- and Intra- Examiner Variability	(2) Navy dress pants	75% polyester, 25% cotton	Weave	
	(3) Navy denim jeans	60% cotton, 22% rayon, 17% polyester, 1% spandex	Weave	
	(4) White short-sleeve dress shirt with blue stripes	100% cotton	Weave	
	(5) Beige tank-top	100% polyester	Weave	
	(6) Navy and white patterned short-sleeve shirt	93% rayon, 7% flax	Knit	
Unicolor Knit	(7) Pink T-shirt	100% cotton	Knit	
	(8) Red T-shirt	100% cotton	Knit	
	(9) Blue T-shirt	100% cotton	Knit	
Multicolor Knit	(10) Navy polo shirt with white pattern	100% cotton	Knit	
	(11) Camouflage T-shirt	100% cotton	Knit	
Unicolor Weave	(12) Grey denim jeans	100% cotton	Weave	
	(13) Navy pants	100% cotton	Weave	
	(14) Light blue denim jeans	100% cotton	Weave	
	(15) Dark denim jeans	100% cotton	Weave	
	(16) Black Denim Jeans	100% cotton	Weave	
Multicolor Weave	(17) Red flannel lounge pants	100% cotton	Weave	
	(18) Pink penguin-patterned lounge pants	100% cotton	Weave	
	(19) Navy dress shirt with stripes	100% cotton	Weave	
	(20) Teal dress shirt with stripes	100% cotton	Weave	
	(21) Tan and black flannel shirt	100% cotton	Weave	

Plastics dataset

Plastic fragments were collected from shattered headlights, taillights, and bumpers from various vehicles from the WVU Crime Scene Complex and local junkyards. The original light assemblies were 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 16** shows an example of one of the headlights in its original form, the different components removed, and some fit pairs prepared for the dataset.

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 falling weight. After fracturing, all fragments are numbered with consecutive numbers, and photographed, then stored in a sealed plastic evidence bag and into a plastic box to preserve the piece.

Since consecutive numbers can induce some bias in the analyst performing examination, another analyst, not participating in the breaking renamed the items with unique identifiers that are selected in a random code. The ID includes differentiation between headlight and taillight, the number of the original polymer sample, the fragment number, and a randomized two-letter code. For example, H2.8-TX represents Headlight 2, Fragment 8, and TX is used as a random ID for that specific fragment. Also, each label placed on the piece will have an arrow pointing at the north orientation of the piece relative to the original assembly.

True fit comparison pairs are created from pieces known to have been joined together. The individuals preparing the samples then compared the fragments to identify fragments that were not joined together but have similar edge characteristics to create convincing non-fitting comparison pairs and a second examiner (not involved in the blind examination) verifies the quality of non-fits selections. Once the true fitting and true non-fitting pairs were curated, north-east-south-west (NESW) direction codes were added to the end of each sample name in the digital templated to assist examiners with the intended comparison directions for each pair. For example, in the bottom image of **Figure 16**, the right-side sample edge will be labeled with an "W" in the digital templates and files as "H2.8-TX-W" representing the comparison of the west edge of that particular fragment. Finally, each pair was photographed together in the orientation that was intended to be analyzed by the examiners. The images were stored in JPEG format without compression.

Over 1,300 fragments were broken, reassembled, and stored in the laboratory. From this set, 445 comparison pairs, including know true fit and known non-fits were created and analyzed (see **Figure 15**), and the remaining pieces are kept in the collection for future studies.

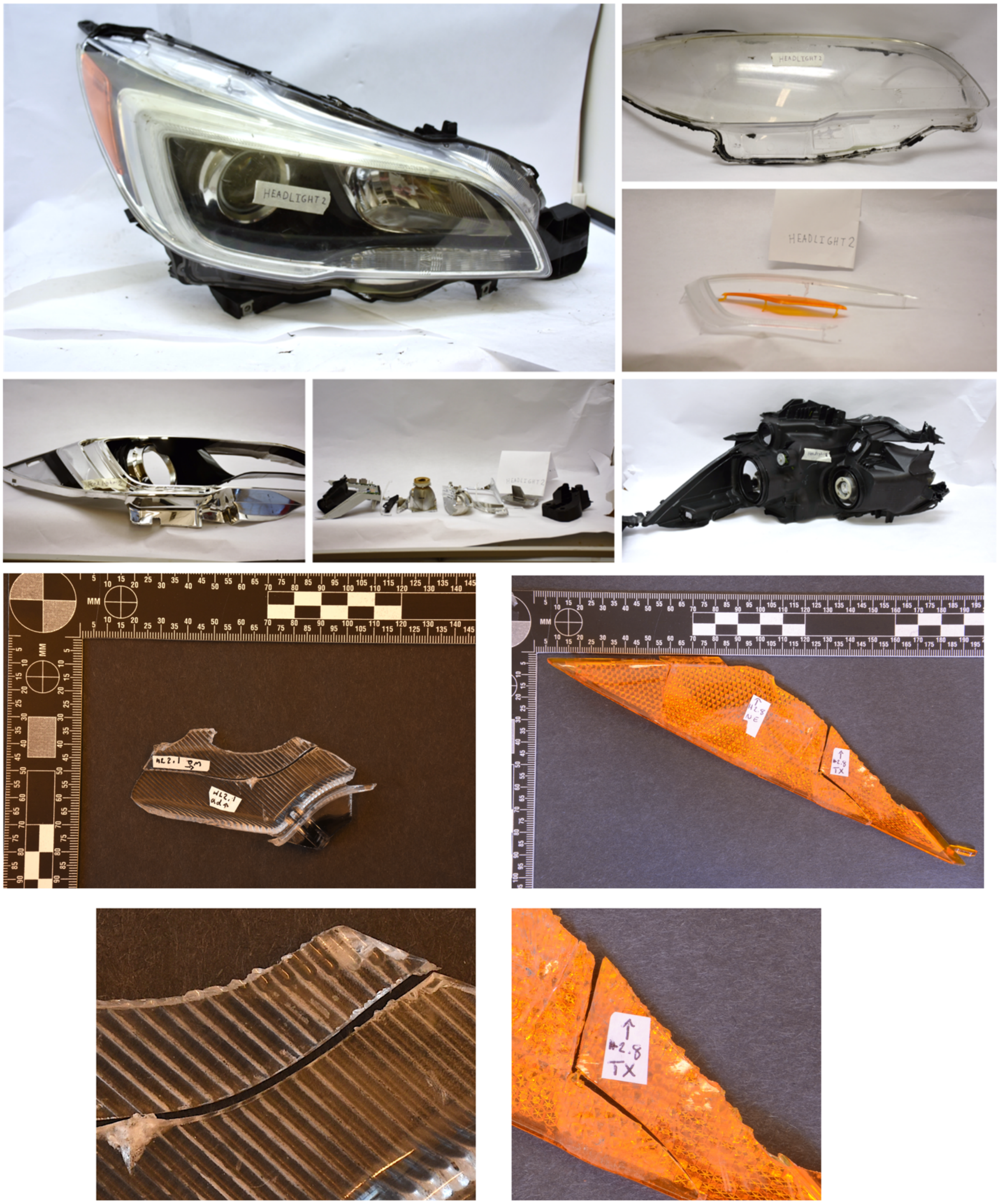


***Figure 16**. Images of an original headlight and the separated components before fracture. The initial assembly is shown in the top left image. The lamp is taken apart to separate the different types of polymers and remove non-polymer parts (metal/glass). The bottom four images are examples of reassembled and relabeled true-fit pairs for the clear and orange portions of the headlight, and a zoomed image of the edges, respectively.*

Task 4 (Objective 2) — Validation of quantitative methods for assessing the probative value of fracture fits.

Several methods are used to evaluate the probative value of a fracture fit using the large datasets described in task 3. The first method considers that fracture fits are reported using categorical conclusions, such as fit, inconclusive, or non-fit using the standardized and systematic methods developed in this research. When examining a fracture fit between two objects from which the ground truth is known (but maintained blind to the analyst), six main outcomes are possible: a) true positive, b) true negative, c) false negative, d) false positive, e) inconclusive when the objects originate from the same source, and f) inconclusive when the objects originate from different sources. By estimating the rate of these six possible outcomes using a dataset of independent pairs of objects that simulate casework samples, we can estimate the overall performance rates of fracture fit decisions, such as sensitivity, selectivity, and accuracy. These are important indicators of the reliability of the method.

To provide measures of the probative value of fracture fits on continuous scales, we leverage the similarity scores developed during Task 1 (ESS and FPS with the systematic manual method) or Task 2 (computer-based method). Empirical probability density functions of the level of similarity in mated and non-mated pairs of objects can provide valuable insights into the capabilities and limitations of the comparison methods through boxplots, Kernel density functions, Receiver Operating Characteristics (ROC), statistical regression models, and score likelihood ratios (SLR). These methods can be used to optimize the comparison algorithms.

A critical aspect of the study was first to identify the most distinctive features in a physical fit that, then develop standardized terminology and a systematic method for documenting those features during examination. Second, study the main factors that can influence the quality of a fit and the quantitative metrics. Finally, develop methods to assess the probative value of the evidence and assess intra and inter-examiner variations.

Task 5 (Objective 3) — Design interlaboratory studies for the evaluation of error rates of the proposed comparison approach among practitioners

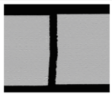
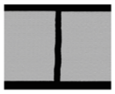
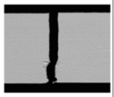
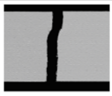
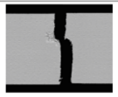
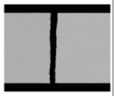
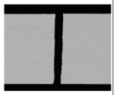
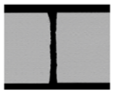
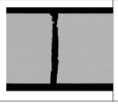
After assessing the methods' accuracies with the large datasets, the overall utility of the methods was tested via inter-laboratory tests. The utility is defined as the "base" consistency rates among examiners using the proposed methodologies. The inter-lab collaborative exercises are anticipated to assist with the standardization of the methods of analysis and interpretation, educate the end-users on the novel protocols, improve the procedures by incorporating the participants' feedback, and facilitate the future adoption of methodologies. In this study, three interlaboratory exercises are completed, two for duct tape examinations and one for textiles. A manuscript describing the findings of the tape interlaboratory has been published a separate manuscript for the textiles is submitted and is under the journal's review.

Duct tape interlaboratory studies

Sample preparation and design of studies. The tape samples utilized in these exercises originate from medium-quality grade duct tape. Each sample consists of a hand-torn, 6-8 cm long strip of the roll. Samples are placed on clear acetate and labeled with unique identification numbers that are traceable to the coordination body but maintain the ground truth unavailable to the participants.[3]

Pre-distribution consensus results are reported by four independent analysts using a blind process, using the protocols described in task 1. Twenty-one (21) pairs resulting in inter-participant ESS relative standard deviations lower than 10% ESS are selected from the sample set. **Table 6** shows that these pairs are rearranged into 3 kits of 7 comparison pairs each, with fracture edge morphology (straight, wavy, or puzzle-like) and ESS as close as possible within each kit. Classification of the seven optimized pairs includes three fits of high confidence (ESS greater than 80%, F+), one fit of lower confidence (ESS below 80%, F-), and three non-fits of ESS lower than 40% (NF+).

***Table 6**. Description of the seven tape pairs selected for each of the three interlaboratory kits, and respective images, ground truth, and consensus values obtained by the pre-distribution examination panel.*[3]

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

The tests are designed as a Round-Robin where each participant independently receives, processes, and returns the kit and documentation to the coordination body. Study kits are distributed so that each kit returns to the coordination body before re-distribution to the next participant. Since only 3 kits can be shipped at a time, and each laboratory is given 3-4 weeks to complete the exercise, each study took nearly one year from design to collection of data. The results include 252 examinations from 38 participants (from 20 and 18 participants in the first and second interlaboratory study, respectively). Only five individuals participated in both studies, however, the experiments are designed to ensure that they receive different kits on each exercise to prevent bias. Participants received training material containing examples of the feature descriptions and an explanation of the ESS method to familiarize themselves with the protocols.

Textile interlaboratory study

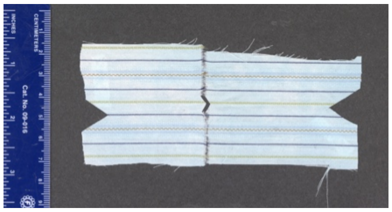
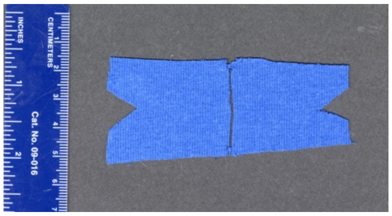
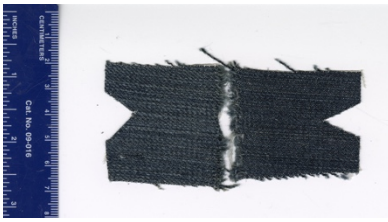
Sample preparation and experimental design

The inter-laboratory study is distributed to 15 participants from ten U.S. laboratories. Each participant is assigned a unique identification code, and the study is conducted anonymously and following a blind approach, meaning participants don't have access to the ground truth and are instructed to complete the study independently of any other analyst.

Prior to distribution, a consensus of results is evaluated by a panel of five independent student analysts. The student analysts examine the images of each comparison pair and document their conclusions using the same techniques the study participants will use. The study consists of three pairs, one true non-fit, and two true fits, with one true fit exhibiting more straightforward features that should lead to a high-confidence fit conclusion, while the other has more challenging features that it is anticipated to assign less confidence in the fit conclusion. The inter-analyst average scores for the two true-fitting comparisons are 93 ± 4.5 % and 89 ± 2.2%, respectively. The average score for the true non-fitting comparison is 27% ± 14%. A summary for each of the three pairs included in this study can be seen in **Table 7.** [7]

To minimize the risk of edge distortions while passing the fabrics across participants, the exercise consists of digital images of the textile pairs instead of the items themselves. Also, because some laboratories verify physical fits through images of the samples, this study is designed to simulate that verification process. One advantage of this design, as compared to the duct tape one, is that the digital version allows faster turn-around times. Each pair is scanned using an Epson Expression 12000XL scanner and four images are created for each pair. The first two images are individual images of each sample comprising the pair. The third image consists of the two samples aligned together as they would be for comparison purposes, while a fourth image was generated by annotating the third image with ten comparison bins, ensuring the participants consistently applied the ESS method to this study.[7] The ILS instructions provide case scenarios for each of the three comparison pairs to provide casework-like context, along with the reporting template with annotated step-by-step instructions.

Table 7. Summary of textile pairs selected for the pilot interlaboratory study, including ground truth and pre-study consensus scores for each pair.[11]

Pair ID	Ground Truth	Consensus ESS score (%)	Example Image of the Comparison Pair
1	Fit (F+)	93 ± 4.5	
2	Fit (F-)	89 ± 4.2	
3	Non-fit (NF+)	27 ± 14	

1.4.2. Data Analysis

Data analysis in this project required using metadata (descriptive nominal), and numerical data (ESS, FPV, FPS metrics, probability outputs) as well as digital images. To assess performance, false-positive rates, false-negative rates, sensitivity, specificity, and accuracy are reported for each duct tape, textile, or plastic dataset.

The data are also analyzed using box plots, a logistic regression model, and score likelihood ratios from the ESS and FPS metrics. The plots can aid in visualizing the spread of the ESS metrics assigned to true-positives and true-negatives in the dataset, as well as any potential overlaps between ground truth sets. Histograms, kernel density distribution plots, ROC curves, and Tippett plots provide insight of the discrimination power and accuracy of the method. A logistic regression model is used to study the effects of certain factors on the ESS values assigned to physical fits. Logistic regression models are used to assign a value between 0 and 1 (the dependent variable) using different predictors. In this case, the dependent variable is the value of the ESS, and the predictors are factors that are thought to influence the quality of a physical fit. For example, in the case of textile the factors of interest are the separation method (hand-torn or stabbed), the construction of the textile (knit or weave), the composition of the textile (cotton, polyester, mixed), and the design of the textile (unicolor or multicolor). Variable selection determines the level of interactions between the predictors.

Several computational algorithms are used throughout the study. Convolutional neural networks are used to evaluate the digital images of known-matted and known-non-matter pairs as explained in task 2 and to create the database. Mutual information algorithms are used to extract bib-by-bin data to evaluate level of importance of each fracture feature and determine the effect of comparing two samples, when one of them has only a portion of the item left. Lastly, this study uses the DecisionTreeClassifier function implemented in the Scikit learn package to provide a computer-based model to classify an image as fit or non-fit based on training from the human-based data. Entropy is used as the criterion for growing the decision tree and quantify the quality of each split.

Performance rates and statistical analysis are performed in Microsoft Excel (Version 19.08), JMP Pro 16 (v.2021, SAS Institute Inc., NC), and mathematical and statistical algorithms created in open access R (version 4.2.2, R studio version 2022.07.2+576). Computational algorithms used open-source Python packages.

To maintain traceability of the data, files are named with pre-determined nomenclature. An inventory master list is created for this databaset, containing the ID number and the respective metadata and descriptors associated with each sample. For the interlaboratory study, the ID of the participants remains anonymous. Each data file collected is stored in a centralized computer following WVU technical support protocols to ensure data security. All the collected data is evaluated separately by at least two independent examiners to assure the integrity and traceability of the data before archiving. The required datasets and associated documentation are submitted to the funding agency at the end of the project for archiving and availability to any government laboratory that requests it; however, the rights for publication of results derived from the data are retained by WVU investigators.

1.5. Expected applicability of the research

Fracture fits are considered the highest degree of association between two trace materials. Still, an objective and statistical assessment of the weight of the evidence is not yet used in current practice. This research has generated, for the first time, a large dataset of fractured duct tapes, textiles, and plastics to provide:

1) Systematic methods of analysis,
2) Quantifiable methods for the evaluation of the quality of a fracture match,
3) Assessment of the accuracy and reliability of the fracture fit comparisons and conclusions,
4) Decision criteria thresholds for human-based and computational-based approaches to assess the evidence,

5) Formal assessment of inter-examiner error rates that can serve as a basis for optimal content on training, proficiency testing, and
6) A model for an effective and traceable peer-review process.

In particular, our study is designed to address the research needs identified by NIST- OSAC [1] and six of the top ten operational requirements specified by the NIJ-TWG on pattern and trace evidence[2]. Moreover, the strategic multi-disciplinary team of researchers and practitioners is critical for transformative and adoptable end-products. Relevant population datasets are used to develop user-friendly automated interfaces to estimate the significance of a given fracture fit and to substantiate the expert conclusions. These computational tools will be available to forensic practitioners and the legal community. The methods developed in this research can further serve as models that can be generalized to other disciplines, expanding impact. As a result, the research offers the criminal justice a valuable body of knowledge to integrate trace evidence information for a broader contribution to the criminal justice system and a more objective assessment of the evidential value in the U.S. courts.

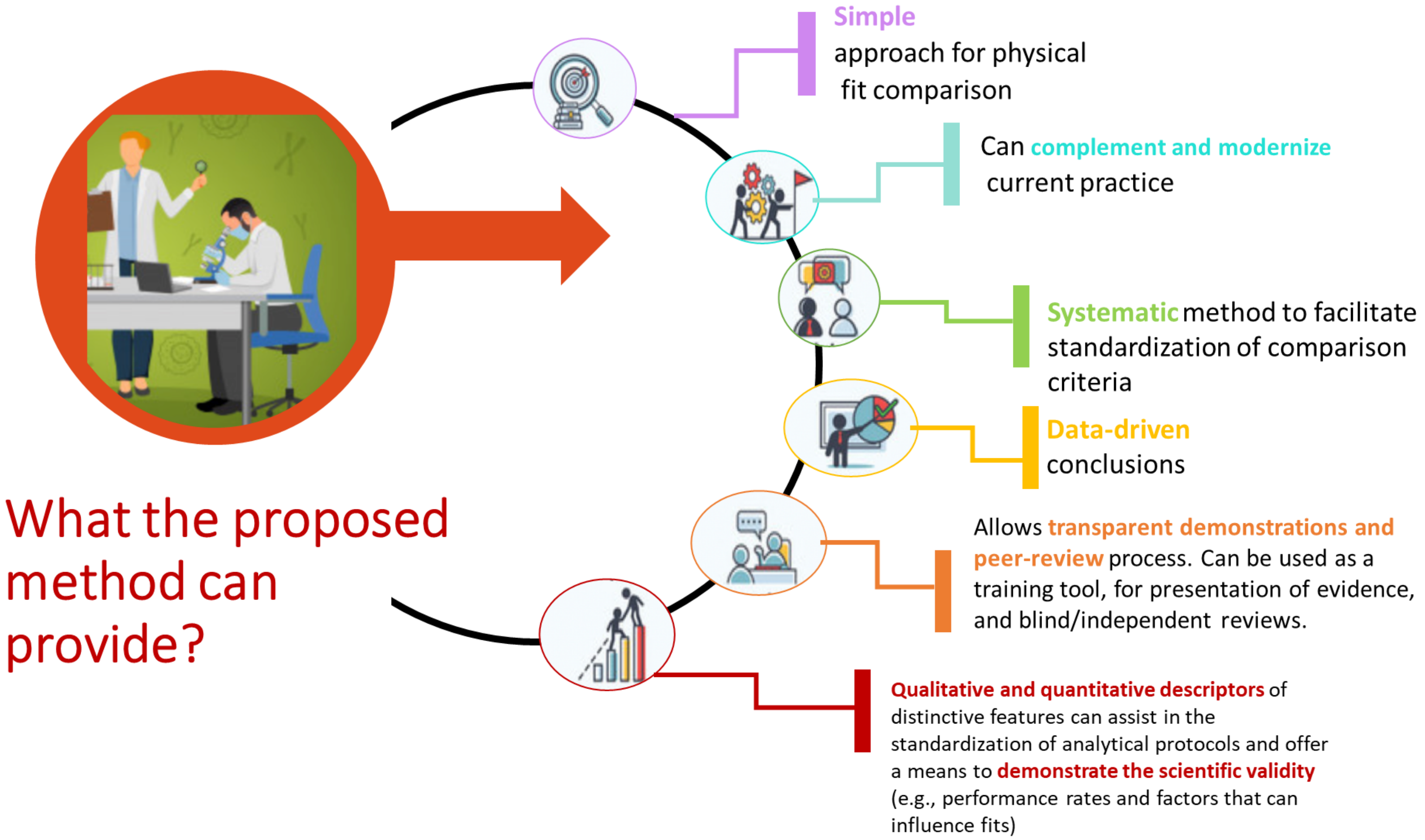
Moreover, the interdisciplinary nature of this study has provided advanced STEM technical training and education to undergraduates, graduate students, and post-doctoral fellows, preparing a future generation of forensic scientists with more robust skills to enhance forensic science practice.

The project is at the stages of method development and validation. Nonetheless, the partnership between the diverse academic team, statisticians, and practitioners has been crucial in disseminating the primary outcomes of this project and envisioning future adoption in the field. In particular, we collaborate with forensic laboratories that provide physical fit examination services. Also, the interlaboratory studies help engage the end-users in assessing the utility of the proposed approach (51 practitioners from 33 US forensic laboratories).

One major advantage of this approach is that the method adoption does not require much investment, other than a microscope (widely available at crime labs), the custom-made Microsoft reporting templates, and the personnel time required to train the practitioners and incorporate the new methods into their quality management system.

This research narrows the current knowledge gap in forensic fit examination and brings several benefits (**see Figure 17)** to the criminal justice system:

1) Provides simple protocols that can be easily adopted at laboratories,
2) Increases the current capacity to demonstrate the thought process and judgment criteria in a physical fit examination to complement and modernize current practice,
3) Access to systematic approaches to aid in the standardization of examination and interpretation criteria for physical fits and increase the consensus among laboratories protocols and practitioners' opinions when conducting physical fit examinations,
4) Improves objectivity with quantifiable data-driven conclusions and probabilistic interpretation of the probative value of the evidence,
5) Increases transparency in the documentation and peer review process, facilitating more independent, objective and blind verification processes, and assisting with training protocols to compare directly the decision criteria between trainers and trainees. This also helps in the incorporation of more stringent quality controls and monitoring of potential bias in the process,
6) Assists practitioners in supporting and informing opinions with protocols and metrics that built the scientific validity in this field.



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 17. *Diagram denoting the main benefits of the physical fit methodologies developed in this research*

II OUTCOMES

2.1. Activities/accomplishments

One of the main goals of this project is to contribute to the preparation of a specialized future workforce in STEM. This project provides unique opportunities for students and faculty to network across several disciplines, including forensic science, physics, mathematics, and statistics. Our research team comprises several researchers; Dr. Trejos and Dr. Romero serve as the PIs, overseeing the project and mentoring graduate students and undergraduates. Also, Dr. Cedric Neumann collaborates as consulting statistician (sub-award). During this project, 20 progress meetings are completed to discuss research results and planning. Also, biweekly research group meetings have been conducted to discuss significant findings and project management. The graduate students' essential milestones in their program directly impact the advance of this research. For example, Zachary Andrews defended his master's thesis in the summer of 2022, and was admitted to the WVU doctoral program in Forensic Science in Fall of 2022. Meghan Prusinowski graduated with a Ph.D. in Forensic Science in Spring 2023 and immediately after joined the forensic science workforce. Also, we hired a postdoc who was a doctoral student in this project from the Physics Department, who defended his dissertation in Spring 2022 and completed his postdoctoral experience in May 2023, helping him to a smooth transition to join academia.

The research team progressed on each of the main five tasks and 77 activities proposed in this award, with the following major accomplishments:

1. Novel methods for comparing fracture fits using human-based and automated algorithm approaches. Comparison methods for the forensic fit examination of duct tapes, textiles, and polymers. The methods include identifying and reporting relevant and distinctive features and an approach to document and quantify the quality of the fit.
2. The creation of the ForensicFit database and access to the package and algorithms.
3. A collection database that consists of 3321 duct tape comparison pairs (various quality grades), 967 textile fit comparisons (various fabrics compositions, textures, and constructions), and 455 comparison pairs from vehicle plastics (headlights, taillights, and bumpers). To simulate samples typically seen in casework, the duct tape edges are created by scissor cut or hand-torn, and further stretched, whereas the textiles are stabbed or hand-torn, and the polymers are fractured by impact to simulate impact force during automotive crashes.
4. Analysis of nearly 5,000 tapes, textiles, and polymers, and a physical collection of around 9,000 samples and digital images. For each set, the results are documented, including data analysis and interpretation.
5. Validation of a quantitative method for assessing the probative value of duct tape fits, which serve as a basis for other materials in this study.
6. A logistic regression model is developed to evaluate the effect of various factors on score metrics for predicting a fit or non-fit for duct tapes and textiles.
7. Design of interlaboratory studies for duct tapes, instructional videos, and training sessions to recruit forensic practitioners. Through the collaboration of 38 forensic practitioners from 23

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

laboratories, the results from 252 examinations are compared across participants and to a consensus ESS established prior to administering the studies by an independent panel.

8. A workshop for 30 practitioners on physical fit examinations at the MAFS/ASTEE 2022 meeting. This helps to disseminate the methods developed within future end-users, receive valuable feedback for improvements, and recruit volunteers for interlaboratory exercises.
9. Design of an interlaboratory study for textile fit examination, instructional videos, and training sessions to recruit forensic practitioners. The inter-laboratory study is distributed to 15 participants across ten laboratories. The results from 45 examinations are compared across participants and to a consensus ESS established prior to administering the studies by an independent panel.
10. A virtual session to discuss the results of the duct tape and textile interlaboratory studies, with 42 participants from forensic agencies, research centers, and academia.
11. Graduate students are trained in statistical packages (R) and programming language (Python), and undergraduate and graduate students in data curation and archiving, sample preparation, and examination of duct tapes, textiles, and polymers.
12. The research findings are disseminated through 1) publishing four manuscripts and 4 more are under journal revisions, 2) presenting the findings at 12 scientific meetings, nationally and internationally, 3) leading one workshop and one informative session with practitioners to familiarize them with the new methods and scope.

2.2. Results and findings

2.2.1. Executive summary of the main findings of the research

This project aimed to develop an effective and practical approach that provides an empirically demonstrable basis to assess the significance of trace evidence fracture fits. We have accomplished this goal by:

1) Developing a systematic method for the comparison of fracture fits of common trace materials such as duct tapes, textiles, and plastics, using both human-based protocols and automated computational algorithms.
2) Developing a relevant extensive database of nearly 5,000 comparison pairs to assess the weight of a fracture fit using similarity metrics, probabilistic estimates, and score likelihood ratios.
3) Evaluating the utility and reliability of the proposed approach through inter-laboratory studies that can establish consistency base rates. The strategic partnership of experienced forensic researchers, computational material science physicists, statisticians, and practitioners has been crucial for planning the adoption of the developed approaches within crime laboratories.

Some of the major findings of this study are:

1) Not every physical fit determination holds the same probative value. There is a wide arrange of factors that can influence the quality of a fit; therefore, our study demonstrates that quantifying the quality of a fit can assist forensic practitioners in informing and supporting their decisions. The study also raises awareness of the importance of assessing the suitability of certain materials for physical fit examinations and conducting a thorough assessment of a fractured edge to substantiate a physical fit opinion.

2) The fracture edge features that are relevant and more individualizing are particular to each material composition, construction, and separation method. The results of this study reveal that the separation of textiles, duct tapes, and plastics impart different features to the fractured edges, and that the influence of various factors on the quality of a fit and error rates vary by material type. Thus, standardized material-specific terminology and criteria are crucial to harmonize and optimize protocols of examination and interpretation.

3) There is a risk of introducing bias and errors when the examination of physical fits is conducted merely based on the judgment of the examiner, particularly in the absence of consensus-based criteria. To minimize those risks, qualitative and quantitative descriptors of the quality of a fit or non-fit can be standardized and documented to demonstrate the basis for conclusions.

4) The methods developed in this study have several benefits: 1) provide a systematic method to utilize qualitative descriptors and quantitative metrics to inform and substantiate the examiner opinion, 2) offer a practical mechanism to document the examiner's thought process, which adds transparency and minimize risks of bias, it also allows for a means to improve peer-review and verification processes, 3) the metrics provide criteria to assess the probative value of a fit and visualization methods to demonstrate a decision further, 4) provide new avenues to evaluate the scientific reliability of fit examinations and identify potential sources of error.

5) This study demonstrates the feasibility of computational algorithms to build physical-fit databases and automated comparisons using deep neural networks, which can be used as a model for other materials. Although the algorithm rates are not as good as the human-based rates, it shows that CNN are a feasible approach to assist practitioners and to understand the most critical features identified by the CNN and supplement decision criteria independently documented by the examiner.

6) Overall, performance rates evaluated in this study through the blind examination of extensive datasets of duct tapes, textiles, and hard polymers representing casework-like items demonstrate that the accuracy of physical fit examinations is high with a very low incidence of false positives. These error rates, however, depend on various factors, including the type of material and conditions of the specimens:

   a. **Duct tapes**: 1) The accuracy of physical fit examinations is generally high (over 98%) except for higher quality grade hand-torn tapes (~84%). 2) No false positives were reported for any of the sets examined (>3,320 pairs examined). Overall, this research demonstrates that the occurrence of observing a physical fit on two duct tape pieces that were not joined together is extremely rare, as no false positives are observed in the various populations evaluated. 3) When evaluating the statistical effect of the experimental factors of interest, different variables have varying impacts on the quality of a fit and edge similarity score. For non-fits, the influence of both the separation method and the quality of tape on the ESS values is negligible, and the ESS trend towards low values regardless. For true fit pairs, however, scissor-cut tapes tend to result in higher scores in comparison to hand-torn pairs. Regarding tape quality, in true fit pairs, medium-quality tapes tend to receive higher scores overall, followed by low-quality, and then high-quality grades. As such, it is critical for examiners to consider the type of tape during physical examinations.

b. **Textiles:** 1) The accuracy of physical fit examinations is generally high (88 to 100%) but generally lower than duct tape, as more variables can cause distortions on fractured textiles. 2) False positive rates are low, but not zero; the observed false positive rate (2% false positives, 10 of 477 total true non-fit pairs) raises a flag and demonstrates the importance of assessing the quality of a physical fit during an examination to minimize risks. 3) A suitability assessment was deemed necessary prior to physical fit examinations of textiles as some fabrics' composition and construction are more prone to distortion. For example, highly deformable fabrics, such as polyesters, show poor unacceptable accuracy (61%). 4) While analysts must consider the composition of the fabric when conducting physical fit comparisons, once suitability is established, a logistic regression model shows that varying factors, such as separation method and construction of the fabric, do not have a substantial effect on the ESS used as an indicator of the quality of a physical fit.

c. **Hard plastics**: 1) The accuracy of physical fit examinations is generally acceptable (85 to 90%) but relatively lower than duct tape and textiles, as more variables cause distortions on fractured plastics and some lack of distinctive features can lead to higher rates of false negatives. 2) 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 (greater than +15). Therefore, ESS and FPS metrics are reliable quantitative metrics to inform and support the practitioner's opinion. 3) Misidentification rates for the comparison set are low, with only one false positive reported (1%). This raises a warning on inspecting the suitability of certain plastic materials for physical fit examinations.

d. **Sample size and suitability**: As it's not unusual for analysts to receive items that are partially damaged or with missing portions, this study answers the question of how small a partial sample can be before it becomes unreliable for physical fit examinations. The results of the models indicate that acceptable accuracies for correctly identifying true fits and non-fits occur when at least 35% of a sample length is present. However, lower accuracies are observed for samples prone to stretching or distortion.

e. **Quantitative metrics of the quality of a fit**: The ESS and FPS metrics demonstrate good performance to assess the quality of a fit and they are very versatile in the sense that they can be used in different ways to assess a fit examination. For example, the metrics are easy to interpret and can be used as a simple criterion based on experimental thresholds of the scores. They also provide a basis to evaluate the scientific validity of the field through performance rates. Additionally, the FPS also provides an additional means to express the weight each feature had in the examiner decision process, and it is recommended to extend the evaluation of this metric to other materials. Another way of assessing that feature importance is also demonstrated with computational algorithms. Lastly, the ESS, FPS, and CNN outputs allows a probabilistic interpretation of the evidence.

7) Interlaboratory studies reveal that inter-examiner agreement rates above 95% are attained when using the proposed examination, documentation, and interpretation methods. Overall, the studies 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.

8) 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 can be adapted to current workflows easily.

The focus of this research is to improve **objectivity, consensus, and scientific validity** in the discipline. This is achieved through various stages of the investigation, and a brief discussion of the primary results is provided below. However, additional information can be found in the cited publications derived from this award effort (see section 3.1 of this report)

2.2.2. Duct tapes physical fit method: evolution through the validation process

Duct tape was the first material of choice for this study as it is one of the primary items submitted for fit examinations.[5.16] The versatility of duct tape makes it a piece of evidence that can be used in many circumstances, such as gaging or restraining a victim, building an improvised explosive device, and smuggling drugs, to mention a few. Thus, its potential value in forensic science is remarkable, as it can provide leads during an investigation and high probative value to link the fractured object to another item found at the scene and to an individual of interest. However, as we have described in this report the scientific foundations of physical fit examinations do not necessarily align with the impact that this evidence can have in decision-making by the trier of fact. Thus, to minimize potential misleading of evidence, our endeavor is to further understand the error rates in this discipline not without first providing standardized means to conduct the examination and present the findings.

Some literature on duct tape fits provides an important foundation for this research [39-44], as well as a preliminary method developed by our group that serves as an important basis[6] More recently, some contemporaneous publications agree with our findings and provide additional validity to the experimental approaches that are used here.[45-46] In this research, the proposed method for examination, documentation, and interpretation has evolved through the feedback provided by analysts and an inter-disciplinary team of researchers. Four major milestones assist with the improvement of the overall strategy for approaching duct tape physical fit examinations.

Milestone 1—Method development and optimization of standardized criteria

First, the main novel aspect of the proposed method is the development of standardized terminology and the identification of relevant features. This is not a trivial task, as it is the central aspect of creating sound criteria for what constitutes an individualizing feature. This is achieved through the analysis of the occurrence of various edge features in known true fits and non-fits sets. Second, we develop quantitative metrics to assess the quality of a fit and serve as a more objective means to interpret the evidence and communicate that to the end-users in a transparent manner. A focus through our experimental designs is to assure we could assess potential bias and error rates with an appropriate sample size to make valid statistical inferences of factors that influence the quality of a fit.

Here, one of the main discoveries is that duct tape tends to fracture in four main patterns that we defined as angled, straight, wavy, and puzzle-like, as illustrated in **Figure 18**. Also, the occurrence of these patterns is dependent on the separation method and quality grade of the tape. **Figure 19** and **Figure 20**, illustrate these findings. For example, for hand-torn sets, the lower scrim count in the low- and medium-quality tapes can cause irregularity and more puzzle-like edges when the tape is torn, while the high scrim reinforcement in high-quality hand-torn (HQ-HT) set results in very straight edges and puzzle-like patterns with less prominent protrusions/cavities. Interestingly, none of the low-quality hand-torn (LQ-HT) 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. 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 20)**.[4,7]

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 19**).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

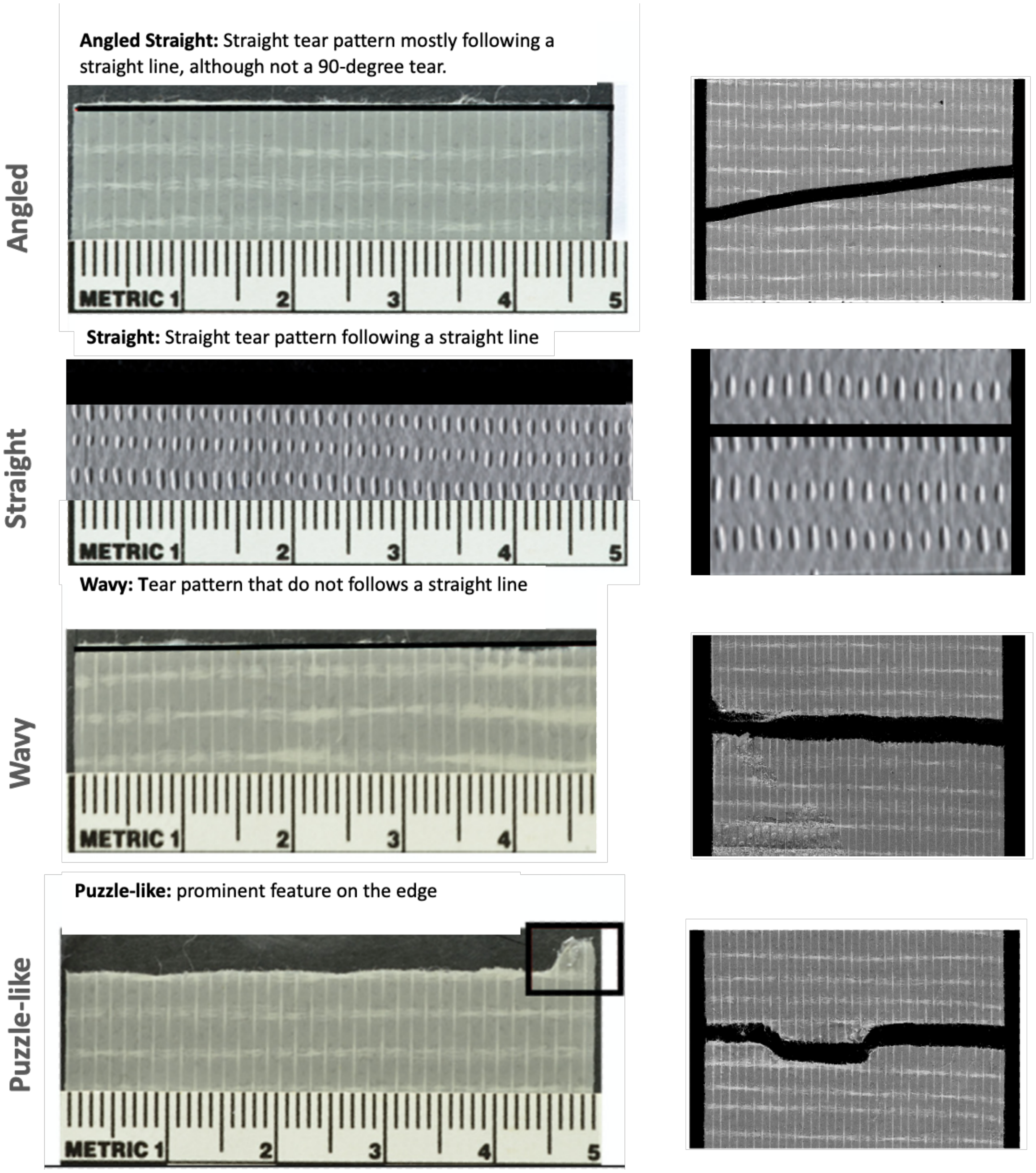

Angled
Angled Straight: Straight tear pattern mostly following a straight line, although not a 90-degree tear.
METRIC 1 2 3 4 5
Straight
Straight: Straight tear pattern following a straight line
METRIC 1 2 3 4 5
Wavy
Wavy: Tear pattern that do not follows a straight line
METRIC 1 2 3 4
Puzzle-like
Puzzle-like: prominent feature on the edge
METRIC 1 2 3 4 5

Figure 18. *Examples of angled, wavy, and puzzle-like patterns observed in duct tape separations.*

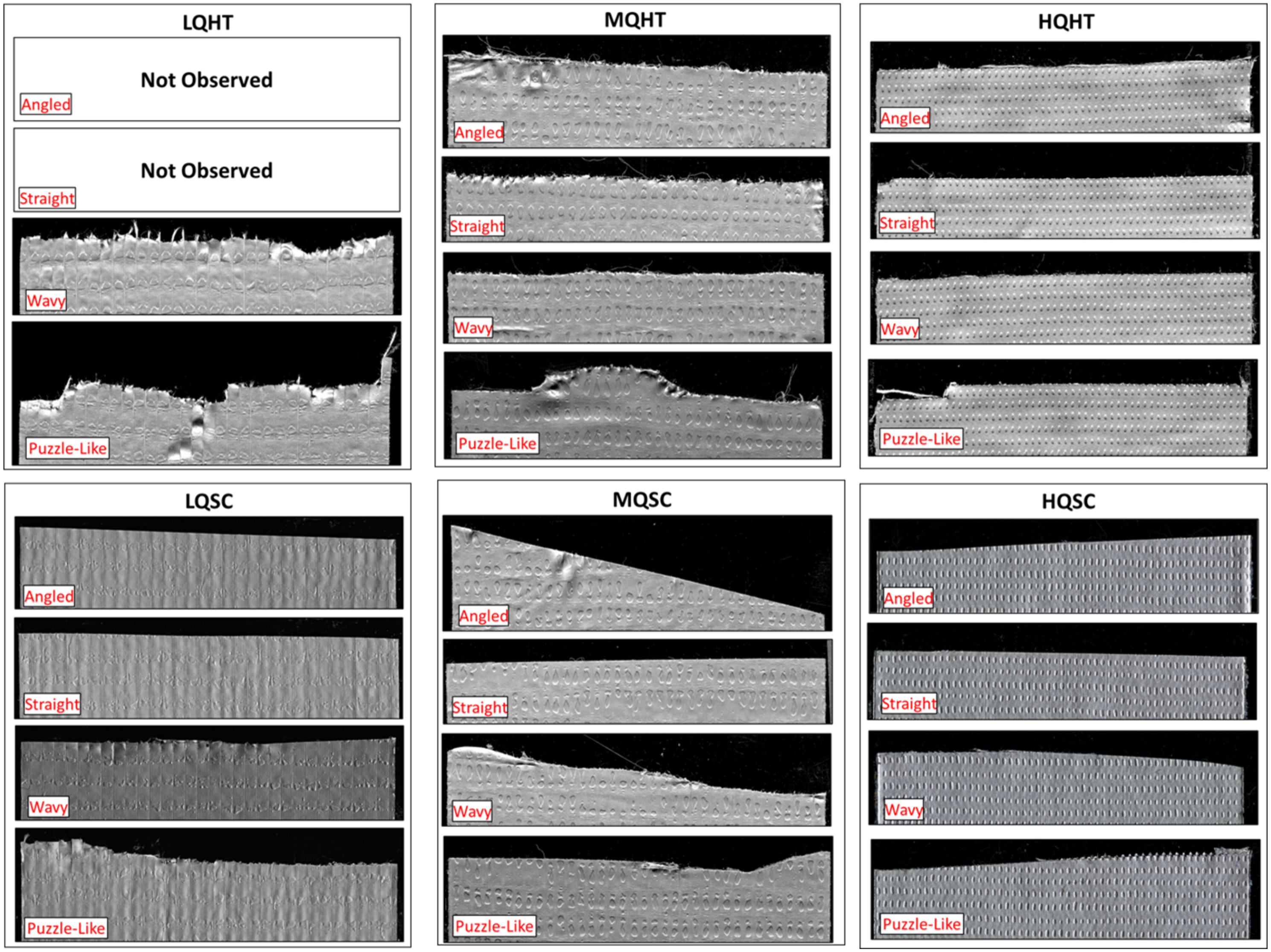


Figure 19. *Examples of edge morphology for each sample set. Straight and angled edges were not observed in the LQ-HT set. Stretched sets shared the same edges as the non-stretched edges, so no additional examples are demonstrated here.* [4]

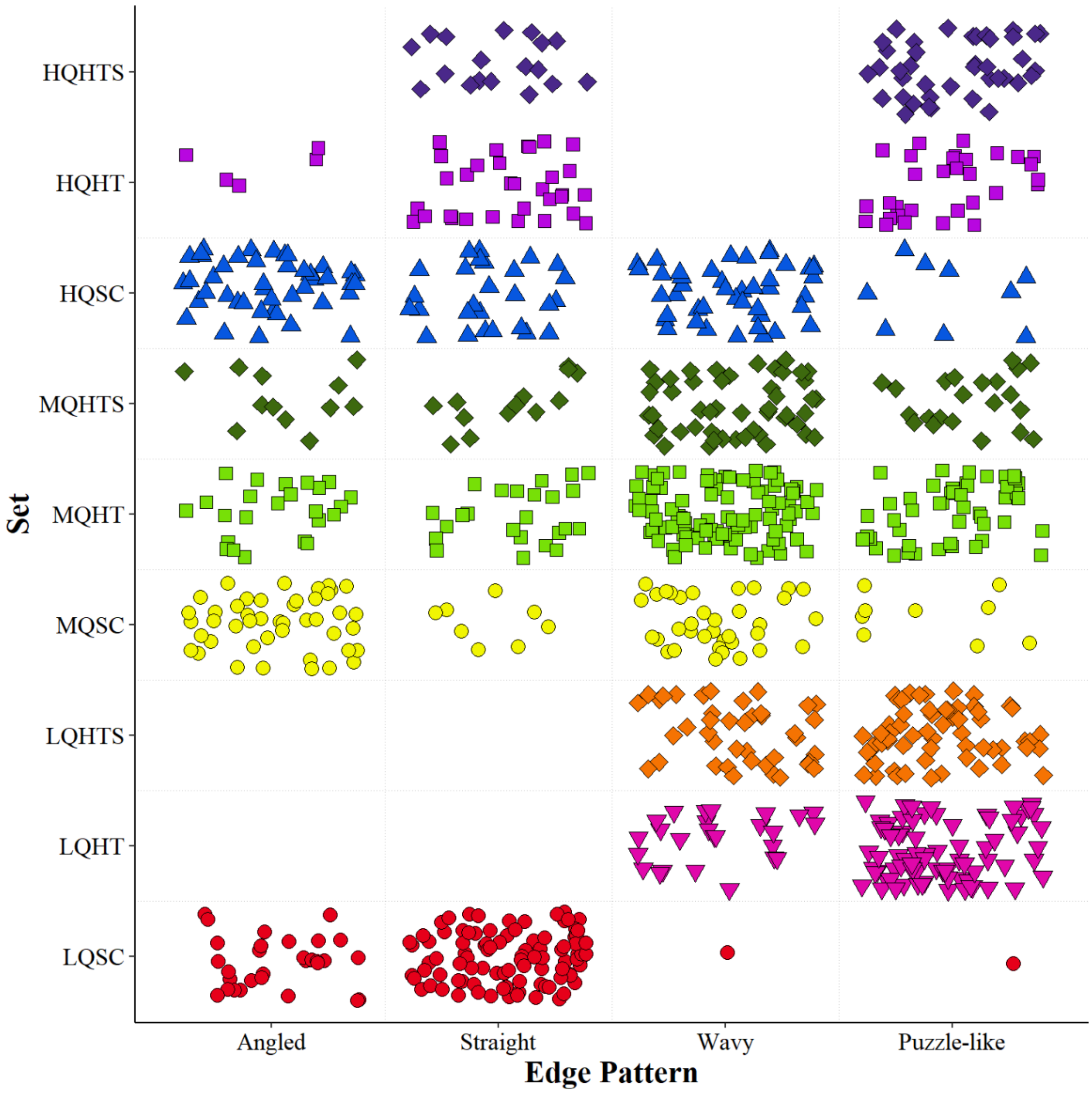


Figure 20*. Edge pattern occurrence trends for true-fit pairs for all compared sets. Overall, puzzle-like edges are more common in hand-torn sets, while straight or angled edges are more commonly observed in scissor-cut sets.* [4]

Also, after separating hundreds of duct tape pieces, and evaluating the features that are indicative of a fit or non-fit, we define eight main features for the examination of duct tape as follows (see **table 8**):

1. **Alignment of severed dimples:** these are severed dimples on tape backings that align from one edge to the other in shape, size and location across the fracture. This feature is only applicable on the backing side. When it is present and aligns across the separated edges, it can provide support to a fit decision because these manufactured-imparted marks have some inherent variability across a single roll and when split through the fracture, those patterns are very unlikely to align by chance. Likewise, when there is a major misalignment of the severed dimples, the feature provides support to the non-fit decision.

2. **Calendaring striations across the edge:** calendaring striations are small scratches or marks left during the manufacturing process on the backing side. When these marks align across fracture edges in their relative position, shape, and depth, they can provide support for a fit. Otherwise, when they show misalignment, they support a non-fit.

3. **Alignment of warp scrim:** warp fibers are an inherent component of duct tape, and they are known to be present and constructed in a reproducible manner across a tape roll. Therefore, when these fibers transverse the fracture and align to the corresponding fiber on the other side, they support a fit decision, and vice versa in a non-fit situation.

4. **Correspondence of protruding warp yarns and the respective pattern gaps in the other edge:** when the separation of the tape lifts warp fibers away from one of the edges, leaving an indentation on the adhesive of their original location, the correspondence of the warp fibers that extend past the edge of one tape piece and the gap of the missing scrim on the opposite side becomes evident. This feature, when present can support a fit decision.

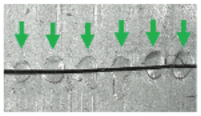
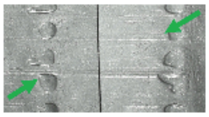
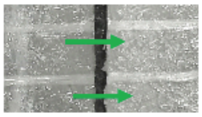
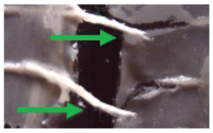
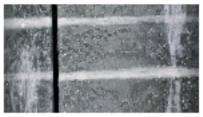
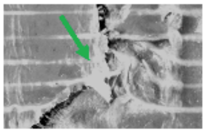
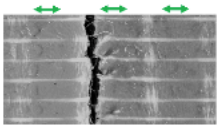
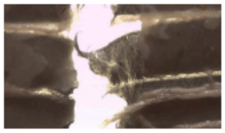
5. **Weft scrim at or near the edge consistent with the overall weft pattern:** another important component in the scrim construction is the yarn that runs across the width of the tape. Because, the separation and construction patterns are reproducible within a single roll, and variable between different roll sources, this feature can be valuable in the examination. When the weft yarns on each edge are consistent with the rest of the weft fibers on the opposing edge, they support a fit decision; otherwise, they support a non-fit.

6. **Continuation of scrim weave pattern:** this feature refers to the consistency, or inconsistency, of the construction pattern of the weave and warp yarns in the separated edge as compared to the expected sequence of the pattern observed in the rest of the tape pieces. And the continuation, or lack of, can aid in the fit or non-fit determination.

7. **Distortion explained by stretching directionality:** stretching inevitably occurs to some extent during a tape separation and this can hide both non-fitting and fitting features. When the alteration to the backing and adhesive morphology coincides with the direction of the tearing, the distortion can be explained although it would not provide a strong support of a fit. Otherwise, when the distortion is not explained by stretching directionality, the feature can lead to a non-fit decision or an inconclusive result.

8. **Missing material:** gaps left on the edge alignment by missing material can provide support for non-fit decisions.

***Table 8**. Table of features extracted from the documentation for tapes and the respective response options for observation of the features at the macroscopic and microscopic level.*[7]

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, 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	

With these bases, a systematic documentation template is created to document the observed features during macroscopic and microscopic stages of the examination, report a score per bin that provides visualization at-a-glance of fit, non-fit and inconclusive areas, and estimate the ESS. A workflow and defined criteria are proposed, proving a path forward to address standardization and consensus within the discipline (see **Figure 21**).[4]

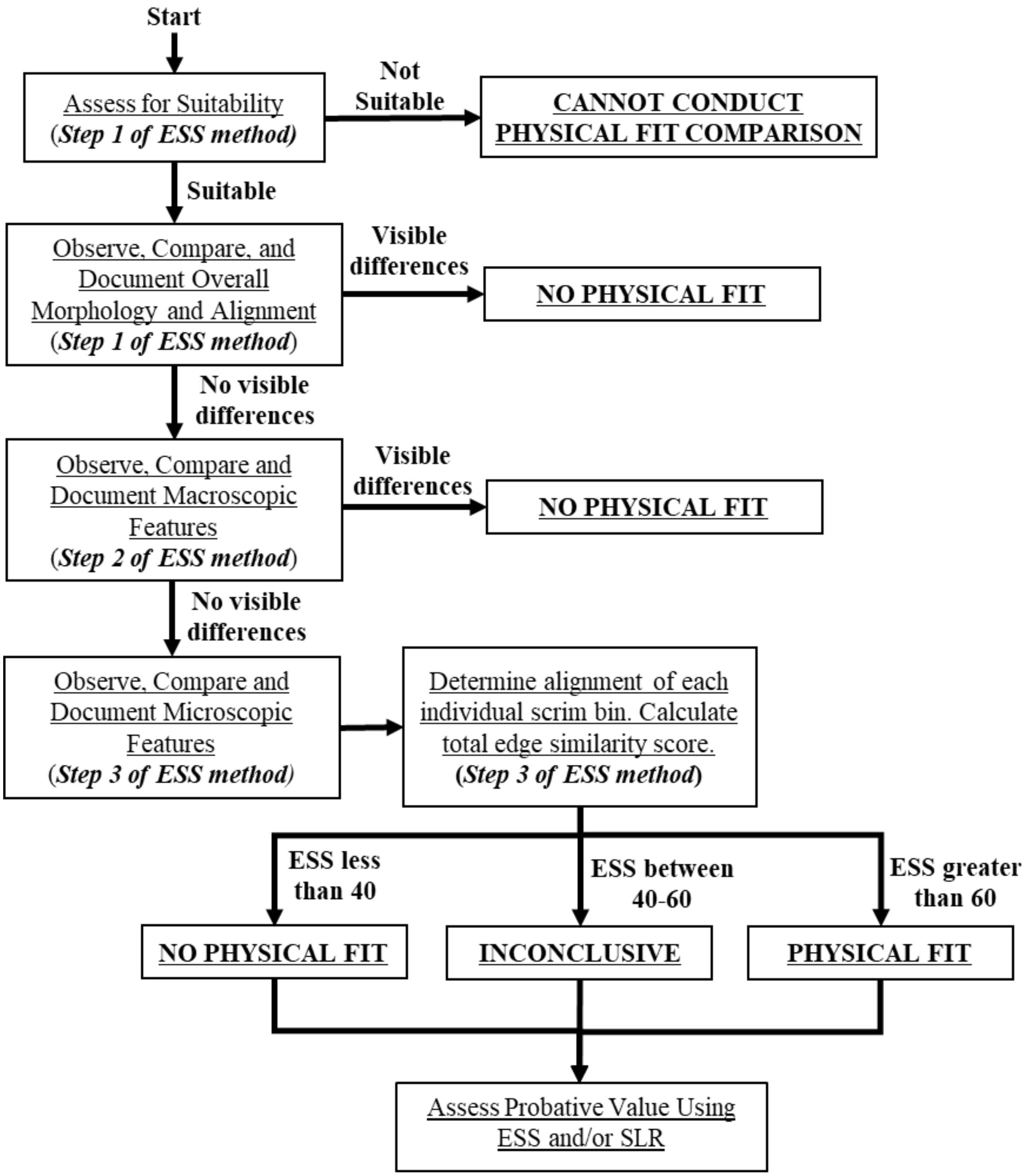

Start
Assess for Suitability (Step 1 of ESS method)
Not Suitable
CANNOT CONDUCT PHYSICAL FIT COMPARISON
Suitable
Observe, Compare, and Document Overall Morphology and Alignment (Step 1 of ESS method)
Visible differences
NO PHYSICAL FIT
No visible differences
Observe, Compare and Document Macroscopic Features (Step 2 of ESS method)
Visible differences
NO PHYSICAL FIT
No visible differences
Observe, Compare and Document Microscopic Features (Step 3 of ESS method)
Determine alignment of each individual scrim bin. Calculate total edge similarity score. (Step 3 of ESS method)
ESS less than 40
ESS between 40-60
ESS greater than 60
NO PHYSICAL FIT
INCONCLUSIVE
PHYSICAL FIT
Assess Probative Value Using ESS and/or SLR

Figure 21. *Proposed examination scheme for physical fit comparisons. If the samples are not suitable for physical fit examination, then other chemical examinations are necessary. If the sample edges demonstrate obvious differences in the comparison features at any stage, the outcome is "no physical fit". Beyond the microscopic comparison (Step 3 of ESS), the outcome 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.*[4]

To illustrate the use of the proposed approach, a kidnapping/homicide mock case is described here. In this case, the mock scene includes a victim who is bound and gagged with duct tape, and the evidence is collected to simulate a high level of difficulty for the tape examiners. For example, some items are stretched during the restraining to the victim and placed on full acetate sheets, while others are crumpled up in plastic bags, as shown in **Figure 22.** A known tape roll is submitted for comparison. Upon examination, the analyst determines the tape was hand-torn and of medium quality grade, therefore the ESS scores from comparing various recovered items (questioned items) with the known tape are calculated, and the score likelihood ratios estimated from the MQ-HT dataset (see **Figure 23**), using a shinny app developed by our group. **Figure 23** shows the use of ESS data from the MQHT datasets to create kernel distributions to build the log score likelihood ratios (SLR) shown on the right side. Thus, when the analysts complete examination, they have two metrics to support and inform their opinion: ESS scores and SLRs.

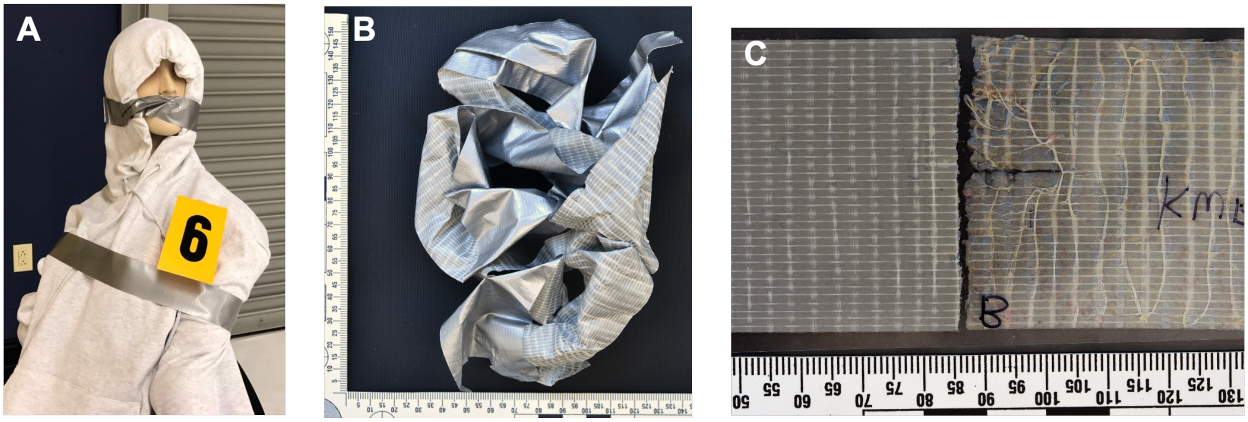


Fig 22. *Images taken of the tape samples collected from the mock crime scene. Image A, shows a tape placed on the mannequin. Image B, shows a tape sample that was received crumpled, and image C shows a false negative example of a distorted tape (right) compared to the known source (left). Adapted from Prusinowski et a.l.* [6]

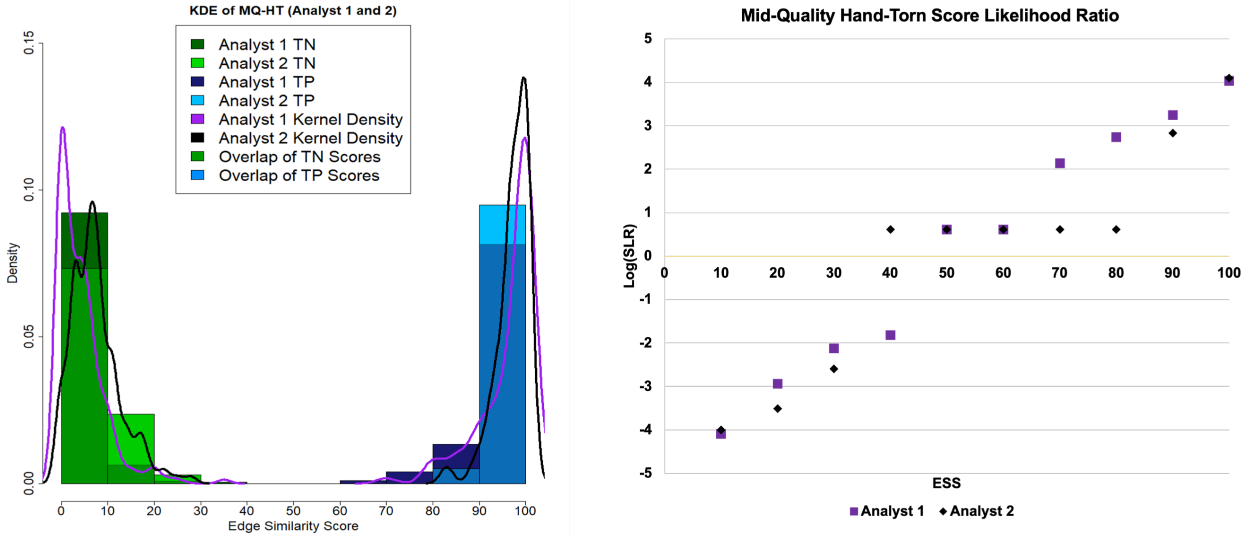


Fig 23. Left: *Score distributions of the true positives (TP, blue) and true negatives (TN, green) of the medium quality tape hand-torn set for both participating analysts.* ***Right:*** *Logarithmic score likelihood distribution for both analysts in the medium quality tape hand-torn set.*

Three analysts independently examine the tape pairs. The criteria they used for ESS are 0-40% fit support, 40-60% inconclusive, and 60-80% weak-moderate support of fit, 80-100% moderate-strong support of fit. The SLRs provide an easier scale to express the opinion as it is not tightly binned as the ESS, but they complement each other. All of the non-fitting tapes are correctly identified in this case, with ESS ranging from 0-30%, and SLR ranging from 0.0001 to 0.01, thus no false positives were reported. Of the nine total true fits, two examiners correctly identify six, and the third examiner identify five. The range of scores between the examiners and the score likelihood ratio values are calculated as seen in **Error! Reference source not found.**.[6] The tapes with high ESS receive correspondingly high SLR values, indicating stronger support for the conclusion of a fit. The tapes that receive lower scores indicate support for the conclusion of a non-fit, rendering three false negatives and one inconclusive. This case illustrates that regardless of the extreme stretching in the samples, there are 6 out of the 9 true fits that render high ESS scores and strong support via SLR.

Table 9. *Range ESS and score likelihood value (n=3) and interpretation for each known match for the casework set (nine questioned items). Adapted from Prusinowski et a.l.* [6]

Case Tape	ESS	SLR	Degree of Support
D1.2	61-70	20-40	Limited support of **same source** **(weak TP to INC)**
E2.1	91-100	2000-10000	Strong support of **same source** **(TP)**
E1.1	81-90	400-2000	Moderately strong support of **same source** **(TP)**
F2.1	81-90	400-2000	Moderately strong support of **same source** **(TP)**
E1.2	1-10	0.0001-0.0003	Strong support of **different source** **(FN)**
F1.1	71-80	40-400	Moderately strong support of **same source** **(TP)**
A2.1	31-40	0.02-0.09	Support of **different source** **(FN)**
A1.2	31-40	0.02-0.09	Support of **different source** **(FN)**
C1.1	81-90	400-2000	Moderately strong support of **same source** **(TP)**

Milestone 2—Method validation through large databases and evaluation of factors that affect performance rates

As described in the "research design, methods and data analysis" section, the duct tape forensic examination method is validated through a large database of over 3,000 comparison pairs of various qualities and separation methods. This validation provides answers to the following research questions:

1) Do all physical fits hold the same probative value? Can quantitative metrics demonstrate the quality of a fit and be used for the probabilistic interpretation of the evidence?
2) What are the performance rates of physical fit examinations?
3) Which factors influence the occurrence of these features and the quality of a physical fit?

4) Can standardized protocols be developed for the examination, documentation, and interpretation of physical fits through the assessment of the method via large datasets and interlaboratory studies?

The answer to the first question is "no", not all physical fits hold the same probative value, as we will demonstrate in the next paragraphs. Indeed, the use of quantitative metrics ESS and SLRs are key to demonstrating this point.

Let's first look at the analysis of method performance and distributions of edge similarity scores on true-fit and true-non-fit populations from an exploratory perspective. For the ESS score (and respective SLR) to help in assessing the quality of a fit, there should be a minimal overlap of the observed scores in true fit and non-fit datasets, and the value of that score should serve as a scaled range of the probative value, rather than a binary decision. In other words, ideally, if the metrics are informative of the quality of the fit, we would expect low similarity scores for non-fits, and high scores for fits, with the larger the ESS or SLR, the stronger the support for the fit.

One simple way of visualizing the distribution of the scores in true fits and true non-fit datasets, is through distribution graphs such as histograms or boxplots. **Figure 24** displays the experimental ESS for each set of tapes based on the ground truth. For most of the sets, there is an observable separation between the ESS obtained for true fit pairs and the non-fit pairs (i.e., minimal overlap, high discrimination power). The score distributions for all the sets for true non-fit pairs are generally consistent at a score of 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.[4,6]

Some trends are helpful to understand the different behavior in some subsets. For example, for hand-torn sets (HT), a broader variability of scores and a shift to slightly lower rates in the distribution of true fitting pairs is observed in the HQ-HT and LQ-HT, resulting from the edge morphology and predominant features observed on those sets. [4]The distribution of scores for stretched true fits in the LQ-HT-S is also consistent with the LQ-HT set. The scissor-cut sets for both low- and high-quality tape have distributions more similar to the medium-quality scissor-cut set, with most fitting pairs having ESS of 90 or higher. These are quite interesting results, as the popular belief that a scissor-cut holds a less probative value than a hand-torn fracture is demystified here.

The distributions of scores for the true fitting pairs in both HQ-HT and HQHT-S are much wider than in any other set. This is explainable when looking back at the most prevalent fracture patterns of HQ-HT. As illustrated before, HQ-HT tends to fracture in straight edges; thus, is prone to contribute fewer features for comparison. The distortion of the samples caused by removing the thick adhesive and the additional stretching compounds the issue with this type of 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 wider distribution of scores for true fit pairs for the HQ-HT, generally reduced the quality of fits, and therefore additional chemical analysis may be warranted even if a fit is observed. [4]

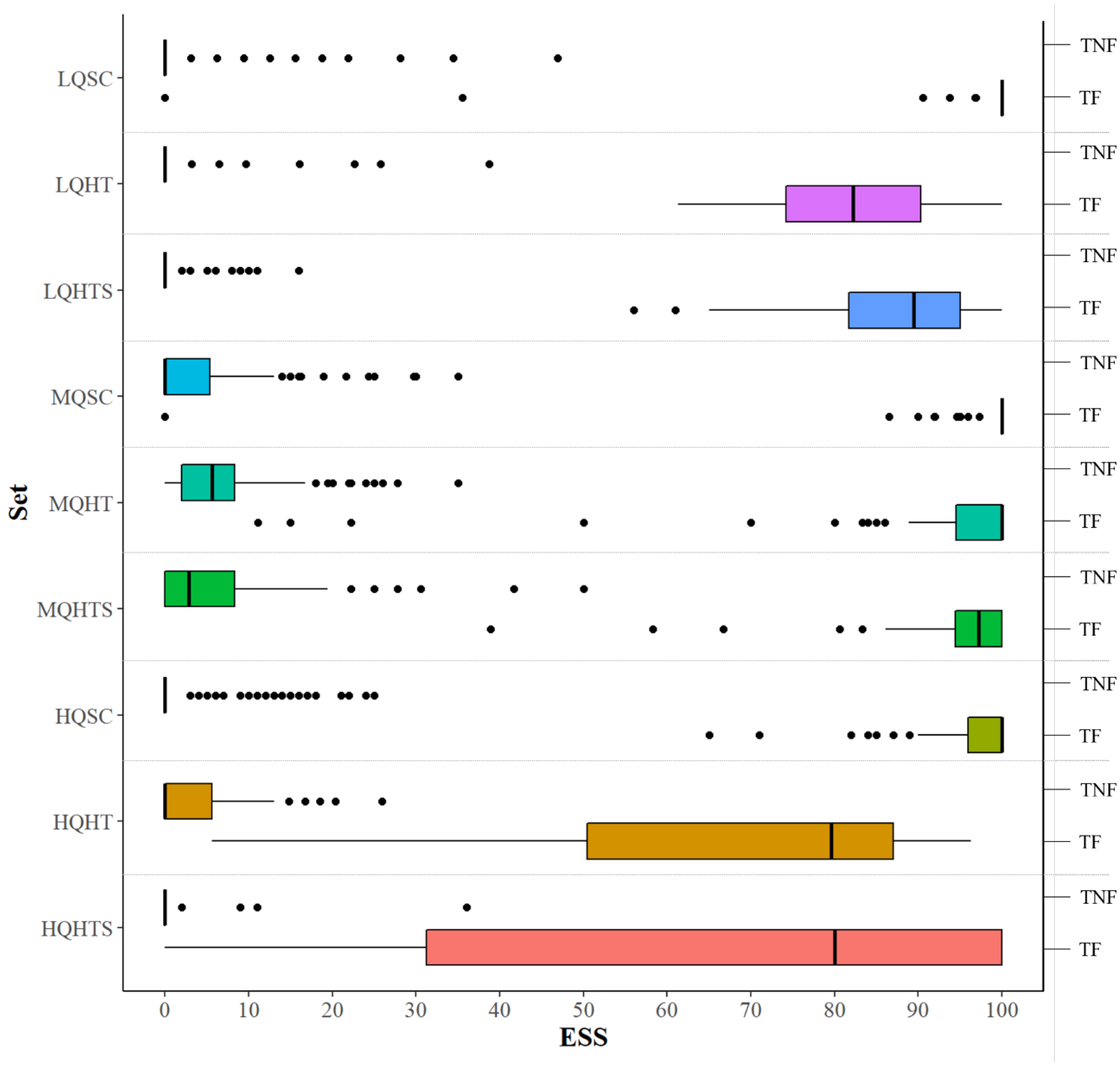


Figure 24. Boxplots showing the ESS distribution of each tape set, with true non-fits (TNF) shown on the top boxplot of each set and true-fits (TF) shown on the bottom boxplot of each set. Generally, there is a separation between the ESS distributions of the true fit and true non-fitting pairs, with higher scores for fits and lower scores for non-fits. [4]

When using the ESS criteria to form opinions, the method demonstrates good performance, with accuracies for all sets at approximately 98% or higher, except for the HQ-HT sets (80-85%). As seen in **Table 10**, the tape subsets in this study did not result in any false positive results. This is a critical finding as it provides scientific support to the general belief that tape separated items exhibit physical features that realign in a manner that is not expected to be replicated by chance. As anticipated from the boxplot ESS distributions, HQ-HT and HQ-HT-S are more prone to distortion or possess fewer distinctive features upon separation and tend to result in more false negative or inconclusive results.

***Table 10**. 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.* [4]

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 further evaluate factors that can affect the ESS scores and quality of a fit, we use a logistic regression generalized linear mixed effect model. This model is only meant to interpret the effect of the different factors through the interpretation of the regression coefficients, and not for predictive purposes. The coefficients of our model are estimated by considering the log-odds transform of the ESS for each comparison as the dependent variable, and 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. Additional details of this model can be found in Prusinowski et al.[4]

The results show that the effect of separation and the quality of the tape have varying effects depending on the ground truth. For example,

1) For non-fits, the ESS trend towards low values, and the influence of the separation method and quality of tape on the ESS values is negligible

2) For true fit pairs, scissor-cut tapes tend to result in higher scores in comparison to hand-torn pairs.

3) Regarding tape quality, in true fit pairs, medium-quality tapes tend to receive higher scores overall, followed by low-quality, and then high-quality.

The observed effect seen in the counterfactual plot shown in **Figure 25** coincides with the exploratory data analysis from the different tape sets. 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.[4]

Figure 25A 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, while the right side of the plot shows kernel density estimated distributions of the empirical ESS data from the analyzed tape pairs. The empirical results indicate that there are statistically different effects for the different levels of the grade factor, as the distributions for the coefficient values for the different grades of tape are very well separated around log odds "0". Nonetheless, the effect of tape grade factor is not particularly pronounced when accounting for the other factors and analyst variability.

When considering the separation method, **Figure 25B** 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.[4]

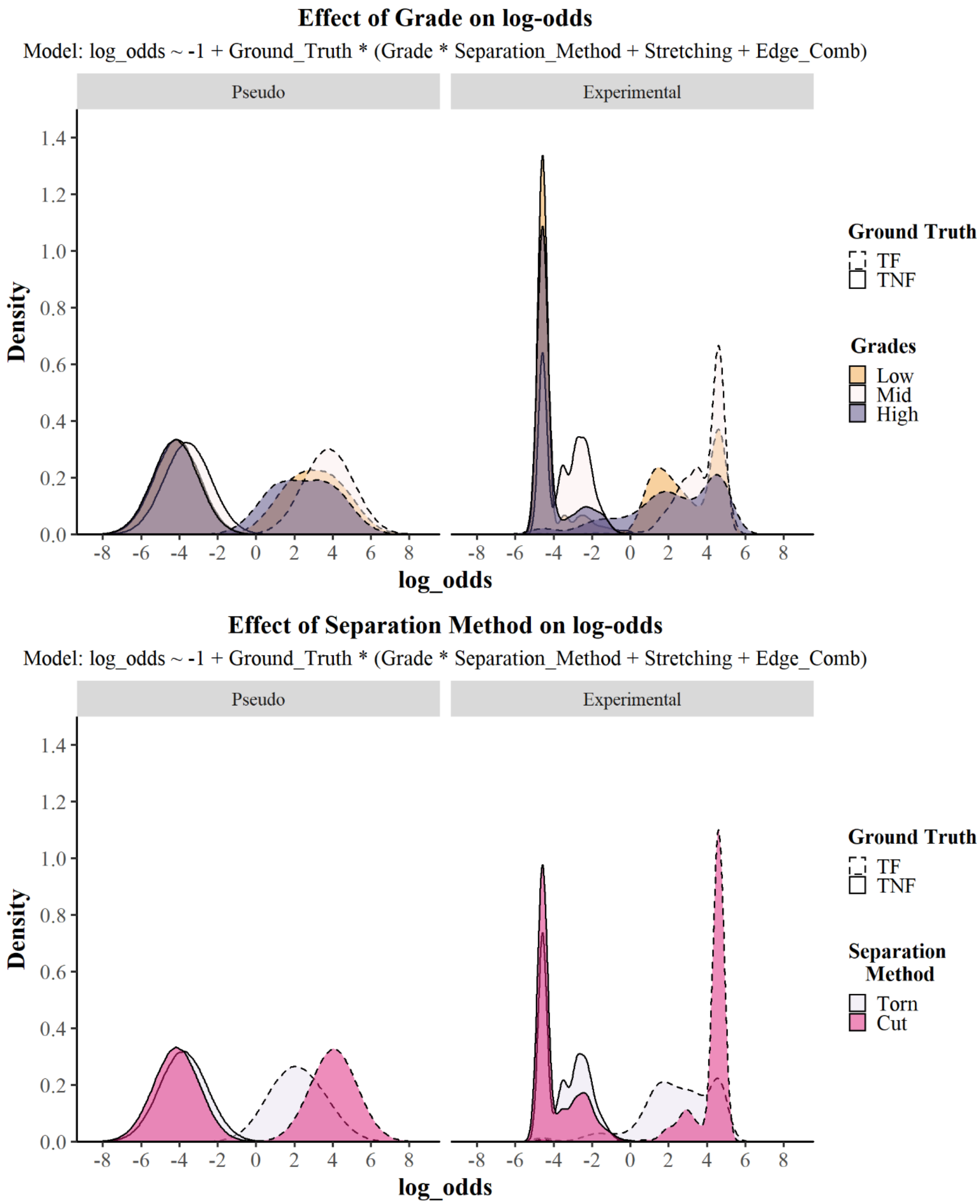


Figure 25. *Counterfactual plot showing the distributions of the expected ESS values for duct tape data. The counterfactual plot shows both the expected ESS values resulting from the model, as well as the experimental data.* ***Figure 25A*** *(top) shows the distributions marginalized for grade of tape. Medium-quality tape generally results in higher ESS than when other types of tape when samples truly originated from the same tape.* ***Figure 25B*** *(bottom) shows the distributions marginalized for 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 method do not seem to provide substantial differences in expected ESS values for true non-fit pairs.* [4]

Milestone 3—Practitioners' contributions: testing and fine-tuning through interlaboratory exercises.

The practitioner's feedback is one of the most critical stages in the assessment of a new method. Here, we conduct two interlaboratory studies to evaluate the performance of the method. A total of 266 pairs are examined by thirty-eight (38) participants across 23 laboratories. Each participant receives a kit with seven questioned-known tape pairs to conduct the physical examination and fit assessment. The participants' responses are compared to a consensus pre-distribution panel and to the mean of the participant's values via the Dunnett's test and z-score statistics, respectively. The results are very encouraging, with overall accuracies ranging from 90-100% and most ESS scores falling within 95% confidence intervals. This is very remarkable considering the participants are asked to use a new method with just a brief instruction session to familiarize themselves with the features terminology, examination method, interpretation criteria, and reporting protocols. We can deduct then, that the method is simple to understand and to apply by practitioners and that inter-examiner variability could be improved as further training is incorporated in technology transfer strategies. Indeed, the inter-participant agreement and accuracy improves from the first to the second study based on the depth of training provided. For example, the first test offers written instructions and an optional virtual session, while the second test requires attendance to a one-on-one virtual training session. A recording is also made available for review as needed.

Several improvements are made to the overall method based on practitioners' feedback. For instance, the process is split into three main defined steps, in the first step the analyst reviews the question sample first before seeing the known, which is an additional effort to minimize bias. The second step and third steps include additional auto-populated cell options to annotate the importance of each of the relevant features. Also, an inconclusive bin-score option is added to the template. Finally, the template incorporates a color-code that assists the user to visualize at a glance the fit quality by bin location.

Figure 26 shows an example of the display of the participants' responses for the second exercise using z-scores. Here, the z-scores are bracketed into three regions, z-scores below 2 are considered satisfactory meaning they agree with the study-mean within that interval, while z-scores between 2 and 3 are considered cautionary, and above 3 the results are insufficient. In this study, most responses were deemed satisfactory, with only 5 responses being cautionary and one insufficient for one participant for one of the samples only. Interestingly, this sample was the more complex fit, that was intended to be a lower confidence fit (i.e., 74±5 ESS by the consensus panel).[3]

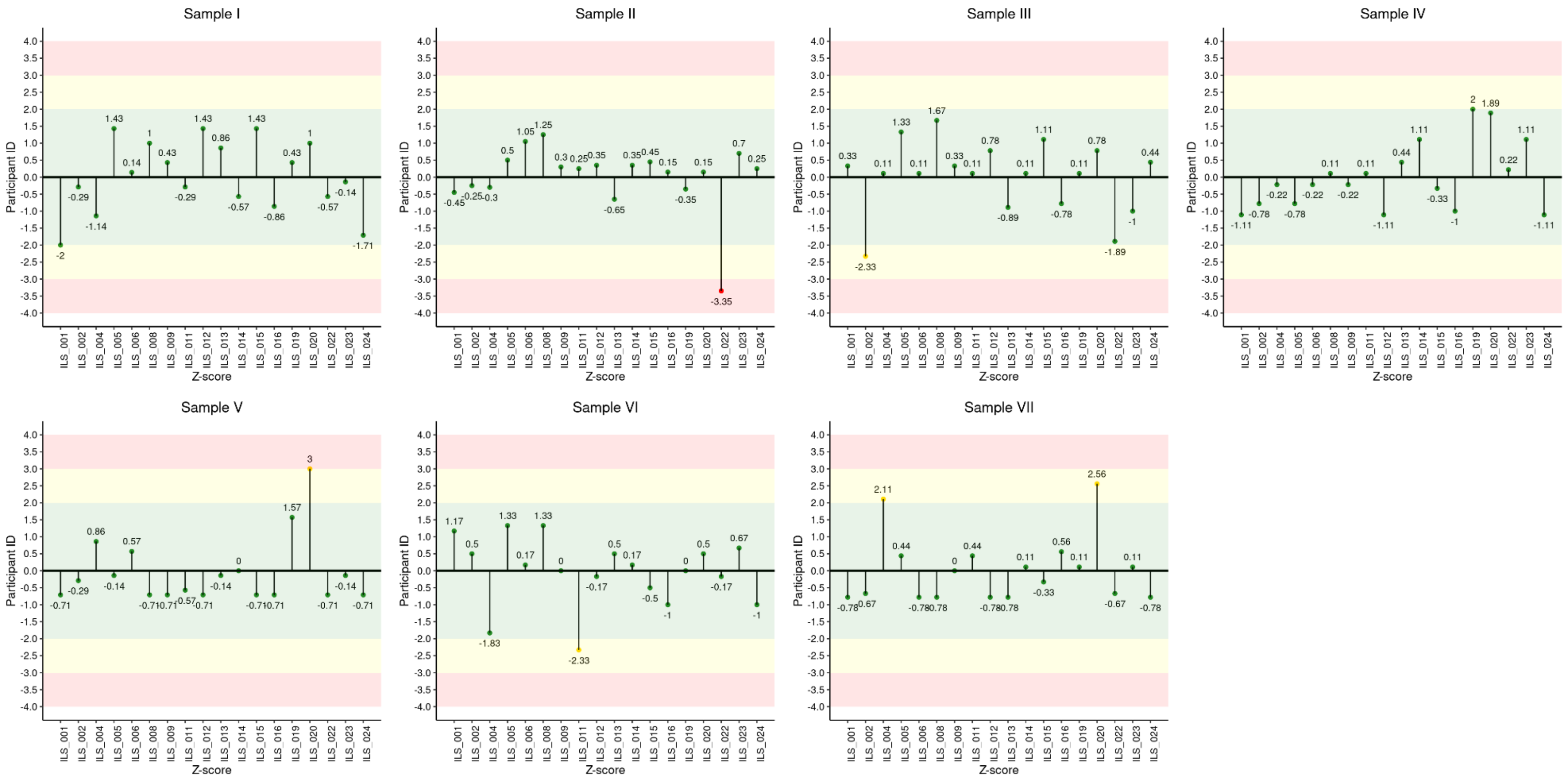


Figure 26. *Z-scores of the reported ESS values for ILS 2 for each participant. The participant IDs are independent of the IDs used in ILS 1. The z-values have been color-coded for visualization. Green bars are considered satisfactory, yellow bars are considered cautionary, and red bars are considered 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+).* [3]

To visualize the utility of the template documentation, **Figure 27** illustrates the color-coded bins for the full width of the compared edges for the more complex fit items by seven participants. It becomes evident, that most of the variability observed in the lower end is caused by edges distortion where various inconclusive bins were reported due to stretching. More extensive discussion of results can be found in the publication by Prusinowski et al. [3]

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

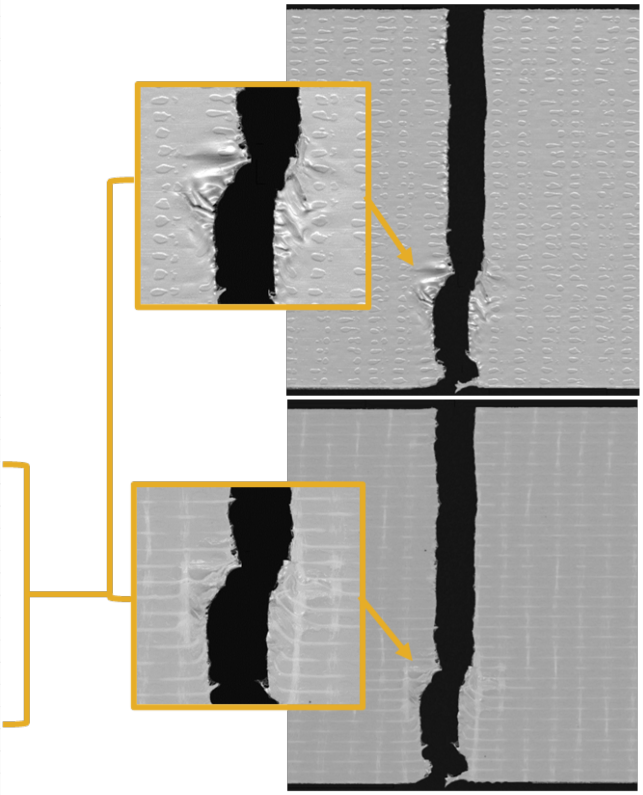


Figure 27. *Example of application of documentation template to a duct tape physical fit examination for the pair F-of Kit 1. The participants in this example have slightly different reported ESS, but the overall conclusion of fit is consistent, and most of the participants report the area of distortion consistently.* [3]

Milestone 4—Complementing human-based approaches with computational algorithms

Often machine learning is referred as computer systems that learn and adapt by using algorithms and statistical models to analyze and draw inferences from patterns in data. Here we loop the process and use machine learning results to "learn from the machine", as the results also provide a further understanding of the decision process in human-based fit comparisons.

ForensicFIT database and CNN approach

The study provides a computational platform for physical fit predictions that can assist analysts in their evaluations. We report the development of an open-source python package, ForensicFit[32], designed to pre-process images obtained for forensic physical fit examination. The package provides pre-processed data for machine learning to train two independent convolutional neural networks —

one on the backing side, and the other on the scrim side. Statistical analysis is performed on the resulting probabilities from the network outputs and the performance on known true-fits and non-fits sets is compared to the quantitative assessment of duct tapes using human-based approaches. High agreement is observed between both methods and therefore demonstrates the potential of machine learning models to provide statistical support to the analyst conclusions.

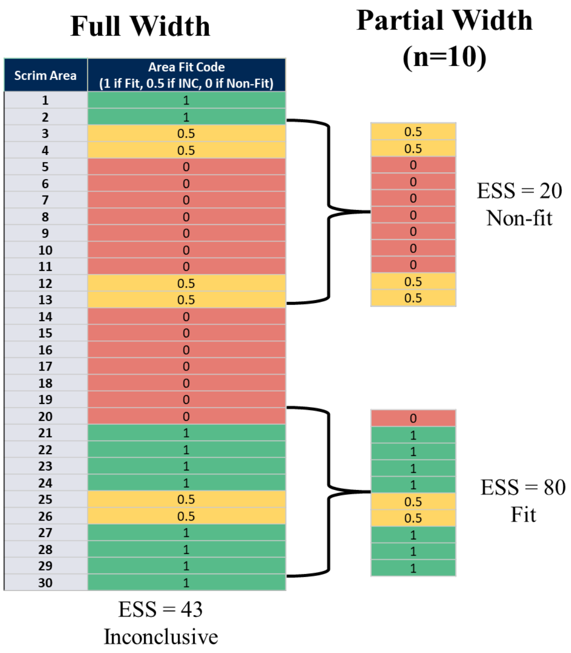
The main findings derived from this CNN study can be summarized as follows:

1) CNN have shown to be an effective mean to compare separated tape edges of various grade qualities and fracture methods using an automated imaging processing platform (ForensicFit),
2) The distribution of human-estimated metrics (ESS) and computer-based CNN-membership probabilities for ground truth fits and non-fits populations shows a minimal overlap between these groups
3) Human-estimated ESS and CNN-membership probabilities yield low rates of misleading evidence and provide a means to employ these metrics for statistic assessment of the probative value of the evidence
3) The boxplots and kernel distributions illustrate that the occurrence of error rates, mostly false negatives, is influenced by the method of separation and quality of the tape and that those effects are similarly captured by analyst-examination and by the computer-based feature recognition,
4) The Layer-wise Relevance Propagation (LRP) analysis can be used to understand the most critical features identified by the CNN and supplement decision criteria independently documented by the examiner.

Therefore, the results demonstrate the feasibility of using CNN to assist analysts to enhance objectivity in their fit examinations. Larger datasets are necessary to strengthen the capabilities and accuracy of the computational models. [47-49]

Algorithms for extracting and interpreting edge feature data for fit examinations using mutual information and decision trees.

This study uses mutual information and decision tree algorithms to support the analyst's decisions in physical fit examinations of duct tapes and textiles. The first question we are interested in answering comes from a request we received from practitioners during a feedback session. They often receive questioned items that have just a partial edge so in these cases, the whole fractured edge on the known item cannot be compared in its totality to the partial questioned item. For instance, only a small portion of a torn fabric or a partial tape piece with missing material is submitted for comparison. In these cases, the analyst must decide how small the questioned item could be before it is no longer suitable for a physical fit. Making those decisions without data-driven criteria is not optimal. Therefore, this research addresses this concern utilizing the data generated in the population set studies. 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. Since analyst records the bin-by-bin data, we can use that information to randomly remove consecutive regions of the edge comparison to simulate partial edges as illustrated in **Figure 28**. In this example, two partial samples of 10 bins each is shown resulting in ESS of 20 and 80, as compared to the ESS of 43 when the complete 30 bins are available. This shows potential risks of misidentifications when partial samples are evaluated.



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
0.5
0.5

0
1
1
1
1
0.5
0.5
1
1
1

Figure 28*. 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).* [9]

Only data collected using the latest template versions is utilized in this study to ensure consistency and minimize variability. 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 smaller sample size available for comparison (see **figure 29**). When considering the high-quality tape 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. For textiles, the hand-torn accuracy suffers more than the stabbed items when decreasing the percent of sample available. 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.

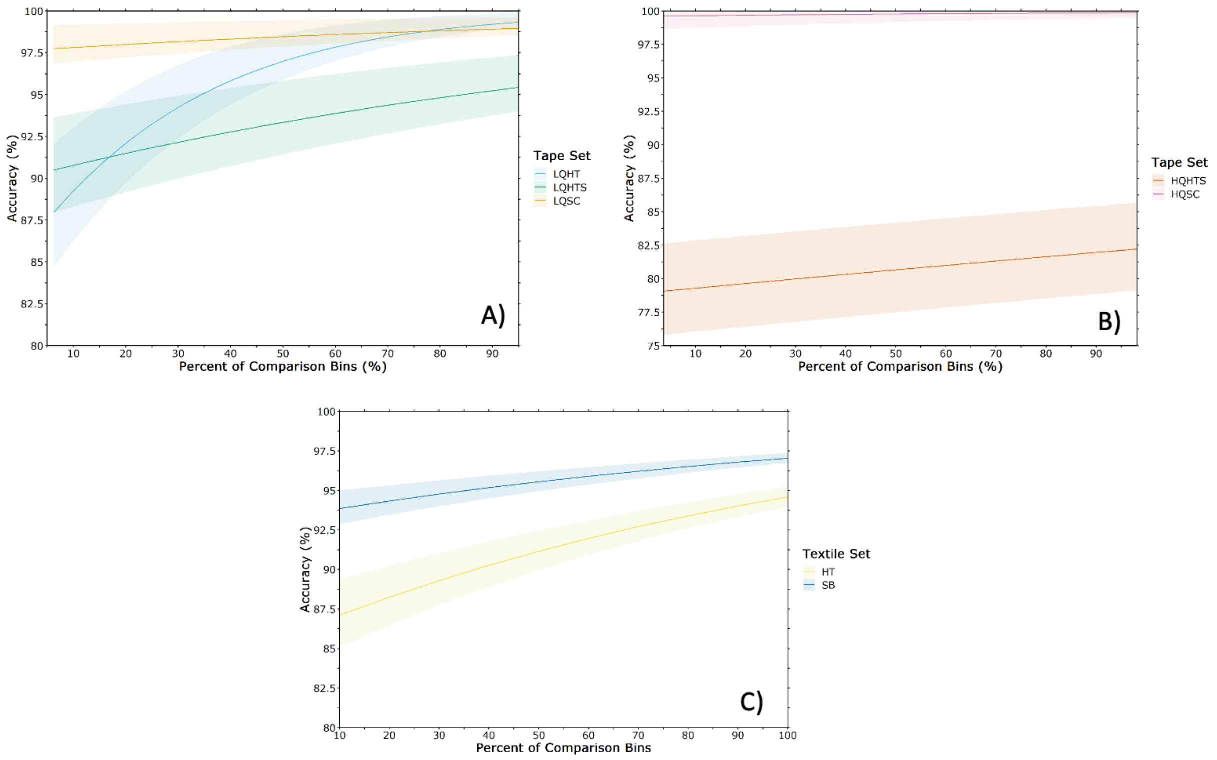


Figure 29. *A) Accuracy of the ESS method as applied to partial widths of the low-quality tape subsets. B) Accuracy of the ESS method as applied to partial widths of the high-quality tape subsets. C) Accuracy of the ESS method as applied to partial widths of the textile samples. HT represents hand-torn samples, while SB represents stabbed samples. The x-axis represents the percent of comparison bins and the y-axis the observed accuracy with the respective uncertainty intervals. Adapted from rom* [9]

The second aspect we address here is what features hold more weight in the analyst's bin decisions. Here, 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 support and add objectivity to the analysts' conclusions.

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 (**Figure 30**). 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/design and construction alignment hold more value for stabbed samples, while the yarn alignment is more informative in hand-torn samples (**Figure 31**).[9] This information provides, for the first time, a more tangible understanding of the relative importance that these features have in a fit or non-fit determination.

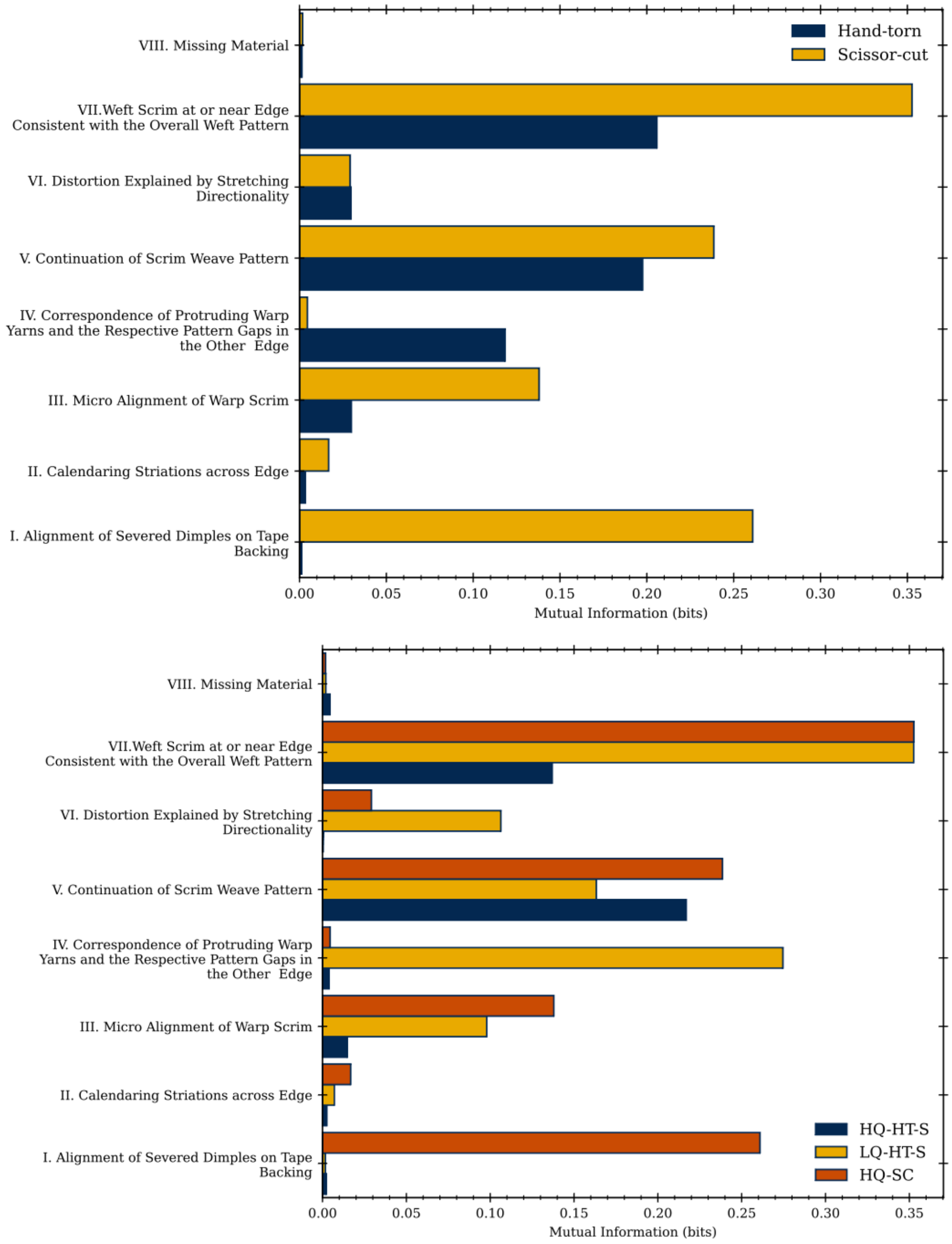


Figure 30. *Barplots representing the mutual information of the tape features by separation method (top) and by sample subset (bottom). The larger the bar, the more value the feature has for comparisons.* [7]

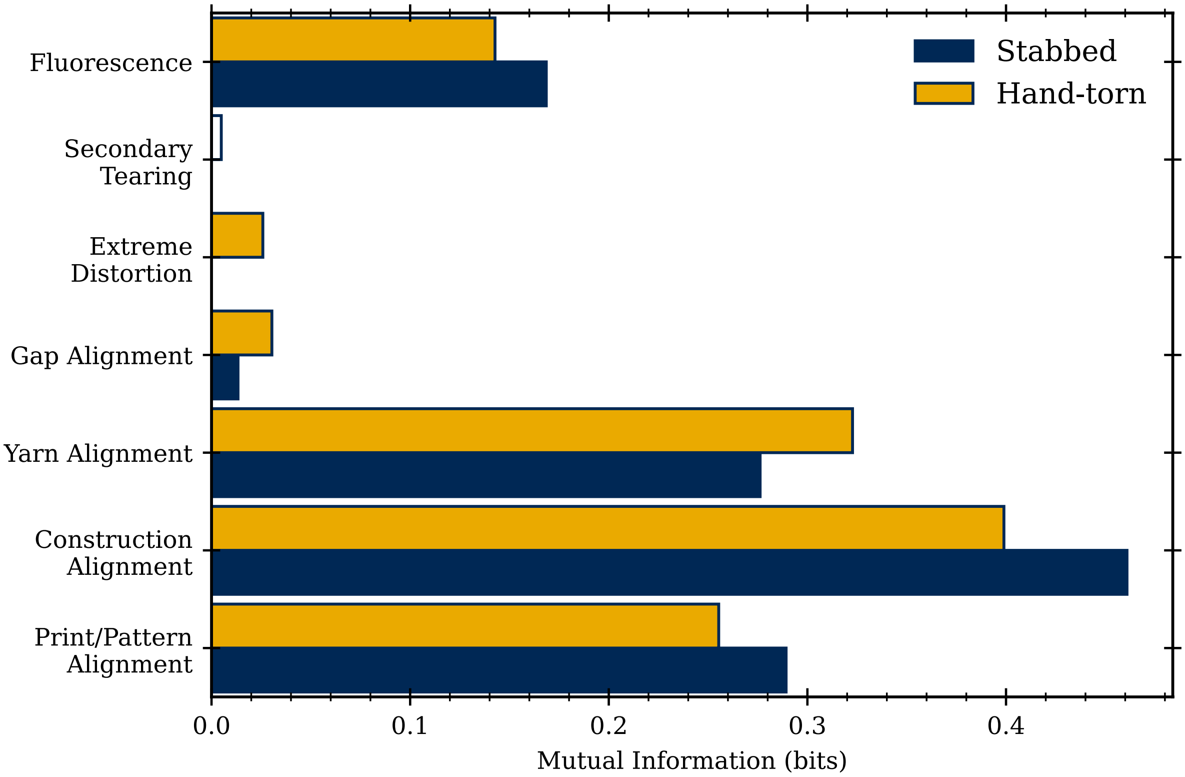
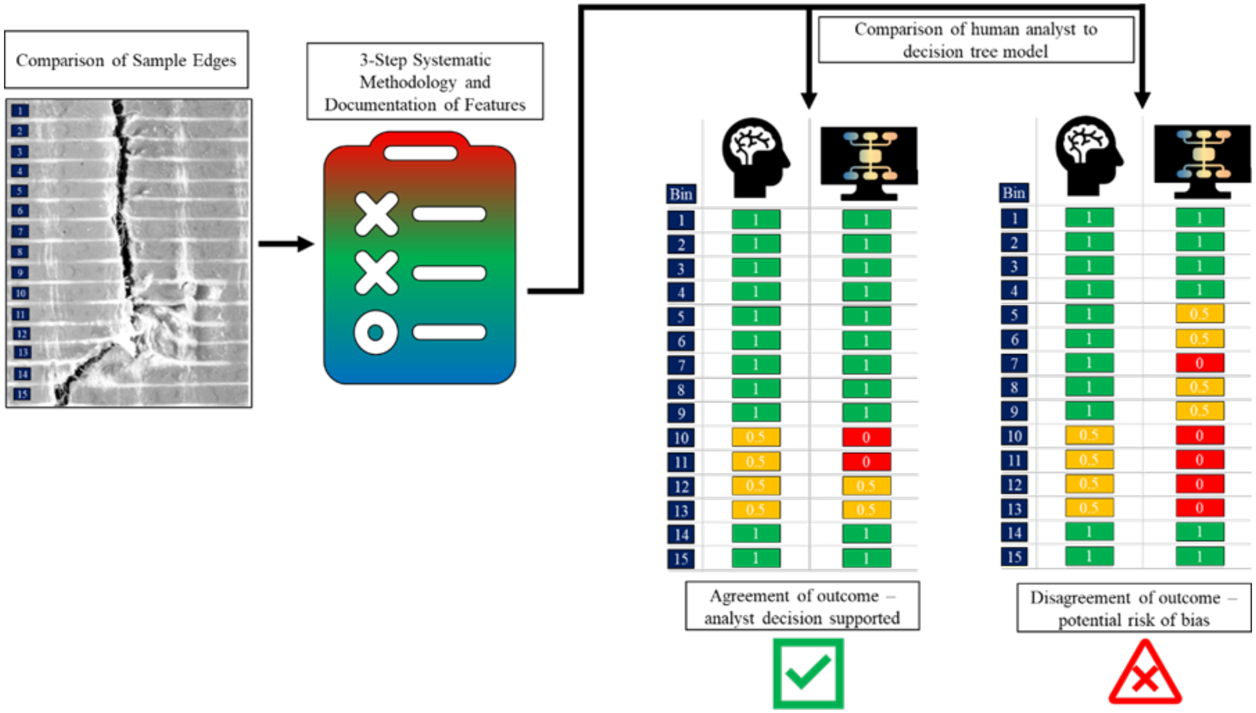


Figure 31. *Barplots represent the mutual information of the textile features by separation method. The larger the bar, the more value the feature has for comparisons* [7]

This importance feature 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. **Figure 32** illustrates this process of comparing human-based results to those of the decision tree to support the examiners opinion.

Figures 33 and 34, show the level of agreement between the computational decision and the human approach. [9] These results indicate that:

1) The decision tree model shows significant potential as a tool to help in the decision-making process for physical fit comparisons.
2) While caution is needed regarding the chance of false identifications, if used in tandem with human-based analysis, the tool could help identify samples where a further examination is recommended if the model outcome disagrees with the analyst.
3) 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.
4) Notably, the algorithm removes the judgment from the decision process and minimizes the risk of bias from the prior information from the analyst's observations.


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 32. 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.[9]

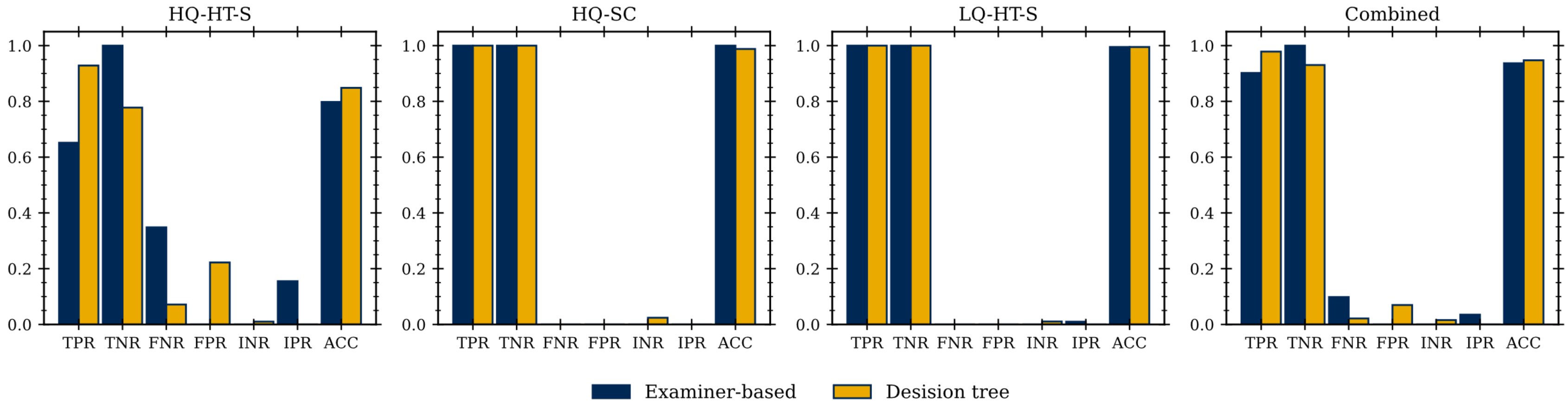


Figure 33. Performance of the decision tree against the human analysis for each duct tape sample set. The performance rates included are the true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR), inconclusive rates for both true non-fits (INR) and true fits (IPR), and the accuracy (ACC). [7,9]

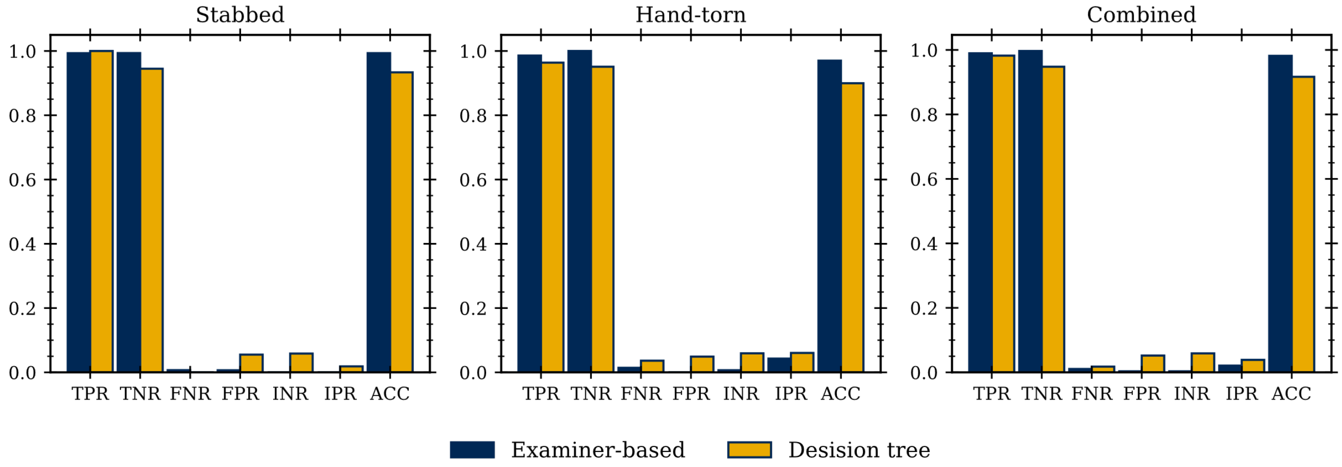


Figure 34. Performance of the decision tree against the human analysis for each textile sample set. The performance rates included are the true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), false positive rate (FPR), inconclusive rates for both true non-fits (INR) and true fits (IPR), and the accuracy (ACC). [9]

2.2.3. Textiles physical fit method

For textiles, we apply the main lessons learned from the physical fit examinations of duct tapes with modifications that are necessary to adapt to the inherent factors that influence the fabric's fractures. Like tape, four main milestones were critical to building knowledge in these materials.

Milestone 1—Method development and optimization of standardized criteria

Much less literature is available about fabrics' fracture fit examinations than tapes. [5, 50-53] Therefore, we compiled information from practitioners' protocols and evaluated the occurrence of fracture features from pilot datasets analyzed by various analysts. These pilot datasets considered the main factors that can play a role in the features imparted, such as separation method (stabbed, torn), compositions (cotton, polyester, rayon, or mixed), color design (unicolor or multicolor), and construction (knit, weave). Although many other factors can be evaluated, we use these four to represent fabric types that are expected in casework. Then, after analysis of the data, we narrow them down to seven features that show to be most informative. These features are listed in **Table 11,** along with a description and the options provided in the reporting template drop-down auto-populated cells to document how that feature influences the bin-decision.

Some of these features are **intrinsic to the fabric** itself and can show continuity across the fractured items when they were once a single object. This includes design alignment, construction alignment, and fiber fluorescence. [8]

1) **Design alignment** refers to the agreement and alignment of fabric patterns, also referred to as multicolor, across the comparison edge of two samples. An example of this feature is the alignment of the camouflage design shown in the table below. These can hold a high influence on the fit or non-fit decision.

2) **Construction alignment** occurs when there is an agreement between two samples regarding the type and direction of the construction of the weave and weft yarns, in addition to the consistency of yarn or thread count in each comparison area. Two true fitting samples woven in a diagonal direction relative to the comparison edge and each possessing the same number of yarns in the comparison area would possess this feature. The alignment of these features typically increases an analyst's confidence in the presence of a physical fit, as they indicate that the two samples could have once been joined together.

3) The **fluorescence** of individual yarns in the fabric can also aid in identifying a physical fit. However, fluorescence is rarely the determining factor for fits or non-fits, as fibers originating from the same common clothing source will likely demonstrate similar fluorescence regardless of the location of the fracture. [8]

Other features are **extrinsic to the fabric and caused by the separation event.**[8] These features include edge alignment, yarn alignment, extreme distortion, and secondary tearing.

4) **Edge alignment** denotes the alignment of the overall edge morphology between two samples. The three common edge morphologies that are identified in this study are straight, wavy, and puzzle-like edges. The overall edge morphology must align between the two fragments for a physical fit to occur, and yarn count consistency is also part of this feature.

5) **Yarn alignment** refers to the alignment of loose yarns that have been pulled out of the fractured edge of a sample. This has been observed to be much more common in hand-torn samples, which are subject to vigorous pulling and tearing. Stabbed samples generally do not demonstrate the same degree of loose yarns.

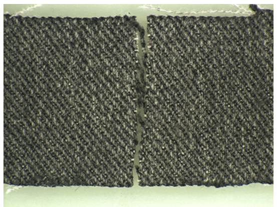
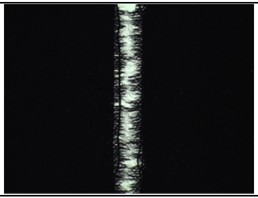
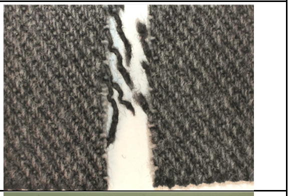
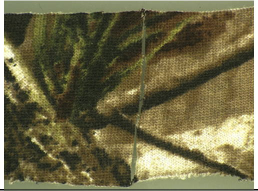
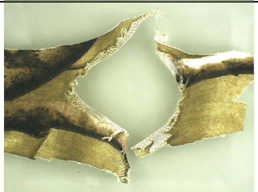
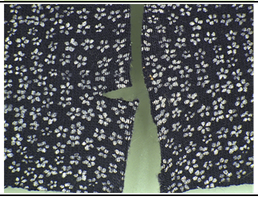
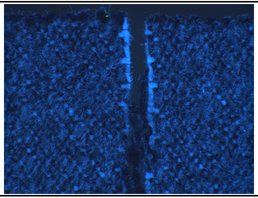
6) **Extreme distortion** is caused by stretching and missing material and can hinder other relevant features in the fractured edges. Hand-torn samples are also much more likely to exhibit extreme distortion.

7) **Secondary tearing** describes a minor fracture, often perpendicular to the comparison edge, that is not the primary fracture of interest between two samples. This feature may cause a "non-fit" conclusion for a given bin, as the fracture will most likely only be present on one edge.

Some features are more common than others. Construction alignment, for example, is applicable for all comparisons from the same clothing article, and statements about design alignment can be made for all cases involving multicolor fabric. On the other hand, secondary tearing is rare. This information is now available to provide a starting point towards the standardization of terminology, distinctive features, and defined criteria on how to use them on fit examinations.

With these features, the reporting template is modified for textiles using a simplified two-step method, in which the first step corresponds to an overall assessment of the edges, and the second step conducts macro and microscopic examination of ten defined bins equally separated across the length of the fracture.

***Table 11**. Table of features extracted for textile comparisons, including a description of the feature, an image of the feature, and the response options available to select for each feature.*[11]

Feature	Description	Options	Image
I. Construction Alignment	Consistency and construction alignment, 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	
II. 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	
III. 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	
IV. 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	
V. 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	
VI. 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	
VII. Fluorescence	Fluorescence of individual yarns can aid in the identification of a physical fit.	• Consistent • Inconclusive • Inconsistent • Cannot be assessed	

Milestone 2—Creation of textiles dataset and validation

The textile population dataset used in this study consists of 967 textile fit comparisons from the examination of 774 paired items by one analyst, and a subset of 193 of those items compared by a second independent analyst. These sets contain known true-fits and true-non-fits that allow the assessment of performance rates as explained in the tape section. **Figure 15** illustrates the main subsets and respective studies.

The main overall findings for textiles are:

1) Not all the textiles' fits hold the same value.
2) Not all textiles are suitable for fit examinations. For instance, knit-polyester fabrics yield unacceptable accuracy and therefore fit examinations are not recommended for these types of textiles.
3) The separation method and the construction and composition of fabrics do not have a significant effect in the observed ESS. However, the combination of some of these factors critically influence the suitability for examinations.
4) Although accuracy of textile fit examinations is relatively high, the occurrence of false positives is possible in textiles, something not observed in duct tapes datasets.
5) Given a physical fit's probative value, and the observed experimental error rates, we recommend reporting a fit only for ESS scores 80 or above for textile materials. For lower scores (80-40), we recommend reporting a non-fit and submitting the items for chemical and physical textile/fiber comparisons, if appropriate. Scores below 40 are reported as non fits, and therefore no further analysis are required.
6) High agreement is observed between analysis in the database population sets and by interlaboratory exercises, indicating the ESS method and reporting criteria can be used effectively to create consensus-based results in textile fit examinations.

One remarkable finding during the validation stage is that not all textiles are suitable for textile fit examinations. For instance, as some fabrics such as jersey knit polyester are more prone to produce an unreasonably large number of misclassifications. This is illustrated in the analysts' performance of the initial preliminary textile set, where almost two-thirds of the total true-fitting pairs are misclassified as non-fits by the two analysts. As soon as the fabric was cut or torn, the edges curl, fibers are missing on the edges, and the construction stretches. The distortion was too overwhelming to arrive at valuable results. However, without knowing the ground truth of these samples, those distortions may not be obvious to the analyst. Moreover, regarding the overall fit or non-fit conclusion, the analysts disagree on almost one-third of the comparisons.

Among the sample sets investigated in this study, only 100% polyester knit sets led to these suitability issues. However, there may be other fabric compositions and constructions not evaluated here that could as well be unsuitable for fit examinations. Thus, it is recommended to first assess the fabric distortion level. If the items are deemed unsuitable for a physical fit examination, the textiles must instead be considered for other chemical and physical comparisons.

In addition to the method's performance rates, the inter- and intra-analyst variation is investigated. For the inter-analysts, at least two analysts separately analyzed the same subsets. We also introduce a blind intra-analyst test, where the analyst was given the same subset, but randomly re-organized and relabeled. The analyst receives this duplicate set several months later with the assumption he was receiving a new subset to minimize potential bias.

Tables 12 to 17 summarize the performance results between different analysts (**table 12**), the same analyst (**Table 13**), the knit unicolor (**Table 14**) and multicolor (**Table 15**), and the weave unicolor (**Table 16**) and multicolor (**Table 17**). Each of these subsets includes error rates for stabbed and hand-torn separations. Accuracy ranged from 87 to 100% depending on the set.

Table 12. *Performance rates for the inter-analyst variability textile set* [11]

Stabbed Subset	**Analyst 1 Reported Fit**	**Analyst 1 Reported Non-Fit**	**Analyst 1 Reported Inconclusive**	**Analyst 2 Reported Fit**	**Analyst 2 Reported Non-Fit**	**Analyst 2 Reported Inconclusive**
True Fit	25 of 26 (96% True Positive)	1 of 26 (4% False Negative)	0 of 26 (0% Inconclusive)	23 of 26 (88% True Positive)	2 of 26 (8% False Negative)	1 of 26 (4% Inconclusive)
True Non-Fit	3 of 24 (12% False Positive)	21 of 24 (88% True Negative)	0 of 24 (0% Inconclusive)	2 of 24 (8% False Positive)	21 of 24 (88% True Negative)	1 of 24 (4% Inconclusive)
Accuracy	92%			88%		
Hand-torn Subset	**Analyst 1 Reported Fit**	**Analyst 1 Reported Non-Fit**	**Analyst 1 Reported Inconclusive**	**Analyst 2 Reported Fit**	**Analyst 2 Reported Non-Fit**	**Analyst 2 Reported Inconclusive**
True Fit	25 of 26 (96% True Positive)	1 of 26 (4% False Negative)	0 of 26 (0% Inconclusive)	21 of 26 (81% True Positive)	3 of 26 (11% False Negative)	2 of 26 (8% Inconclusive)
True Non-Fit	0 of 24 (0% False Positive)	24 of 24 (100% True Negative)	0 of 24 (0% Inconclusive)	1 of 24 (4% False Positive)	23 of 24 (96% True Negative)	0 of 24 (0% Inconclusive)
Accuracy	98%			88%		

Table 13. Performance rates for the intra-analyst variability textile set[11]

Stabbed Subset	Replicate 1 Reported Fit	Replicate 1 Reported Non-Fit	Replicate 1 Reported Inconclusive	Replicate 2 Reported Fit	Replicate 2 Reported Non-Fit	Replicate 2 Reported Inconclusive
True Fit	20 of 23 (87% True Positive)	2 of 23 (9% False Negative)	1 of 23 (4% Inconclusive)	22 of 23 (96% True Positive)	1 of 23 (4% False Negative)	0 of 23 (0% Inconclusive)
True Non-Fit	2 of 24 (8% False Positive)	21 of 24 (88% True Negative)	1 of 24 (4% Inconclusive)	1 of 24 (4% False Positive)	22 of 24 (92% True Negative)	1 of 24 (4% Inconclusive)
Accuracy	87%			94%		
Hand-torn Subset	**Replicate 1 Reported Fit**	**Replicate 1 Reported Non-Fit**	**Replicate 1 Reported Inconclusive**	**Replicate 2 Reported Fit**	**Replicate 2 Reported Non-Fit**	**Replicate 2 Reported Inconclusive**
True Fit	19 of 23 (82% True Positive)	2 of 23 (9% False Negative)	2 of 23 (9% Inconclusive)	23 of 23 (100% True Positive)	0 of 23 (0% False Negative)	0 of 23 (0% Inconclusive)
True Non-Fit	0 of 24 (0% False Positive)	24 of 24 (100% True Negative)	0 of 23 (0% Inconclusive)	2 of 23 (9% False Positive)	21 of 23 (91% True Negative)	0 of 23 (0% Inconclusive)
Accuracy	89%			96%		

Table 14. Performance rates for the unicolor knit textile set[11]

Unicolor Knit Stabbed Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive
True Fit	30 of 30 (100% True Positive)	0 of 30 (0% False Negative)	0 of 30 (0% Inconclusive)
True Non-Fit	0 of 31 (0% False Positive)	31 of 31 (100% True Negative)	0 of 31 (0% Inconclusive)
Accuracy	100%		
Unicolor Knit Hand-torn Textile Set	**Reported Fit**	**Reported Non-Fit**	**Reported Inconclusive**
True Fit	29 of 30 (97% True Positive)	0 of 30 (0% False Negative)	1 of 30 (3% Inconclusive)
True Non-Fit	0 of 29 (0% False Positive)	29 of 29 (100% True Negative)	0 of 29 (0% Inconclusive)
Accuracy	98%		

Table 15. Performance rates for the multicolor knit textile set[11]

Multicolor Knit Stabbed Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive
True Fit	20 of 20 (100% True Positive)	0 of 20 (0% False Negative)	0 of 20 (0% Inconclusive)
True Non-Fit	0 of 20 (0% False Positive)	20 of 20 (100% True Negative)	0 of 20 (0% Inconclusive)
Accuracy	100%		
Multicolor Knit Hand-torn Textile Set	**Reported Fit**	**Reported Non-Fit**	**Reported Inconclusive**
True Fit	19 of 20 (95% True Positive)	1 of 20 (5% False Negative)	0 of 20 (0% Inconclusive)
True Non-Fit	0 of 20 (0% False Positive)	20 of 20 (100% True Negative)	0 of 20 (0% Inconclusive)
Accuracy	98%		

Table 16. *Performance rates for the unicolor weave textile set*[11]

Unicolor Weave Stabbed Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive
True Fit	50 of 51 (98% True Positive)	1 of 51 (2% False Negative)	0 of 51 (0% Inconclusive)
True Non-Fit	0 of 49 (0% False Positive)	49 of 49 (0% True Negative)	0 of 49 (0% Inconclusive)
Accuracy	99%		
Unicolor Weave Hand-Torn Textile Set	**Reported Fit**	**Reported Non-Fit**	**Reported Inconclusive**
True Fit	48 of 49 (98% True Positive)	1 of 49 (2% False Negative)	0 of 49 (0% Inconclusive)
True Non-Fit	0 of 51 (0% False Positive)	50 of 51 (98% True Negative)	1 of 51 (2% Inconclusive)
Accuracy	98%		

Table 17. Performance rates for the multicolor weave textile set[11]

Multicolor Weave Stabbed Textile Set	Reported Fit	Reported Non-Fit	Reported Inconclusive
True Fit	47 of 47 (100% True Positive)	0 of 47 (0% False Negative)	0 of 47 (0% Inconclusive)
True Non-Fit	1 of 53 (2% False Positive)	52 of 53 (98% True Negative)	0 of 53 (0% Inconclusive)
Accuracy	99%		
Multicolor Weave Hand-torn Textile Set	**Reported Fit**	**Reported Non-Fit**	**Reported Inconclusive**
True Fit	48 of 48 (100% True Positive)	0 of 48 (0% False Negative)	0 of 48 (0% Inconclusive)
True Non-Fit	0 of 52 (0% False Positive)	52 of 52 (100% True Negative)	0 of 52 (0% Inconclusive)
Accuracy	100%		

Most misleading rates in the experimental datasets originate from false negative or inconclusive results on true fits. However, unlike tapes, textiles can be more prone to false positives. When false positives are observed, they range from 2 to up to 9%. **Figure 35** shows two examples of false positive comparisons, Fabric ID 3, denim and Fabric ID 18, 100% cotton. The comparison of the blue denim sample 3 is assigned an edge similarity score of 70 by the analyst, who notes construction alignment as a particularly influential feature for this comparison, leading to a false positive. The multicolor cotton pair #18 receives an edge similarity score between 60 and 70. Both of these cases are reported as weak fits. In sample 18, the magnification boxes highlight two interesting areas of multicolored design alignment, one of the critical features cited by the analyst as influential in their assessment. The analyst also notes the consistency in the weave direction as another feature of influence. There is a slight difference in edge shape in this comparison, however. These false positives bring an important example of the utility of using the ESS to inform the analyst's opinion. The population data shows that an ESS score below 80 does not provide strong support for a fit; therefore, an F- was reported by the analyst in these samples. In a real case, this can be clearly expressed when using the proposed criteria (ESS, or SLR) and prevent misleading evidence. Here, for purposes of the performance rates evaluation, we used the threshold of 0-40 (non-fit), 40-60 (inconclusive), and 60-100 (fit) to encompass worst-case scenarios. However, given a physical fit's probative value, and the observed experimental error rates, we recommend a more conservative approach and report a fit only for scores 80 or above for textile materials.

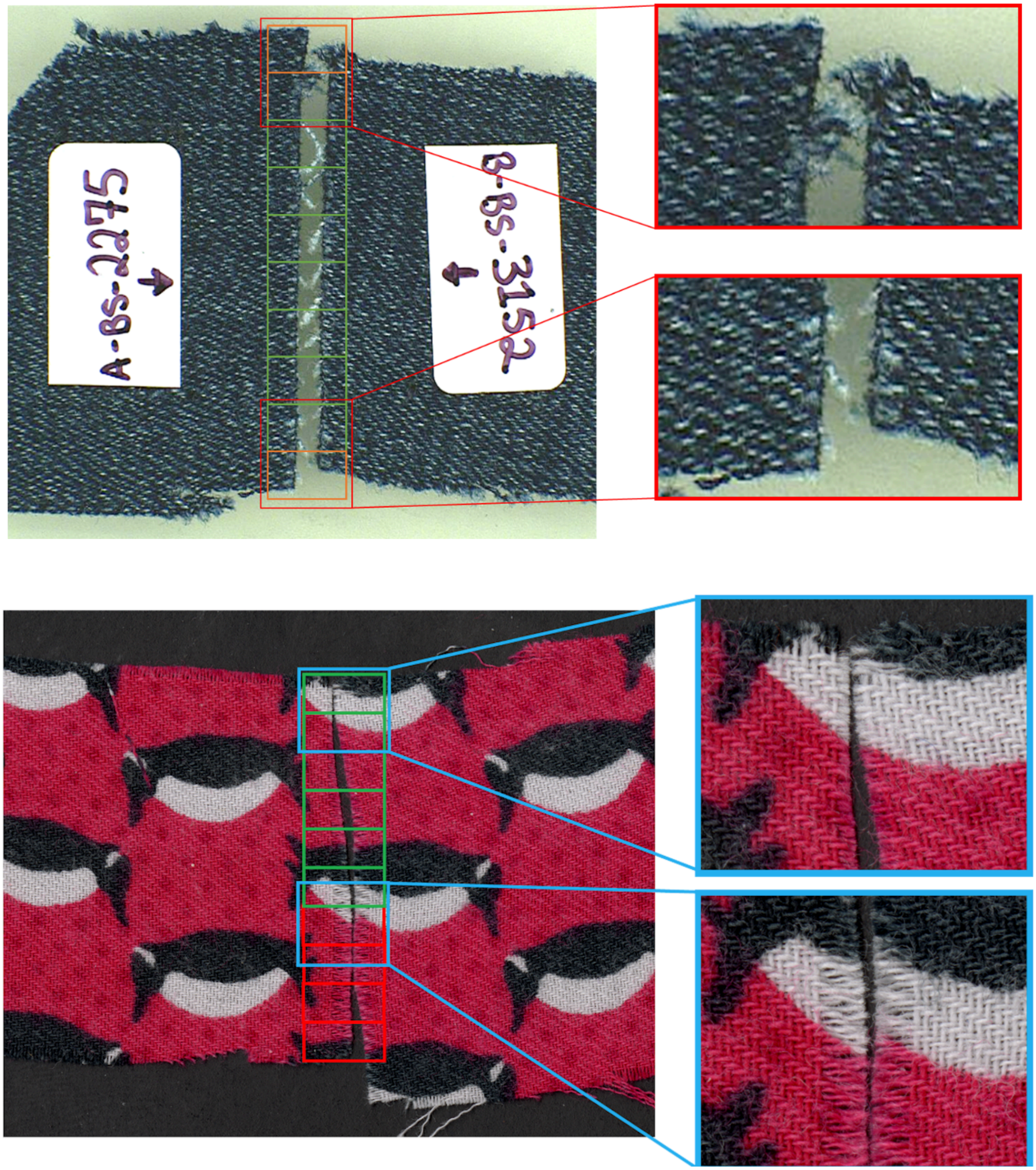


***Figure 35**. Top: Example of a true non-fitting comparison classified as a "fit" by the analyst for replicate 2 (Fabric ID 3, denim). Areas outlined in green were considered a fit for replicate 2 only, while areas outlined in orange were classified as a non-fit for both replicates. Areas of interest are showcased in red magnification boxes. Bottom: False-positive comparison of a non-fit pair identified as a fit (Fabric ID 18, 100% cotton).* (Adapted from Andrews et al. [11])

When considering each of the subsets and potential effects on the ESS and the quality of a physical fit, it is difficult to make any inferences using performance rates solely, as the rates are similar between subsets. However, differences appear when the edge similarity scores are analyzed more closely using boxplots and logistic regression.

Figure 36 shows the spread of the scores for true fits and true non-fits for each subset. Overall, true fits appear to produce a broader range of scores (70-100) than true non-fits, which cluster at lower scores (0-10). This indicates that the analyst was more comfortable classifying a comparison as a non-fit. In contrast, for true fits, certain features influenced the score assigned to the physical fit identified by the analyst, producing a wider range of scores.

When considering specific features that could affect the ESS, the separation method is a prominent one. Interestingly, hand-torn pairs presented more variability of scores than stabbed pairs from the same subset, likely due to distortions. However, stabbed comparisons can produce more false-positive classifications than torn comparisons, as the stabbing mechanism produces less distinctive edge patterns than the tearing process, which may make the identification of a non-fit slightly more difficult in some cases.

Fabric construction is another element of interest for a physical fit. A knit fabric generally produces lower scores than woven fabric because the knit fabric is more likely to unravel, stretch, and deform Indeed, the average score for the hand-torn multicolor knit fabric was approximately 70, while the average score for the hand-torn unicolor knit fabric was about 90. On the other hand, the woven fabric may be prone to producing false positives at a higher rate than knit fabric. [11]

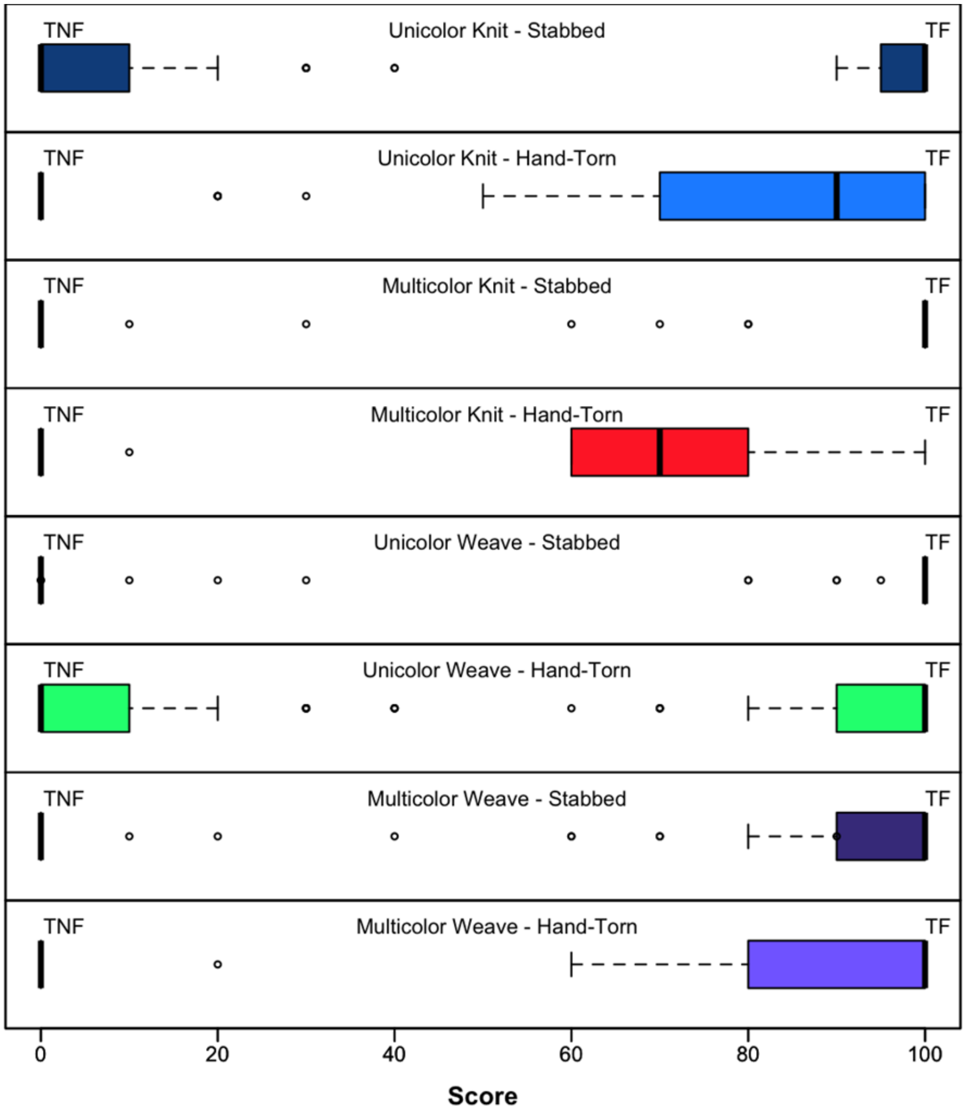


Figure 36. *Boxplots showing the distribution of edge similarity scores for each subset of textile comparisons. Scores for true non-fits are shown on the left, and scores for true fits are shown on the right.*[11]

A logistic regression model is also used to complement boxplots information and show the effect of each factor on the resulting edge similarity score. After evaluating several possible models, one was selected that includes an interaction between construction and separation method as this potential interaction is observed in the experimental data.

While analysts must consider the composition of the fabric when conducting physical fit comparisons, the logistic regression model shows that varying factors, such as separation method and construction of the fabric, do not have a substantial effect on the ESS used as an indicator of the quality of a physical fit. **Figure 37** illustrates the effect on textiles shown in the counterfactual plots is minimal, as compared to the effects on tapes, previously discussed.

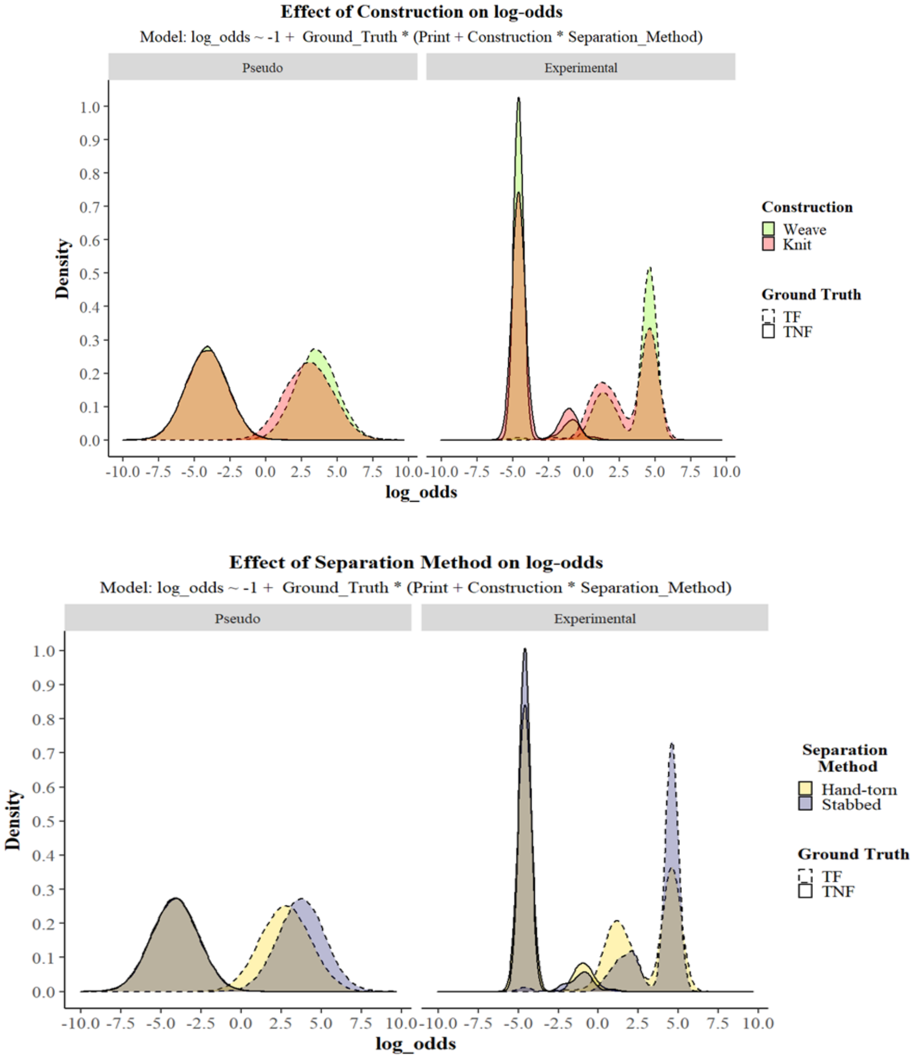


***Figure 37*. *Top*.** *Counterfactual plot demonstrating the effect of construction (weave or knit) on edge similarity scores. True fits (TF) are presented in dotted lines, while true non-fits (TNF) are in solid lines. Bottom: Counterfactual plot demonstrating the effect of separation method (hand-torn or stabbed) on edge similarity scores. True fits (TF) are presented in dotted lines, while true non-fits (TNF) are in solid lines.* [11]

Finally, score-based likelihood ratios (SLRs) are calculated from ESS of the population sets as a proxy for the probative value of the evidence when compared to a relevant population. Because the logistic regression model did not demonstrate that any of the tested factors substantially affected the similarity scores on textile edges, the knit and woven hand-torn and stabbed subsets are combined to evaluate the textile set as a whole. **Figure 38** shows the distribution of the SLR, represented in the log scale, where positive log SLR values at a specific ESS value indicate that the score provides support for a fit conclusion, with stronger support as the larger log SLR is. For example, scores above 60 provide some support for a fit (log SLR 0 to 1, or SLR 1 to 10), but ESS values higher than 80 provide strong support for a fit conclusion (log SLR ranging from ~1 to 2.7, or SLR 10 to 500). Conversely, negative log SLR values observed at a specific ESS value indicate that the score provides support for a non-fit conclusion. In this dataset, scores below 10 support a non-fit, with log SLR ranging from 0 to -3 (SLR

1 to 0.001). An edge similarity score of 0 results in a log SLR of approximately -3, which indicates that observing a score of 0 is about 1000 times more likely if the pieces were not once joined than if they were once part of the same object. On the other hand, a score of 100 results in a log SLR of approximately 2.7, which indicates that a score of 100 is about 500 times more likely if the pieces were once joined than if they were not part of the same object. It is important to note that because the dataset is somewhat limited in size, few or no values are observed in experimental data at scores ranging from 10 to 60. Therefore, this range of scores should be carefully considered as larger datasets are needed before generalizing results.[8]

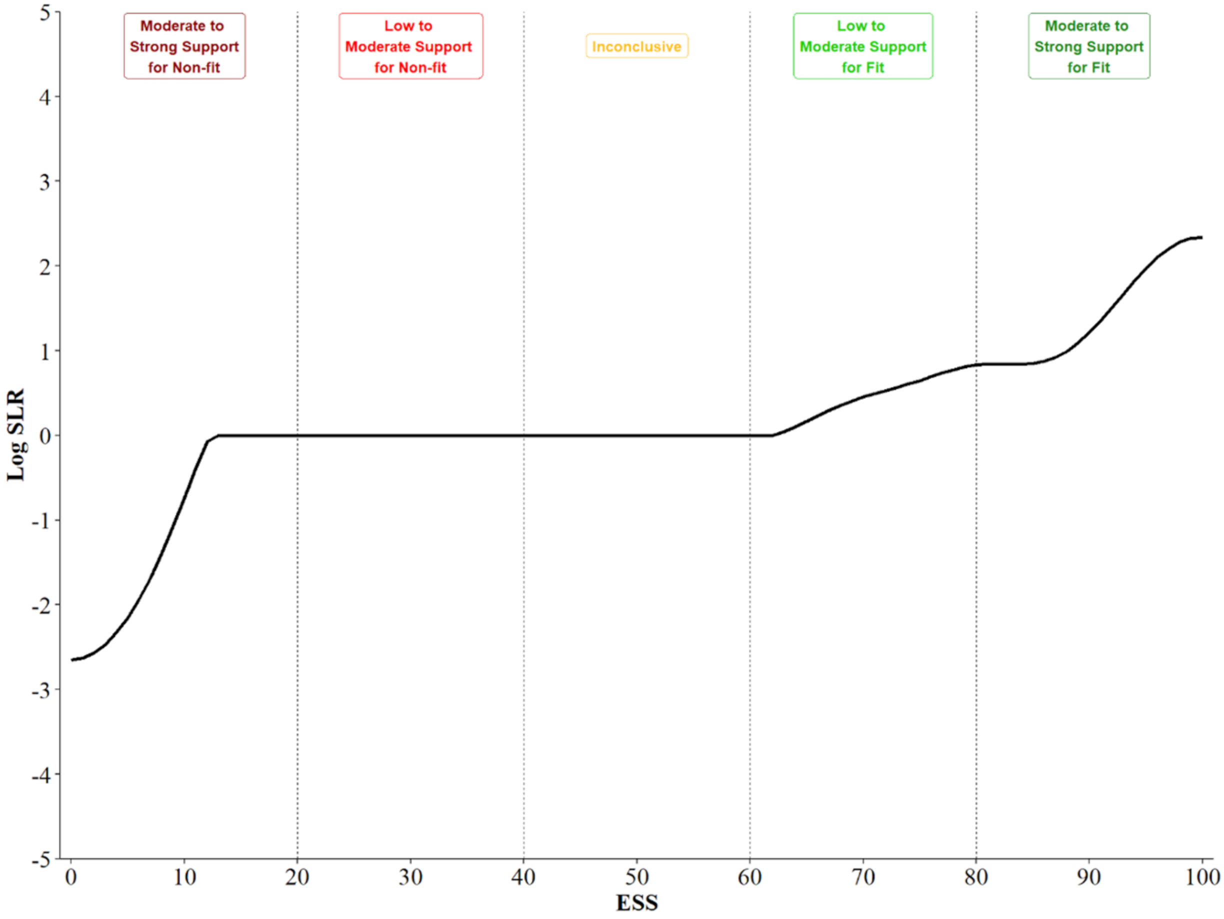


Figure 38. *Plot displaying log score-based likelihood ratios versus the ESS for the 100% cotton textile dataset.*[11]

Milestone 3—Practitioners' contributions: testing and fine-tuning through interlaboratory exercises.

The textile interlaboratory exercise assist with improvement of the method. Unlike tapes, textiles can't be immobilized in clear acetate sheets to preserve the edges when submitted for examination to various analysts. As a result, a decision was made to conduct the study utilizing high resolution images that the examiner can zoom on each bin at the equivalent observation level as the microscope. This approach not only prevent distortion of fracture features but also saves turn-around times and allows to mimic a verification process using our template format.

This interlaboratory study involves 15 participants conducting physical fit comparisons of images of three textile sample pairs. Participants are familiarized with the examination protocol, the terminology and criteria, and the reporting template and interpretation z-score. Using this method only one false negative conclusion is reported (3%), while inconclusive results ranged from 8 to 11%, and no false positives are observed in this set (**Table 18**)

Table 18. *Performance rates calculated using participant-reported conclusions and ESS thresholds of fit, inconclusive, or non-fit. Inconclusive conclusions are counted as "errors" for the sensitivity and specificity calculations. TPR= true positive rate, TNR= true negative rate. Adapted from Andrews et al.* [11]

	Performance rate by participant reported conclusion	**Performance rate by ESS threshold**
Sensitivity (TPR)	90% (27/30)	90% (27/30)
Specificity (TNR)	87% (13/15)	87% (13/15)
False Positive Rate	0% (0/15)	0% (0/15)
False Negative Rate	0% (0/30)	3% (1/30)
Inconclusive Rate	11% (5/45)	8% (4/45)
Accuracy	89% (40/45)	89% (40/45)

Inter-participant agreement between scores is also generally high. **Figure 39** shows that z-scores calculated for each participant show that 42 of 45 total comparisons were within the average range of ESS values for their respective pairs. The remaining three comparisons were deemed cautionary, while no comparisons in this study were deemed unsatisfactory.

The survey responses gathered by this study show that most participants find the ESS approach easy to follow and useful for describing their physical fit examinations, especially for verification or peer review purposes. The standardized terminology and descriptors used in this study also offer an opportunity to improve the consistency of reporting language used by practitioners, as seen in **figure 40** that displays the median and variation quartiles of the results reported by all participants per each pair set. Remarkably, all answers of ESS fell within the expected ranges seen by the consensus pre-distribution panel.

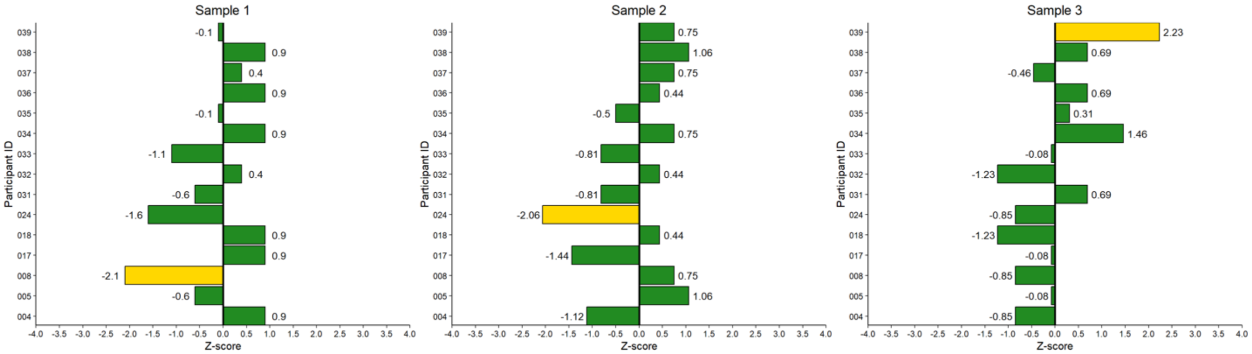


Figure 39. *Z-scores for the reported ESS value from each participant for each comparison pair. The z-scores have been color-coded for enhanced visualization, where green bars indicate satisfactory scores, yellow bars are cautionary, and red bars indicate unsatisfactory scores that fall outside the bounds of the confidence interval.* [8]

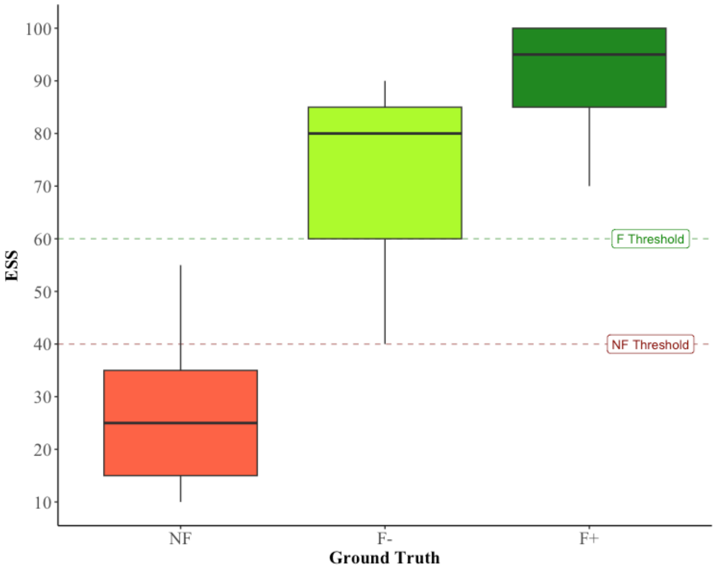


Figure 40. *Boxplots showing the distribution of scores for the three pairs in the interlaboratory study. The true non-fit pair is shown in tomato red (NF), the challenging true fit pair is shown in lime green (F-), and the true fit pair is shown in forest green. The thresholds for fits (60%) and non-fits (40%) are shown in green and red, respectively.* [11]

Milestone 4—Complementing human-based approaches with computational algorithms

The use of mutual information and decision tree algorithms has shown valuable for textiles and for tapes, and the results were previously discussed in the respective tape section. However, the use of machine learning (ML) models to make predictions of whether two input images of textiles are fit or non-fit was not as straightforward for textiles. Much of the methodology was like the tape prediction process, as the problem is very similar besides from a few subtilties in the preprocessing. The main challenge however with textiles is the occurrence of artifacts produced by loose fibers that are inherently created during a separation process, particularly when hand-torn. Also, another issue is the fact that unlike duct tapes, the size of a known and questioned textile comparison vary largely in their fractured edge size.

The CNN model is optimized to a point where it can be trained on reasonably processed textile pair images. In the textile digital datasets there are 294 Fit pairs and 305 non-fit pair examples. During the preprocess we remove textile pairs that are not able to be processed with the current method. This is either caused by bugs in the edge detection algorithm or long strands that extend far in the x-dimension. Some examples of these issues can be seen in **Figure 41.** In total 93 Fit pairs were removed such that the preprocessing scripts run without issue and the textiles look reasonable. Thus, the model only considers 201 fit pairs.

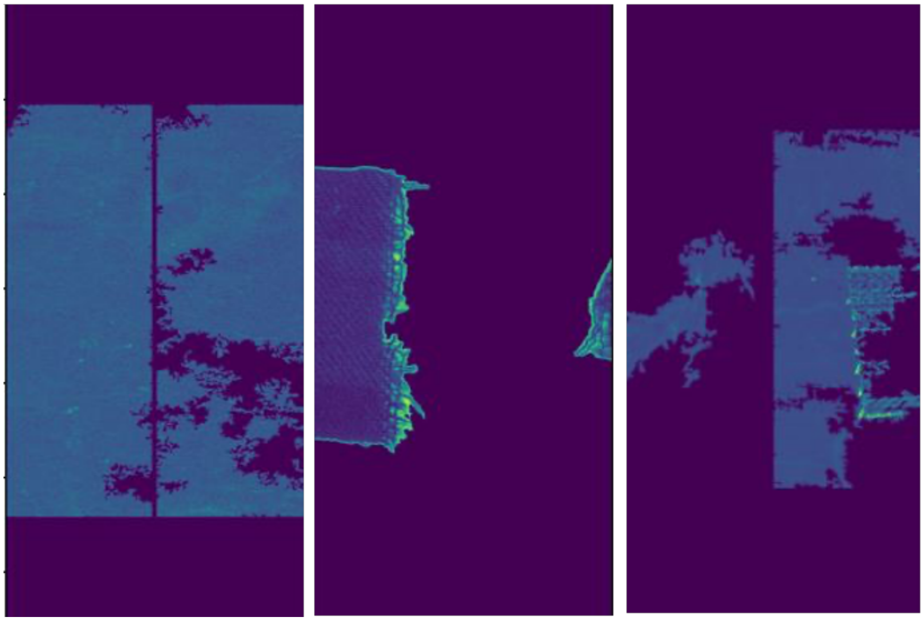


Figure 41. *Examples of issue textiles. These textiles failed the preprocessing due to artifacts produced.*

A 5-fold cross-validation is applied. The textile fit pairs are split into five folds, with four folds acting as the training set and the last fold as the validation set. This is performed five times, switching out each fold for the testing set. This will judge how well the model generalizes to random datasets. Next, the non-fits are generated from the fits for the training and validation up to a certain fit/non-fit ratio. Here, are fit/non-fit ratio was chosen to be 0.3 fit/non-fit as this turned out to be optimal or the textiles pair training. The metrics are displayed in **Table 19** with accuracies ranging from 74 to 86%. Although these rates are inferior to the human-based examination, they provide a basis for future improvements as the database increases in sample size. Following Layer-wise Relevance Propagation (LRP), the method highlights important pixels for the decision of a prediction. In **Figure 42**, LRP is demonstrated on a true fit pair.

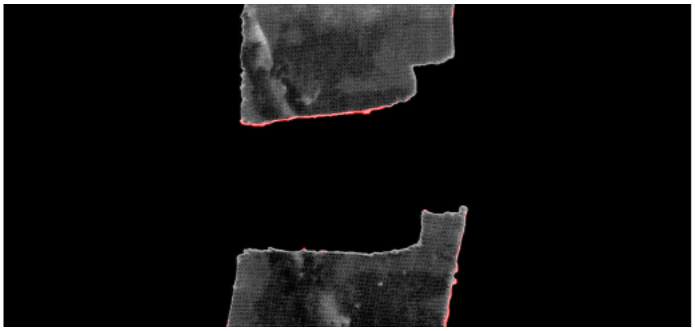


Figure 42. *Layer-wise Relevance Propagation (LRP) analysis is compared to human comments on Fits. Important LRP pixels are colored. LRP identifies the most important features in accordance with the examiners bin-by-bin respective annotations.*

Table 19. *Metrics from 5-fold cross validation for the textiles comparisons using CNN*

	Accuracy	Precision	Recall	Specificity	F1
Training	85.6	82.7	72.0	92.4	76.6
Validation	74.3	62.2	64.2	79.4	62.8

It should be noted the results are achieved with a relatively low number of fit pairs. Increasing the amount of textile fit pairs can improve the results. To do this, the preprocessing algorithm must be improved to take care of the issues highlighted before. Along with improving the preprocessing algorithm, more test needs to be done regarding the model architecture. The preliminary model was trained using grayscale image; however, the textile images contain color information that might prove useful in classification. In the future, a model will be created that can include all color channels. On the other hand, it is expected that the model accuracy will also depend on the number of images used for training.

2.2.4. Automotive plastics physical fit method

The third material evaluated in this study is hard brittle polymers; we focus on automotive parts as they occur frequently in vehicle-related offenses. The results are anticipated to be applicable to similar polymers in other applications.

Milestone 1—Method development and optimization of standardized criteria

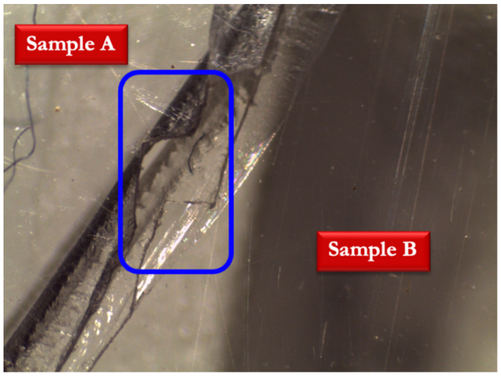
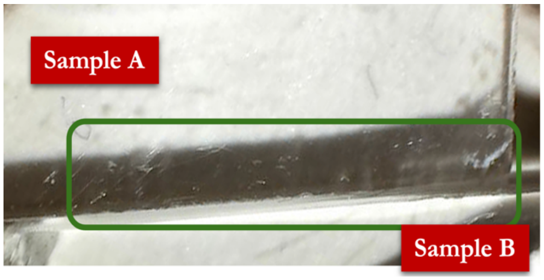
Major modifications were necessary to develop the method of examination and to identify relevant features on automotive plastics as their fracture characteristics and multi-dimensions and planes of the pieces make it very different to handle than flat and thinner materials like duct tape and textiles. Also, literature on brittle plastics' fits is relatively scarce.[54-59] Therefore, the method development 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 also designed to guide analysts through the comparison and process documentation. 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.

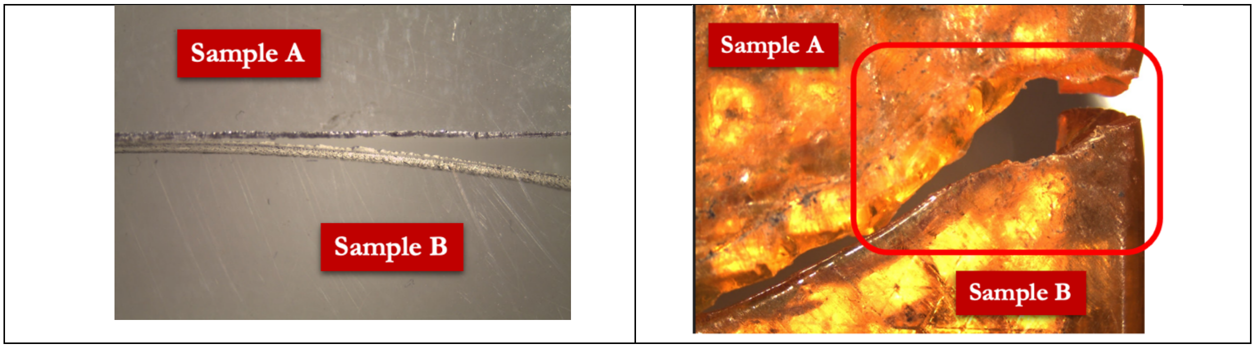
Ten main features are established for this material and described in the methodology section. In addition to the ESS similarity score, another metric is implemented in this method to estimate the influence of each feature on a given decision. To evaluate the features quantitatively, each response option 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 either more positive or negative FPV for fits or non-fit, respectively. **Table 20** shows an example the assigned values for a feature and respective report options in an arbitrary scale of -2 to +2, and **Table 21** illustrates examples of 3D edge alignment and some FPV descriptors. The, the FPV of all features is summed for each bin and then across all bins to provide the feature prominence sum (FPS). Finally, the final overall decision, ESS, and FPS are reported in each comparison.

Table 20. Example of documented features, response options, and respective feature prominence value (FPV) for the brittle polymer comparisons.

Features	Options for a response to these features	Feature Prominence Value
1. 3D Edge Alignment	**Present and Highly Distinctive (Indicative of Fit)**	2
	Present and Highly Distinctive (Indicative of Fit)	1
	Inconclusive	0
	Present but Misaligns (Indicative of Non-fit)	-1
	Present but Misaligns (Highly Indicative of Non-fit)	-2

Table 21. Description of polymer 3D alignment feature and examples of the weight on the decision of a fit or non-fit.

3D Alignment Feature	
A. **Present and Highly Distinctive (Indicative of Fit) -** This polymer is highly distinctive due to how these 3D alignments are puzzle-like. Since the pieces are puzzle-like, the odds of the fracture quickly changing directions in the exact same zig-zag pattern increases the rarity making it highly distinctive. This is an example of a bin with an FPV value of +2. 	B. **Present (Indicative of Fit) -** This polymer is 3D alignments are "smooth", with no distinct waviness or "puzzle-like" areas. Although these pieces align, it is likely that these pieces could align with other pieces that are flat, translucent, and straight-edged. This is an example of a bin with an FPV value of +1 
C. **Present but Misaligns (Indicative of Non-fit) -** In this image, Sample A and B align towards the left. However, there is a gap that grows wider as the viewer goes more right. The gap is notable enough to make it seem as though these pieces do not fit together, but not necessarily big enough to have an examiner think there may not be an intermediate piece that could fit in between Sample A and B. This is an example of a bin with and FPV of -1.	D. **Present but Misaligns (Highly Indicative of Non-fit) -** In this image, Sample A and B has edges that do not align. In the top right area, there is a gap that starts to go in opposite directions. This is an example of a bin with and FPV of -2.



Milestone 2—Method validation through large databases and evaluation of factors that affect performance rates

The 445 pairs of polymer samples originating from automotive headlight and taillight assemblies are compared by multiple analysts to evaluate the newly developed method. The polymer sources are grouped into three classes based on the main polymer composition and morphology: translucent clear, translucent colored, and opaque colored.

Table 22 shows that misidentification rates of the initial comparison set are relatively low and an overall accuracy of 86%, with only one false positive reported in this dataset. 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 misidentifications 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).

Table 22*. Performance rates of the analysis of the polymer pairs. This overall set includes a mixture of polymer types, compositions, and morphologies.*[10]

Performance Rates	Overall (n = 445)
# of True Fits/# of True Non-fits	347/98
True-Positive Rate (%)	83.9
True-Negative Rate (%)	96.0
False-Negative Rate (%)	8.9
False-Positive Rate (%)	1.0
Inconclusive Rate (True Fits) (%)	7.2
Inconclusive Rate (True Non-fits) (%)	3.0
Accuracy (%)	86.5

As shown in **Figures 43 and 44,** the method demonstrates that most true non-fit polymers receive low ESS (0-10%) and low FPS (less than -5). A worrisome exception is observed for the false positive that received a score of 90%. 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. Therefore, the proposed approach is expected to facilitate the implementation of consensus-based protocols at forensic laboratories and provide scientific foundations for data-driven opinions. [7,10]

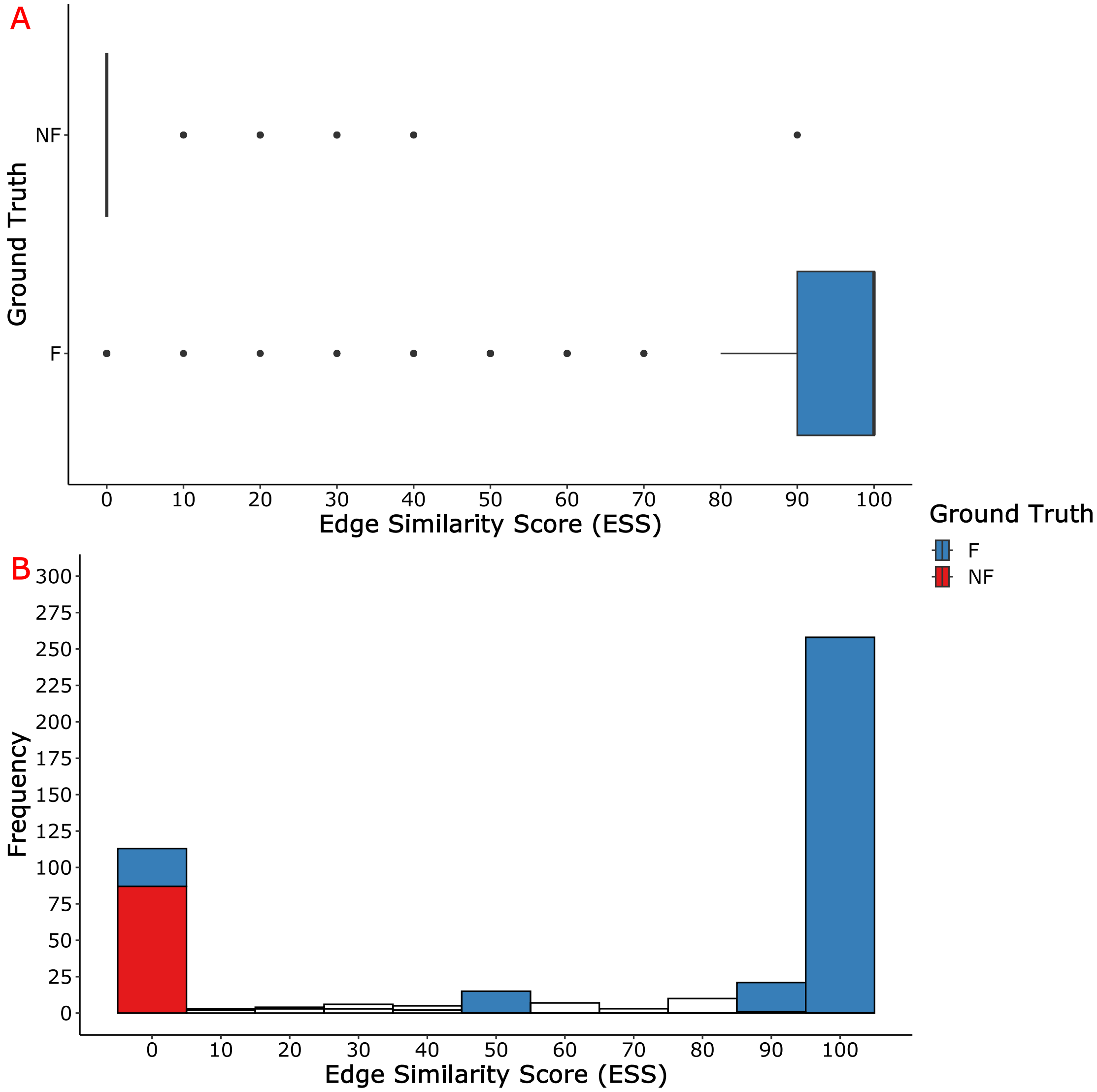


***Figure 43**. Distribution of ESS for true fit (TF) and true non-fit (TNF) pairs from the polymer analysis (n=385 pairs). One false positive is reported with an ESS of 90, and only a few TNF pairs are reported as inconclusive. **Figure 43A, top** shows boxplot distributions of the ESS values of the 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 43B, bottom** 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.* [10]

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

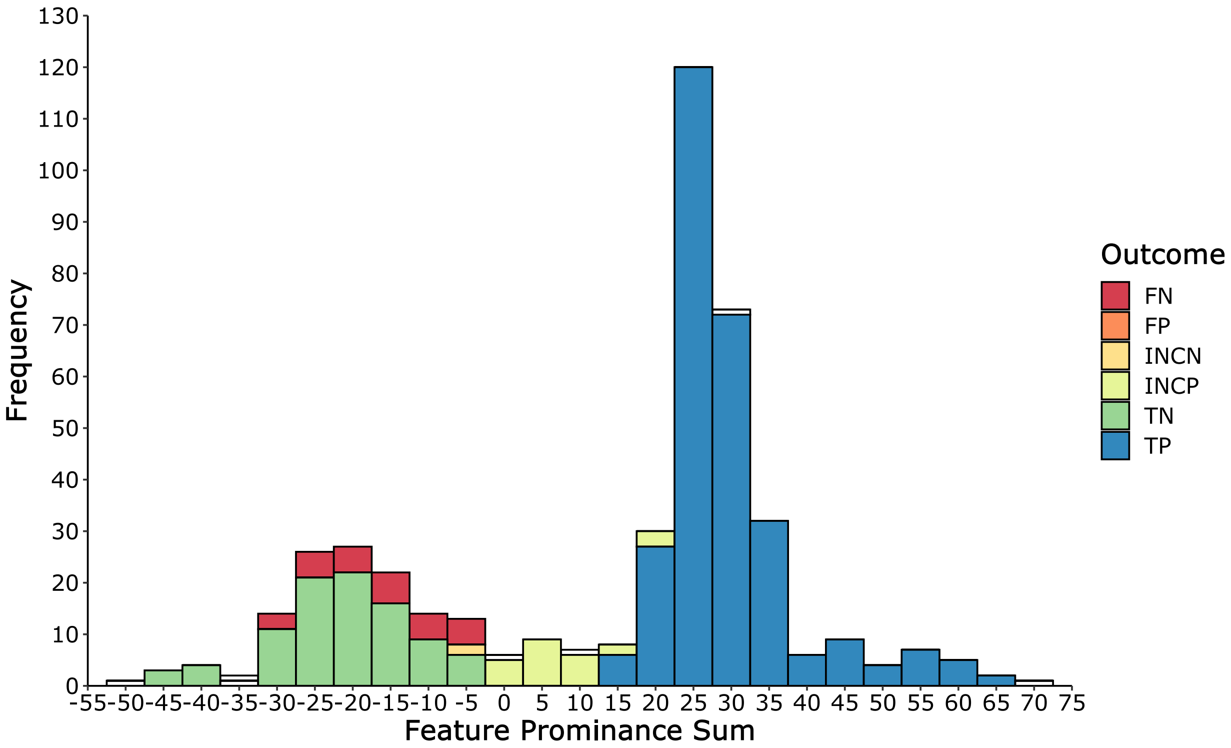


Figure 44. Frequency distribution of FPS for the initial polymer set. The histogram is color-coded by outcome: True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), Inconclusive-True-Non-fit (INCN), and Inconclusive-True-Fit (INCP). One FP is reported (FPS of 31), and a few 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. [10]

Table 23 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 translucent colored and then the opaque colored.

Of the three sets, the translucent-colored set generated a higher false negative rate and a high inconclusive rate for true fits, as well as the only false positive reported in the set. The misidentifications in this set reduce the accuracy for these polymers. During the comparisons, the analysts note that these polymer fragments, particularly the orange and red fragments originating from a headlight, tend to distort more substantially than some other polymers (see **Figure 45 and 46**). 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. **Figure 47** shows the boxplot distributions of pair ESS based on polymer class and composition. The distribution of the ESS for true fits and true non-fits is similar for both the samples in the translucent clear and translucent colored sets that are made of polycarbonate.[7]

Table 23*. Performance rates of the initial analysis of the polymer pairs, grouped by polymer class of translucent clear, translucent color, or opaque color.* [10]

Performance Rates	Opaque Colored (n=40)	Translucent Clear (n=314)	Translucent Colored (n=90)
# of True Fits/ # of True Non-fits	31/9	253/62	63/27
True Positive Rate (%)	90.3	83.8	81.0
True Negative Rate (%)	88.9	96.8	96.3
False Negative Rate (%)	3.2	7.5	12.7
False Positive Rate (%)	0.0	0.0	3.7
Inconclusive Rate (True Fits) (%)	6.5	8.7	6.3
Inconclusive Rate (True Non-fits) (%)	11.1	3.2	0.0
Accuracy (%)	90.0	86.3	85.6

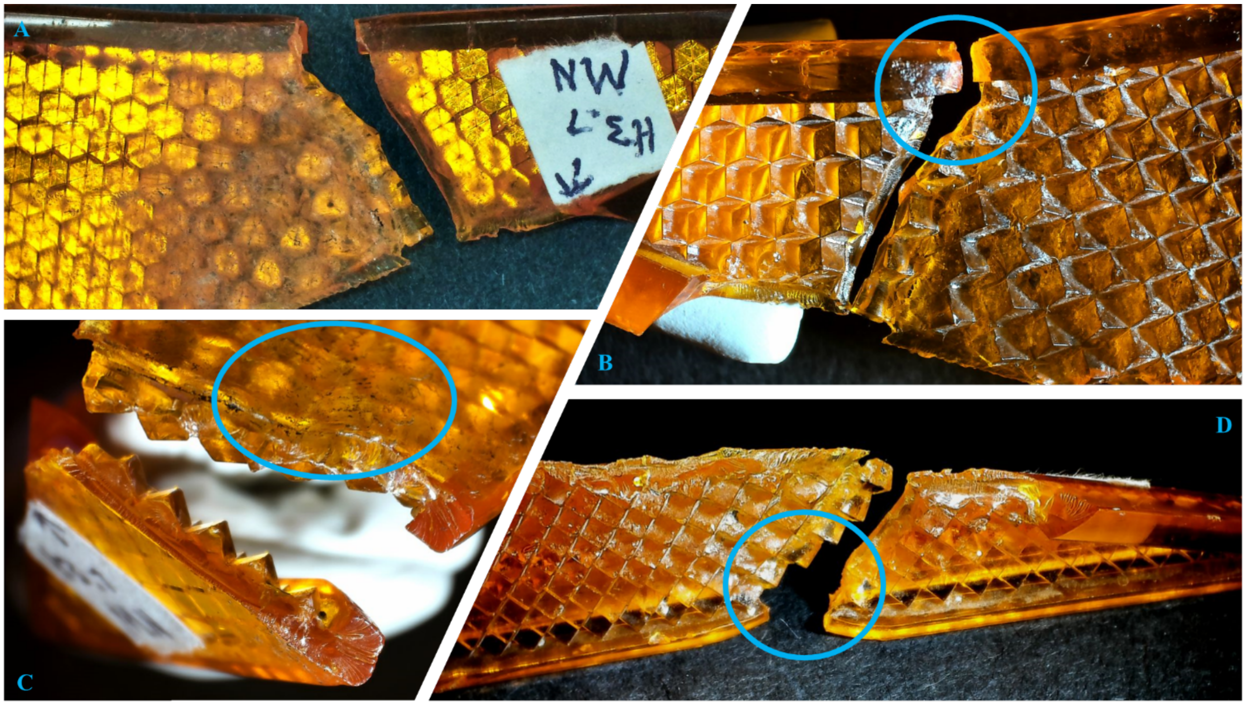


Figure 45. *Example pair from translucent color polymers. The two fragments have visible surface damage, and the respective comparison edge has a noticeable gap between the material of the two edges in a manner that indicates they do not fit together.* ***Figures 45A, 45B,*** *and* ***45D*** *show the gap between the edges caused by missing material and distortion of the edges. In contrast,* ***45C*** *shows an indent in the surface of top sample that causes misalignment of the pattern (texture) and the edge of the samples.* [10]

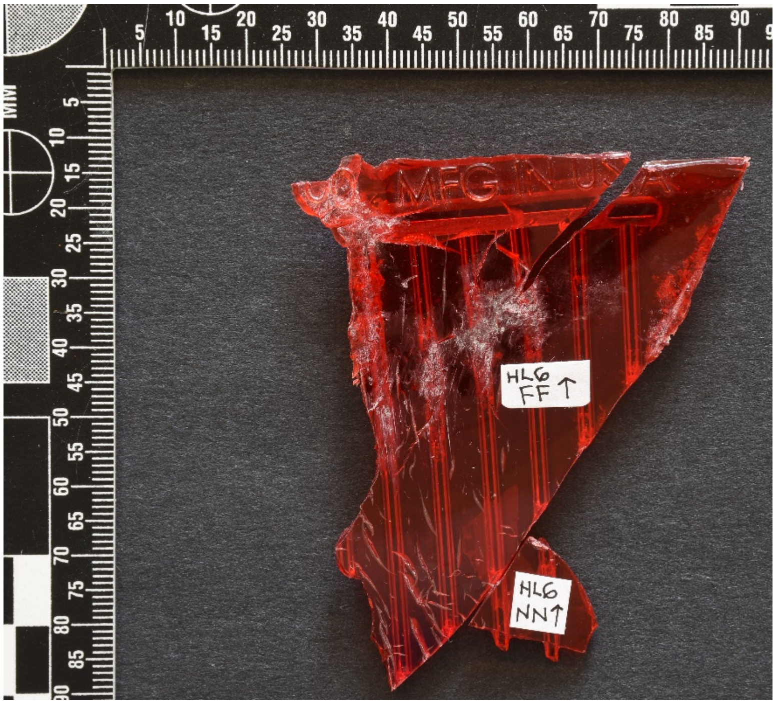


Figure 46. *Image of the one false positive pair reported in the set. The sample edge is relatively small, and has consistent patterning across the fracture edge, along with additional features noted by the analyst that indicated the pieces had substantial similarities* [10]

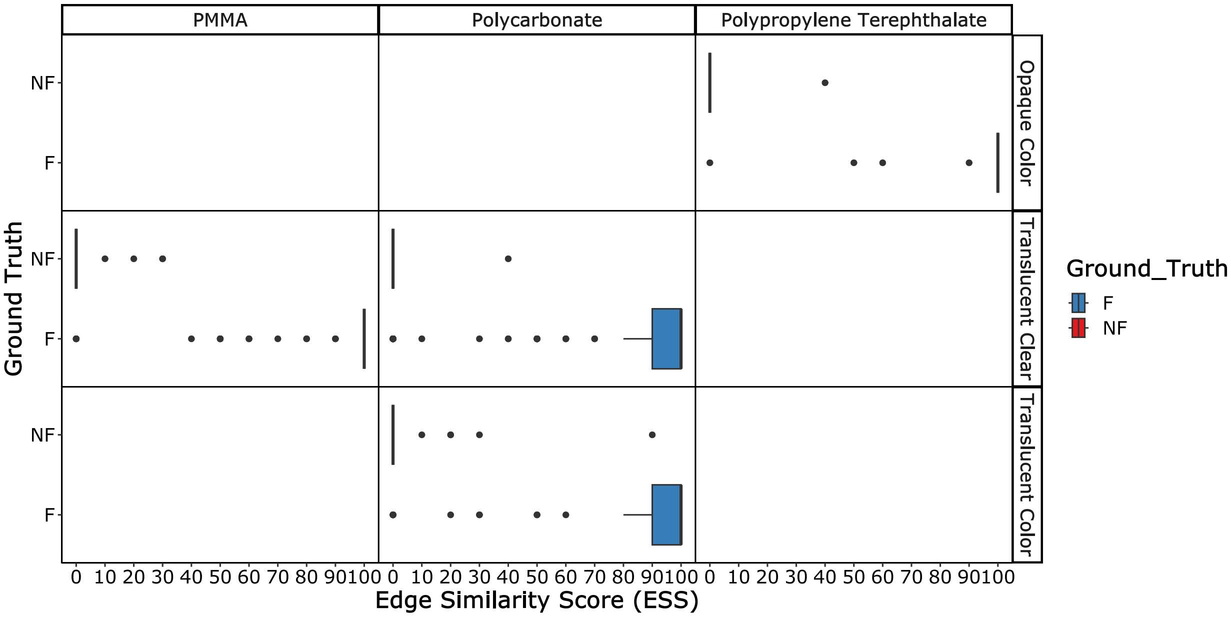


Figure 47*. Boxplot distributions of reported ESS grouped by polymer class (translucent clear, translucent colored, and opaque colored) and polymer composition (PMMA, polypropylene terephthalate, or polycarbonate). The separation between the ESS of true fits and true non-fits for the translucent clear and opaque color sets is relatively strong, and most pairs have ESS of 100 or 0 for fits and non-fits, respectively* [10]

Inter-analyst 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 24.** Overall, the accuracy is comparable between the two analysts. 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. One conclusion that is derived from this assessment is that because there are 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.[7]

Table 24. *Performance rates of the inter-analyst examination. This set is a subset of the initial set and contains a mixture of the polymer classes.*[10]

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 48**).

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

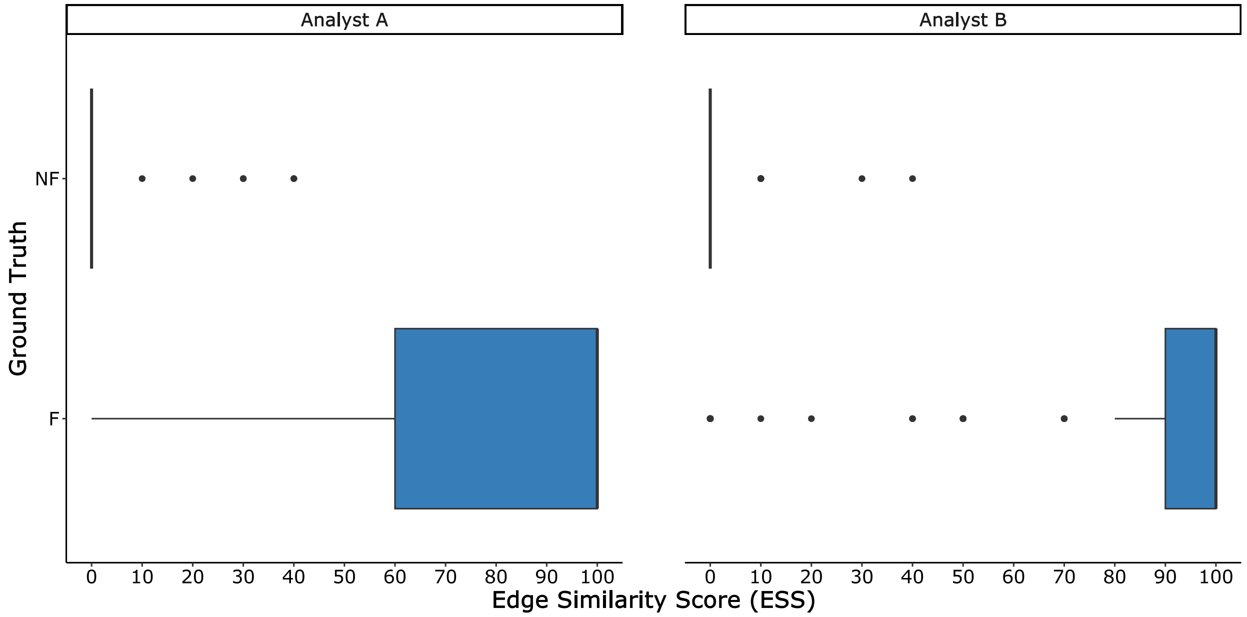


Figure 48. Boxplot distribution of ESS for true fit (TF) and true non-fit (TNF) pairs from the inter-analyst polymer analysis. No false positives are reported, and only a few TNF pairs are reported as inconclusive for both analysts.[10]

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 49** 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.

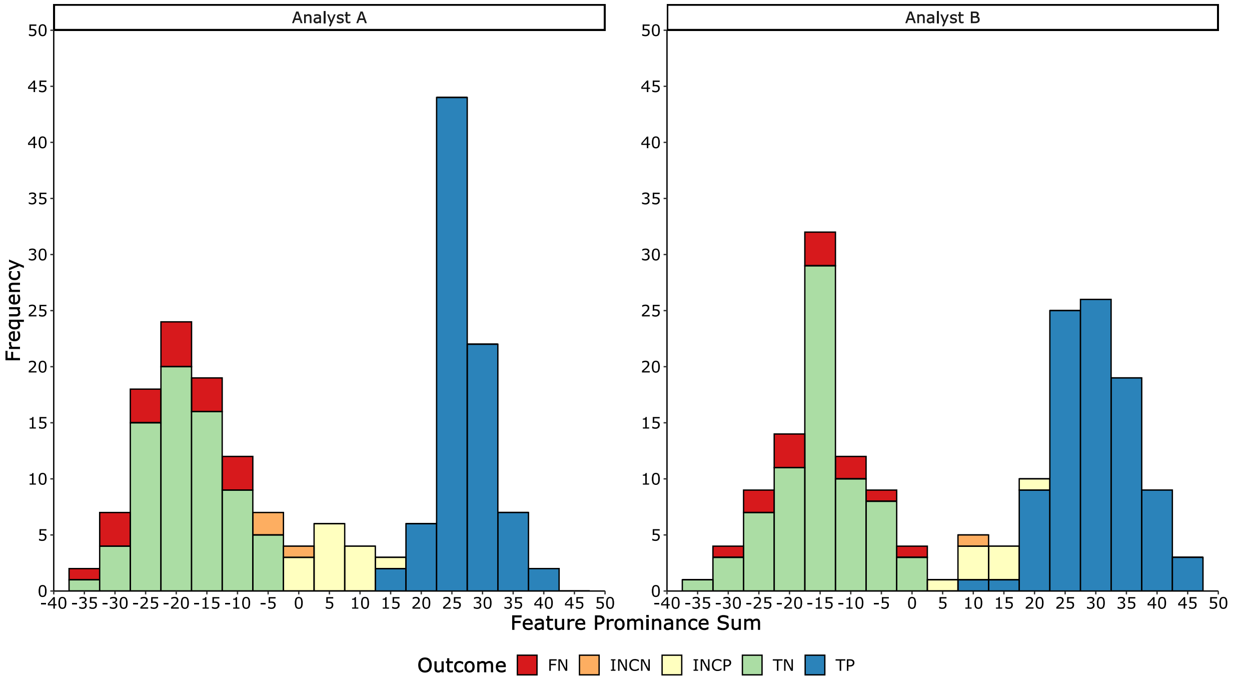


Figure 49. *The frequency distribution of FPS for the 187 inter-analyst pairs. True Positive (TP), True Negative (TN), False Negative (FN), False Positive (FP), Inconclusive-True-Non-fit (INCN), and Inconclusive-True-Fit (INCP) are shown for each analyst. The distributions of FPS are similar for both analysts, with only minor variations in the distributions of the inconclusive pairs.* [10]

Overall, the main lessons learned in the polymer set are summarized as follows:

1) The ESS scores provide a quantitative assessment of the quality of a fit. For this population set, ESS scores below 10 support non-fits, while scores above 90 were typically observed for true fits.
2) The novel supplement 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.
3) The qualitative features along with the quantitative ESS and FPS metrics demonstrates good overall performance for physical fit examinations of brittle automotive polymers, The initial comparison set of 445 comparisons result in an overall accuracy of 86.5%.
4) Most error rates originate from false negative misidentifications or inconclusive caused by distortions during the breaking process. The distortion significantly masks distinctive features.
5) False positives in this dataset are low (1%) but when present, the ESS and FPS values were high, which raises awareness that brittle polymers could produce misleading fit results.
6) Inter-analyst performance shows consistency, with analysts demonstrating similar overall accuracies, ESS, and FPS distributions, indicating the method can further assist the discipline with a standardized approach for brittle polymers.

7) Most informative features occur at the microscopic level. Unfortunately, imaging of 3D features are complex and therefore an image database was created at the macroscopic level, but microscopic images were not appropriate for the ForensicFit package or CNN networks.
8) 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.

2.3. Limitations

The main limitation encountered in this study is the implementation of computational CNN algorithms for textiles and polymers. The main challenge for polymers is to capture in an image the microscopic three-dimensional features that the analysts observe under the microscope. Light reflection and refraction, and different focal planes and depths within a polymer broken edge are some of the issues that limited the imaging automated comparisons. For that reason, 3D molds and more sophisticated 3D scanning technology may be needed to deal with the database imaging of brittle plastics. In contrast, tape and textile features are more easily stored in a 2-dimensional matrix where each element represents a pixel intensity.

Although the performance of CNN for automated assessment of tapes and textiles is very promising, we acknowledge that the dimensionality of the data requires of much larger datasets. Our image collection contains about 9000 images, while CNN computational algorithms for image feature recognition. CNN often require more than ten times the size of these sets. However, the collection, imaging and cross-validation with examiner-based results is time consuming and unpractical at that level of sample size.

III ARTIFACTS

3.1 List of products

3.1.1 Publications at scientific peer-reviewed journals and dissertations

1. **Published products in scientific peer-reviewed journals**

1) M Prusinowski, E Brooks, C Neumann, T 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) M Prusinowski, Z Andrews, C Neumann, T Trejos. Assessing significant factors that can influence physical fit examinations – Part I. Physical fits of torn and cut duct tapes. Forensic Science International. 2023, 343, https://doi.org/10.1016/j.forsciint.2023.111567
3) E Brooks, M Prusinowski, S Gross, T Trejos. Forensic Physical Fits in the Trace Evidence Discipline: A Review, Forensic Science International, 313, 2020, https://doi.org/10.1016/j.forsciint.2020.110349 (served as a basis for this research)
4) M Prusinowski, E Brooks, T Trejos. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. Forensic Science International, 307, February 2020, https://doi.org/10.1016/j.forsciint.2019.110103 (served as

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

a basis of this research)

2. **Published thesis and dissertations.**

5) Meghan Prusinowski, Ph.D. WVU Department of Forensic and Investigative Science, Enhancing the forensic comparison process of common trace materials through the development of practical and systematic methods. Graduate Theses, Dissertations, and Problem Reports. 2023, 11644. https://researchrepository.wvu.edu/etd/11644
6) Zachary Andrews, MSFS (Summer 2022), WVU Department of Forensic and Investigative Science, "Evaluating the Validity and Reliability of Textile and Paper Fracture Characteristics in Forensic Comparative Analysis" Graduate Theses, Dissertations, and Problem Reports. 2022. 11373. https://researchrepository.wvu.edu/etd/11373

3. **Submitted under journal review.**

7) M Prusinowski, P Tavadze, Z Andrews, L Lang, Divyanjali Pulivendhan, C Neumann, AH. Romero, T. Terjos. Experimental results on data analysis algorithms for extracting and interpreting edge feature data for duct tape and textile physical fit examination. Under review, submitted Journal of Forensic Science, June 2023
8) M Prusinowski, P Tavadze, Z Andrews, C Dolton, C Vogler. Development of a systematic comparison method for forensic physical fit analysis of automotive polymers. Under review, submitted Forensic Chemistry, June 2023
9) Z Andrews, M Prusinowski, E Nguyen, C Neumann, T Trejos. Assessing physical fit examinations of stabbed and torn textiles through a large dataset of casework-like items and inter-laboratory studies. Under review, Journal of Forensic Sciences. Submitted May 2023
10) P Tavadze, L Lang, M Prusinowski, Z Andrews, T Trejos, and AH. Romero. Using convolutional neural networks to support examiners in duct tape physical fit comparisons. Under review, Forensic Science International. Submitted January 2022

3.1.2. Presentations at Scientific Meetings

1) February 2023, Zachary Andrews, Meghan Prusinowski, Tatiana Trejos. Assessment of a novel method for physical fit examinations using an extensive database of casework-like samples and interlaboratory studies. AAFS meeting, Orlando, FL (poster presentation)
2) September 14th, 2022. T Trejos, A Quigley-McBride, M Prusinowski, Z Andrews. Workshop: Forensic Examinations of Physical Fits—Past, Present, and Future. MAFS 51st Annual Fall Meeting A Joint Meeting with ASTEE, Des Moines, Iowa. (workshop organizer and instructor, full day workshop)
3) September 15th, 2022. Meghan Prusinowski, Zachary Andrews, Tatiana Trejos. Development of systematic and practical documentation templates for tape and textile physical fit comparisons. MAFS 51st Annual Fall Meeting A Joint Meeting with ASTEE. Des Moines, Iowa. (Oral presentation)
4) September 16th, 2022. Zachary Andrews, Colton Diges, Tatiana Trejos. Evaluating the use of microfiber alignment in office paper and postage stamps to identify physical fits. MAFS 51st Annual Fall Meeting A Joint Meeting with ASTEE. Des Moines, Iowa. (Oral presentation)

5) June 1st, 2022. Meghan Prusinowski, Zachary Andrews, Cedric Neumann, Tatiana Trejos. Assessing significant factors that can influence physical fit examinations of tape and textiles. European Academy of Forensic Sciences (EAFS) conference, Stockholm, Sweden (Poster)
6) February 2022. Meghan Prusinowski, Evie Nguyen, Tatiana Trejos. Validation of a Systematic Method for Duct Tape Physical Fits Through Inter-Laboratory Studies. 2022 AAFS Conference, Seattle, WA. (Poster, Virtual)
7) February 2022. Zachary Andrews, Colton Diges, Tatiana Trejos. Feature Occurrence and Error Rates in Textile Physical Fit Comparisons. 2022 AAFS Conference, Seattle, WA. (Poster)
8) October 2021. Meghan Prusinowski, Zachary Andrews, Tatiana Trejos. Development of systematic methods for the physical edge comparison of trace materials. 2021 Brazil Winter 3rd School of Forensic Sciences (Virtual, Oral Presentation)
9) July 29th, 2021. Colton Diges, Zachary Andrews, Meghan Prusinowski. Microfiber Alignment in Stamp Edges for Physical Fit. 13th Annual summer undergraduate research symposium, Morgantown, WV https://www.youtube.com/watch?v=tdt-TiiNtXM
10) July 28th, 2021. Zachary Andrews, Colton Diges, Meghan Prusinowski, Tatiana Trejos. Assessing the Value of Microfiber Alignment Between Stamp Edges for Physical Fit Comparisons. Current Trends in Forensic Trace Analysis 2021 Online Forensic Symposium. (poster)
11) June 2nd, 2021. Tatiana Trejos, Meghan Prusinowsli, Zachary Andrews. Forensic Examination of Duct Tape Physical Fits: Interlaboratory Results, NIST-OSAC Trace Subcommittee (oral)
12) February 2021, Meghan Prusinowski, Zachary Andrews, Evie Nguyen, Tatiana Trejos. Development of Systematic Approaches for Physical Fit Comparisons of Trace Materials. Presented at Virtual AAFS Conference (E-Poster)

3.1.3. Website(s) or other Internet site(s)

Development of the package ForensicFit. Tavadze P, Lang L. romerogroup/ForensicFit: First release of ForensicFit Package [Internet]. Zenodo; 2022. Available from: https://doi.org/10.5281/zenodo.7435058

3.2. Data sets generated

According to our data management plan, the data resulting from this research was curated and compiled into a centralized dataset repository. The dataset generated in this study consists of a physical collection of about 9,000 fractured items, which is maintained at the Trejos' laboratory, and from these samples, a total of 4,773 pairs were generated for analysis. The overall composition of the datasets is shown in **Figure 15** of this report. The digital dataset contains the archived data, and includes:

a) A master inventory with the sample unique identifier (no personal identification information) and the ground truth of the items (i.e., known true fit, known true non-fit)
b) Images of the fractured edges: tape and textile scans, automotive plastic photographic images.

c) Microsoft Excel reporting templates with the sample's unique identifier and the examiner's observations of the physical fit examinations, including qualifiers and similarity scores.
d) Microsoft Excel files with the ground truth and respective examiner's conclusions and performance rate estimations.
e) Materials such as templates, presentations, and instructions submitted for the interlaboratory studies
f) Fourier Transform Infrared Spectroscopy (FTIR) analysis for the polymer study project.

Data Storage and File Descriptions
A data drive folder is named Physical Fits NIJ 2020-DQ-BX-0012 archiving, which contains four sub-folders, one with the master inventory of all physical samples and their respective unique identifier, and three other folders containing each data per type of material, 1) tape, 2) textile, and 3) automotive plastic (figure 1)

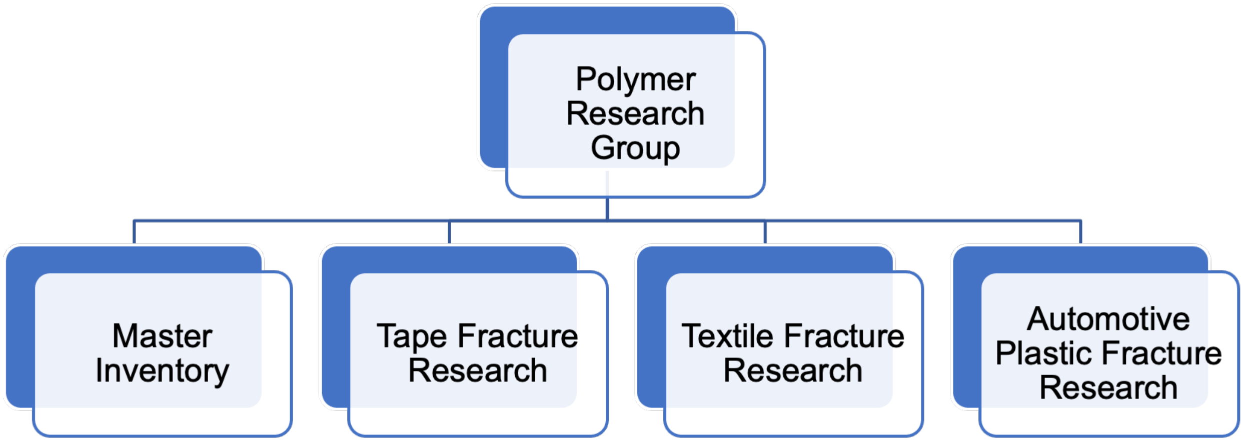

Polymer Research Group
Master Inventory
Tape Fracture Research
Textile Fracture Research
Automotive Plastic Fracture Research

Figure 51: *Polymer Research Group overall folder structure.*

Each of the Fracture Research subfolders contains 1) the reporting template(s) used during the study, 2) the data, split by subsets, each containing an Excel file with all reported results, and one file compiling the ground truth, 3) the photos or scans, and 4) the interlaboratory results when applicable.

3.3. Dissemination activities

To date, the main dissemination routes have been the publication of manuscripts in scientific journals and the presentation of research results at scientific meetings, as described in 3.1. An in-person workshop was organized at the 2023 MAFS/ASTEE joint meeting, with 30 practitioners and a virtual session was organized in Spring 2023 to discuss the result with the interlaboratory participants and the invitation was also extended to other agencies of interest, with a total of 42 attendees.

IV PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS

This research has provided a robust platform for training the next generations of forensic scientists in trace evidence, physical fits, and experimental design in forensics. This research has provided research opportunities for undergraduate students and graduate students (Master and Doctoral). **Table 25** lists the main participants and collaborators.

Moreover, this project's resources and research settings have provided all undergraduate and graduate students the unique opportunity to present their results at scientific venues. The opportunities provided to undergraduate researchers, some of the first-generation university students or minority students, have served as an essential foundation to their professional development. Two of our PhD students joined the workforce, and the Master's student completed his degree and started in the doctoral program. One of our undergraduates was hired at a forensic laboratory, two of our undergraduate researchers joined graduate school, and another one joined law school, and the three remaining undergraduates continue conducting research in our group. These student's achievements and STEM professional preparation are, in our opinion, the most valuable product of NIJ-funded efforts like this one.

This project also allowed a valuable collaboration across disciplines, and between academia and practitioners at state and federal laboratories, exposing the students, faculty, and practitioners to an enriching multi- and inter-disciplinary environment to develop solutions for our criminal justice system.

Table 25. *List of main participants and collaborating organizations*

Participant Name	**Affiliation**	**Role**	**Contributions**
Tatiana Trejos	West Virginia University	Principal investigator, Associate Professor	Managed the project and directly supervised students on experimental designs, sample collection, method development, and statistical interpretation of the data. Supervised dissemination plans, data curation and management plans.
Aldo Romero	West Virginia University	Co-Principal investigator, Associate Professor	Supervised research related with computational algorithms and digital database. Assisted with reports and manuscripts.
Cedric Neumann	Battelle Memorial Institute	Statistician Collaborator (subaward)	Collaborated as expert in statistical analysis and interpretation of the data and as co-author of manuscripts.
Meghan Prusinowski	West Virginia University	Graduate Student (PhD)	PhD graduate student working at the Trejos's group. Meghan was the lead student researcher in the tapes and polymers materials, contributed with collections, physical and digital database, the data acquisition, analysis and interpretation. She has been a primary contributor to the

Participant Name	Affiliation	Role	Contributions
			manuscripts and dissemination of results.
Zachary Andrews	West Virginia University	Graduate Student (MSFS and PhD)	Graduate student working at the Trejos's group. Zach was the lead student researcher in the textiles, contributed with collections, physical and digital database, the data acquisition, analysis and interpretation. He has been a primary contributor to the manuscripts and dissemination of results.
Pedram Tavadze	West Virginia University	PhD student (2021-2022), postdoctotal fellow (May 2022-May 2023)	Pedram was a PhD student at WVU-Physics Department under Romero's group, who completed his degree in Spring 2022. Then, he joined the team as postdoctoral fellow under Dr. Trejos supervision. His main contributions were the development of computational algorithms.
Paige Smith	West Virginia University	Undergraduate student	Paige's main contribution was assisting with sampling collections and imaging tapes and textiles. Paige graduated in spring 2021
Elizabeth Hanley	West Virginia University	Undergraduate student	Elizabeth's main contribution was assisting with sampling collections and imaging tapes and textiles. Elizabeth graduated in spring 2022
Colton Diges	West Virginia University	Undergraduate student	Colton's main contribution was assisting with sampling collections, imaging, and analysis of polymers and textiles, and database curation. Colton graduated in December 2022
Katelin Radonovich	West Virginia University	Undergraduate student	Katelins's main contribution was assisting with sampling collections, imaging, and preparation of the physical and digital collection of polymers.
Divyanjali Pulivendhan	West Virginia University	Undergraduate student	Divyanjali Pulivendhan's main contribution was assisting with sampling collections, imaging, and preparation of the physical and digital collection of polymers and tapes. Divyanjali graduated in spring 2023
Claire Dolton	West Virginia University	Undergraduate student	Claire's main contribution was assisting with sampling collections, imaging, analysis, and preparation of the physical and digital collection of polymers.

Participant Name	Affiliation	Role	Contributions
Charlotte Vogler	West Virginia University	Undergraduate student	Charlotte's main contribution was assisting with sampling collections, imaging, analysis, and preparation of the physical and digital collection of polymers.

We would like to thank members of the NIST-OSAC trace subcommittee and the physical fits task group for their input during the research surveys and feedback sessions. Also, we would like to thank the many forensic practitioners and agencies that participated in the interlaboratory studies, their names are not listed as we maintained the study anonymous.

V CHANGES IN APPROACH

Nothing to report.

VI REFERENCES

1. OSAC Research Needs Trace Materials Subcommittee. Development of Quantitative Assessment and Evaluation of Error Rates in Physical Fit Determinations of Trace Materials. 2018. https://www.nist.gov/sites/default/files/documents/2019/02/01/osac_research_needs_assessment_form_materials_sc_validation_of_physical_matches.pdf
2. Forensic Science TWG Operational Requirements. National Institute of Justice, 2019. https://nij.ojp.gov/sites/g/files/xyckuh171/files/media/document/2019-11-forensic-twg-table.pdf
3. M Prusinowski, E Brooks, C Neumann, T 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
4. M Prusinowski, Z Andrews, C Neumann, T Trejos. Assessing significant factors that can influence physical fit examinations – Part I. Physical fits of torn and cut duct tapes. Forensic Science International. 2023, 343, https://doi.org/10.1016/j.forsciint.2023.111567
5. E Brooks, M Prusinowski, S Gross, T Trejos. Forensic Physical Fits in the Trace Evidence Discipline: A Review, Forensic Science International, 313, 2020, https://doi.org/10.1016/j.forsciint.2020.110349 (served as a basis for this research)
6. M Prusinowski, E Brooks, T Trejos. Development and validation of a systematic approach for the quantitative assessment of the quality of duct tape physical fits. Forensic Science International, 307, February 2020, https://doi.org/10.1016/j.forsciint.2019.110103 (served as a basis of this research)
7. Meghan Prusinowski, Ph.D. WVU Department of Forensic and Investigative Science, Enhancing the forensic comparison process of common trace materials through the development of practical and systematic methods. Graduate Theses, Dissertations, and Problem Reports. 2023, 11644. https://researchrepository.wvu.edu/etd/11644

8. Zachary Andrews, MSFS (Summer 2022), WVU Department of Forensic and Investigative Science, "Evaluating the Validity and Reliability of Textile and Paper Fracture Characteristics in Forensic Comparative Analysis" Graduate Theses, Dissertations, and Problem Reports. 2022. 11373. https://researchrepository.wvu.edu/etd/11373

9. M Prusinowski, P Tavadze, Z Andrews, L Lang, Divyanjali Pulivendhan, C Neumann, AH. Romero, T. Terjos. Experimental results on data analysis algorithms for extracting and interpreting edge feature data for duct tape and textile physical fit examination. Under review, submitted Journal of Forensic Science, June 2023

10. M Prusinowski, P Tavadze, Z Andrews, C Dolton, C Vogler.. Development of a systematic comparison method for forensic physical fit analysis of automotive polymers. Under review, submitted Forensic Chemistry, June 2023

11. Z Andrews, M Prusinowski, E Nguyen, C Neumann, T Trejos. Assessing physical fit examinations of stabbed and torn textiles through a large dataset of casework-like items and inter-laboratory studies. Under review, Journal of Forensic Sciences. Submitted May 2023

12. P Tavadze, L Lang, M Prusinowski, Z Andrews, T Trejos, and AH. Romero. Using convolutional neural networks to support examiners in duct tape physical fit comparisons. Under review, Forensic Science International. Submitted January 2022

13. National Academy of Sciences (NAS). National Academy of Sciences, Strengthening Forensic Science in the United States: A Path Forward. 2009. doi: 0.17226/12589.

14. 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/ostp/PCAST/pcast_forensic_science_report_final.pdf

15. American Statistical Association Position on Statistical Statements for Forensic Evidence [Internet]. American Statistical Association (ASA). 2019. Available from: https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf. [cited 2019 Jan 30].

16. 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.

17. RC Gonzalez, RE Woods, SL Eddins. Digital image processing using MATLAB. Pearson Prentice Hall. New Jersey. 2004.

18. T. Pavlidis. Algorithms for Graphics and Image Processing. Computer Science Press. Rockville, Maryland. 1982.

19. R Pradhan, S Kumar, R Agarwal, MP Pradhan, MK Ghose. Contour line tracing algorithm for digital topographic maps. Int J. Image Processing (IJIP). 2010. 4(2): 156-63.

20. D Peng, B Merriman, S Osher, H Zhao, M Kang. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulation. J. Comp. Physics. 1999. 155: 410-38

21. Martín~Abadi et al., "{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems." 2015, [Online]. Available: https://www.tensorflow.org/.

22. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490.

23. C. R. Harris et al., "Array programming with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

24. P. Virtanen et al., “SciPy 1.0: fundamental algorithms for scientific computing in Python,” Nat. Methods, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

25. J. D. Hunter, “Matplotlib: A 2D graphics environment,” Comput. Sci. Eng., vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

26. G. Bradski, “The OpenCV Library,” Dr. Dobb’s J. Softw. Tools, 2000.

27. Itseez, “Open Source Computer Vision Library.” 2015.

28. S. Van Der Walt et al., “Scikit-image: Image processing in python,” PeerJ, vol. 2014, no. 1, p. e453, 2014, doi: 10.7717/peerj.453.

29. D. Goodger and G. van Rossum, “PEP 0257 - Docstring Conventions.” 2001, [Online]. Available: https://www.python.org/dev/peps/pep-0257/.

30. G. van Rossum, J. Lehtosalo, and Ł. Langa, “PEP 484 – Type Hints.” 2015, [Online]. Available: https://peps.python.org/pep-0484/.

31. S. Suzuki and K. Be, “Topological structural analysis of digitized binary images by border following,” Comput. Vision, Graph. Image Process., vol. 30, no. 1, pp. 32–46, 1985, doi: https://doi.org/10.1016/0734-189X(85)90016-7.

32. Tavadze P, Lang L. romerogroup/ForensicFit: First release of ForensicFit Package [Internet]. Zenodo; 2022. Available from: https://doi.org/10.5281/zenodo.7435058

33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.

34. Agarap AF. Deep Learning using Rectified Linear Units (ReLU). 2018; Available from: http://arxiv.org/abs/1803.08375

35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.

36. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015. p. 1–15.

37. Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. In: Lecture Notes in Electrical Engineering. 2020. p. 209–21.

38. van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided Diagnosis: How to Move from the Laboratory to the Clinic. Radiology [Internet]. 2011 Dec;261(3):719–32. Available from: http://pubs.rsna.org/doi/10.1148/radiol.11091710

39. MJ Bradley, RL Keagy, PC Lowe, MP Reckenbach, DM Wright, MA LeBeau. A validation study for duct tape end matches. J. For. Sci. 2006. 51(3): 504-508.

40. MJ Bradley, JM Gauntt, AM Mehltretter, PC Lowe, DM Wright. A validation study for vinyl electrical tape end matches. J. For. Sci. 2011. 56(3): 606-611.

41. F Tulleners, J Braun. The statistical evaluation of torn and cut duct tape physical end matching. National Institute of Justice. 2011. Jul. Report No. 235287.

42. KR McCabe, FA Tulleners, JV Braun, G Currie, EN Gorecho. A quantitative analysis of torn and cut duct tape physical end matching. J. For. Sci. 2013. 58(S1): S34-S42.

43. W Ristenpart, F Tulleners, A Alfter. Quantitative algorithm for the digital comparison of torn and cut duct tape; Final Report to the National Institute of Justice Grant 2013-R2-CX-K009. University of California at Davis. 2017.

44. Y Yekutieli, Y Shor, S Wiesner, T Tsach. Physical matching verification; Final Report to United States Department of Justice on Grant 2005-IJ-R-051. National Criminal Justice Reference Service. 2012.

45. C.D. van Dijk, A. van Someren, R. Visser, M. Sjerps, Evidential value of duct tape comparison using loopbreaking patterns, Forensic Sci. Int. 332 (2022) 111178. https://doi.org/10.1016/j.forsciint.2022.111178

46. JS. Spaulding, G.M. Picconatto, Characterization of fracture match associations with automated image processing, Forensic Sci. Int. 342 (2023) 111519. https://doi.org/10.1016/j.forsciint.2022.11151

47. Deng L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Process Mag. 2012;29(6):141–2.

48. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2014;8693 LNCS(PART 5):740–55.

49. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2010. p. 248–55.

50. VR Matricardi, MS Clark, FS DeRonja. The comparison of broken surfaces: a scanning electron microscopy study. J. For. Sci. 1975. 20(3): 507-523.

51. WR Pelton. Distinguishing the cause of textile fiber damage using the scanning electron microscope (SEM). J. For. Sci. 1995. 40(5): 874-882.

52. SE Kemp, DJ Carr, J Kieser, BE Niven, MC Taylor. Forensic evidence in apparel fabrics due to stab events. For. Sci. Int. 2009. 191: 86-96.

53. K. Sloan, M. Fergusson, J. Robertson. Textile damage examinations on the cutting edge - an Australian perspective. Aus. J. For. Sci. 2018. 50(6): 682-688

54. T Gummer, K Walsh. Matching vehicle parts back to the vehicle: a study of the process. For. Sci. Int. 1996. 82: 89-97.

55. 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.

56. 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.

57. 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

58. Thompson GZ, Dawood B, Yu T, Lograsso BK, Vanderkolk JD, Maitra R, Meeker WQ, Bastawros AF. Fracture mechanics-based quantitative matching of forensic evidence fragments. arXiv preprint arXiv:2106.04809. 2021 Jun 9.

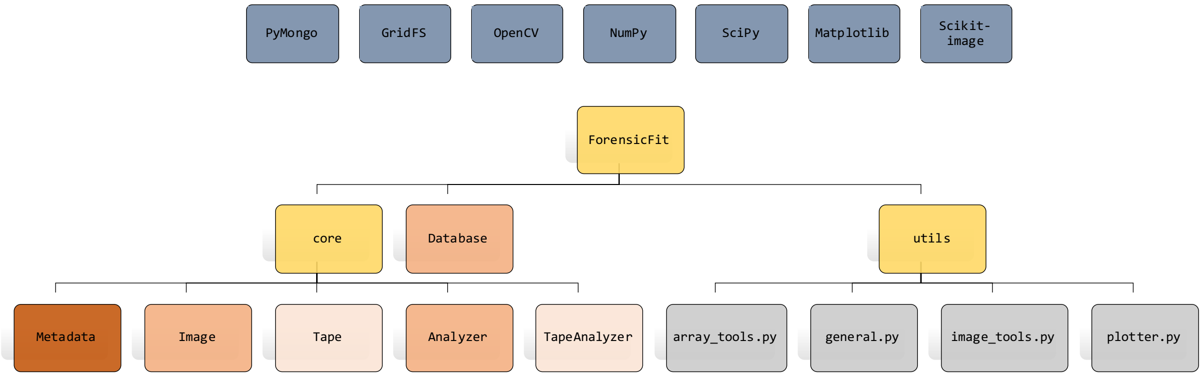
59. Dawood B, Llosa-Vite C, Thompson GZ, Lograsso BK, Claytor LK, Vanderkolk J, Meeker W, Maitra R, Bastawros A. Quantitative matching of forensic evidence fragments utilizing 3D microscopy analysis of fracture surface replicas. Journal of Forensic Sciences. 2022 May;67(3):899-910

VII SUPPLEMENTAL

Supplementary information: Using convolutional neural networks to support examiners in physical-fit comparisons in duct tape

1. ForensicFit

As mentioned in the main manuscript, the most important step in developing a successful machine-learning model is data preparation. The first step is to develop a well-controlled and efficient database where the user can store, query, analyze, and use the data created for a particular application. ForensicFit uses state-of-the-art image processing methods to analyze and store the generated data. The data is compatible with well-known machine-learning packages such as TensorFlow[1], PyTorch, and SciKit-learn[2]. It utilizes NumPy[3], SciPy[4], matplotlib[5], OpenCV[6], [7], scikit-image[8], PyMongo, and GridFS. For ease of use and future development, the package follows PEP-257[9] and PEP-484[10] for documentation and type hints, respectively. ForensicFit uses state-of-the-art image processing methods to analyze and store the generated data. The package is organized into three main sub-packages, *core*, *database*, and *utils*. *core,* as the name suggests, contains the most important functionalities within the package. It contains python classes that manage the read/write, analysis, and metadata storage. These classes provide a skeleton for the data structure used in the package. Moreover, they define the standards for future implementations used for different types of materials. *database*, provides an efficient and flexible method for storing and retrieving the raw and preprocessed data. The functionality of the rest of the package does not depend on this sub-package. It was added merely to simplify the storage and query process of the data. One can still store and access the raw or analyzed data using the traditional image storage approaches. Lastly, *utils*, contains all the image manipulation, plotting, and memory access tools that are used in different sections of the package.



The package is organized into three main sub-packages, *core*, *database*, and *utils*. Each sub-package contains python classes and their methods (functions). Additional information on a selected number of important methods is also provided. For more information on the usage and functionality of each method please refer to the package documentation. Additionally, three stand-alone python scripts accompany the package for batch processes, *create_metadata.py*, *preprocess_bin_based.py*, *store_on_db.py*. The following is an introduction to the package structure, modules, and their purposes in the preprocessing.

1.1.Core

As the name suggests, this sub-package contains the most important functionalities within this package. The fundamental classes are *Metadata*, *Image*, and *Analyzer*. These parent classes provide a skeleton for the data structure used in the package. Moreover, they define the standards for future implementations used for different types of materials. *Metadata* is used as an attribute in both classes *Image* and *Analyzer*. For application to duct tapes, two children classes were defined: *Tape* and *TapeAnalyzer*. *Tape* inherits all the methods and attributes of the class *Image*, while *TapeAnalyzer* inherits those of the class *Analyzer*. Thus, we will introduce the child and parent classes together.

1.1.1.Metadata

Metadata is a mapping that stores details about their objects. These details include but are not limited to, tape name, file location, image resolution, tape quality, and separation method. The metadata is stored in the database with the object and can be used as filters for querying.

1.1.2.Tape (Image)

The *Tape* class contains data from scanned images, and each instance of this class represents a tape sample. It also contains methods that handle the read/write as well as some basic image manipulation tasks. This class can be instantiated directly by providing the image data as a NumPy array. Additionally, the instantiation can occur by one of the following classmethods: *from_file, from_buffer,* and *from_dict*. The methods included in this class to help image processing are *isolate*, *crop*, *convert_to_gray*, *convert_to_rgb*, *rotate*, *resize*, *flip_h*, *flip_v*, *plot*, *exposure_control*, *apply_filter*, and *split_v*. *Split_v* divides the image vertically into two sections and retains the requested side. This is especially important for images of tapes because only one edge of is relevant at a time. Because of the size of the tape and the placement of the tape in the scanner the location in which the splitting must happen varies from one image to the other. This is taken care of by finding the location of the boundaries of the tape using the *TapeAnalyzer* and splitting the image at the midpoint of the boundary.

As mentioned before *Metadata* is used as an attribute in this class. It carries out the task of storing information about the tape and any further analysis performed on it.

1.1.3. TapeAnalyzer (Analyzer)

This class receives the *Tape* (*Image*) class as input and performs further processing to prepare input data for machine-learning. In addition to instantiation directly by providing the *Tape* class, it can use its classmethod *from_dict*. The method *from_dict* uses a python dictionary as an input. This method is very useful when retrieving previously saved data from the database. The most important methods that are called upon instantiation in this class are *preprocess*, *get_image_tilt*, *auto_crop_y*.

- *Preprocess* finds the boundary of the tape. It first binarizes the image and uses the algorithm introduced by Suzuki and Abe[11] implemented in the OpenCV package to find all the contours in the image. The largest contour is assumed as the boundary of the tape.
- *Get_image_tilt* finds the angle of the scanned image with the horizontal line. This is done by dividing the top and bottom edges of the boundary of the tape into multiple segments. Using a linear fit the slope and standard deviation (in the *y*-direction) of each segment are calculated. The two segments with the least standard deviation are selected (one from the top and one from the bottom). The angle is calculated from the average slope of these two lines.

- *auto_crop_y* crops the image in the *y*-direction based on the information gathered in the *preprocess* and *get_image_tilt*. This is done carefully to avoid any additional noise introduced by protruding weft or warp fibers.

Now that the boundaries of the tape have been located, the data can be prepared for the machine-learning process. This task is handled by one of three methods in this class, *get_coordinate_based*, *get_bin_based*, *get_max*_contrast. After calling any of the aforementioned methods, the data can be accessed like a python mapping (*e.g., TapeAnalyzer[‘bin_based’]*) or a class attribute (*e.g., TapeAnalyzer.bin_based*). The following is a brief description of the functionality of each method.

- Coordinate-based quantifies the most important area of the tape, the comparison edge, into a collection *(x, y)* of coordinates. This is done by first identifying the comparison edge by dividing the boundary of the tape in *x* direction into *n* sections and only keeping the leftmost section. The comparison edge is now described by a collection of points in the *x-y* plane. The number of points describing this edge varies by a great deal because of the ragged nature of the comparison edge. To provide more consistency in the machine-learning input data, a necessity in neural networks, this method asks for the number of output points. It divides the comparison edge into small windows and stores the following three values for each window: 1) average values of the *x* and *y* coordinates of the points in the window; 2) standard deviation of the points in the *x* direction; 3) slope of the linear fit to collection of the points.
- Bin-based represents the comparison by many smaller images selected from the area around the edge. This can be visualized by picturing a rectangular window that sweeps over the comparison edge and stores the images inside the window. The width (*y*-direction) of this window is defined automatically. By providing the number of bins, *n_bins*, it will divide the width of the tape into *n* bins. The length, however, must be defined by the user. The tape edge divides the image in the window into two parts, the tape, and the background. The method receives two variables *window_tape* — number of pixels from the tape edge towards the background — and *window_background* — number of pixels from the tape edge towards the tape. The length of the window, therefore, is the addition of the two variables, *window_background,* and *window_tape*.
- Max-contrast represents the comparison edge by maximizing the contrast between the edge and the rest of the image. This is done by setting the values of the pixels on the comparison edge to their maximum (255) and assigning the minimum value (0) to the remaining pixels.

Similar to the *get_bin_based* method, *get_max_contrast,* receives the two arguments *window_background* and *window_tape*.

It is worth mentioning that these methods can be combined. For example, if one wants to retrieve the coordinate-based data of each bin in the bin-based analysis, one can use the mapping *TapeAnalyzer['bin_based+coordinate_based']*. Other methods used in image processing are *flip_v*, *flip_h*. The data generated for this study was through the *get_bin_based* method by selecting only one bin to represent the whole tape as a big-picture examination. The *window_background* and *window_tape* was selected at 10 and 400 pixels, respectively.

1.2.Database

This sub-package provides an efficient and flexible method for storing and retrieving the raw and preprocessed data. The functionality of the rest of the package does not depend on this sub-package. It was added merely to simplify the storage and query process of the data. One can still store and access the raw and analyzed data using the traditional image storage approaches. There are three different types of data in this study, raw data (scanned images), analysis data (*e.g.*, bin-based, or maximum contrast) of the tape, and array-like data (*e.g.*, coordinate-based). As the types of these data are very different from each other, we use the flexible document-oriented database, MongoDB. The raw data (classes *Tape* and *Image*) is stored using the GridFS storage specification of MongoDB, while the rest of the data types are stored using the standard JSON-like documents used in MongoDB. The type of data is referred to as *mode* in this database. The GridFS is efficient as it divides documents (files larger than 16 MB) into multiple "chunks". This is beneficial when accessing portions of the file content, such as a small bin of the comparison edge of a tape. Moreover, the "chunks" are accompanied by a "metadata" collection. This approach is very suitable for this study as the statistical analysis is often categorized by different types of materials. For instance, information like high-quality, scissor-cut, scrim side, etc., can be stored in the metadata. This will improve the efficiency of the query process. This sub-package contains two classes *ClassMap* and *Database.*

1.2.1. ClassMap

This class is simply a mapping from the stored information about the data type to its corresponding python object. This mapping helps the *Database* class recognize which core class needs instantiation. As mentioned before there are different types of data in this study. This class was introduced to help generalize the code, *i.e.,* to avoid writing a database class for each data type. Each data entry contains information about its data type (*mode*) prior to storage. *ClassMap* links this information, a python string, to a python object.

1.2.2.Database

The class database is instantiated by providing the name, host, and port of the database. The instantiation does not create a database it will only connect to the MongoDB server. The database is created, if it does not exist, the moment a document is provided for storage. The two necessary methods in any database class are means to store (put) and retrieve (find) data on the database. These operations are performed using the *insert* and *find* methods.

- *Insert* stores any object from the sub-package *core* in the database. This has the option to overwrite, skip, or create a new document if the same object already exists. It is worth mentioning that overwriting (update) does not exist in GridFS. For practicality, we have added this option which simply finds the duplicate and removes it before inserting a new document. After storage, each document is given a unique id.

- *find* queries for all documents matching the provided filter. It then returns them as a list of objects from the core sub-package. There are two additional methods that perform this task, *find_one,* and *find_with_id*. Both methods return an object instead of a list of objects.

Other useful methods from this class include *map_to*, *filter_with_metadata*, *count_documents*, *export_to_files*, *drop_collection*, *delete*, and *delete_database.*

This sub-package contains two important functions, *dump* and *restore* which export and import the database, respectively.

1.2.3. Utils

This sub-package is a collection of tools for manipulating images and arrays as well as plotting. These tools are not developed in this work, they have been gathered under this sub-package for ease of access. The names of each function have been selected such that the functionality would be self-explanatory. This sub-package contains three main modules: *array_tools.py, image_tools.py,* and *plotting_tools.py*. The following is a list of functions in each module.

- *Array_tools.py* contains *vote_calculator*; *read_bytes_io*; and *write_bytes_io.*
- *Image_tools.py* contains *rotate_image; gaussian_blur; split_v; to_gray; to_rbg; contours; largest_contour; remove_background; get_masked; resize; exposure_control; apply_filter; binerized_mask,* and *imwrite.*
- *Plotting_tools.py* contains *get_figure_size; plot_coordinate_based; plot_pairs; plot_confusion_matrix; plot_kde_distribution; plot_hist_distribution;* and *plot_metrics.*

2. Fit to non-fit ratio determination

First, the best **Error! Reference source not found.**fit/non-fit ratio for training the model is determined. For this test, a form of 5-fold cross-validation[12] is applied. The tape fit pairs are split into five folds, with four folds acting as the training set and the last fold as the validation set. This is performed five times, switching out each fold for the testing set. This will judge how well the model generalizes to random datasets. Next, the non-fits are generated from the fits for the training and validation up to a certain fit/non-fit ratio. For the validation set, this ratio is kept constant at 0.5. The results of the test are shown in **Error! Reference source not found.**. The ratio 0.3 performs the best. This is not surprising since as the ratio decreases, the number of non-fits increases. This will cause the model to be more certain when predicting fits, thus subsequently higher positive predictive value and lower true positive rate.

Table 1 5-fold cross validated test for fit to non-fit ratio of 2.5:10.

2.5:10	False-positive rate	False-negative rate	True-negative rate	True-positive rate	Accuracy
HQHT	0.024 (0.057)	0.710 (0.837)	0.976 (0.943)	0.290 (0.163)	0.899 (0.786)
HQSC	0.073 (0.078)	0.269 (0.363)	0.927 (0.922)	0.731 (0.637)	0.872 (0.806)
MQHT	0.042 (0.048)	0.538 (0.621)	0.958 (0.952)	0.462 (0.379)	0.863 (0.766)
MQSC	0.043 (0.067)	0.346 (0.439)	0.957 (0.933)	0.654 (0.561)	0.902 (0.818)
LQHT	0.029 (0.037)	0.426 (0.531)	0.971 (0.963)	0.574 (0.469)	0.894 (0.806)

LQSC	0.094 (0.142)	0.311 (0.406)	0.906 (0.858)	0.689 (0.594)	0.857 (0.761)
HQ	0.048 (0.066)	0.381 (0.482)	0.952 (0.934)	0.619 (0.518)	0.885 (0.798)
MQ	0.042 (0.060)	0.442 (0.528)	0.958 (0.940)	0.558 (0.472)	0.883 (0.791)
LQ	0.062 (0.091)	0.363 (0.454)	0.938 (0.909)	0.637 (0.546)	0.875 (0.782)
Total	0.051 (0.071)	0.393 (0.486)	0.949 (0.929)	0.607 (0.514)	0.881 (0.791)

Table 2 5-fold cross validated test for fit to non-fit ratio of 3:10.

3:10	False-positive rate	False-negative rate	True-negative rate	True-positive rate	Accuracy
HQHT	0.070 (0.123)	0.367 (0.513)	0.930 (0.877)	0.633 (0.487)	0.890 (0.793)
HQSC	0.092 (0.094)	0.152 (0.244)	0.908 (0.906)	0.848 (0.756)	0.889 (0.836)
MQHT	0.054 (0.098)	0.222 (0.328)	0.946 (0.902)	0.778 (0.672)	0.910 (0.835)
MQSC	0.046 (0.041)	0.071 (0.105)	0.954 (0.959)	0.929 (0.895)	0.949 (0.937)
LQHT	0.032 (0.062)	0.186 (0.394)	0.968 (0.938)	0.814 (0.606)	0.933 (0.836)
LQSC	0.146 (0.216)	0.125 (0.193)	0.854 (0.784)	0.875 (0.807)	0.859 (0.789)
HQ	0.081 (0.109)	0.207 (0.316)	0.919 (0.891)	0.793 (0.684)	0.890 (0.818)
MQ	0.050 (0.062)	0.147 (0.217)	0.950 (0.938)	0.853 (0.783)	0.929 (0.891)
LQ	0.090 (0.140)	0.152 (0.281)	0.910 (0.860)	0.848 (0.719)	0.895 (0.811)
Total	0.074 (0.104)	0.169 (0.273)	0.926 (0.896)	0.831 (0.727)	0.904 (0.840)

Table 3 5-fold cross validated test for fit to non-fit ratio of 3.5:10.

3.5:10	False-positive rate	False-negative rate	True-negative rate	True-positive rate	Accuracy
HQHT	0.062 (0.115)	0.430 (0.593)	0.938 (0.885)	0.570 (0.407)	0.882 (0.788)
HQSC	0.102 (0.122)	0.128 (0.216)	0.898 (0.878)	0.872 (0.784)	0.889 (0.839)
MQHT	0.093 (0.147)	0.329 (0.423)	0.907 (0.853)	0.671 (0.577)	0.849 (0.763)
MQSC	0.071 (0.069)	0.214 (0.316)	0.929 (0.931)	0.786 (0.684)	0.895 (0.852)
LQHT	0.085 (0.157)	0.168 (0.275)	0.915 (0.843)	0.832 (0.725)	0.895 (0.801)
LQSC	0.169 (0.211)	0.149 (0.178)	0.831 (0.789)	0.851 (0.822)	0.837 (0.805)
HQ	0.082 (0.118)	0.206 (0.311)	0.918 (0.882)	0.794 (0.689)	0.886 (0.818)
MQ	0.082 (0.110)	0.272 (0.372)	0.918 (0.890)	0.728 (0.628)	0.872 (0.806)
LQ	0.126 (0.185)	0.157 (0.221)	0.874 (0.815)	0.843 (0.779)	0.866 (0.803)
Total	0.096 (0.138)	0.208 (0.299)	0.904 (0.862)	0.792 (0.701)	0.875 (0.809)

Table 4 5-fold cross validated test for fit to non-fit ratio of 4:10.

4:10	False-positive rate	False-negative rate	True-negative rate	True-positive rate	Accuracy
HQHT	0.077 (0.159)	0.475 (0.658)	0.923 (0.841)	0.525 (0.342)	0.853 (0.736)
HQSC	0.103 (0.121)	0.137 (0.202)	0.897 (0.879)	0.863 (0.798)	0.884 (0.841)
MQHT	0.072 (0.132)	0.336 (0.394)	0.928 (0.868)	0.664 (0.606)	0.859 (0.787)
MQSC	0.078 (0.049)	0.240 (0.260)	0.922 (0.951)	0.760 (0.740)	0.879 (0.886)
LQHT	0.058 (0.131)	0.235 (0.340)	0.942 (0.869)	0.765 (0.660)	0.894 (0.797)
LQSC	0.167 (0.196)	0.191 (0.263)	0.833 (0.804)	0.809 (0.737)	0.825 (0.784)
HQ	0.090 (0.143)	0.223 (0.316)	0.910 (0.857)	0.777 (0.684)	0.871 (0.797)

MQ	0.075 (0.091)	0.288 (0.327)	0.925 (0.909)	0.712 (0.673)	0.869 (0.837)
LQ	0.114 (0.162)	0.209 (0.305)	0.886 (0.838)	0.791 (0.695)	0.858 (0.790)
Total	0.093 (0.131)	0.238 (0.315)	0.907 (0.869)	0.762 (0.685)	0.866 (0.808)

3. Artificial Neural networks:

Artificial neural networks are machine learning methods that were inspired by the workings of biological brain. In the brain, neurons are connected through synapses, which are junctions where electrical or chemical signals are transmitted from one neuron to another. The strength of these connections, or synaptic weights, determines how the signals are transmitted and processed. Similarly, in artificial neural networks, the connections between artificial neurons are represented by weights. These weights determine how the input signals are combined and processed as they pass through the network. The weights are adjusted during the training process, allowing the neural network to learn patterns and make predictions based on the input data[13].

3.1. Architecture

The architecture of a neural network defines how the nodes are connected to each other.

3.2 Single layer (Perceptron)

The perceptron is the simplest form of an artificial neural network. It is often considered as the foundation of more complex machine learning algorithms. It comprises an input layer, which is a collection of nodes, and a single output node.

The input layer receives values that represent various features or characteristics of the data we want to analyze. Each input value is multiplied by its corresponding weight, which indicates the importance of that input in making a decision or prediction. The weighted inputs are then combined, and the sum is fed to the output node.

At the output node, the sum is processed through a function called the activation function. This function converts the sum into a final output value, which represents the perceptron's decision or prediction[13]–[15].

3.2. Fully connected neural networks

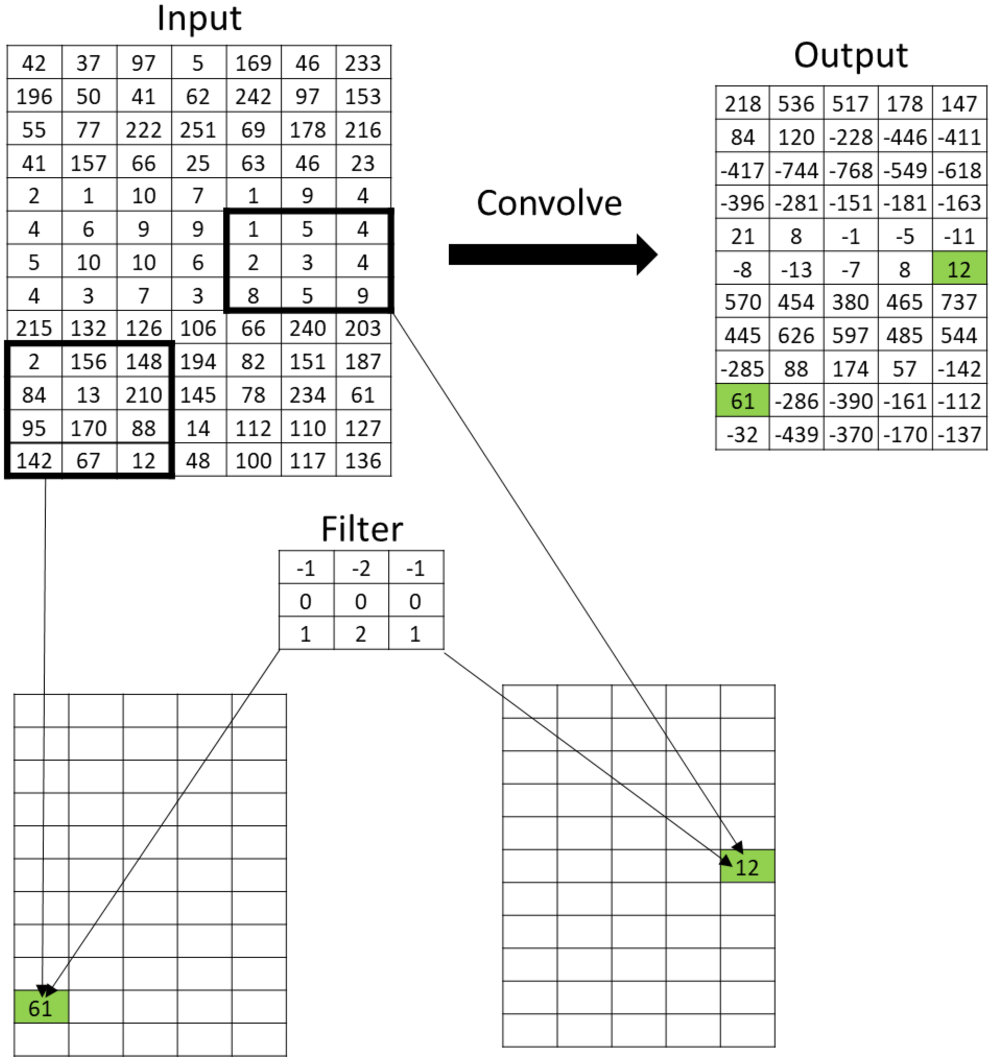
Fully connected neural networks, also known as dense neural networks, are an extension of the basic perceptron model (Section 3.1.1.1). This extension takes place in two main stages, enhancing the capabilities of the network for more complex tasks.

The first stage involves expanding the output layer to consist of multiple nodes instead of just one. In this case, each input node is connected to every output node, with each connection having its own unique set of weights. This creates a one-layer network with multiple input nodes and multiple output nodes, allowing for a broader range of decisions or predictions.

The second stage of expansion introduces intermediate layers, known as hidden layers, between the input and output layers. The output nodes from the previous layer serve as inputs for the next layer. This multi-layer structure enables the network to learn more complex patterns and representations from the input data, ultimately leading to more accurate predictions and better performance [13]–[15].

3.4. Convolutional neural networks:

Convolutional neural networks (CNNs) are specialized neural networks designed for handling data with grid-like structures, where data points on the grid exhibit spatial dependencies with their neighbors. A prime example of such data is a 2-dimmensional image, where the pixels representing a specific feature or pattern share strong spatial connections. A features of CNNs is their ability to capture the translational symmetry in data, such as an image. This means that the importance of a particular pattern or feature (*e.g.* comparison edge of a duct tape) remains unchanged regardless of its location within the image, whether it is at the center or in one of the corners. The common layers used in CNNs are convolution, pooling.

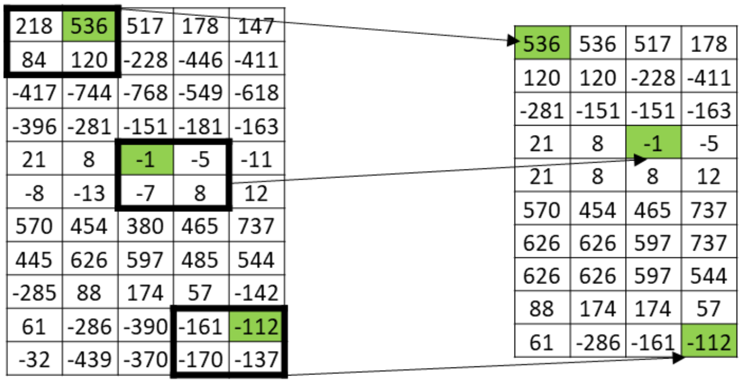


42	37	97	5	169	46	233
196	50	41	62	242	97	153
55	77	222	251	69	178	216
41	157	66	25	63	46	23
2	1	10	7	1	9	4
4	6	9	9	1	5	4
5	10	10	6	2	3	4
4	3	7	3	8	5	9
215	132	126	106	66	240	203
2	156	148	194	82	151	187
84	13	210	145	78	234	61
95	170	88	14	112	110	127
142	67	12	48	100	117	136

-1	-2	-1
0	0	0
1	2	1

218	536	517	178	147
84	120	-228	-446	-411
-417	-744	-768	-549	-618
-396	-281	-151	-181	-163
21	8	-1	-5	-11
-8	-13	-7	8	12
570	454	380	465	737
445	626	597	485	544
-285	88	174	57	-142
61	-286	-390	-161	-112
-32	-439	-370	-170	-137

Figure 1 An example of convolution applied with a 3×3 filter and a stride of 1×1. This figure was inspired by a figure from reference [aggarawal]



218	536	517	178	147
84	120	-228	-446	-411
-417	-744	-768	-549	-618
-396	-281	-151	-181	-163
21	8	-1	-5	-11
-8	-13	-7	8	12
570	454	380	465	737
445	626	597	485	544
-285	88	174	57	-142
61	-286	-390	-161	-112
-32	-439	-370	-170	-137

536	536	517	178
120	120	-228	-411
-281	-151	-151	-163
21	8	-1	-5
21	8	8	12
570	454	465	737
626	626	597	737
626	626	597	544
88	174	174	57
61	-286	-161	-112

Figure 2 An example of max-pooling with a 2x2 window and a stride of 1x1.

- Convolution layer: In CNNs, rectangular filters are used to scan the image, examining all pixels to detect various patterns within the filter's window. This window is often referred to as the filter or kernel. The algorithm scans the image using a predefined number of steps, called the stride, which determines how many pixels the window will move across the image after each convolution calculation. The stride is typically chosen by the designer of the CNN architecture and helps balance the trade-off between computational efficiency and the ability to capture finer details. Figure 1 shows an example of this process.

- Pooling layer: To reduce the spatial dimensions of the feature map while retaining essential information a scheme is employed to map a window of pixel values to one value. The most common approach is called max-pooling where the highest value within a rectangular window is selected. Figure 2 demonstrates this process.

3.5. Activation functions

Activation functions play a crucial role in neural networks, as they process the inputs (from the nodes of the previous layer) using the corresponding weights to produce a single output value. The choice of an activation function is a critical aspect of neural network design [13]. Several activation functions are available, such as sign, sigmoid, tanh, ReLU, and Hard Tanh. In this appendix, we will explain the functions used in this study.

- Rectified Linear Unit (ReLU): ReLU is one of the most popular activation functions. It is defined as $f(\mathbf{x}) = \max(0, \sum w_i x_i)$. In other words, it calculates the weighted sum of inputs and returns it if it's larger than zero, otherwise it returns zero ReLU functions have gained popularity in recent years due to their computational efficiency and scale invariance.
- Sigmoid (logistic) function: The sigmoid function is a widely-used activation function defined as $f(x) = \frac{1}{1+\exp(-\sum w_i x_i)}$. The output of a logistic function always falls between zero and one, making it a good choice for binary classification tasks.

3.6. Loss function, optimizers and learning rates

In practical machine learning problems, the goal is to find a mapping that takes a set of input variables and generates output results by identifying patterns in the available input and output data (*i.e.*, training data). The loss function is a tool used to evaluate the effectiveness of the mapping in predicting known data. Often, the best mapping is achieved by optimizing the loss function.

A well-known example of a loss function is the squared loss: $L(y, \hat{y}) = \sum (y_i - \hat{y}_i)^2$. By minimizing this loss function, we encourage the predictions $(\hat{y}_i)$ to be closer to the actual data (y_i).

Another common loss function is binary cross-entropy, also known as log loss. It measures the difference between the true labels and the predicted probability of an instance belonging to a certain class. Binary cross-entropy is defined as $L(y, p) = -\sum (y_i \ln p_i + (1 - y_i) \ln(1 - p_i)$ [14], [16]–[18]. In this formula y_i represents the actual label of the i-th instance, and p_i, the predicted probability of that instance belonging to one of the classes.

The process of minimizing the loss function involves adjusting the model parameters until the desired accuracy is achieved. For instance, in the case of a best-fit line, these parameters are the slope and the y-intercept of a line, while in a neural network, the parameters are the weights of the nodes. The minimization of the loss function is performed using various computationally efficient algorithms that have been developed over time, such as gradient descent. These algorithms are called optimizers.

Gradient descent iteratively updates the model parameters based on the gradient of the loss function. By using the gradient, the algorithm can identify the direction in which the loss function decreases most rapidly (steepest decline) and adjust the parameters to minimize the loss and improve the predictions' accuracy with respect to the training data. A crucial aspect of this process is determining the step size for modifying the parameters. The algorithm must balance between making changes that are not too large, which may cause it to overshoot the minimum, and not too small, which would compromise computational efficiency. This step size is referred to as the learning rate.

An extension of gradient descent commonly used in neural networks is the Adaptive Moment Estimator (ADAM). ADAM dynamically adjusts the learning rate for each parameter, enabling faster convergence and improved performance across various problems. It achieves this by computing adaptive learning rates for each parameter using the first and second moments of the gradients, resulting in more efficient and effective optimization.

3.7. Dropout

Dropout is a method to overcome the challenges of overfitting. In this method at a certain layer of the fully connected network, a percentage of the nodes (usually from the hidden layers) are randomly removed. If a node is dropped, all the incoming and outgoing connections from that node need to be dropped as well [13].

3.8. Convolutional neural network

Figure 3 demonstrates the CNN used in this study.

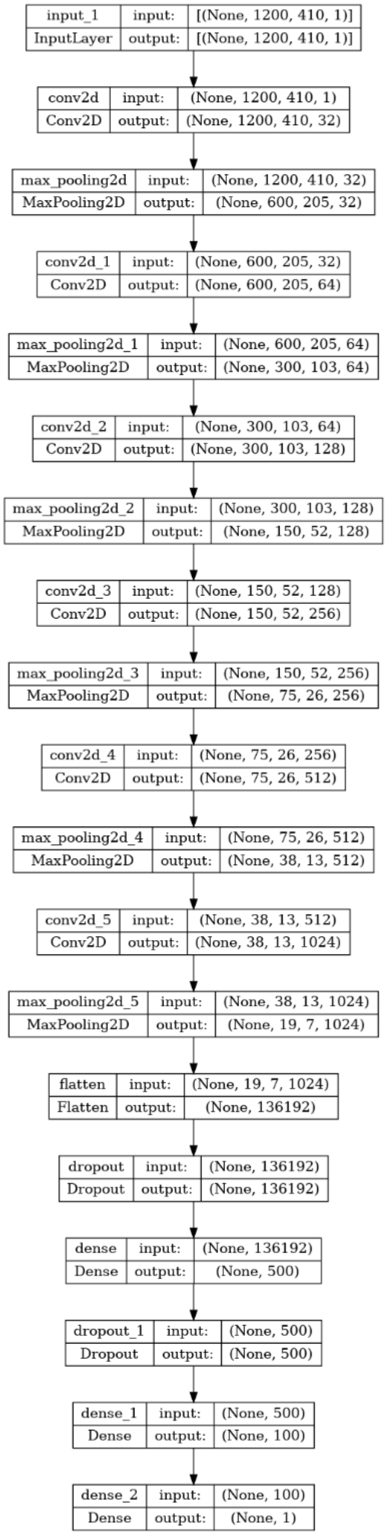

input_1 | input: | [(None, 1200, 410, 1)]
InputLayer | output: | [(None, 1200, 410, 1)]
conv2d | input: | (None, 1200, 410, 1)
Conv2D | output: | (None, 1200, 410, 32)
max_pooling2d | input: | (None, 1200, 410, 32)
MaxPooling2D | output: | (None, 600, 205, 32)
conv2d_1 | input: | (None, 600, 205, 32)
Conv2D | output: | (None, 600, 205, 64)
max_pooling2d_1 | input: | (None, 600, 205, 64)
MaxPooling2D | output: | (None, 300, 103, 64)
conv2d_2 | input: | (None, 300, 103, 64)
Conv2D | output: | (None, 300, 103, 128)
max_pooling2d_2 | input: | (None, 300, 103, 128)
MaxPooling2D | output: | (None, 150, 52, 128)
conv2d_3 | input: | (None, 150, 52, 128)
Conv2D | output: | (None, 150, 52, 256)
max_pooling2d_3 | input: | (None, 150, 52, 256)
MaxPooling2D | output: | (None, 75, 26, 256)
conv2d_4 | input: | (None, 75, 26, 256)
Conv2D | output: | (None, 75, 26, 512)
max_pooling2d_4 | input: | (None, 75, 26, 512)
MaxPooling2D | output: | (None, 38, 13, 512)
conv2d_5 | input: | (None, 38, 13, 512)
Conv2D | output: | (None, 38, 13, 1024)
max_pooling2d_5 | input: | (None, 38, 13, 1024)
MaxPooling2D | output: | (None, 19, 7, 1024)
flatten | input: | (None, 19, 7, 1024)
Flatten | output: | (None, 136192)
dropout | input: | (None, 136192)
Dropout | output: | (None, 136192)
dense | input: | (None, 136192)
Dense | output: | (None, 500)
dropout_1 | input: | (None, 500)
Dropout | output: | (None, 500)
dense_1 | input: | (None, 500)
Dense | output: | (None, 100)
dense_2 | input: | (None, 100)
Dense | output: | (None, 1)

Figure 3 Convolutional neural network architecture.

4. Decision trees

Decision trees, also known as classification and regression trees (CART), are a type of supervised machine learning algorithm. They simplify complex decision-making processes by breaking them down into smaller, more manageable steps. This is achieved by recursively partitioning the feature space into a set of rectangles and assigning a constant value to each. This process can be visualized as a tree with multiple leaves, each representing a distinct region in the feature space [19].

Decision trees have various hyperparameters that need to be fine-tuned for optimal performance. The most important hyperparameters include the criterion used to measure the quality of a split, the splitting method applied to partition the data, and the maximum depth of the tree. Adjusting these hyperparameters can help achieve the right balance between model complexity and accuracy.

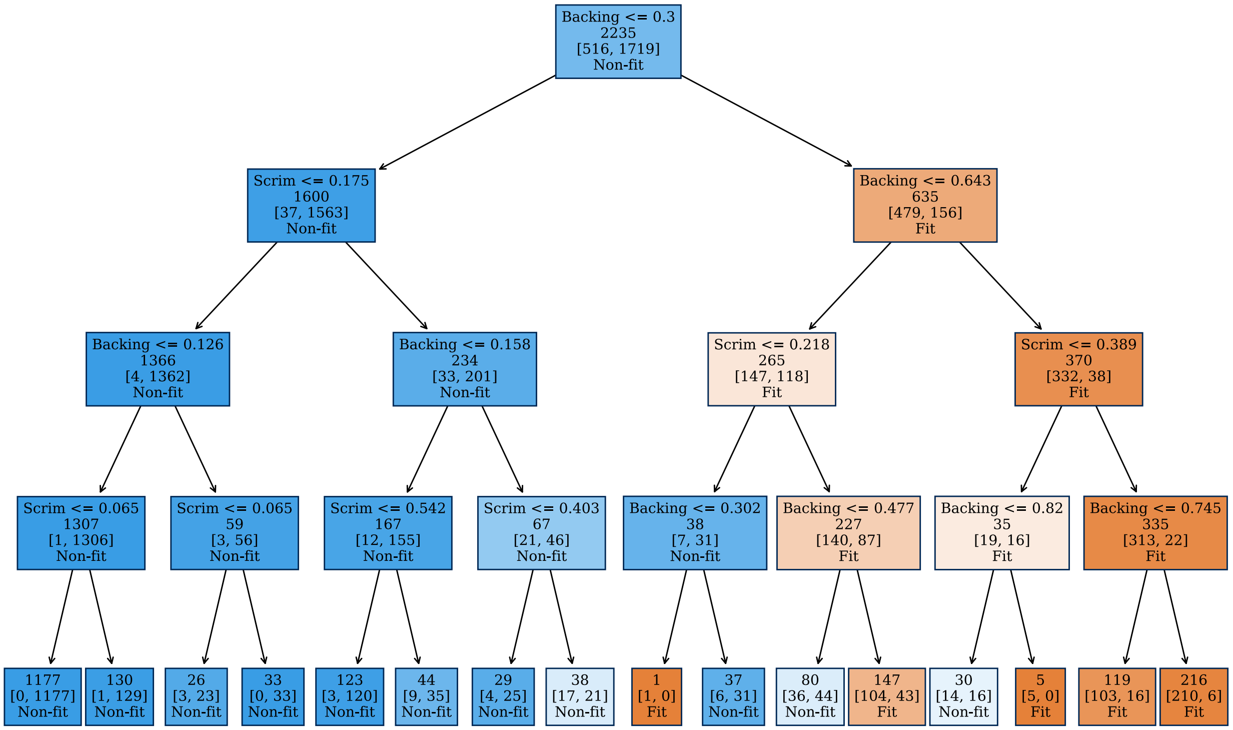


Figure 4 The decision tree used to combine outputs of the scrim and backing convolutional neural networks.

5. Merging results of Scrim and Backing CNNs

As described in the main manuscript, two separate CNNs were trained on scanned images of the scrim and backing sides of duct tapes. To combine the results of these two CNN models, various supervised learning approaches were evaluated. These algorithms include Gradient Boosting Classifier (GBC), K-nearest neighbors (KNN), Decision tree (DT), Support vector machine (SVM), logistic regression (LR), random forest (RF), and AdaBoost (ADA).

The decision tree algorithm was chosen, taking into account the separation of the distribution of the membership probabilities assigned to true fits and true non-fits, as well as the statistical accuracy. Figure 5 displays the kernel density estimation of the distribution of these probabilities. Table 4 presents the performance of these classifiers, as measured by various statistical metrics. By selecting the decision tree method, we aim to achieve a balance between model complexity and accuracy in merging the results from the two CNN models.

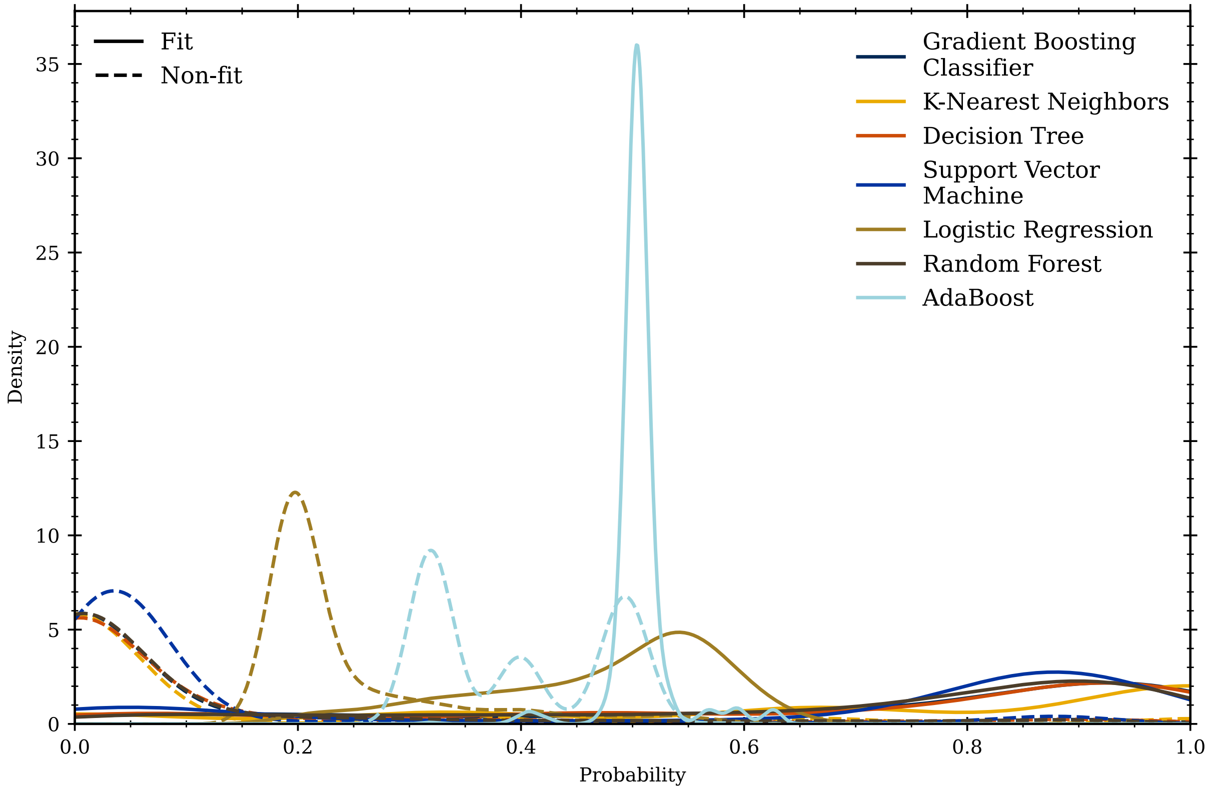


Figure 5 Kernel density estimation of the membership probabilities assigned to true fit and true non-fit by different classifiers.

Table 5 Performance of classifiers used to combine the results of the scrim and backing CNNs. The color map for true-positive rate (TPR), true-negative rate (TNR), and accuracy (ACC) ranges from purple to blue (low to high). Conversely, the color map for false-negative rate (FNR), false-positive rate (FPR), inconclusive-negative rate (INR), and inconclusive-positive rate (IPR_ transitions from blue to purple (low to high). In essence, blue signifies improvement, while purple indicates a decline. The classifiers evaluated include Gradient Boosting Classifier (GBC), K-nearest neighbors (KNN), Decision tree (DT), Support vector machine (SVM), logistic regression (LR), random forest (RF), and AdaBoost (ADA).

	GBC	KNN	DT	SVM	LR	RF	ADA

TPR	0.734 ± 0.036	0.716 ± 0.045	0.729 ± 0.077	0.726 ± 0.044	0.491 ± 0.073	0.737 ± 0.053	0.223 ± 0.062
TNR	0.916 ± 0.013	0.918 ± 0.023	0.914 ± 0.018	0.926 ± 0.011	0.962 ± 0.009	0.919 ± 0.015	0.756 ± 0.045
FNR	0.249 ± 0.031	0.284 ± 0.045	0.268 ± 0.080	0.271 ± 0.042	0.444 ± 0.086	0.254 ± 0.050	0.055 ± 0.008
FPR	0.077 ± 0.015	0.082 ± 0.023	0.085 ± 0.017	0.074 ± 0.011	0.033 ± 0.007	0.079 ± 0.015	0.012 ± 0.007
INR	0.007 ± 0.006	0.000 ± 0.000	0.001 ± 0.002	0.000 ± 0.000	0.005 ± 0.003	0.002 ± 0.003	0.232 ± 0.050
IPR	0.017 ± 0.011	0.000 ± 0.000	0.003 ± 0.007	0.003 ± 0.004	0.066 ± 0.023	0.009 ± 0.007	0.723 ± 0.063
ACC	0.855 ± 0.017	0.851 ± 0.018	0.853 ± 0.023	0.859 ± 0.016	0.805 ± 0.024	0.858 ± 0.022	0.578 ± 0.048

6. References for the supplemental section

[1] Martín~Abadi *et al.*, "{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems." 2015. [Online]. Available: https://www.tensorflow.org/

[2] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490

[3] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[4] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[5] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[6] G. Bradski, "The OpenCV Library," *Dr. Dobb's J. Softw. Tools*, 2000.

[7] Itseez, "Open Source Computer Vision Library." 2015.

[8] S. Van Der Walt *et al.*, "Scikit-image: Image processing in python," *PeerJ*, vol. 2014, no. 1, p. e453, 2014, doi: 10.7717/peerj.453.

[9] D. Goodger and G. van Rossum, "PEP 0257 - Docstring Conventions." 2001. [Online]. Available: https://www.python.org/dev/peps/pep-0257/

[10] G. van Rossum, J. Lehtosalo, and Ł. Langa, "PEP 484 – Type Hints." 2015. [Online]. Available: https://peps.python.org/pep-0484/

[11] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vision, Graph. Image Process.*, vol. 30, no. 1, pp. 32–46, 1985, doi: https://doi.org/10.1016/0734-189X(85)90016-7.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY:

Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[13] C. C. Aggarwal, *Neural Networks and Deep Learning*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-94463-0.

[14] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1996.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[16] J. J. Hopfield, "Learning algorithms and probability distributions in feed-forward and feed-back networks," *Proc. Natl. Acad. Sci.*, vol. 84, no. 23, pp. 8429–8433, Dec. 1987, doi: 10.1073/pnas.84.23.8429.

[17] E. B. Baum and F. Wilczek, "Supervised Learning of Probability Distributions By Neural Networks.," in *Neural Information Processing Systems*, 1987, vol. 0, p. 20. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1987/file/eccbc87e4b5ce2fe28308fd9f2a7baf3-Paper.pdf

[18] S. A. Solla, E. Levin, and M. Fleisher, "Accelerated learning in layered neural networks," *Complex Syst.*, vol. 2, no. 6, pp. 625–639, 1988.

[19] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.