Forensic Science International: Digital Investigation 34 (2020) 301016

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

The role of evaluations in reaching decisions using automated systems supporting forensic analysis

Timothy Bollé*, Eoghan Casey, Maëlig Jacquet

School of Criminal Justice, Faculty of Law, Criminal Justice and Public Administration, University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 11 January 2020
Received in revised form
1 June 2020
Accepted 14 June 2020
Available online xxx

Keywords:
Automated systems
Evaluation
Decision-making
Machine learning
Explainability
Forensic science
Forensic-by-design
Understandability

ABSTRACT

Automated systems allow forensic practitioners to perform analysis tasks that would otherwise be infeasible. However, unless the outputs of such systems are critically evaluated using a scientifically-based framework, there is a risk of undetected errors or bias resulting in wrong decisions. Furthermore, decisions based on automated system outputs that are not well understood or clearly explainable could violate fundamental human rights. These risks can apply to any automated system that supports forensic analysis, and are raised when machine learning is involved. This work presents a framework based on principles of scientific interpretation, and provides an evaluation hierarchy for automated systems, including machine learning approaches, to strengthen forensic conclusions. Specifically, three levels of evaluation are presented: performance, understandability and forensic evaluation. Approaches to clearly conveying the weight of forensic evaluations are discussed. Each level of evaluation is demonstrated in relation to actual automated systems. Finally, requirements for designing automated systems supporting forensic analysis are proposed, and future work is discussed.

1. Introduction

The increasing volume, variety, velocity, distribution and complexity of information are overwhelming forensic scientists, crime analysts and security professionals in various contexts, including criminal investigations, national security, and international disputes. As a result of information overload, these practitioners are reaching decisions without sufficient support mechanisms, increasing the risk of incorrect conclusions. At the same time, there is a growing demand for forensic analysis of traces with probative value and scientific validity. The opportunity costs of not being able to rely on results of automated systems for forensic purposes are growing as more investigations involve large quantities of digital evidence from a multitude of sources. Missing or misinterpreting inculpatory or exculpatory digital evidence can result in innocent individuals being falsely accused, victims being denied justice, and criminals remaining at liberty to commit offenses. To deal with these challenges and opportunities, there is a pressing need for automated systems that integrate scientific principles and processes to address forensic questions.

In the context of this work, an automated system is defined as *any system that performs a process instead of a person to address forensic questions (authentication, classification, identification, reconstruction, evaluation),* as defined in Pollitt et al. (2018). Automated systems, including but not limited to machine learning approaches, are being developed to help humans find valuable insights more effectively and efficiently in massive amounts of information. The following are just a few examples of automated systems that support practitioners for forensic and security purposes:

- Extraction of information from various data sources on a computer or mobile device (Metz and White, 2020).
- Semantic file recovery, reassembly, repair and validation of carved content to increase the amount of renderable content and reduce the number of false positives (Casey and Zoun, 2014).
- Child Sexual Exploitation Material (CSEM) classification on the basis of age and sexual activity to increase accuracy and efficiency, while reducing forensic examiner exposure to stress (Anda et al., 2020; Sanchez et al., 2019)
- Face recognition to support verification (1-vs-1) and identification (1-vs-N).

* Corresponding author.
E-mail address: timothy.bolle@unil.ch (T. Bollé).

https://doi.org/10.1016/j.fsidi.2020.301016

- Detect grooming activity in online discussion forums (Meyer, 2015)
- Detect links between related crimes (Bollé and Casey, 2018)

Automated systems can present the outputs of complex analyses in ways that appear quite simple. For instance, the open source forensic framework called Plaso transforms digital evidence into event-like structures to enable temporal analysis, adding a description of the associated (probable) activities (Metz and White, 2020). An exemplar web history event-like record produced by Plaso presents a URL with the description:

`Visit from: https://accounts.google.nl/[edited for length] (resurrection stone price - Google zoeken) Type: [GENERATED - User typed in the URL bar and selected an entry from the list - such as a search bar] (URL not typed directly - no typed count).`

Such descriptions of activities in Plaso are, in fact, inferred from the data using encoded expert knowledge, which has some level of uncertainty. There is a risk that forensic practitioners will not understand that such output from an automated tool is an inference that requires evaluation. In this instance, the ability to review source code provides some transparency and understandability, provided the forensic practitioner knows Python. However, the output of an automated system such as Plaso must also be evaluated in light of the forensic question of interest in the investigation, which typically takes into account circumstances of the case and may involve experiments to test alternative hypotheses (e.g., caused by malicious code, not the user).

Automated file recovery (carving) tools that codify expert knowledge and employ best-match algorithms can salvage content that is not attainable using other methods (Durmus et al., 2019). However, it can be difficult for a forensic practitioner to verify or explain a specific file recovery output, and there is a risk of advanced carving and reassembly methods incorrectly combining two different files. As a result of the difficulty detecting erroneous output and explaining successful recovery, practitioners are slow to adopt such powerful capabilities for forensic purposes.

As another example, a system for detecting the age of people in digital photographs can allow forensic practitioners to look only for children under ten years old. However, hidden complexity and bias can cause problems, including incorrect classification (false positives), missing relevant evidence (false negatives) and misinterpretations (Anda et al., 2019). The risk of making mistakes and missing relevant evidence motivates forensic practitioners to place more trust in their own observation of photographs and videos than automated classification methods (Sanchez et al., 2019). Reviewing source code is typically ineffective for finding more subtle problems in complex automated systems, making it necessary to perform evaluation of their outputs (Taylor et al., 2017).

Another risk is that forensic practitioners will not formally evaluate the output of automated systems in light of the specific question(s) under consideration. For instance, consider an automated system intended for object/face classification, but that was not designed to compare similar objects or people. In this situation, the output of the system would be useful to address the question of whether a photograph/video contains an object or a person but it would not be suitable to answer the question of whether it contains a specific object/person. In other words, a forensic practitioner could misinterpret the level of similarity in light of the question "Are these the same object" rather than "Are these the same class of object". Many automated tools are not crystal clear about when they are fit for purpose, and do little to help forensic practitioners assess outputs critically. As a result, forensic practitioners can incorrectly assume that the outputs of automated systems are reliably addressing their needs, unless they are guided through a structured evaluation process.

This work considers the core challenges of forensic practitioners understanding and interpreting the output of automated systems, and more specifically machine learning systems, for forensic purposes. A central premise of this work is that forensically sound decisions can still be made using outputs from automated systems, provided proper evaluation is performed.

In order to support decisions in a forensic setting, the design of software should abide by forensic principles and practices (Rahman et al., 2016). From both a scientific and legal perspective, any automated system that helps a forensic practitioner reach conclusions for forensic purposes must be transparent and reproducible (Margagliotti and Bollé, 2019). Furthermore, forensic practitioners must be able to understand and explain the results of such automated systems in a clear, complete, correct and consistent way (Berger, 2019; Casey, 2020). Many existing machine learning approaches lack sufficient transparency and reproducibility for forensic purposes, and are not designed in a way that helps forensic practitioners evaluate and explain the outputs of automated systems effectively.

The solution to this problem is not only technical, but involves a structured process for exploring and evaluating that abides by the principles of scientific reasoning and interpretation. In addition to output from automated systems being transparent, reproducible and understandable, it is necessary to shepherd users through a structured process of scientific interpretation. This process is generally referred to as evaluation in forensic science, but the term *forensic evaluation* is used throughout this work to differentiate it from other forms of evaluation applied to automated systems. This work proposes design requirements for automated systems supporting forensic analysis, which includes supporting consideration of alternatives and weight of evidence assignment. This work considers systems applied to any form of evidence, including face comparison on identity cards/passports, fingerprints and shoemarks in criminal investigations, and digital evidence in any type of investigation. However, the examples discussed will be taken from the field of expertise of the authors.

This work makes an inventory of existing terms and definitions surrounding the evaluation of automated systems and the resulting decision making process. In order to support forensic analysis, we propose the following general recommendations when designing an automated system:

- Performance evaluation results should be sufficiently detailed to determine if it is fit-for-purpose for a given forensic question.
- Understandability and transparency should be a requirement in the design of automated systems supporting forensic analysis.
- The system should support contextual analysis by keeping the context of the information at every stage
- The system should guide users through the forensic evaluation and decision making steps to be sure that they can understand and explain the result in a clear, complete, correct and consistent manner.
- The hypotheses should be explicitly formulated, even if it is automated and always the same. If some steps are automated, they still should be explicitly described.

The paper will start with the definition of automated systems (Section 2) and the different levels of evaluation we consider to be involved when using automated systems (Section 3). In Section 4, we present the role of these evaluations in the decision making process. In Section 5, we will rapidly discuss how it is possible to

express weight, in regards to multiple hypotheses, during the forensic evaluation. We will then present some use case examples to discuss the strengths and weaknesses of the different approaches, in the light of the proposed model (Section 6). In Section 7, we give some recommendations to design systems that support evaluation. We conclude with the challenges that future work need to address (Section 8).

2. Automated systems

The approach presented in this paper can be applied to any automated system as defined above, including file recovery tools, expert systems and machine learning systems. File recovery tools apply codified knowledge of file characteristics to automatically classify and authenticate files in deleted state. Modern file carving tools encode expert knowledge about data formats, characteristics, structures, components and their arrangements in order to find, reassemble and repair content. Expert systems use a knowledge base that contains some facts and an inference engine that applies logical rules to solve problems. Digital forensic tools are increasingly exploiting codified expert knowledge to make automated inferences from digital evidence to help investigators address forensic questions (Henseler and Hyde, 2019).

Machine learning systems are becoming more widely used for forensic purposes to support classification, regression and clustering, but also for data manipulation (preprocessing, dimensionality reduction, model selection, etc.). The learning can be supervised, meaning that the rules are derived from a labeled dataset considered as ground truth, or unsupervised, meaning that the rules are learned from unlabeled data.

A generalized depiction of an automated system is presented in the top-left part of Fig. 1, delimited in the "System development" square. The system by itself is represented in the "System" box. A system will usually be composed of an algorithm or a set of algorithm that will handle data, produce a result and output it.[1] In Fig. 1, we separated the "Selected features" and "output" box from the "System" to insist on those particular elements in other sections of the paper. As stated previously, the system can apply codified knowledge and/or statistically learned knowledge on a particular case data. Those two approaches are represented in Fig. 1 with the boxes "Knowledge & Rules" and "Train or Test system". The statistically learned knowledge is produced by training the system on labeled data, for instance with systems based on supervised machine learning algorithms. In both cases, the system can be tested on known data in order to measure the performance of the system. Finally, the system can be used in a particular case on unknown data. The rest of the diagram will be detailed in the following sections, concentrating on evaluation.

3. Different levels of evaluation

A core challenge when using automated systems for forensic purposes is to evaluate the output, particularly when machine learning is involved. Evaluation is generally defined as *a process producing a value that can be fed into a decision process* (Pollitt et al., 2018). The value produced by an evaluation can be binary (e.g., 0 or 1 relating to a decision of "false" or "true"), continuous (e.g., between 0 and 1 representing a probability), or a mark of a discrete nature (e.g., "low", "medium", high" representing strength). The evaluation of automated systems can occur on multiple levels: performance evaluation, understandability evaluation, and forensic evaluation.

3.1. Performance evaluation

Before forensic practitioners use a particular automated system to help them reach forensic conclusions, they need to have a basic level of trust. It is thus necessary to measure the performance of the system for a given purpose. For instance, performance measurement is different for automated systems that detect faces in photographs versus systems that find a specific face of interest. Different performance measurements are required for traditional file carving tools versus specially designed methods targeting fragments of partially overwritten files. Depending on the type of system, different approaches to measuring performance are needed.

The aim of the following paragraphs is to present some commonly used performance metrics for some systems and to show that the usage and comprehension of these metrics are not trivial.

3.1.1. Classification

Classification systems, for instance, can be used to determine in which class a sample belongs. The classes are chosen and the system is trained to differentiate samples from each class. The training is done on samples for which the class is known. When testing the system, known samples are classified and the predicted class can be compared with the true class of the samples. A confusion matrix displays for each class, how many items were correctly classified. An example of confusion matrix is provided in Table 1.

Common metrics are the accuracy, precision, recall and F1-score. Accuracy is defined as the percentage of correctly classified samples (Jeni et al., 2013). Precision is defined as the "ability of the classifier not to label as positive a sample that is negative" (Scikit-Learn Developers, 2019). The recall is the "ability of the classifier to find all positive samples" (Scikit-Learn Developers, 2019). The *FN*-score "can be interpreted as a weighted harmonic mean of the precision and the recall" and when using the F1-score, precision and recall have the same weight (Scikit-Learn Developers, 2019). The F1-score is more adapted to measure performance when the classes are imbalanced. Concretely the F-score combines the precision and the recall which ease the comparison between systems. However, systems that have similar precision and recall will have a better F-score. In some situations, we would prefer to have a system with a good precision (or recall) score and not focus on the recall (or precision). The commonly used F1-score will not be adapted to measure the performance in these situations.

In two-class classification, the definition of these metrics are as follows:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn},$$

$$precision = \frac{tp}{tp + fp},$$

$$recall = \frac{tp}{tp + fn},$$

$$F_1 = 2\,\frac{precision \times recall}{precision + recall}.$$

With *tp, tn, fp, fn* being the number of true positive, true negative,

[1] It includes all the algorithms needed to transform or cluster data, predict classes or values, or visualize elements. For instance, a system could be composed of the algorithm needed to clean the data, extract the features, classify the samples (using neural networks or linear regression) and finally display the results.

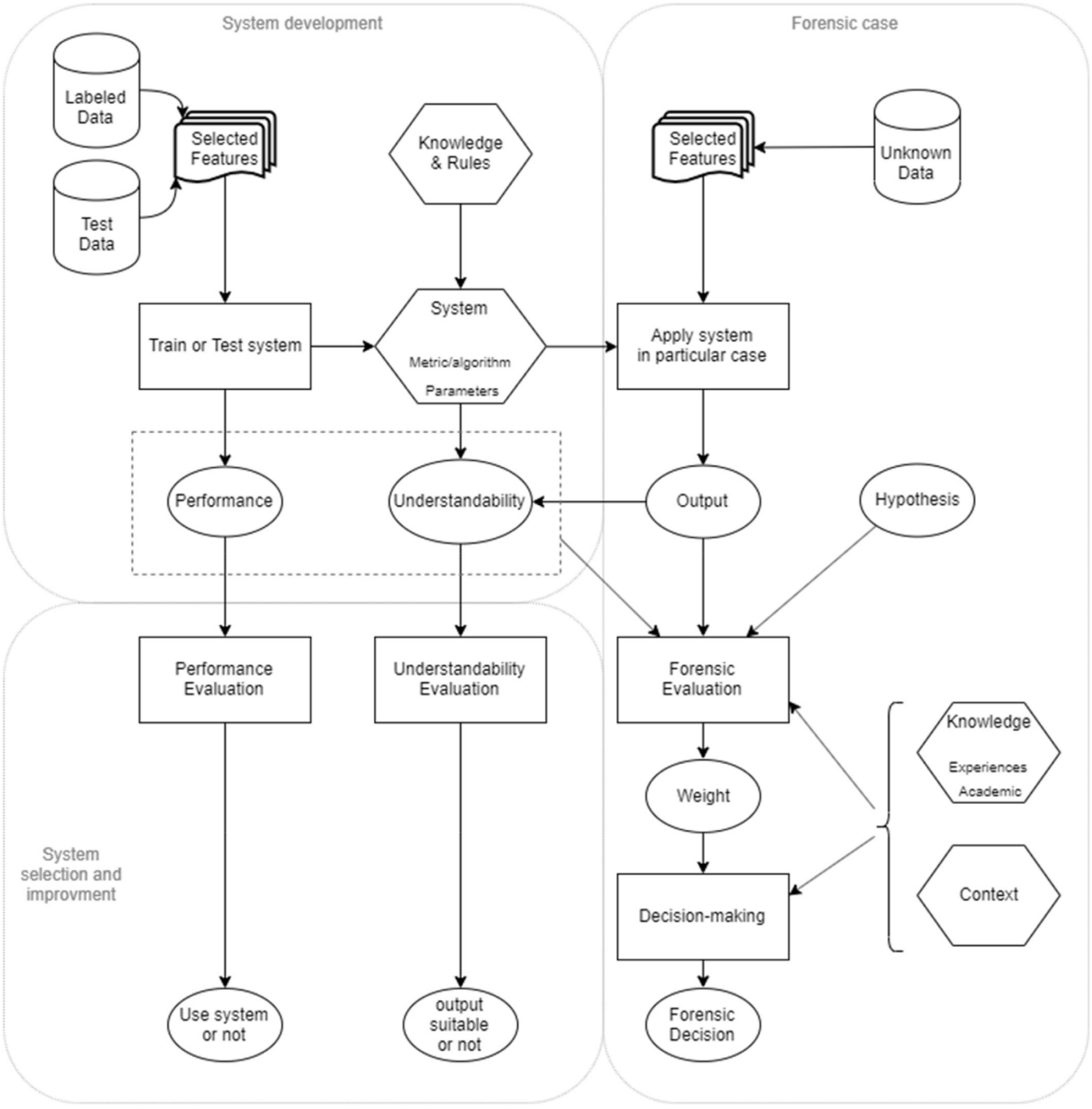

Fig. 1. General workflow of an automated system implementation, its use for forensic purposes, and the different levels of evaluation.

Table 1
Confusion matrix for a two-class classification. "True class" represents reality and "Predicted class" represents the result of the classification (class predicted by the classifier). If we consider Class A as the positive result, and Class B as the negative, *tp, tn, fp, fn* are the number of true positive, true negative, false positive and false negative, respectively.

		Predicted class	
		Class A	Class B
True class	**Class A**	*tp*	*fn*
	Class B	*fp*	*tn*

Table 2
Representation of web history as event-like (Metz and White, 2020).

type	datetime	user	url	description	extra
Last Visited Time	11/30/2018 8:42	Default	https://www.google.nl/search?q=elder+wand&oq=elder+wand&aqs=chrome..69i57j0l3.6043j0j7&client=ms-android-samsung&sourceid=chrome-mobile&ie=UTF-8	(elder wand - Google Search) [count: 0] Type: [GENERATED - User typed in the URL bar and selected an entry from the list - such as a search bar] (URL not typed directly - no typed count)	page_transition_type: 5 schema_match: False url_hidden: False

false positive and false negative, respectively.

Practical example

For example, results of file recovery operations are evaluated using precision, recall and F-score (NIST Software Quality Group, 2017). A problem with evaluating file carving tools only using these metrics is that it rewards production of junk files, i.e. data that are not renderable in a useable form for forensic purposes. In this context, renderable refers to the ability to display recovered content in a human viewable form. For example, some file carving tools produce thousands of picture or video fragments that cannot be viewed, which can create time waste and confusion when each (false positive) result must be reviewed and explained. File carving tools that eliminate junk files and/or repair damaged files produce more useful results for forensic purposes and reduce the amount of time forensic practitioners spend reviewing junk files. Another problem when evaluating with these metrics is that they are not well suited to specialized file carving methods designed to recover fragments of partially overwritten files (Durmus et al., 2019).

From these definitions, one can already see that it might be difficult to understand exactly the meaning of these metrics.

3.1.2. Regression

Regression systems are used to predict a value (continuous data) instead of a class, leading to an infinity of possible outcomes. Thus, it is not possible to use a confusion matrix. However, multiple metrics can be used to measure how close the output of the algorithm is from the ground truth. Common metrics are to measure the error between the real value and the predicted value. Many variations exist, such as Max Error, Mean and Median Absolute Error, Root Mean Square Error (RMSE). The choice of one of the metrics compared to another will be made based on what type of errors we want to evaluate. For instance, the Max Error will detect the worst case error but the Median Absolute Error is not affected by outliers (Scikit-Learn Developers, 2019). Another common metric is the R^2 score that measures the proportion of variance explained by the regression model. Scikit-Learn Developers (2019) gave a good overview of the metrics that can be used to evaluate the different types of models.

[2] It can be defined as the relation between the predicted probability of an event and the real hypothetical occurrence of the event (Ramos and Gonzalez-Rodriguez, 2013; Meuwly et al., 2017).

Further considerations

The list of metrics provided here is not exhaustive and the definitions may change depending on the context. For instance, when using Likelihood ratios (LR) and bayesian statistics, performance can be measured with accuracy and calibration[2] among others (Ramos and Gonzalez-Rodriguez, 2013). Meuwly et al., (2017) give the definition of the term *calibration* and *accuracy* and underline the fact that the definition in LR-based methods are different from the one used in metrology for analytical methods. The point of this quick example is to emphasise that the definitions of terms can vary depending on the context.

3.1.3. Challenges

The objective of this quick overview of metrics is to show that it requires a basic knowledge of statistics to understand these performance measurements and that it might be difficult for a forensic practitioner or a judge to assess it. It also shows that the terminology can change depending on the context. This adds difficulties to rely only on such metrics to justify the use of a particular system.

Another limit of such methods, which allows to evaluate the performance of an algorithm, is that they can only be used when "ground truth" data are available to test the system. Even if such data are available, they must be representative of the diversity of the data on which the model will run in real conditions. This condition might be hard to satisfy in a criminal context where events might be rare and diverse, and when you cannot predict or have an influence on the data you will have.

For instance, if a face recognition system performance is measured using good quality images of adult males (i.e identity documents), but the system is then used on low quality images such as surveillance footage, or on young girls identity documents, the performance could be completely different. It would be problematic to apply the general performance measures in these particular cases, and it would undermine the robustness of the system when doing the forensic evaluation.

Another example of the difficulty to assess performance is when the user of the system needs to make sense of the value of the performance and make a decision based on it. Imagine testing a classification tool that classifies images as containing a minor or not. The testing gives a f1-score of 0.84. Is this score sufficient for such a tool for the given purpose? Should it be retrained? Will it meet the operational needs? Answering such questions might be hard for someone that is not familiar with performance scores.

A performance evaluation will help forensic practitioners decide whether the system is actually adapted to the forensic question. In other words, whether or not a system is suitable for a given purpose and situation. However, when performing forensic analysis, it is also necessary to evaluate the specific output of the system and to understand why the system produced the specific result.

3.2. Understandability evaluation

To evaluate the specific output of an automated system, forensic practitioners need a mechanism to study features and factors used to produce this result. Such insights support understandability, explainability and transparency.

The distinction between those terms is narrow. For further discussion on the definitions of these terms, the reader can refer to the text box "Definitions & Discussion".

In the rest of this paper, we will use the term *understandability*.

Understandability[3]: The ability of a human to understand the functioning of a system, and in particular its purpose, along with the output results, the features used and the inferences made by the system.

We chose this term because it is focused on the user of the system and it is in our opinion the important aspect of automated systems: *they have to be understood by their users*.

Definitions & Discussion

The distinction between the different terms is not easy to understand and even the literature is ambiguous.

Typically, the terms *explainability* and *interpretability* are often used in literature related to explainable artificial intelligence. Doshi-Velez and Kim (2017) and Gilpin et al. (2018) propose a distinction between the two terms. They define *interpretability* as the "ability to explain or present in understandable terms to a human" (Doshi-Velez and Kim, 2017). A system is considered *explainable* if it is "able to summarize the reasons for a behavior, gain the trust of users or produce insight about the causes of their decisions" (Gilpin et al., 2018). They also note that the *explainability* is a trade-off between *interpretability*, which suppose the understanding by a human, and *completeness*, which could be seen as the exact mathematical representation of the system.

Our understanding of these two aspects are the *interpretability* being how easily the system can be understood by a human and the *explainability* being all the mechanisms that have been put in place in the system in order to be more interpretable.

Other research considers the two terms to be synonyms as in Beaudouin et al. (2020), where both terms are defined as "the ability, inclination or suitability to make plain or comprehensible, or explain the meaning of, an algorithm". The authors indicate that the term *interpretability* is preferred in the data science community and that the term *explainability* is preferred in policy documents.

For clarity, we present the following definition of the different terms:

Understandability: The ability of a human to understand the functioning of a system, and in particular its purpose, along with the output results, the features used and the inferences made by the system.

Explainability[4]: The ability of a system to make its functioning and its purpose clear to a human. This includes the output results, the features used and the inferences made. As one can see, the definition is similar as the one of *understandability* except that the focus is made on the system.

Concretely, we reach the conclusion that both terms, *explainability* and *understandability* serve the same purpose and that *understandability* requires *explainability*. As explained previously in the paper, we will mainly use the term *understandable* in this paper because we want to insist on the fact that it is important that the user or the decider understand the system.

Interpretability[5]: We consider this term as a synonym of *understandability*. To avoid any confusion with the *forensic interpretation*, we will completely avoid this term.

Completeness: The exact representation of the system. It can be its exact mathematical or algorithmic definition, or the exact set of rules used by the system.

Transparency: The ability for the user to have access to the detail of functioning of system. A fully open-source system is completely transparent, as opposed to black-box systems.

The discussion around the definitions of the different terms is very open and we acknowledge its importance, in order for the scientific community to speak a common language. In this paper, our approach was to clearly define the terms as we intended them and to present some challenges regarding their use.

The transparency of the system can help improve its understandability but does not necessarily imply it. For instance, for neural networks or very complex algorithms, having access to the details of the implementation and to the precise algorithm will not help understanding it. For complex and black box systems, it is necessary to add some other mechanism to improve the understandability of the system.

Many concerns are raised by various institutions about the usage of black box systems. For instance, Campolo et al. (2017), stated in the 2017 annual report of the AI Now Institute that "*Core public agencies, such as those responsible for criminal justice, healthcare, welfare, and education (e.g "high stakes" domains) should no longer use black box AI and algorithmic systems*". The DARPA also started a program that encourages the development of explainable artificial intelligence - XAI (Gunning, 2017). At the European Union level also, a COST Action project titled "CA17124 - Digital Evidence: evidence analysis via intelligent systems and practices (DigForAsp)" is conducting research on the synergies between digital forensics and automated systems along with the challenges of reliability, verifiability and understandability of the results in a legal context. Fig. 2 presents an example of a system that would add mechanisms improving the understanding of the user.

Practical example

As an example, a document analysis visualization tool can be useful for efficiently analyzing large quantities of communications extracted from a smartphone, including SMS, chat, instant messages, and e-mail messages. Such an automated system is useful for finding themes, trends, and relationships in textual data, and performing link analysis between people or numbers on a single device or across multiple devices. Although the specific algorithms used

[3] The word understand can be defined as: "to know or realize how or why something happens, how it works or why it is important" ("Understand" n.d.).

[4] The word explain can be defined as: "To tell somebody about something in a way that makes it easy to understand" ("Explain" n.d.).

[5] The word interpret can be defined as: "to explain the meaning of something" ("Interpret" n.d.).

within the system are not known, the system shows how much specific terms contributed. For instance, grouped together within a theme related to purchasing certain items narrows focus to 166 messages. Checking the terms that contribute to this buying-selling theme-based cluster reveals the primary terms relate to selling a generator, rifle and bullet proof vest, with some ancillary terms related to buying and selling drugs as shown in Fig. 3. The system allows the user to drill down, i.e., select specific terms to see the surrounding context, and delve deeper into the data source to examine the complete contents of any message in order to understand the context and evaluate the digital evidence directly in order to address specific forensic questions.

Multiple techniques can be used to improve the explainability of the models. For instance, LIME[6] allows understanding how the features impact the result by running multiple times the system, each time "hiding" a feature, to understand how it influences the results (Guestrin et al., 2016). DeepLift[7] is also a method that can be used to know which are the important features in a deep learning system (Shrikumar et al., 2019). The difference between those two approaches is that the first one "changes" the input, by changing the features, and observe the impact on the output, and the second one starts from the output and propagates the contribution of each neuron in the neural network back to the input.

A final example of an approach that helps to understand the system is the use of Nearest Neighbors (Lau and Biedermann, 2020). When using a trained system on an unknown sample, the authors propose to also look at the nearest sample in the training set. It is then possible to concretely determine if the system has the same "sense of proximity" that the user of the system. They indicate that this method is compatible with existing systems and that it is particularly suitable for judges and lawyers.

Understandability is thus an important element to integrate in modern automated systems. Understandability evaluation will help the forensic practitioner to determine if the system is understandable enough and if the system works properly.

Understandability evaluation can be done generally and aims at determining if the system is understandable or if it provides sufficient explanations. In some cases, the forensic practitioner could decide that the system will not be used because it is a black box and that there is no explanation at all. In some cases, the system could be quite simple and perfectly understandable, meaning that the user will know exactly what the system is doing and why. A last example of such evaluation is when the system is very complicated or a black box but sufficient explanations were added to understand the output.

Understandability evaluation should also be done when using the system on a specific case. If it has some explainability mechanisms, the user should be able to understand what the system did and on which feature it based the output. At this stage, the user can determine (e.g., with some level of confidence) whether the system is using correct features, independently of the output. For instance, if we look again at the example in Fig. 2, if the user sees that the system is using the background of the image to produce the output, he or she might decide that the system is not doing what it is supposed to do, and the output is invalid/incorrect. Conversely, if the user sees that the ears of the animal are used, he or she will proceed to the next step of the forensic evaluation of the output in itself.

3.3. Forensic evaluation

For forensic purposes, practitioners ultimately need to assign a (relative) likelihood of observed information in light of a working hypothesis, relative to competing hypotheses. A hypothesis is a claim that is evaluated against competing claims (Pollitt et al., 2018). At the court level, the term "proposition" is usually preferred, whereas "hypothesis" is used during an investigation. In this context, the evidence is the output of the automated system.

It is important to note that changing the hypotheses can alter the results of the evaluation, even when considering the same automated system output (e.g. similarity score). For instance, a system could find similar malware in multiple intrusion cases. The result of the forensic evaluation might be different if we are trying to determine whether the same group of hackers is behind the attacks or not, or whether all attacks are using the same or different modus operandi.

It is based on this forensic evaluation that a decision is taken. In their article, Lau & Biedermann differentiate between evaluation and decision:

> "*Evaluation is the assessment of the strength of evidence regarding competing propositions and their relative plausibility or probability. After evaluation comes decision, that is, the acceptance of a proposition as a conclusion.*" (Lau and Biedermann, 2020, p. 7)

Although forensic evaluation is most often discussed in relation to court testimony, it is applicable at any time during an investigation (Casey, 2020).

As presented in Fig. 1, the performance and the understandability will also be taken into account in the forensic evaluation. The former conveys the fact that the system can make errors in its recommendations. The latter mitigates the possibility of errors by making the output understandable to the user.

The final step, which is to support a decision related to a forensic question(s)[8] under consideration, is based on this forensic evaluation as well as on other contextual data, knowledge and experience. The automated system generally only provides a part of the answer as it cannot take everything into consideration. This aspect will be particularly important in an intelligence framework where information from multiple sources and knowledge about the general context are used to make a decision. For instance, a system may classify a burglary into a specific phenomenon, based on time and entry point, but the forensic practitioner can also consider other contextual elements, such as the lighting of the entry point, to classify the burglary differently. The classification of burglaries is used for statistics and prevention, but also to detect series of crimes.

Contextual information and general knowledge could be included formally in either the automated system or the evaluation steps, but could also be taken into consideration during the decision-making process. This ability depends on the possibility to formalize the knowledge and information and to include it in either the system or the evaluative process.

4. Role of evaluations in supporting decisions

As we saw, from the output of automated systems to the

[6] Local Interpretable Model-Agnostic Explanations.
[7] Deep Learning Important FeaTures.
[8] The propositions/hypothesis will be the possible answers to the forensic question.

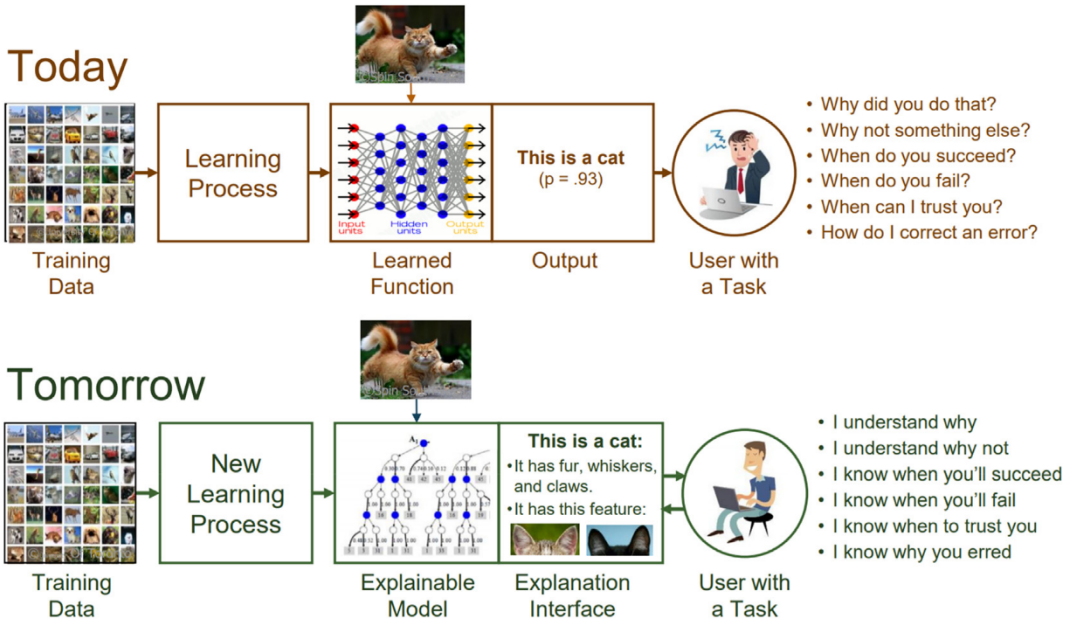

Fig. 2. Example of a system that would integrate explainable models in order to be understandable by the user (Gunning, 2017).

Group	Count
ay, wat	
generator, rifle	
hard, clientell	
talked, call	
bud	
-- REMNANTS --	
bad, girl	
half, gotta half peezie	
bru, ha ha ha coo	
powder	
getta quarter	
wanna	
geni, sell	

Term	Contribution	Docs
ay		151
buy		35
don		31
generator		24
rifle		19
wanna		19
wanna buy		17
bullet proof vest		14
hard		14
nate		9
clientell		10

Fig. 3. Theme based grouping of messages (left) with level of contribution of terms (right).

decision, there are various elements to consider and multiple levels of evaluation. Some of these factors are taken into account naturally by forensic practitioners, but automated systems currently used for forensic analysis provide limited assistance for assessing the uncertainty underlying a specific output.

We propose guidelines and open the discussion on how automated systems could help forensic practitioners understand outputs from automated systems and assess the underlying uncertainty in order to reach reliable, explainable decisions.

The output given by an automated system is a *recommendation*[9] in the sense that it will correlate elements to assign it to a class (classification), extrapolate a value (regression) or create a group of elements (clustering). Forensic practitioners must interpret the output, giving it meaning in the context of the forensic observations and overall investigation.

There are three decisions that can be supported by forensic practitioner's interpretations of outputs from an automated system:

1) Performance evaluation supports a forensic practitioner's decision whether or not to use an automated system to perform a given task.
2) Understandability evaluation supports a forensic practitioner's decision whether a specific output from an automated system is incorrect/invalid and should be disregarded, or is suitable for forensic evaluation.
3) Forensic evaluation supports a forensic practitioner's opinion regarding the weight of evidence in light of given propositions, and is conveyed to others involved in a case to enable them to assign appropriate level of confidence to the evidence when reaching a decision (e.g., investigator, attorney, judge)

5. Expressing the weight of forensic evaluation

When forensic practitioners communicate their expert opinions to factfinders, they must do so in a clear, complete, correct and consistent manner (Berger, 2019). Satisfying all of these requirements while still being understandable to factfinders is an ongoing challenge in forensic science, and there is no ideal or preferred approach (Thompson, 2017). Thompson notes that "*the reporting formats that are easiest for lay people to understand are difficult to justify logically and empirically, while reporting formats that are easier to justify logically and empirically are more difficult for lay people to understand.*" Although simply stating the opinion that a proposition is true (or false) might seem clear and understandable, it does not provide the required transparency and is logically incorrect because, potentially transposing the conditional aspect of the proposition and failing to consider opposing hypotheses. In addition, stating an opinion about a proposition does not deal directly with forensic observations and typically requires additional information outside of the expertise of the forensic practitioner, which is ultimately the responsibility of the factfinder. Furthermore, stating opinions about a proposition raises the risks of

[9] Or a decision directly but we will discuss this point later.

confirmation bias.

There is widespread consensus that the weight of evidence approach is most suitable for forensic evaluation (Aitken and Taroni 2004; Marquis et al., 2016). However, weight of evidence can be expressed in different ways. First, as a likelihood ratio (LR), i.e., the ratio of numerical likelihoods assigned to the evidence by forensic practitioners in light of each proposition on the basis of their personal experience, knowledge from published studies, surveys, experiments and peer review. Alternatively, a verbal scale[10] can be used to express the weight of evidence, which can be aligned with ranges of likelihood ratios (Champod et al., 2016). Another approach is to use a strength of evidence scale to assign a value to forensic observations in light of each proposition, thus conveying their relative strengths (Casey, 2020).

6. Strengths and weaknesses of approaches

This section presents case examples demonstrating the usefulness and the limitations of existing methods for evaluating the results of automated systems supporting forensic analysis. These examples illustrate how the three levels of evaluation can be applied to automated systems, and demonstrate the need for forensic evaluation.

6.1. Automatic extraction and labeling of digital traces

Web history can be represented as event-like, as shown in Table 2 using Plaso, with the addition of labeling such as “User clicked a link” and “User reloaded the page” and “User typed in the URL bar and selected an entry from the list - such as a search bar”.

Treating such digital evidence as an event blurs the distinction between a trace and an associated (probable) action. Most traces with a timestamp can be transformed into event-like entries for analysis purposes, but it is important to differentiate between the digital evidence and the associated (probable) action. Because of their vivid and immediacy, these types of digital evidence can easily be conflated with the associated actions. Jumping to the conclusion that web history is an indication of a particular user action can lead to problems as demonstrated in the case of Connecticut v Julie Amero (Pollitt, 2008). To reduce the risk of misinterpretation, it is crucial to clearly differentiate between a trace and an associated (probable) action.

Performance of Plaso varies depending on the data source and what information is being sought. The developers of Plaso concentrate on security incidents on networked computers, which leads to the best performance in this type of investigation. Some support exists for mobile devices, but support for 3rd party communication applications is limited. In addition, the performance of Plaso could be measured based on the completeness and correctness of its description of inferred (probable) actions (Metz and White, 2020).

Plaso generally has better performance parsing Windows systems than Linux/Mac. However, even on a Windows system, it is not safe to assume that all Web history is parsed. Some browsers are not supported in Plaso. Regarding understandability, some tools report unparsed (new) applications and present the unknown data for the user to decide.

As noted earlier in the paper, a forensic practitioner with knowledge of Python programming can look at the source code to understand the logic of a given result. However, some additional research might be required to determine whether the interpretation is correct, or there could be other explanations for a given digital trace.

From a forensic evaluation perspective, the output of an automated system such as Plaso must also be evaluated in light of the question of interest in the investigation, which typically takes into account circumstances of the case and may involve experiments to test alternative hypotheses.

6.2. Salvaging renderable content from data sources

The second example deals with automated systems for salvaging renderable content such as DC3 Advanced Carver (www.dc3.mil/tools), which provides for recovery, repair, and rendering of content fragments that are not obtainable using traditional file carving methods. In addition to producing more renderable content, salvaging renderable content produces fewer false positives that traditional file carving. Fewer false positives saves forensic practitioners time reviewing junk files, which is an important performance consideration.

Performance evaluations of systems for automatically recovered content often use measures of accuracy, precision and recall as discussed earlier. However, in practice, the performance of a carving tool can vary significantly depending on the data being analysed (Casey and Zoun, 2014). Determining whether a specific automated system is well-suitable to a given dataset is not a simple matter of looking up published performance measures. Although the NIST Computer Forensic Tool Testing (CFTT) program provides results of performance evaluations of various automated systems for different types of content (NIST Software Quality Group, 2017), these controlled tests might not be applicable for the actual data being analysed. For instance, when forensic practitioners are interested in obtaining deleted photographs (e.g., JPEG) they can review the NIST CFTT reports to determine which automated system(s) had the best performance for that type of file. However, when the actual data being analysed contains partially overwritten photographs, many of the available systems will not salvage the remaining content fragments. In such cases, a decision must be made whether using a more advanced carving tool to obtain the fragments of partially overwritten files would be beneficial, or whether forensic questions can only be addressed using fully recovered files.

Understanding why certain content is not salvaged by a given automated system can be challenging, and there is a need for research and development to support this type of evaluation. It is generally easier to understand why incorrect outputs are produced by these automated systems, because forensic practitioners can clearly see when content is partially overwritten or unrelated fragments have been incorrectly combined (e.g., one person's head on another person's body). However, automated systems for salvaging renderable content do not necessarily explain the decision of classifying a group of bytes as related content. Even if the forensic practitioner understands the data structures being salvaged, the mechanisms to reassemble specific fragments may not be clear or reproducible. Detailed audit logs can provide some insights, but there is a need for more advanced approaches that help practitioners understand specific outputs. In particular, a practitioner may be interested in knowing the cases where a potential candidate was classified as not being associated with similar fragments, and therefore excluded from the results. Understanding the automated exclusion of such a fragment can help practitioners verify the correctness of the process and explain the output. In addition, such insights provide additional information to support forensic evaluation, in particular authentication. Other elements in the case could reveal false negatives in the automated system, raising the need for further development to enhance performance.

[10] For instance: the evidence provide [weak/strong/extremely strong] support to the first proposition rather than the alternative (Willis et al., 2015).

Forensic evaluation of outputs from automated systems for salvaging renderable content is usually limited to evaluating the hypotheses "The original contents of a file are (or are not) actually recovered, fully or partially." (Casey et al., 2019). However, there are circumstances in which forensic evaluation of the outputs can include associated filenames and metadata found relating to the content.

6.3. Face recognition

The field of face recognition currently lacks a standardized and validated model for evaluating the outputs of automatic systems. In Jacquet and Champod (2019), the authors propose a general workflow for the implementation of a probabilistic model to evaluate scores generated by automatic systems. In that purpose, the assignation of a likelihood ratio is a recommended and used approach for forensic applications using automatic systems, such as fingerprints and speaker recognition (see Meuwly, 2000 and Egli, 2009 for examples), and has been more recently introduced for face recognition (Ali, 2014; Jacquet and Champod, 2020).

To use automated face recognition systems, the overall quality of trace and reference images must correspond to feasibility criteria of the system (most algorithms cannot operate properly in cases of very low resolution, high shooting angle, ageing, etc.). The operator then determines the hypotheses to be answered, taking into account case information along with the needs of the investigation. The calculus of the 'trace-vs-reference' score as well as its form (similarity score or distance score, range of values, etc.) vary from one system to another. For example, the open-source OpenFace toolkit (based on Google FaceNet algorithm) generate distance scores specific to each comparison with a value between 0 and 1, whereas the Idemia MorphoFace Investigate system gives similarity scores with values from 0 to 50′000, that vary if images are deleted from or added to the searched dataset.

Assigning a score-based likelihood ratio (SLR) allows the operator to gather an analysis output and uncertainties into a single, specific, and standardized piece of information. As mentioned previously, the SLR may vary if case information (hence associated hypotheses), are modified, as it should. However, the model must provide a SLR that does not depend on the system used nor on the operator. This is an essential problematic that currently needs to be thoroughly addressed to provide a method robust enough to be used for court purposes.

One of the main obstacles to using automated systems in forensic face recognition is that most algorithms are operated as black boxes, which means that both training sets and extracted features templates remain undisclosed by developers. Studies enlightened the importance of training sets on systems performances and thus recommend to use training sets that best fit operational case data - in terms of pose, luminosity, resolution and quality (Ali, 2014; Peng, 2019). However, most systems currently do not allow operators to have control over the algorithm training, which means overall performance and understandability are fixed once a system is released.

In the latest NIST tests (Grother et al., 2019), the authors evaluate the performance of 203 automatic face recognition algorithms through 1-vs-N identification tasks and report their accuracy through false acceptance and false rejection rates. The authors also reported that features templates, although unavailable to the operator, showed a large size range, from 100 bytes to more than 4 kilobytes. This suggests substantial variations in features extraction and template building processes among current systems.

In identity verification tasks (1-vs-1, e.g. to unlock electronic devices or pass customs control at the airport), the automatic face recognition system compares two images and calculates a score that translates the similarity between the features templates extracted from both faces. Depending on the value of the score in relation to the value of the threshold set for this task, the system concludes that the same person is visible in the two images compared or, conversely, that they show too many discrepancies to belong to the same person. Within this specific framework, thresholds are set to have false positive rates as close as possible to zero, while limiting false negative rates in order to guarantee a rapid and optimised use of the device. Such a process is entirely automated and requires no human participation. On the other hand, in forensic face recognition tasks (1-vs-N), the output is a list of scores associated with potential candidates, i.e. individuals from the dataset which faces were the closest to the query face, according to the system. The operator then reviews the score list and sorts possible hits, effectively reaching a decision based on understandability, in order to use this information to infer on the ongoing investigation. Using an evaluative model like the one described by Jacquet and Champod (2020) will extend the automatisation of the process to forensic evaluation, but it still needs human-based supervision and guidance.

In forensic science, it is essential to understand and control the intrinsic functioning and performances of the systems. Human-based analysis and comparison may lack objectivity and transparency as the observations and results of expertise as well as their reporting vary from one expert to another, depending on their own experience and training. In cases where the operator chooses to compare faces automatically, the lack of transparency is also caused by the lack of control over the training and outputs of the algorithm.

6.4. Detecting non-obvious link between cybercrimes

Bollé and Casey (2018) presented the performance evaluation of an automated system for detecting links between online frauds by computing near similarity of email addresses. The outputs of the system include a link between two email addresses and the associated similarity score between 0 and 1, which is an important aspect of this approach. This study found the best performance evaluation with Levenshtein distance. They adopt an approach where, in a crime analysis context, it is best to risk increasing false positive rates to reduce the risk of missing relevant links.

Although this automated system does not have an explicit explainability mechanism, it remains completely transparent[11]. The use of Levenshtein distance to compute near similarity of usernames within email addresses is understandable to a technical analyst, but might not be clear to a non-technical user of system. A user would at least know that the link had been established specifically on basis of email addresses, instead of any other types of traces, and would understand the approximate level of similarity between them. If the score is too low for the user's purpose in a given situation, the outputs are not used.

What is not immediately evident from the outputs of this automated system is the meaning of links with strong similarity scores. Such interpretation of the outputs requires forensic evaluation, as demonstrated in Fig. 4, which shows links between pairs of addresses, both with approximately the same similarity score. The absolute value of the score is not particularly important in this example. One can only note that, if not familiar with such systems, it may be hard to make sense of this particular value. The important aspect in this example is the fact that both scores are close to each other. In the first situation, the email addresses p.novo@ and

[11] The system by itself does present in full detail what is happening when computing the similarity, as it only displays the similarity score, but the details are available in the associated paper (Bollé and Casey, 2018).

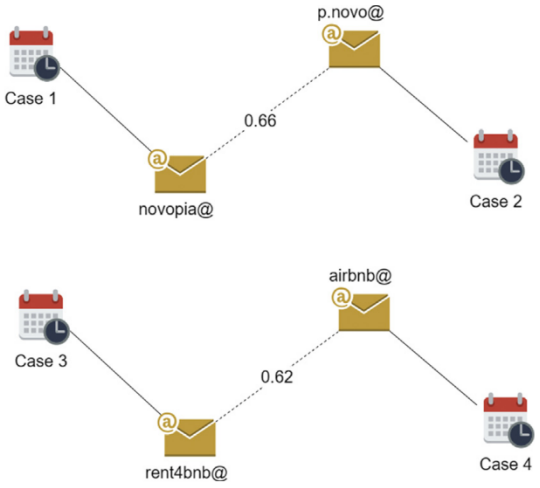

Fig. 4. Example of two links output by the system. The number between the email addresses is their similarity score.

novopia@ are quite specific, so the hypotheses could be "The same person made both email addresses" *versus* "Different persons made both email addresses". The forensic evaluation strongly supports the conclusion that it is the same person who made the addresses and that the two cases are a crime series.

In the second situation, the email addresses rent4bnb@ and airbnb@ are very common. Keeping the same hypotheses as before, forensic evaluation leans more towards the conclusion that different people made these addresses. However, changing the hypotheses to "Both cases use the same modus operandi" *versus* "The cases use different modus operandi", and factoring past experience with online frauds, forensic evaluation leans more towards it being the same modus operandi. In this example, we could detect frauds that send email in the name of AirBnB to scam people looking to rent apartments. This decision would increase understanding about this phenomenon. This clearly shows that based on the same outputs from an automated system, forensic evaluation and formulation of the hypotheses is an important step that can completely change the type decision.

The forensic evaluation made in this example was made in Switzerland. The context here is important as the name "Novo Pia" is not a common name in Switzerland, and that is why it would be more probable to observe such email addresses if the same person created both addresses. However, this name is common in Portugal and in such context, the conclusion would have been different. This illustrates how, given the same automated system output, different conclusions and decisions can be reached by taking into account knowledge and context.

7. Designing automated systems to support evaluations

This section proposes requirements for the design of automated systems supporting forensic analysis to help users perform the three levels of evaluation.

In particular, the principles of accountability, reliability, transparency and scientific reasoning should be at the heart of such automated systems.

7.1. Fit-for-purpose performance evaluation

Automated systems supporting forensic analysis should provide sufficient details about performance metrics for forensic practitioners to determine if the system is fit-for-purpose.

To evaluate the performance of the system, it is necessary to consider a performance metric that was computed in the same situation as the system will be used for forensic purposes. The performance computed has a meaning only for the data on which the system was trained and/or tested. If the system is used on dissimilar data or to address different forensic questions, it is impossible to evaluate how the system will perform. The evaluation of the performance requires at least a description of the data used for training and/or testing. Such transparency could raise ethical problems because it could violate data protection policies. An example could be a system trained on child pornography, which has stirct access restrictions. Models that refer to the training set, such as the Nearest Neighbour, may be not useable in such situations.

7.2. Understandability evaluations

Automated systems supporting forensic analysis should provide forensic practitioners with a way to understand how a given output was produced.

An important point to note is that understandability is subjective. A forensic practitioner conversant with programming languages will have a better understanding of an open source automation framework than a non-programmer. An expert in machine learning will have a better understanding of a system than a forensic practitioner, who will have a better understanding than a judge. It is thus important that a strong communication exists between the developer of the system and the users who will need to understand and use the results.

To improve the communication, it would be beneficial to have feedback loops during the development process that enable users to express their needs and that allow the developer to express the feasibility of the expectations. Then a discussion between them can start where the user can express how confident he or she is in understanding the results of the system and the developer can work on enhancing the understandability of the system.

In general, the trade-off between understandability and completeness in understandability we discussed previously, raises ethical challenges. Indeed, for a non-technical person to understand how the system works, one might be tempted to oversimplify the explanations or to select the aspects that will only give more confidence and trust in the system, and to omit the limitations or risks of the system. This may be particularly true in the justice system where people prefer unfailing systems and definitive answers. This same challenge applies to expressing the weight of forensic evaluation as discussed earlier.

7.3. Forensic evaluation

Automated systems supporting forensic analysis should guide forensic practitioners through the process of setting alternative hypotheses and assessing the strength of observed output in light of each hypothesis to produce a relative strength of evidence.

As stated previously, the forensic evaluation requires a set of competing hypotheses. Concretely, these hypotheses will be related to the forensic question that is being addressed with the assistance of the automated system. The output may be combined with general knowledge, taken from academic research and past experience, and with the specific context of the case. An open question here is how much a system supporting forensic analysis can automate the whole process of selecting hypotheses, taking into account knowledge and context, and making the evaluation.

There is no categorical response to this question. Given the role of expert knowledge in forensic evaluation, it might only be possible to automate certain aspects of the process under certain

circumstances. With this in mind, some thought should be accorded to the possibility of combining an expert system with the machine learning system to take into account rules derived from domain specific knowledge. The information must also be structured in a way that can support contextual analysis. This means that forensic practitioners should be able to observe the context of each piece of information within the automated system.

It may be hard to predict the set of all the reasoning that will be performed by a forensic practitioner in this entire process but the ones that can be identified should be included in the system and should be fully or partially automated. At the same time, we recommend leaving some space for forensic practitioners to customise the different processes to their needs in a specific case. First of all, it will give more transparency to the whole process and if errors occur, it will be possible to identify exactly which step was erroneous (machine learning part, evaluation or decision). Secondly, it leaves room for future improvement, without having to redesign the whole system. Finally, it helps the forensic practitioner to formalize his/her thinking and to learn.

Earlier, we discussed the possibility of a system that will not produce a recommendation but that will instead make a decision. In the end, it is a case where the whole process has been automated. The hypotheses are the different possible outputs but often, such systems do not make a formal evaluation. We still recommend that such systems formalize the different steps in order for a human to later understand how the decision was made.

In summary, when designing an automated system to support forensic analysis, performance evaluation results should be sufficiently detailed to determine whether or not it is fit-for-purpose in a specific case. In addition, transparency and understandability should be a requirement during the design of automated systems supporting forensic analysis. The system should also support contextual analysis by keeping the context of the information at every stage. It may also be useful to keep the possibility of integrating an expert system to reinforce forensic evaluation. The system should guide users through the forensic evaluation steps to be sure that they understand what was done and to stay transparent. We recommend that the hypotheses are explicitly formulated, even if it is automated and always the same. If some steps are automated, we recommend they still should be explicitly described.

8. Conclusions & future work

This section provides a synthesis of the results and considers their significance, potential challenges and future work.

As we saw, automatic system usage supporting forensic analysis can raise many questions in terms of reliability and decision making. We proposed a formalization of three levels of evaluation when using such systems.

Firstly, performance evaluation should be conducted to determine if the system is suitable for a specific purpose. Secondly, understandability evaluation should be conducted to verify that the forensic practitioner can understand the system and that the system is working as expected. Thirdly, it is necessary to evaluate the output of the system in light of multiple hypotheses in order to reach a decision. The possibility of errors and the meaning of the output will also be taken into account at this stage, as well as knowledge and context.

In some situations, this entire evaluation process can be performed completely by the system, or partially assisted by the user. In any case, it is important that the system clearly formalizes each step in order to be transparent and reliable when the final forensic conclusion is being discussed. The system should be designed to support all levels of evaluations. Some recommendations on how to design such systems are provided in this work, but many challenges remain.

The primary challenge is the variety of systems. It would not be possible to have an exhaustive list of possible systems and recommendations for each. The approach presented in this paper defines a more general and conceptual evaluation framework evaluation that could fit any system. The general framework and recommendations in this work can be effective when designing a particular automated system supporting forensic analysis.

Another challenge that will require future research is how to express the results of an understandability evaluation. As stated earlier, understandability evaluation supports a forensic practitioner's decision whether a specific output from an automated system is incorrect/invalid and should be disregarded, or is suitable for forensic evaluation. The challenge partially resides in the fact that understandability is subjective. Two different users may take completely different decisions when determining if the output is suitable for forensic evaluation.

This subjectivity aspect raises a potential ethical challenge concerning understandability. Is it ethical to over simplify the mechanisms behind a system, so that anyone can understand, but with the risk of missing an important aspect? Such oversimplification might also lead the user to have a blind trust in the system and the consequences could be the same as using a black-box. The transparency of the data used to train and test the system might also not be compatible with laws on data protection.

The complexity of evaluating automated systems supporting forensic analysis, and the number of elements to take into account makes this a ripe area for research and development on how to represent and structure information, knowledge and context in such systems.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

Aitken, Colin G.G., Taroni, Franco, 2004. Statistics and the Evaluation of Evidence for Forensic Scientists, first ed. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470011238.

Ali, Tauseef, 2014. Biometric score calibration for forensic face recognition.

Anda, Felix, David Lillis, Kanta, Aikaterini, Becker, Brett, Bou-Harb, Elias, Le Khac, Nhien An, Scanlon, Mark, 2019. Improving the accuracy of automated facial age estimation to aid CSEM investigations. Digit. Invest. 28 (April), S142. https://doi.org/10.1016/j.diin.2019.01.024.

Anda, Felix, Nhien-An Le-Khac, Scanlon, Mark, 2020. DeepUAge: improving underage age estimation accuracy to aid CSEM investigation. In: Forensic Science International: Digital Investigation, March.

Beaudouin, Valérie, Bloch, Isabelle, Bounie, David, Clémençon, Stéphan, d'Alché-Buc, Florence, James, Eagan, Maxwell, Winston, Mozharovskyi, Pavlo, Parekh, Jayneel, 2020. Flexible and context-specific AI explainability: a multidisciplinary approach. Available at SSRN 3559477.

Berger, Charles, 2019. Criminalistics Course. University of Leiden.

Bollé, Timothy, Casey, Eoghan, 2018. Using computed similarity of distinctive digital traces to evaluate non-obvious links and repetitions in cyber-investigations. Digit. Invest. 24 (March), S2–S9. https://doi.org/10.1016/j.diin.2018.01.002.

Campolo, Alex, Sanfilippo, Madelyn, Whittaker, Meredith, Crawford, Kate, 2017. AI Now 2017 Report. *AI Now Institute at New York University*.

Casey, Eoghan, 2020. Standardization of forming and expressing preliminary evaluative opinions on digital evidence. Forensic Sci. Int.: Digit. Invest. 32 (March), 200888. https://doi.org/10.1016/j.fsidi.2019.200888.

Casey, Eoghan, Zoun, Rikkert, 2014. Design tradeoffs for developing fragmented video carving tools. Digit. Invest. Fourteenth Annu. DFRWS Conf. 11 (August), S30–S39. https://doi.org/10.1016/j.diin.2014.05.010.

Casey, Eoghan, Nelson, Alex, Hyde, Jessica, 2019. Standardization of file recovery classification and authentication. Digit. Invest. 31 (December), 100873. https://doi.org/10.1016/j.diin.2019.06.004.
Champod, Christophe, Biedermann, Alex, Vuille, Joëlle, Willis, Sheila, De Kinder, Jan, 2016. ENFSI guideline for evaluative reporting in forensic science: a primer for legal practitioners. Crim. Law Justice Wkly. 180 (10), 189–193.
Doshi-Velez, Finale, Kim, Been, 2017. Towards a Rigorous Science of Interpretable Machine Learning. " ArXiv Preprint ArXiv:1702.08608.
Durmus, Emre, Korus, Pawel, Memon, Nasir, 2019. Every shred helps: assembling evidence from orphaned JPEG fragments. IEEE Trans. Inf. Forensics Secur. 14 (9), 2372–2386. https://doi.org/10.1109/TIFS.2019.2897912.
Egli, Nicole M., 2009. Interpretation of Partial Fingermarks Using an Automated Fingerprint Identification System. *Faculty of Law and Criminal Sciences*.
Explain. Oxford Advanced Learner's Dictionary. n.d, Accessed. https://www.oxfordlearnersdictionaries.com/definition/english/explain?q=explain. (Accessed 31 May 2020).
Gilpin, Leilani H., Bau, David, Yuan, Ben Z., Bajwa, Ayesha, Michael Specter, Kagal, Lalana, 2018. Explaining explanations: an overview of interpretability of machine learning. In: In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp. 80–89.
Grother, Patrick, Ngan, Mei, Hanaoka, Kayee, 2019. Face Recognition Vendor Test (FRVT) Part 2: Identification.
Guestrin, Tulio Ribeiro, Marco, Singh, Sameer, Carlos, 2016. Local Interpretable Model-Agnostic Explanations (LIME): an Introduction. O'Reilly Media. August 12, 2016. https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime.
Gunning, David, 2017. Explainable artificial intelligence (xai). In: Defense Advanced Research Projects Agency (DARPA), Nd Web 2. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.
Henseler, Hans, Hyde, Jessica, 2019. Technology Assisted Analysis of Timeline and Connections in Digital Forensic Investigations.
Interpret. In *oxford advanced learner's dictionary*. n.d. https://www.oxfordlearnersdictionaries.com/definition/english/interpret?q=interpret. (Accessed 31 May 2020).
Jacquet, Maëlig, Champod, Christophe, 2020. Automated face recognition in forensic science: review and perspectives. Forensic Sci. Int. 307 (February), 110124. https://doi.org/10.1016/j.forsciint.2019.110124.
Jeni, László A., Cohn, Jeffrey F., De La Torre, Fernando, 2013. Facing imbalanced data recommendations for the use of performance metrics. In: International Conference on Affective Computing and Intelligent Interaction and Workshops : [Proceedings]. ACII (Conference) 2013, pp. 245–251. https://doi.org/10.1109/ACII.2013.47.
Lau, Timothy, Biedermann, Alex, 2020. Assessing AI output in legal decision-making with nearest Neighbors. Penn State Law Rev. forthcoming https://papers.ssrn.com/abstract=3459870.
Margagliotti, Giulia, Bollé, Timothy, 2019. Machine learning & forensic science. Forensic Sci. Int. 298, 138–139.
Marquis, Raymond, Biedermann, Alex, Cadola, Liv, Champod, Christophe, Gueissaz, Line, Massonnet, Geneviève, Williams, David Mazzella, Franco, Taroni, Hicks, Tacha, 2016. Discussion on how to implement a verbal scale in a forensic laboratory: benefits, pitfalls and suggestions to avoid misunderstandings. Sci. Justice 56 (5), 364–370.
Metz, Joachim, White, David, 2020. Plaso. Github Repository. https://github.com/log2timeline/plaso. (Accessed 20 July 2020).
Meuwly, Didier, 2000. Reconnaissance de Locuteurs En Sciences Forensiques: L'apport d'une Approche Automatique. PhD Thesis. Université de Lausanne, Faculté de droit et des sciences criminelles.
Meuwly, Didier, Ramos, Daniel, Haraksim, Rudolf, 2017. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. Forensic Sci. Int. 276 (July), 142–153. https://doi.org/10.1016/j.forsciint.2016.03.048.
Meyer, Maxime, 2015. Machine Learning to Detect Online Grooming.
Nist Software Quality Group, 2017. Forensic file carving. Text. NIST. https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt/cftt-technical-0. (Accessed 8 May 2017).
Peng, Yuxi, 2019. Face Recognition at a Distance: Low-Resolution and Alignment Problems.
Pollitt, Mark, 2008. Digital orange juice. J. Digit. Forensic Pract. 2 (1), 54–56. https://doi.org/10.1080/15567280701721921.
Pollitt, Mark, Casey, Eoghan, Jaquet-Chiffelle, David-Olivier, Gladyshev, Pavel, 2018. A framework for harmonizing forensic science practices and digital/multimedia evidence. In: Organization of Scientific Area Committees for Forensic Science. https://doi.org/10.29325/OSAC.TS.0002.
Rahman, Ab, Hidayah, Nurul, Glisson, William Bradley, Yang, Yanjiang, Kim-Kwang, Raymond Choo, 2016. Forensic-by-Design framework for cyber-physical cloud systems. IEEE Cloud Comput. 3 (1), 50–59.
Ramos, Daniel, Gonzalez-Rodriguez, Joaquin, 2013. Reliable support: measuring calibration of likelihood ratios. Forensic Sci. Int. 230 (1–3), 156–169. https://doi.org/10.1016/j.forsciint.2013.04.014.
Sanchez, Laura, Grajeda, Cinthya, Baggili, Ibrahim, Hall, Cory, 2019. A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). Digit. Invest. 29 (July), S124–S142. https://doi.org/10.1016/j.diin.2019.04.005.
Scikit-Learn Developers, 2019. 3.3. Metrics and scoring: quantifying the quality of predictions. In: Scikit-Learn 0.22.1 Documentation. 2019. https://scikit-learn.org/stable/modules/model_evaluation.html.
Shrikumar, Avanti, Greenside, Peyton, Kundaje, Anshul, 2019. Learning important features through propagating activation differences. ArXiv:1704.02685 [Cs], October. http://arxiv.org/abs/1704.02685.
Taylor, Duncan A., Jo-Anne, Bright, John, Buckleton, 2017. "Commentary: a 'source' of error: computer code, criminal defendants, and the constitution. Front. Genet. 8 https://doi.org/10.3389/fgene.2017.00033.
Thompson, William C., 2017. How should forensic scientists present source conclusions. Seton Hall Law Rev. 48, 773.
Understand. Oxford advanced learner's dictionary. n.d. https://www.oxfordlearnersdictionaries.com/definition/english/understand?q=understand. (Accessed 21 May 2020).
Willis, S.M., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C.E.H., Sjerps, M.J., Lucena-Molina, J.J., 2015. ENFSI Guideline for Evaluative Reporting in Forensic Science. *European Network of Forensic Science Institutes*.