Forensic Science International 353 (2023) 111884

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Using convolutional neural networks to support examiners in duct tape physical fit comparisons

Logan Lang [b], Pedram Tavadze [a], Meghan Prusinowski [a], Zachary Andrews [a], Cedric Neumann [c], Tatiana Trejos [a], Aldo H. Romero [b,*]

[a] West Virginia University, Department of Forensic and Investigative Science, Morgantown, WV 26506, USA
[b] West Virginia University, Department of Physics and Astronomy, Morgantown, WV 26506, USA
[c] Battelle Memorial Institute, Columbus, OH, USA

ARTICLE INFO

Keywords:
Physical fit
Duct tape
Machine learning
Convolutional neural networks
Decision tree

ABSTRACT

This paper describes the construction and use of a machine-learning model to provide objective support for a physical fit examination of duct tapes. We present the ForensicFit package that can preprocess and database raw tape images. Using the processed tape image, we trained a convolutional neural network to compare tape edges and predict membership scores (*i.e.,* fit or non-fit category). A dataset of nearly 2000 tapes and 4000 images was evaluated, including various quality grades: low, medium, and high, as well as two separation methods, scissor-cut and hand-torn. The model predicts medium-quality and high-quality scissor-cut tape more accurately than hand-torn, whereas for low-quality tape predicts the hand-torn tapes more accurately. These results are consistent with previous studies performed on the same datasets by analyst examinations. A method of pixel importance was also implemented to show which pixels are used to make the decision. This method can confirm some fit features that correspond with analyst-identified features, like edge morphology and backing pattern. This pilot 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.

1. Introduction

Computational algorithms are increasingly common across many forensic science disciplines. Some uses of these models include instrumental library database searches for the identification of unknown substances, image processing, comparison for fingerprint and footwear evidence, and data mining to recognize trends across data from a variety of sources [1]. The value of computational algorithms, when used in forensic science, is multi-fold. They improve efficiency in analysis by identifying potential sample sources and assist in the standardization of the interpretation of forensic data across laboratories. In addition, models can be used to address complex problems in forensic interpretation (*e.g.*, speech recognition), and provide additional objective support to the opinions of an analyst. Importantly, computational results are repeatable and can be reproduced by third parties, which adds transparency and allows the prosecution and the defense to have access to the results and the process used to analyze those results, if needed.

The approach of using automated comparison models to assist in forensic examination and interpretation shortens analysis time and increases efficiency by identifying potential sources for an item. This is particularly important for pattern recognition disciplines, such as tool mark, friction ridge evidence, impression evidence, questioned documents, and trace evidence. However, when narrowing down the potential sources, the risk of false positives must be carefully assessed. It is worth noting that in many disciplines, the use of a computational model in interpretation does not mean that the answer provided by the model is used as-is without additional validation. In forensic science this computational and human interaction is used in fingerprints, DNA, face and voice recognition, and glass, to mn. For example, Automated Fingerprint Identification Systems (AFIS) [2] can automatically assess the minutiae features from millions of fingerprints stored in its database and perform a comparison search to identify potential sources of an unknown fingerprint. However, a skilled practitioner is still required to participate in the process of identification. The fingerprint practitioner often independently mark minutiae features on an unknown fingerprint before running it through the AFIS system. Then, they perform the

* Corresponding author.
E-mail address: aldo.romero@mail.wvu.edu (A.H. Romero).

https://doi.org/10.1016/j.forsciint.2023.111884
Received 19 January 2023; Received in revised form 30 October 2023; Accepted 7 November 2023
Available online 13 November 2023

comparison of the unknown against the individuals the AFIS system returns as potential sources [3].

This paper focuses on the application of a computational comparison model to trace evidence, through the comparison of edge features of fragmented materials. The commission of violent crimes often results in objects present at the scene becoming separated or fragmented. Examples include broken windows during robberies or assaults, fragmented taillights during a hit-and-run, or pieces of tape recovered from assault or kidnapping cases. The comparison of fragments of evidence items can provide valuable information for investigations and reconstruction of the crime scene. Comparing the edges of two items to determine if they were once joined together is referred to as a "physical fit examination". A physical fit indicates the items share not only class characteristics (*e.g.*, color, width, thickness, and layer structure) but also distinctive characteristics that demonstrate they were once the same object (*e.g.*, edge appearance, manufacturer markings, external damage). In trace evidence, a physical fit is considered the highest level of association between evidence items [4]. While other kinds of analytical techniques can provide information that indicates whether two items potentially share a common source, a physical fit is the only indication that two items originated from the same individual source. However, the scientific consistency of these types of examinations has been subject to scrutiny in recent years [5,6].

While physical fit comparisons are performed on a variety of materials, their scientific validity is rather unexplored. The studies that have estimated performance metrics have demonstrated relatively low error rates (*i.e.*, less than 5 %), and recent work has focused on the development of more systematic and quantitative methods for performing physical fit comparisons [4,7–10]. This study incorporates data produced by the method developed by Prusinowski et al. [7,8]. The material used in this study is duct tape, as this type of tape is a common material submitted to forensic trace evidence laboratories and has several characteristics that allow for physical fit examination. Duct tape is highly variable in physical features between different rolls but has low intra-roll variability within a single roll of the same physical features [11,12]. The reinforcement layer of cloth (known as scrim) provides additional support to duct tape, resisting extensive distortion and tearing, especially at the edges. As such, duct tape is more likely to retain edge morphology and features after separation than other tapes, such as electrical tapes. Moreover, its strength contributes to its use in binding and gagging victims, as packaging material in drug trafficking, and improvised explosive devices.

The method proposed by Prusinowski et al. [7,8] for practitioner examination of duct tapes incorporates feature identification and an edge similarity score (ESS) to describe the quality of a fit between a given edge of duct tape samples. These studies report low error rates, with no misidentification of non-fitting edges as fits. In addition, estimated thresholds for ESS and additional statistical assessment through score-based likelihood ratios demonstrate that pairs with ESS above 80 % provide strong support for a reported fit conclusion. These interpretation criteria are consistent regardless of the quality of the roll of tape used, as well as whether the tape was torn by hand or was cut. While the results of this human-based approach are encouraging, physical fit examinations are inherently subjective and are more likely to be challenged in court. This calls for more objective computer-based algorithms that can aid the practitioner in their decision-making. Moreover, comparison methods open an opportunity to create validated databases that can be shared across operational forensic laboratories and researchers to build additional knowledge foundations for estimating rates of misleading evidence. Different approaches use a combination of edge detection and analytic approaches to attempt to provide a computational solution [13,14]. These studies by McCabe et al. and Spaulding et al. serve as independent research examples of the feasibility of computational models for duct tape fit examination. However, some limitations in these studies are the need for validation across more different types of tape, improvement of the algorithms to provide better separation between true non-fit and true-fit classes, and the use of both backing and adhesive sides rather than just documenting edge features in one of the layers.

The primary aim of this study is to supplement the ESS method by introducing a computational comparison model for duct tape edges to provide additional objective support for a physical fit examination. To achieve this, we utilize a machine-learning model to process the acquired tape data [7,8]. We first developed an open-source Python package designed for image analysis tasks such as edge detection, background noise-reduction, and image filtering for materials of interest in the field of forensics science [15]. 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 score output. We apply the CNN approach to the scrim and the backing of the scan images separately and combine the results using logistic regression. The supplementary information includes a brief introduction to neural networks and convolutional neural networks. In the Results and Discussion section of this manuscript, we compare the machine learning model's performance to the outcomes of the examinations of the same samples by human analysts who followed a standardized protocol. The results are used to draw conclusions about how the computational and machine-learning models can assist examiners in duct tape physical fit comparisons.

2. Methods

2.1. Dataset preparation

2.1.1. Digitization

The dataset of images for this study had been created using the samples made by Prusinowski et al. [7]. The full set included tapes generated from three different tape qualities, designated as low quality (LQ), medium quality (MQ), and high quality (HQ). The brand name and physical characteristics of the samples are described in Prusinowski et al. [7,8]. The edges are either hand-torn (HT) or scissor-cut (SC). 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. The tapes were placed on transparent acetate film sheets, to allow the scrim to be seen through the adhesive. The HQ tape, however, had adhesive that was too obscuring so a small strip (~ 3 *mm*) of adhesive was removed from each comparison edge to allow the scrim fibers to be viewed.

We recognize the limitations of sample size and requirements for continuous updating a collection set. However, when developing this study's dataset, with the respective physical and digital collection of the fractured material's images, we consulted with forensic practitioners and statisticians to purposely include samples that are closely representative of casework. For instance, we included factors such as separation methods (scissor cut and hand-torn), quality grade of the tape (low, medium, and high quality), and level of deformation (stretching) to represent types of samples commonly received at the laboratory for examination [4]. In separate studies, we have also addressed the effect of these factors on the performance of physical fit examinations. [8] Moreover, as part of the feedback from practitioners, we also addressed more complex situations in which the comparison sample (questioned item) was stretched and only consisted of a small proportion of the tape width. Interlaboratory and mock case studies have been also utilized as part of the method evaluation [7,10].

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 images were collected using an EPSON 12000XL scanner using SilverFast 8, version 8.8.0r14, interface at a resolution of a minimum of 600 dots-per-inch (see Supplementary information for more details). A black posterboard background was used to accentuate the features of the tape and improve the contrast. Minor corrections to the image 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. Additional corrections were performed using Adobe Photoshop on some images to address specific issues. These corrections include 1) the preprocessing of scanned images featuring a very long warp fiber extending away from the comparison edge, and 2) the preprocessing of scanned images displaying protruding adhesive near the comparison edge, typically caused by the adhesive removal process. For the former, the preprocessing algorithm may fail to select the critical comparison edge, resulting in an image that contains only the extended warp fiber and omits the edge entirely. To address this issue, the warp fiber was manually removed from the image using Photoshop. With regard to the latter, the adhesive materials were identified and manually removed from the image. These corrections were primarily aimed at removing artifacts such as fingerprints, residual adhesive from the adhesive removal, long protruding fibers, and sample labels (example shown in Supplementary Fig. 6). 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. To generate the dataset that the CNN can learn from, one needs to generate a list of known fits and non-fits. Because there are many different combinations of tapes that are non-fits, one should answer the question, how many non-fits are needed to strike a balance between reality and high CNN performance?

2.2. *Data balance and appropriate statistical metrics*

It is necessary to obtain a holistic view of the fit-to-non-fit ratio problem because of the implications of metrics to be used and the inherent sampling bias that may occur. Intrinsically, the dataset is imbalanced because there are substantially more non-fit pairs than there are fit pairs. Imbalanced datasets have been studied in the machine-learning field [16–18]. One of the consequences is that the typical accuracy measure will become unreliable. Therefore, special care must be taken in interpreting the statistical metrics and more suitable metrics such as true-positive rate (TPR), true-negative rate (TNR), false-positive rate (FPR), and false-negative rate (FNR) must be used. For this study, a fit-to-non-fit ratio of 3:10 was selected. This value was selected after investigating the performance of the model for different ratios. For more details, see the Supplementary Table 1.

2.3. *Image Preprocessing*

The most important step in developing a successful machine-learning model is data preparation [19–22]. Data preparation consists of many steps such as data cleaning, transformation, feature extraction, and reformatting. The data preparation to a great extent depends on the kind of machine-learning model to be used. For example, this study uses a neural network, where each data entry, one pair of tapes, is represented by an array of numbers with a fixed length. The trivial approach is to concatenate the two matrices (each representing a tape image, see section ↱2.1) and flatten the resultant into one large array of numbers. However, for applications such as duct tape examination, where the small details are crucial, one needs to use images with high-resolutions leading to computer memory issues. Moreover, this can lead to the so-called “curse of dimensionality” [23–25] problem. This problem arises when the number of dimensions (neural network input nodes) in the problem is in the same order as the number of data points (total number of tape pairs in the dataset). A simplified analogy to this problem is if one tries to find the best fit line using only a small number of points, *e.g.*, two or three. Additionally, by flattening the image, the information about the neighboring pixels will be practically lost. To address this issue, often the image is passed through a convolutional network before the neural network (described in more detail in section ↱2.3↱). Before setting up the architecture of the network the image dimensions can be reduced by; 1) using the smallest image resolution where the tape surface details are still visible; 2) focusing only on the important part of the tape — the comparison edge. For this, we have developed a Python package, ForensicFit [15], that bridges 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 have been explained in the Supplementary Information. The source code is hosted on GitHub[15]. 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*. To maintain consistent inputs, the images were cropped on the borders of the tape and resized to 2400 *px*. Resizing can cause small alterations 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. Fig. 1 shows the output of the reduction. At this point, the two images of the tape pair were concatenated resulting in two images (scrim and backing) of size 410 × 1200 px^2 ready to be passed on to the CNN. An example of the input is shown in Fig. 2.

2.4. *Convolutional neural network configuration*

This model uses a convolutional neural network (CNN) followed by a fully connected neural network as implemented in TensorFlow [26] to train on the prepared images. A brief description of the CNN terminology used here is provided in Section 3 of the Supplementary information. 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 tape pairs are a fit or non-fit. In Fig. 2, it is of note that the order of the location (left or right) of the tapes in the concatenating process is arbitrary — a tape can be located on the left or the right side of the image, in other words, if the image is mirrored with respect to the *y* axis the CNN must lead to the same result. To force the CNN to recognize this inherent symmetry, the input images were randomly mirrored during the training.

2.4.1. *Architecture*

The network was built from a series of convolution layers, where filters with a small kernel of 3 × 3 px^2 window (smallest size capable of capturing the notion of left/right, up/down, and center [27]) and strides of 1 × 1 were used. The convolutional layers used Rectified Linear Unit (ReLU) [28] activation functions, and were followed by 2 × 2 pixel

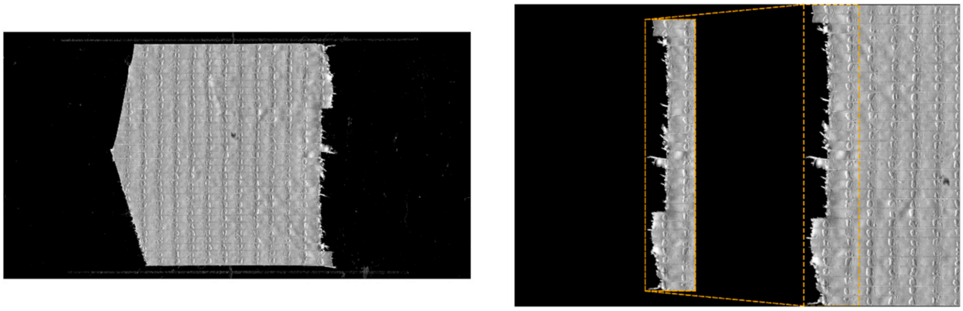


Fig. 1. Left: Scanned image of a low-quality grade tape. Image shows the backing side of the tape. One of the edges has been cut into an arrow shape, representing a non-comparison edge. For this publication this image was manually cropped. Right: Preprocessing of tape image by ForensicFit. The image 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.

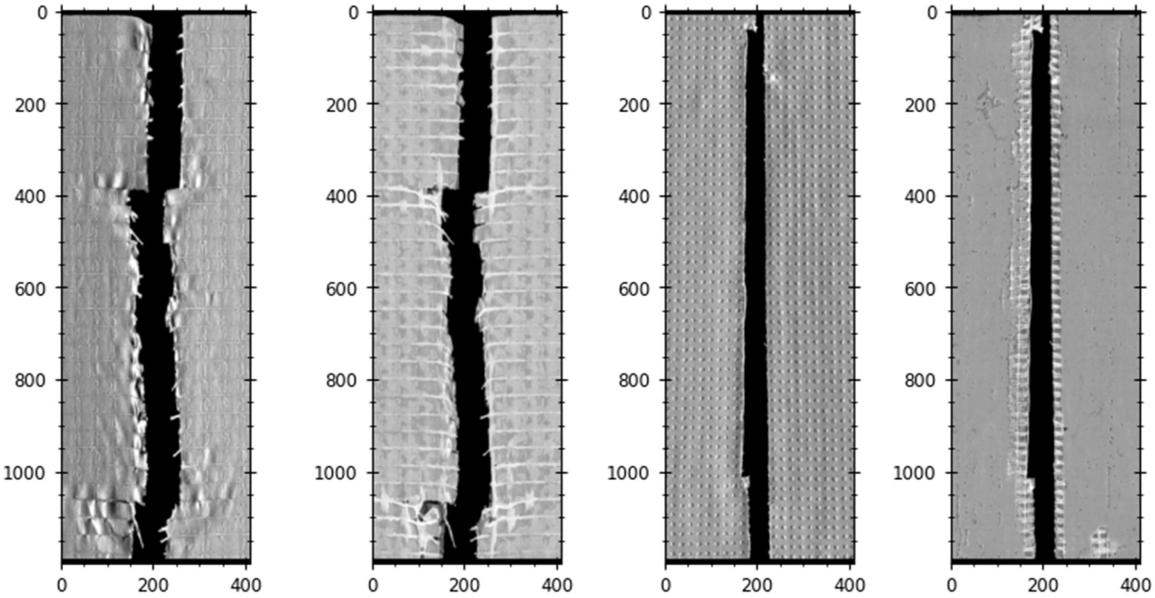


Fig. 2. Examples of the convolutional neural network image inputs. Left: Concatenated image of two tape edges on the backing side for a Known Fit. Center-Left: Concatenated image of two tape edges on the scrim side for a Known Fit. Center-Right: Concatenated image of two tape edges on the backing side for a Known Non-Fit. Right: Concatenated image of two tape edges on the scrim side for aKnown Non-Fit. The tapes on the left originated from the low-quality grade roll off tape. The tapes on the right originated from the high-quality grade roll off tape. Each image has a size of 1200 × 410 px^2.

Table 1
Convolutional neural network architecture. The network consists of a series of consecutive convolutional filters followed by a fully connected neural network.

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	-	-	-	-	136,192
	Dropout	-	-	-	-	136,192
	Dense	ReLU	-	-	500	500
	Dropout	-	-	-	-	500
	Dense	ReLU	-	-	100	100
	Dense	Sigmoid	-	-	1	1

window Max-pooling layers to handle the dimension reductions.

The CNN architecture was inspired by the popular VGG-16 [27], 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 ($1200 \times 405 = 492{,}000$), 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 [28], 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 [29]. A summary of the architecture of the CNN is provided in Table 1, as well as Supplementary Fig. 1.

2.4.2. *Training*

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, the 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.

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 [30]. 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. The weights of both CNN models are provided in the ForensicFit GitHub repository [15].

2.5. *Combining scrim and backing CNNs*

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 the tape pairs. Logistic Regression was ultimately selected due to its straightforward nature and easily understood results. The regression provides a continuous output between 0 and 1, which can be interpreted as a Fit Membership score of the two edges of the duct tape. Its ability to differentiate between the distribution of membership scores for accurate fits and non-fits, along with its performance also influenced our choice. By selecting the logistic regression method, we aim to achieve a balance between model complexity and accuracy in merging the results from the two CNN models.

The logistic regressor was trained using the same five-fold cross-validation method applied to the CNNs. The performance metrics presented in this study represent the average values across the cross-validation folds, with minimums and maximums reported the alongside these averages. The logistic regressor implementation utilized in this study is part of the machine learning Python package Scikit-learn [31]. Additionally, the trained logistic regressor can be downloaded from the ForensicFit GitHub repository [15].

3. Results and discussion

3.1. *Evaluation of computational models through comparison to human-based analysis of tape samples*

In this study, the physical tape samples were examined by trained analysts following a standardized protocol and the respective images of the same samples were evaluated through the computational approach. The outcomes of the human-based analysis of the tape samples is used as a comparison point to the results from the computations model to evaluate common trends in their performance by tape factor (*i.e.*, quality or separation mode) and to assess if they can complement each other. As mentioned before, the basis for the development of the computational model was a method developed by Prusinowski et al. [7] for the comparison of duct tape edges by analysts. The method identified major features for comparison of edges, and incorporated a quantitative metric defined as an edge similarity score. This metric was determined by the analyst by assigning a score of 0, 0.5, or 1 – corresponding to non-fit, inconclusive, or fit, respectively – for each bin area between the scrim fibers along the edge of the tape. The total score was then reported as a relative percentage to represent the similarity between the two tape edges, also referred to as the quality of the fit. This study utilized the same samples as used in Prusinowski et al. [7] and the human based-results were used to compare the performance of the computational model. Thus, it is worthwhile to briefly report the results obtained in that paper. A summary of the statistical metrics of the analyst-based approach can be found in Table 2.

Overall, the examination of the samples indicated that quality grade of the tape has a substantial influence on the appearance of the tape edge features, and the occurrence of features. The HQ-HT set produces the most false-negative pairs of all the sets. The HQ tape roll, when torn by hand, tends to produce more straight edges than are observed in the other tape rolls. In addition, the adhesive on this tape obscures the scrim fibers, requiring removal before examination. While performed as carefully as possible, the adhesive removal process contributes uncertainty as alignment of warp fibers could be affected, leading to higher false negative and inconclusive rates than other sets. Low quality tape samples, however, experience more distortion as the tape tears, reducing the quality of many of the edges. The medium quality tape has a good balance, where there are enough features for comparison while also commonly resulting in more distinct edge morphologies.

Table 2
Summary of the duct tape method sample breakdown and performance rates for the analyst-based examination method. LQ, MQ, and HQ, represent low-quality, medium-quality, and high-quality grade of the tape, respectively. While SC and HT represent, scissor-cut and hand-torn separation methods, respectively. No false positives are reported for any set. Note: MQ-HT* in this study represents the stretched MQ-HT set discussed in Prusinowski et al.[7,8]. Scanning of the images took place after the set had been stretched, so there are no original images of the set. TPR, TNR, FPR, and FNR denote true-positive rate, true-negative rate, false-positive rate, and false-negative rate, respectively. Meanwhile, IPR represents inconclusive rate with a positive (fit) ground truth and INR represents inconclusive rate with a negative (non-fit) ground truth.

Name	LQ-HT	LQ-SC	MQ-HT	MQ-SC	HQ-HT	HQ-SC
No. Comp	200	250	508	500	199	250
No. Fit/Nonfit	104/96	130/120	99/409	99/401	98/101	124/126
TPR	1.000	0.985	0.980	0.990	0.684	1.000
TNR	1.000	0.983	1.000	1.000	1.000	1.000
FPR	0.000	0.000	0.000	0.000	0.000	0.000
FNR	0.000	0.015	0.010	0.010	0.214	0.000
IPR	0.000	0.000	0.010	0.000	0.102	0.000
INR	0.000	0.017	0.000	0.000	0.000	0.000
ACC	1.000	0.984	0.996	0.998	0.844	1.000

3.1.1. Performance of the computational model

In this section, the performance of the model is discussed. Important evaluation metrics are shown in Table 3, including the percent of correct classifications (true positive and true negative rates), percent of misclassifications (false positives and false negative rates), and accuracy. The classification results are from choosing a decision threshold of 0.2 for the output of the logistic regression model. The threshold was chosen by analyzing the Receiver Operating Characteristic curve (ROC) in Supplementary Fig. 4. This curve shows the relationship between the true positive rate and the false positive rate at various decision thresholds. For a good classifier, one wants to maximize the true positives and minimize the false positives. The ROC curve suggests 0.2 is an optimal value for the model.

The results are subdivided into corresponding tape quality and separation methods, *i.e.,* HQ-HT, HQ-SC, MQ-HT, MQ-SC, LQ-HT, and LQ-SC. Similar to the analyst examination, the metrics model shows the quality grade as well as the separation method has a major influence on the outcome decision. The overall accuracy ranges from 80 % to 95 %, depending on the subset. The model performs best when comparing MQ tape pairs across all metrics, while LQ and HQ tape pairs perform slightly worse. More specifically, the evaluation reveals false-positive rates below 10 % for the LQ-HT and HQ-SC, and for the MQ regardless of the separation method (MQ-SC, MQ-HT), while the false positive rates are higher for LQ-SC and HQ-HT. On the other hand, the LQ-HT and HQ-HT are the sets that produce the worst false negative rates. This trend is somehow consistent with the findings observed during the analyst's examination of the corresponding quality-separation samples (Table 2). While the decreased performance on HQ samples is not surprising due to fewer observable features, the assumption would be that LQ pairs should perform better than MQ pairs as they have more distinctive features such as puzzle-like edges. However, as observed during the initial comparison by the analysts, LQ samples tend to distort during separation when torn by hand [8]. The increased distortion and number of extraneous artifacts in the tape samples contribute to lower performance in the computational model. This brings up a limitation of the computational models, as the extraneous artifacts can be moved by an analyst during an examination to find the best possible orientation for a comparison, however, each instance is supplied to the CNNs only in one orientation. In addition to the observations made about tape quality, the model also shares similar prediction trends about separation method. For MQ and HQ tape pairs, the model more accurately predicts SC compared to HT pairs, whereas for LQ tape the HT pairs are more accurately predicted. This agrees with the results seen in the analysis in Table 2. This is an interesting result for the MQ and HQ tape pairs because general assumptions about scissor cut tape suggest that they would not provide enough features for physical fit comparison. However, both, the analyst examination and computation model demonstrate that scissor-cut edges retain sufficient features for comparison, and in many cases are predicted better than hand-torn edges because cutting the edges reduces the amount of distortion [8].

To better understand the trends in the observed membership scores for tape comparisons, violin plots (Fig. 3) and kernel density estimation (Fig. 5) were employed. The violin plots shown in Fig. 3 explore the distributions of reported ESS (assigned by the analyst) and fit membership score (assigned by the ML model) for each set of tapes and are organized based on the ground truth (fit or non-fit). The data suggests that the separation between fits and non-fits is well-defined for both ESS and ML model scores, regardless of tape quality or separation method. In general, true non-fit pairs receive an ESS of 30 or lower, with the majority scoring below 10. A similar pattern is observed for the ML model scores, where most true non-fits are assigned low fit-membership scores (below 0.20). Conversely, true fits typically exhibit higher ESS scores (often greater than 80) and fit-membership scores (greater than 0.8). Notably, the scissor-cut sets demonstrate greater separation compared to hand-torn sets, as the latter exhibit a wider range of scores and ESS for true fits, evident in the medians displayed in the violin plots. In general, the analyst performs better than the ML model for the hand torn sets.

The values assigned to true-fit pairs in the HQ-HT set are quite dispersed in both human-based and computer-based approaches. This can be attributed to the fact that the HQ samples contain fewer observable features, particularly the HT set, which does not benefit from the presence of severed dimples like the SC samples. Furthermore, the range of the predictions with a fit ground truth, as illustrated by the distributions of the violin plots, is quite extensive for all sets except for MQ-SC, spanning across the entire score output range. MQ-SC does not follow this trend, as the medium quality set has nearly twice the amount of training data available. For the MQ-HT set, despite having a large volume of training data, the range of the whiskers is broader than ideal. This could be attributed to the fact that the MQ-HT set underwent stretching before the digitization process.

Table 3 and Fig. 4 show the performance of the model assessed using a range of metrics. These results demonstrate that the scissor-cut sets exhibit lower false-negative rates. It is important to note that the ML results depicted in the bar plots of Fig. 4 represent predictions made by the ML model for the entire dataset. These predictions are not derived from the training stage but rather from the aggregated testing sets across all five folds of the cross-validation. Additionally, it is worth mentioning that the statistical analysis conducted during the analyst examinations was performed on the entire tape dataset, whereas the ML approach analyzed only twenty percent of the dataset, reserving the remaining portion for training. The results presented in Table 3 display the mean values of the statistical metrics obtained from the five-fold cross-validation, with standard deviations shown as uncertainty values. This raises the biggest challenge of a machine learning approach to any problem, the size of the dataset. The high-performing CNNs are often trained on image datasets with orders of magnitudes larger and predicting sets of classes much larger than this study. For example, the datasets MNIST [33], MS COCO [34], and ImageNet [35] contain 60

Table 3
Summary of the duct tape method sample breakdown and performance rates for the CNN-LR model. The metrics represent the average values obtained across the cross-validation folds, with minimums and maximums reported in the parenthesis. A 0.2 decision threshold was used to evaluate the classification. The abbreviations, LQ, MQ, and HQ, refer to low quality, medium quality, and high-quality tape, respectively. SC and HT refer to scissor-cut, and hand-torn separation methods, respectively. TPR, TNR, FPR, FNR, and ACC denote true-positive rate, true-negative rate, false-positive rate, false-negative rate, and accuracy, respectively.

Name	LQ-HT	LQ-SC	MQ-HT	MQ-SC	HQ-HT	HQ-SC
TPR	0.729 (0.38–0.95)	0.925 (0.84–1.00)	0.801 (0.67–0.91)	0.943 (0.85–1.00)	0.796 (0.67–0.94)	0.912 (0.87–0.95)
TNR	0.895 (0.83–0.96)	0.750 (0.67–0.86)	0.900 (0.80–0.97)	0.949 (0.88–1.00)	0.791 (0.74–0.92)	0.887 (0.80–0.96)
FPR	0.105 (0.04–0.17)	0.250 (0.14–0.33)	0.100 (0.03–0.20)	0.05 (0.00–0.12)	0.208 (0.08–0.26)	0.113 (0.04–0.20)
FNR	0.271 (0.05–0.62)	0.075 (0.00–0.16)	0.199 (0.09–0.33)	0.057 (0.00–0.15)	0.204 (0.06–0.33)	0.088 (0.05–0.13)
ACC	0.845 (0.71–0.95)	0.816 (0.75–0.87)	0.873 (0.80–0.95)	0.946 (0.91–1.00)	0.792 (0.73–0.89)	0.898 (0.84–0.96)

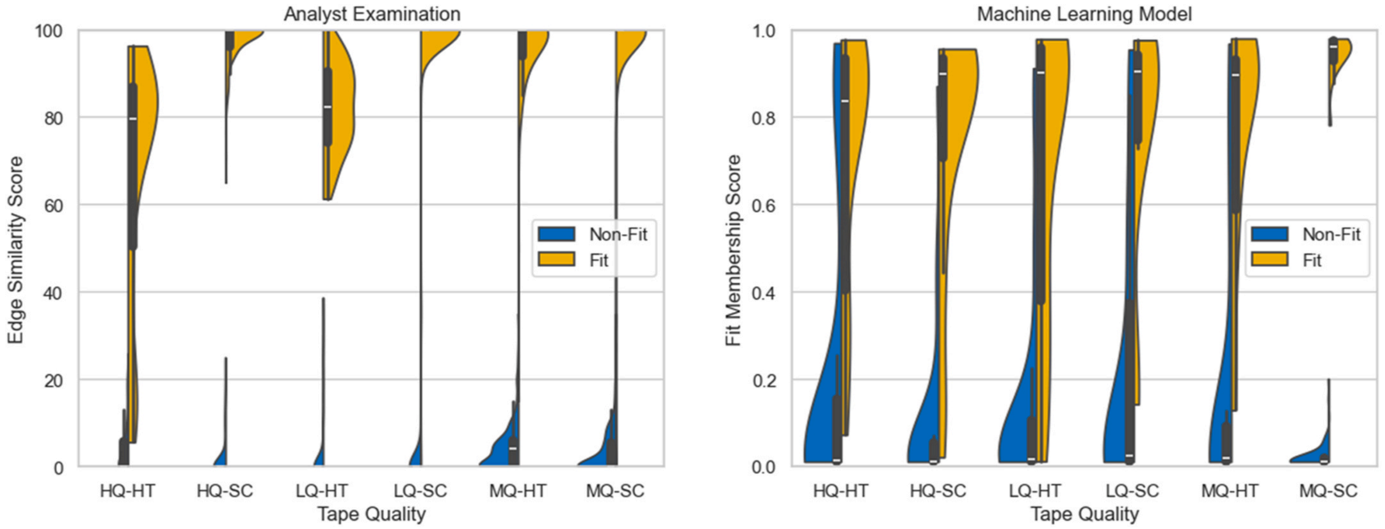


Fig. 3. (left) Violin plot of edge similarity score for all the sets from analyst examination. (right) violin plot of fit membership score from the machine learning model. The data was obtained by combining the output of two convolutional neural networks, one analyzing the scrim side of the tape and the other analyzing the backing side, using a logistic regressor.

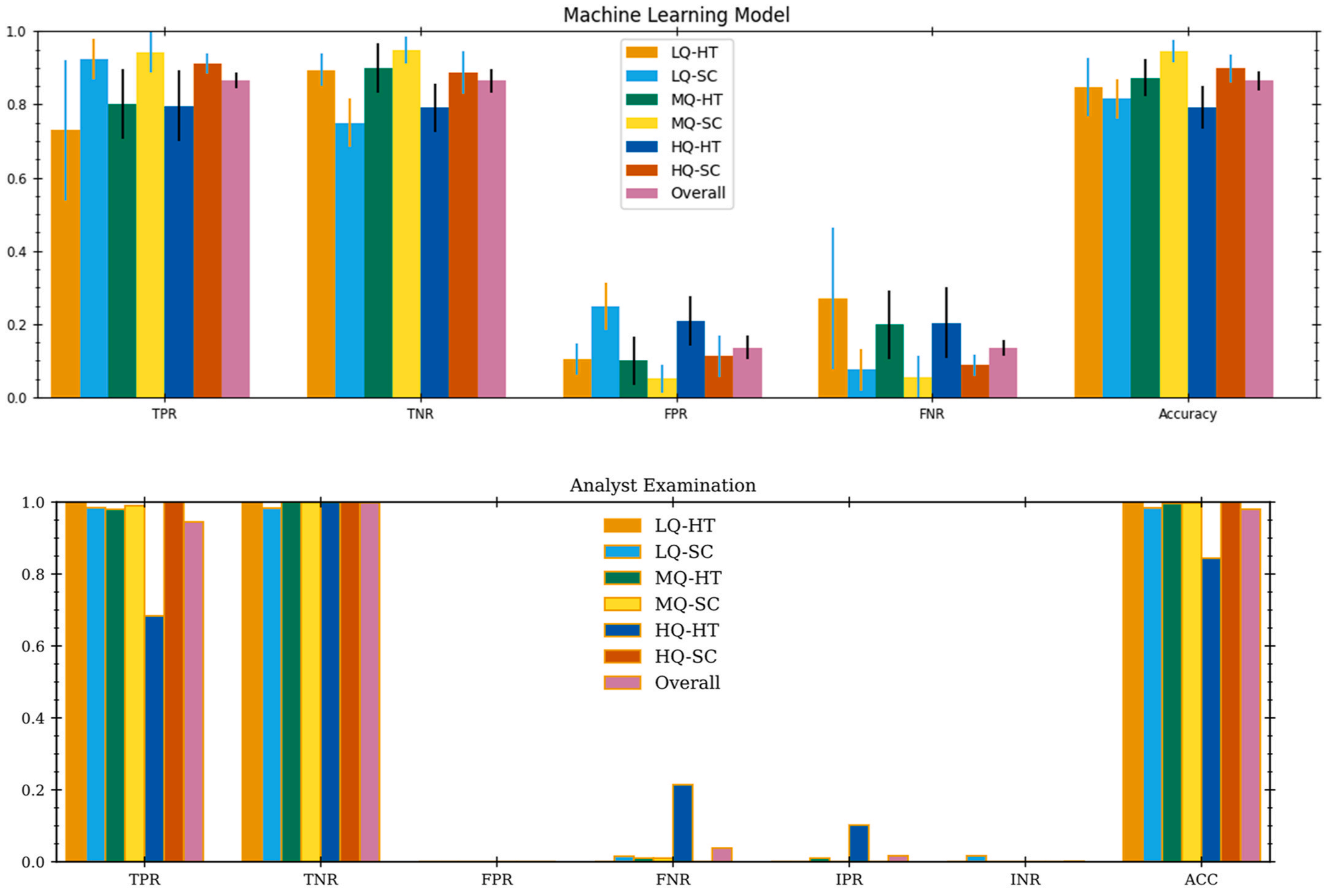


Fig. 4. Model evaluation metrics. The abbreviations LQ, MQ, and HQ denote low quality, medium quality, and high-quality tapes, respectively. Similarly, SC and HT refer to scissor-cut and hand-torn separation methods, respectively. TPR, TNR, FPR, and FNR denote true-positive rate, true-negative rate, false-positive rate, and false-negative rate, respectively. Meanwhile, in the case of analyst examinations, IPR represents inconclusive rate with a positive (fit) ground truth and INR represents inconclusive rate with a negative (non-fit) ground truth. (Top) Performance of the machine learning model. The bar heights represent the mean values obtained from the five-fold cross-validation, while the error bars indicate the corresponding standard deviations. (Bottom) Performance of analysts examination categorized by different duct-tape quality and separation.

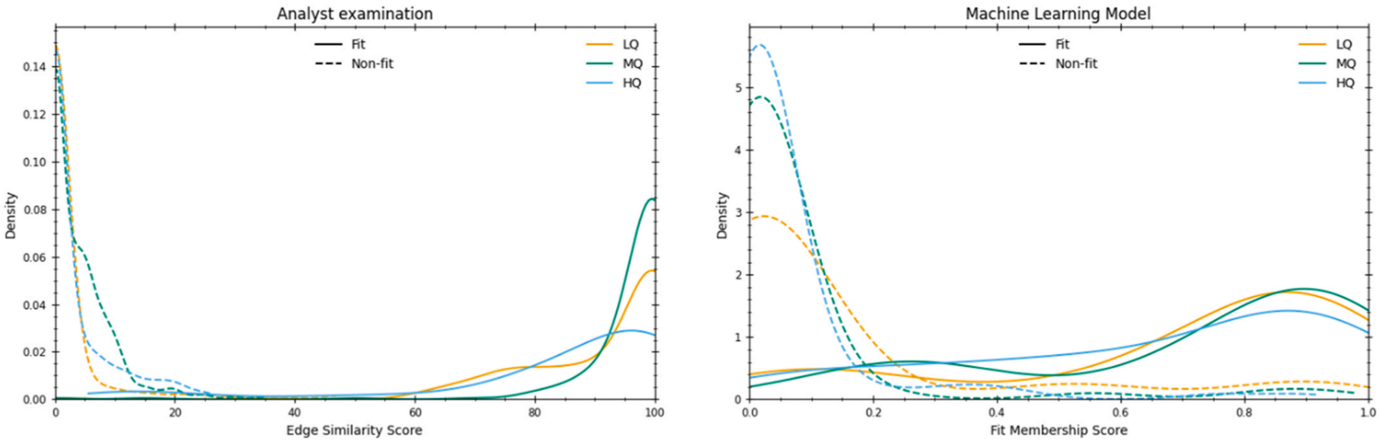
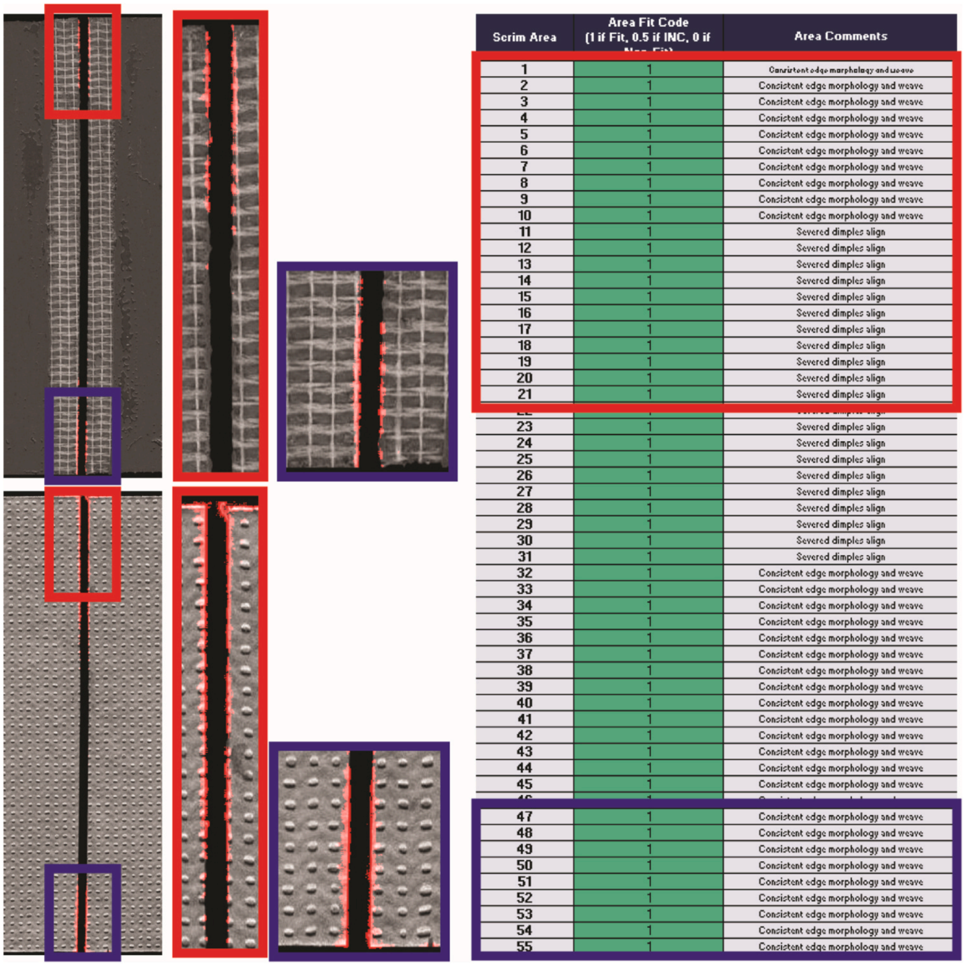


Fig. 5. (left) Kernel density estimation of the edge similarity score assigned by the analyst's examination (right) Kernel density estimation of the model scores assigned to the tape pairs. The data was obtained by combining the output of two convolutional neural networks, one analyzing the scrim side of the tape and the other analyzing the backing side. The kernel densities were constructed using the Scott method [32].

thousand, 330 thousand, and 14 million, images, respectively. However, the size of the data set of this study still permits to establish the feasibility of deep learning CNN to assist with automated quantification of the similarity/dissimilarity of duct tape edges and provides a basis for further expansion.

To further explore the comparison between analyst-based and computer-based approaches, the distribution of ESS and CNN-scores are estimated using kernel density estimation (Fig. 5). In both approaches,



Scrim Area	Area Fit Code (1 if Fit, 0.5 if INC, 0 if Non-Fit)	Area Comments
1	1	Consistent edge morphology and weave
2	1	Consistent edge morphology and weave
3	1	Consistent edge morphology and weave
4	1	Consistent edge morphology and weave
5	1	Consistent edge morphology and weave
6	1	Consistent edge morphology and weave
7	1	Consistent edge morphology and weave
8	1	Consistent edge morphology and weave
9	1	Consistent edge morphology and weave
10	1	Consistent edge morphology and weave
11	1	Severed dimples align
12	1	Severed dimples align
13	1	Severed dimples align
14	1	Severed dimples align
15	1	Severed dimples align
16	1	Severed dimples align
17	1	Severed dimples align
18	1	Severed dimples align
19	1	Severed dimples align
20	1	Severed dimples align
21	1	Severed dimples align
22	[illegible]	Severed dimples align
23	1	Severed dimples align
24	1	Severed dimples align
25	1	Severed dimples align
26	1	Severed dimples align
27	1	Severed dimples align
28	1	Severed dimples align
29	1	Severed dimples align
30	1	Severed dimples align
31	1	Severed dimples align
32	1	Consistent edge morphology and weave
33	1	Consistent edge morphology and weave
34	1	Consistent edge morphology and weave
35	1	Consistent edge morphology and weave
36	1	Consistent edge morphology and weave
37	1	Consistent edge morphology and weave
38	1	Consistent edge morphology and weave
39	1	Consistent edge morphology and weave
40	1	Consistent edge morphology and weave
41	1	Consistent edge morphology and weave
42	1	Consistent edge morphology and weave
43	1	Consistent edge morphology and weave
44	1	Consistent edge morphology and weave
45	1	Consistent edge morphology and weave
46	[illegible]	[illegible]
47	1	Consistent edge morphology and weave
48	1	Consistent edge morphology and weave
49	1	Consistent edge morphology and weave
50	1	Consistent edge morphology and weave
51	1	Consistent edge morphology and weave
52	1	Consistent edge morphology and weave
53	1	Consistent edge morphology and weave
54	1	Consistent edge morphology and weave
55	1	Consistent edge morphology and weave

Fig. 6. Human/Model comparison. Layer-wise Relevance Propagation (LRP) analysis is compared to human comments on Fits. Important LRP pixels are colored in red. The Examiner's remarks are on the right, they denote what are the main features to determine a fit. LRP identifies the most important features found at the top and bottom of the tape. On the top LRP and the examiner make note of the severed dimples and the edge morphology. (top left) LRP overlay for scrim side of the duct tape. (bottom left) LPR overlay for the backing side of the duct tape. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the datasets have been combined by quality grade to generate the kernel density plots. Fig. 5 shows that for true fit pairs, the MQ tapes more commonly receive high scores, followed by LQ and then HQ. The effect is not as pronounced in the computational model, but it is still consistent with the analyst examination results. A similar trend is observed for true non-fits, in which the majority tape pairs have low values (ESS 20 or less). In addition, the overlap of the densities between true fits and true non-fits is limited in both approaches, indicating good discrimination power between the populations of interest (known fits and non-fits). More importantly, both scores (ESS and CNN outputs) can be used as the basis to estimate score-based likelihood ratios to estimate the probative value of the evidence [7,8]. This is an advantage afforded by both approaches, as it provides the opportunity to use a quantitative interpretation of the evidence instead of merely depending on human judgment to demonstrate a fit or non-fit conclusion. While the use of score-based likelihood ratios merits caution as they are dependent on the sample set and only utilize a portion of the information available about the evidence, they provide an intuitive means by which to present the probative value of a physical fit. Moreover, as the databases permit expansion, the larger and more representative the population of casework-like specimens, the more confidence can be established in the examination of physical fits, especially as the samples collected on a crime scene and casework-like specimens are not as pristine as samples generated in the laboratory environment. These approaches provide a venue to estimate rates of misleading evidence that can assist the community in having better tools to support its scientific foundations. The findings presented here raise a flag that physical fit examinations, whether conducted by an analyst or by a computer-algorithm, are not error-free and it is therefore critical to gain knowledge of potential error sources and factors that can influence the accuracy of physical fit examinations.

Often, neural networks are viewed as “black boxes”, which give predictions without knowing how the network came to the decision. However, for convolutional neural networks, schemes have been developed to give a pixel importance score depending on the output prediction. The technique used in this study is called Layer-wise Relevance Propagation (LRP) [36], where the method will highlight important pixels for the decision of a prediction. In Fig. 6, LRP is demonstrated on a true fit pair from the HQ-SC subset. The model predicted this tape pair as a fit with a score of 0.87. In addition to the model prediction, Fig. 6 also shows the ESS calculation and documentation provided by the analyst for the same tape pair.

Important LRP pixels are highlighted in cyan in Fig. 6. Overall, the most important pixels come from the edges, which correspond with where many critical features observed by the human eye are found. The model not only places importance on the edge morphology but texture information as well. The top portion of the pair has severed dimples close to the edge highlighted by the model. This makes sense as in the human-based analysis these dimples are noted as critical for use in the decision-making process on those same top comparison bins. In the bottom section, although severed dimples are absent, other features, such as micro alignment of edges and spacing between dimple markings were noted by the analyst as the basis for the fit decision. For the central bins not highlighted by the algorithm, the alignment of the warp scrim (not seen in this image side) were reported by the analyst as a dominant fit-favored feature.

4. Conclusion

This study provides a computational platform for the tape physical fit problem that can assist analysts in their evaluations. We report the development of an open-source Python package, ForensicFit [15], designed to pre-process images obtained for forensic physical fit examination. The package has been used to provide data for machine learning to train two independent convolutional neural networks — one on the backing side, and the other on the scrim side. The results were tested for model calibration and then combined using logistic regression. The performance rates to classify images as fit or non-fits is presented using the combined score. The proposed computational model performed well with low false-positive rates and high true-negative rates.

Moreover, this work compares the quantitative assessment of duct tapes using human-based and computer-based approaches, with encouraging results that indicate a high agreement between both methods and therefore demonstrate the potential of machine learning models to provide statistical support to the analyst conclusions. This study confirms the previous findings that the scissor-cut tapes indeed contain sufficient features for comparison examinations, while high-quality hand-torn tapes increase false negative occurrences [7,8].

To summarize, the main findings derived from this study are: 1) CNN have shown to be an effective mean to compare separated tape edges using an automated imaging processing platform, 2) The distribution of metrics associated with known fits and non-fits (ESS for human-based and CNN-membership scores for computer-base) shows a minimal overlap between these groups, indicating relatively low rates of misleading evidence and the feasibility to employ them for statistic assessment of the probative value of the evidence, 3) The violin plots 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.

Even though this study used a relatively small dataset (*ca.* 4000 images), it shows reasonably accurate results and the potential to help analysts deliver more objective examinations. This proof of principle study shows that the approach has great potential for improvement in addition to the need to generate larger datasets. This goal is only achievable by exposing the learners to more samples. The presented approach is intended to be a starting point in the seldom-explored area of machine learning in physical fit examinations. The model presented here opens opportunities to build databases that can be further developed for a user-friendly platform that requires minimal human intervention and be expanded to other materials of forensic interest. Finally, computational methods could be utilized as a supportive tool for practitioners, and the results generated by these methods should not be taken at face value. Instead, they should be considered in conjunction with the practitioner’s judgment. The notion of familiarity differs significantly between humans and ML models. The human brain possesses the ability to adapt its knowledge to novel situations, while current ML models, in their current stage, have limited capabilities in this regard. In future research, it would be interesting to explore various data augmentation techniques and assess the performance of ML models in unfamiliar scenarios. Another method that could be employed in future research is an image pyramid [37,38] approach, where the comparisons are initially conducted at a lower resolution. If the algorithm predicts a high fit membership score, the resolution is increased, and the comparison is performed once more. This iterative process has the potential to yield more accurate results.

CRediT authorship contribution statement

Logan Lang: Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Pedram Tavadze:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Meghan Prusinowski:** Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Zachary Andrews:** Data curation. **Cedric Neumann:** Methodology, Formal analysis, Writing – review & editing. **Tatiana Trejos:** Conceptualization, Methodology, Supervision, Project administration, Funding acquisition. **Aldo H. Romero:**

Conceptualization, Methodology, Formal analysis, Supervision, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is sponsored by NIJ Award #2020-DQ-BX-0012 to West Virginia University. The opinions, findings, and conclusions are those of the authors and do not necessarily reflect those of the Department of Justice.

This study used the Thorny Flat Cluster at West Virginia University (WVU), which is funded in part by the NSF Major Research Instrumentation Program (MRI) Award (MRI-2117575). The authors would like the work of WVU students Colton Diges, Hannah Good, and Divyanjali Pulivendhan for their help with sampling and imaging. Figures in this publication were generated using the Matplotlib Python packages. Python packages, Numpy [44], SciPy [45], and Pandas [46] were used for post-processing of the results.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.forsciint.2023.111884.

References

[1] Franke K., Srihari S.N. , Computational Forensics: an Overview. In: Computational Forensics (Internet), Springer, Berlin Heidelberg, 1–10. 〈http://link.springer.com/10.1007/978–3-540–85303-9_1〉.
[2] Moses K.R., Higgins P., McCabe M., Probhakar S., Swann S. , Automated fingerprint identification system (AFIS), in: Fingerprint Sourcebook (Internet). 2011. 〈https://www.ojp.gov/ncjrs/virtual-library/abstracts/fingerprint-sourcebook-chapter-6-automated-fingerprint〉.
[3] VanderKolk J.R. , Examination process, in: Fingerprint Sourcebook (Internet), 2011. 〈https://www.ojp.gov/ncjrs/virtual-library/abstracts/fingerprint-sourcebook-chapter-9-examination-process〉.
[4] E. Brooks, M. Prusinowski, S. Gross, T. Trejos, Forensic physical fits in the trace evidence discipline: a review, Forensic Sci. Int. 313 (2020), 110349.
[5] PCAST, Report to the president - forensic science in criminal courts: ensuring scientific validity of feature comparison methods (Internet), Executive Office of the President President's Council of Advisors on Science and Technology. Executive Office of the President President's Council of Advisors on Science and Technology, 2016, 1–160. 〈www.whitehouse.gov/ostp/pcast〉.
[6] Strengthening forensic science in the United States: a path forward (Internet), Strengthening Forensic Science in the United States: a Path Forward, National Academies Press, Washington, D.C., 2009, 1–328. 〈http://www.nap.edu/catalog/12589〉.
[7] 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 Sci. Int. 307 (2020), 110103.
[8] 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 Sci. Int. 343 (2023), 111567.
[9] 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.
[10] 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 Chem. 34 (2023), 100487.
[11] A.H. Mehltretter, M.J. Bradley, Forensic analysis and discrimination of duct tapes, Jastee 3 (1) (2006) 2–20.
[12] K. LaPorte, R. Weimer, Evaluation of duct tape physical characteristics: part I – within-roll variability, Jastee 7 (1) (2017) 15–34.
[13] K.R. McCabe, F.A. Tulleners, J.V. Braun, G. Currie, E.N. Gorecho, A quantitative analysis of torn and cut duct tape physical end matching, J. Forensic Sci. 58 (2013) S34–S42.
[14] J.S. Spaulding, G.M. Picconatto, Characterization of fracture match associations with automated image processing, Forensic Sci. Int. 342 (2023), 111519.
[15] P. Tavadze, L. Lang, romerogroup/ForensicFit: first release of ForensicFit Package, Zenodo (2022), https://doi.org/10.5281/zenodo.7435058.
[16] H. Kaur, H.S. Pannu, A.K. Malhi, A systematic review on imbalanced data challenges in machine learning, ACM Comput. Surv. 52 (4) (2020) 1–36.
[17] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, Experimental perspectives on learning from imbalanced data. in: Proceedings of the Twenty Fourth International Conference on Machine learning - ICML '07, ACM Press, New York, New York, USA, 2007, pp. 935–942 (Available from:).
[18] S. Tyagi, S. Mittal, Sampling approaches for imbalanced data classification problem in machine learning, Lect. Notes Electr. Eng. (2020) 209–221.
[19] S. García, S. Ramírez-Gallego, J. Luengo, J.M. Benítez, F. Herrera, Big data preprocessing: methods and prospects, Big Data Anal. 1 (1) (2016) 9.
[20] J. Huang, Y.-F. Li, M. Xie, An empirical analysis of data preprocessing for machine learning-based software cost estimation, Inf. Softw. Technol.] 67 (2015) 108–127, https://doi.org/10.1016/j.infsof.2015.07.004.
[21] S. Zhang, C. Zhang, Q. Yang, Data preparation for data mining, Appl. Artif. Intell 17 (5–6) (2003) 375–381.
[22] A. Famili, W.-M. Shen, R. Weber, E. Simoudis, Data Preprocessing and Intelligent Data Analysis, Intell. Data Anal. 1 (1) (1997) 3–23.
[23] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning. Springer Series in Statistics, Springer New York, New York, NY, 2009.
[24] R.E. Bellman, Dynamic programming, Math. Sci. Eng. 40 (1967) 101–137.
[25] E.S. Page, R. Bellman, Adaptive control processes: a guided tour, J. R. Stat. Soc. Ser. A (Gen.) 125 (1962) 161.
[26] Martín A., Ashish A., Paul B., Eugene B., Zhifeng C., Craig C., et al. {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems ([Internet), 2015. https://www.tensorflow.org/〉.
[27] Simonyan K., Zisserman A. , Very deep convolutional networks for large-scale image recognition, in: Proceedings of the Thirrd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.
[28] Agarap A.F. , Deep Learning using Rectified Linear Units (ReLU). 2018.〈http://arxiv.org/abs/1803.08375〉.
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn Res. 15 (2014) 1929–1958.
[30] Kingma D.P., Ba J.L. , Adam: a method for stochastic optimization. in: Proceedings of the Thirrd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015, 1–15.
[31] F. Pedregosa, R. Weiss, M. Brucher, G. Varoquaux, A. Gramfort, V. Michel, et al., Scikit-learn: machine learning in python, J. Mach. Learn Res. 12 (2011) 2825–2830.
[32] D.W. Scott, Multivariate density estimation and visualization, Handb. Comput. Stat. Concepts Methods.: Second Ed. N. Y. (2012) 549–569.
[33] L. Deng, The MNIST database of handwritten digit images for machine learning research, IEEE Signal. Process Mag. 29 (6) (2012) 141–142.
[34] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.) (2014) 740–755, 8693 LNCS(PART 5).
[35] Deng J., Dong W., Socher R., Li L.-J., Kai L., L. Fei-Fei , ImageNet: a Large-scale Hierarchical Image Database. In 2010, 248–55.
[36] A. Binder, G. Montavon, S. Lapuschkin, K.R. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.) 9887 LNCS (2016) 63–71.
[37] J.L. Crowley, A.C. Parker, A representation for shape based on peaks and ridges in the difference of low-pass transform, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6 (2) (1984) 156–170.
[38] P. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.