KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-022-00763-9

DISCUSSION

Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques

Abiodun A. Solanke[1] · **Maria Angela Biasiotti**[2]

Received: 31 January 2022 / Accepted: 26 April 2022

Abstract
The impact of AI on numerous sectors of our society and its successes over the years indicate that it can assist in resolving a variety of complex digital forensics investigative problems. Forensics analysis can make use of machine learning models' pattern detection and recognition capabilities to uncover hidden evidence in digital artifacts that would have been missed if conducted manually. Numerous works have proposed ways for applying AI to digital forensics; nevertheless, scepticism regarding the opacity of AI has impeded the domain's adequate formalization and standardization. We present three critical instruments necessary for the development of sound machine-driven digital forensics methodologies in this paper. We cover various methods for evaluating, standardizing, and optimizing techniques applicable to artificial intelligence models used in digital forensics. Additionally, we describe several applications of these instruments in digital forensics, emphasizing their strengths and weaknesses that may be critical to the methods' admissibility in a judicial process.

Keywords Digital forensics · AI · Machine learning · Evaluation · Standardization · Optimization · Evidence mining

1 Introduction

The advancement of research and development of methodologies for big data mining [1] powered by Artificial Intelligence (AI) [2, 3], which seeks to discover meaningful and explorable patterns in data, has enabled/motivated its application in digital forensics (DF) investigation.[1] Digital artifacts are collections of digital data that are frequently large, complex, and heterogeneous. Despite concerns about the ability of "black-box" AI models [4] to generate reliable and verifiable digital evidence [5], the assumption that cognitive methodologies used in big data analysis will succeed when applied to DF analysis has fueled a decade-long surge of research into the application of AI in DF. Note that, our reference to AI methods in this paper includes machine learning (ML) [6, 7] and deep learning (DL) [170] methods; with distinctions made where necessary.

To begin, a misunderstanding exists regarding the colloquial use of the terms "Forensics AI" and "AI Forensics" within the forensics community (and beyond), with some using the phrases interchangeably as referring to the application of AI in DF. While both phrases are self-explanatory, it is vital to clarify common misconceptions and distinguish the two concepts. On the one hand, according to [8], a word preceding 'forensics' in the DF domain denotes the target (tool or device) to be analyzed (e.g., cloud forensics, network forensics, memory forensics, etc.). As a result, the author refers to "AI Forensics" as a forensic analysis of AI tools or methods, rather than forensic investigation applying AI techniques. In the same vein, the authors in [9], refers to AI Forensics as *"scientific and legal tools, techniques, and protocols for the extraction, collection, analysis, and reporting of digital evidence pertaining to failures in AI-enabled systems."* To summarize their definition, AI Forensics is the analysis of the sequence of events and circumstances that led to the failure of an intelligent system, including assessing

✉ Abiodun A. Solanke
abiodun.solanke@unibo.it

Maria Angela Biasiotti
mariangela.biasiotti@igsg.cnr.it

[1] CIRSFID Alma-AI, University of Bologna, Bologna, Italy

[2] Institute of Legal Informatics and Judicial Systems (IGSG-CNR), Firenze, Italy

[1] an investigation of a case (criminal or civil) enabled by computing devices, with the primary goal of establishing facts or finding admissible digital evidence in court.

Published online: 12 May 2022

whether or not the failure was caused by malicious activity and identifying responsible entity(ies) in such scenario.

In contrast to the previously described concept, a comprehensive review of research databases such as Google Scholar, IEEE Explore, and Scopus for the terms "Forensics AI" or "Digital Forensics AI" reveals that the majority of resources are based on DF analysis methods assisted by AI techniques. However, in this paper, we refer to Digital Forensics AI (hereafter referred to as DFAI), *as a generic or broader concepts of automated systems that encompasses the scientific and legal tools, models, methods; including evaluation, standardization, optimization, interpretability, and understandability of AI techniques (or AI-enabled tools) deployed in digital forensics domain.* Also, we refer to "digital evidence mining" as *the process of automatically identifying, detecting, extracting, and analyzing digital evidence with AI-driven techniques.* The phrase "mining" is borrowed from the notion of data mining, which embodies procedures and components that can be applied in the analysis of digital evidence.

Importantly, as accurate and precise as most AI algorithms are; owing to numerous research focus and resources dedicated to them of recent, their applications to digital forensics require significant cautions, and consideration for domain-specific intricacies. Clearly, the results of a business-oriented AI task will be evaluated differently from those of a forensic investigation. Additionally, the bulk of AI algorithms are based on statistical probabilities, which commonly results in non-deterministic outputs. Thus, the challenge would be to establish the correctness of the outcomes and to communicate the probabilistic conclusion of a forensic examination in the simplest and most understandable manner possible in order for it to be admissible in legal proceedings.

As a result, in this work, we emphasize the importance of three scientific instruments in the application of AI in digital forensics: evaluation; standardization; and optimization of the approaches used to accomplish the tasks. In subsequent sections of this work, we will discuss the significance of these instruments and their components.

This paper makes the following contributions:

- We present various AI model evaluation approaches, emphasizing their importance for both DFAI methodologies and the forensic tasks to which they are best suitable.
- We propose a confidence scale (C-Scale) for the evaluation of strength of evidence that is adaptive to an AI generated probabilistic results.
- We discuss numerous optimization techniques that may be appropriate for certain forensic analysis, as well as a comparison of their strengths and drawbacks, including their time complexity for DFAI tasks.

The subsequent parts of the paper are organized as follows. Section 2 covers the methods for evaluating DFAI techniques. In Sect. 3, the methods for standardizing DFAI techniques are discussed, while Sect. 4 elaborates on the techniques optimization. Finally, in Sect. 5, we discussed the future direction and conclusions.

2 Methods for Evaluating DFAI Techniques

During a forensic investigation, examiners develop an initial hypothesis based on observed evidence. Following that, the hypothesis is evaluated against all other competing hypotheses before final assertions are made [10]. The issue is that, as highlighted in [11], in an attempt to make sense of what they observe (sometimes coercively to ensure that it fits the initial assumption), investigators subconsciously: (1) seek findings that support their assertions; (2) interpret relevant and vague data in relation to the hypothesis; and (3) disregard or give less weight to data that contradict the working hypothesis. Numerous factors may contribute to this bias, including but not limited to: confidence (as a result of the presumption of guilt), emotional imbalance, concern about long-term consequences (e.g., loss of prestige), and personality characteristics (e.g., dislike for uncertainty or a proclivity to over-explore various scenarios) [12]. Consequently, before a forensic investigation can reach a conclusion, each component of the initial hypothesis must be independently and thoroughly tested (or evaluated) to ascertain the degree of confidence in the methodology that produced the fact. Evaluation, therefore, is the process of judging the strength of evidence adducing opposing assertions, as well as their relative plausibility and probability [13].

Expert examiners can evaluate forensic examination data using a variety of techniques, some of which are based on predefined scientific standards and others on logical deductions supported by experience or subjective reasoning. However, in the context of DFAI, forensic evaluation is performed by evaluating the AI algorithms deployed in the forensic investigation. This deployment requires metrics and measurements that are compatible with AI model evaluation. The evaluation of DFAI models can be carried out on the algorithm's functional parameters (i.e., individual modules) or on their outputs. Unlike conventional approaches for evaluating ML or DL models, which apply standard metrics associated with the task or learning algorithm, gaining confidence in the outcome of a DFAI research may require additional human observation of the output. Numerous studies in DF have revealed that forensic practitioners frequently issue inconsistent or biased results [13, 14]. In addition, the majority of AI-based approaches lack the necessary clarity and replicability to allow investigators to assess the accuracy

of their output [15]. Thus, a forensically sound process[2] , is one that integrates automated investigative analysis—evaluated through scientific (accuracy and precision) metrics—with human assessments of the outcome. For example, a DF investigation into Child Sexual Exploitation Material (CSEM) [16, 17] may seek to automatically detect and classify images of people found on a seized device as adult or underage (based on automatic estimated age). Because of possible misrepresentation in the dataset, misclassification (i.e., false positive), misinterpretation of features, and missing of critical features during the classification process that could have served as evidence (false negative; e.g., an underage wearing adult facial makeup) may occur [18]. In this case, merely addressing bugs in algorithmic codes may not be sufficient, as the classification errors may be subconsciously inherited and propagated through data. Similarly, the work described in [19] is a temporal analysis of e-mail exchange events to detect whether suspicious deletions of communication between suspects occurred and whether the deletions were intended to conceal evidence of discussion about certain incriminating subjects. One significant drawback of that analysis is the model's inability to thoroughly investigate if the suspicious message(s) were initiated or received by the user or were deliberately sent by an unauthorized hacker, remotely accessing the user's account to send such incriminating message. To reach a factual conclusion in this case, various other fragmented unstructured activity data (unrelated to e-mail, perhaps) must be analyzed and reconstructed. Depending on the design, a robust AI-based system can uncover various heretofore unrecognized clues. If these new revelations (even though relevant) are not properly analyzed and evaluated, they may lead investigators to believing that the outputs dependably fulfil their needs [15]. As a result, an extensive review of the output of DFAI will be required (supposedly provided by human experts) to arrive at a factually correct conclusion. This has also been highlighted as an important instrument for examining digital evidence in [10]. Additionally, expert knowledge that has been codified as facts (or rules) in a knowledge base can be used in place of direct human engagement to draw logical inferences from evidence data.

As with the output of any other forensic tool capable of extracting and analyzing evidence from digital artifacts, which frequently requires additional review and interpretation that are compatible with the working hypothesis, the results of forensic examinations conducted using DFAI should be viewed as "recommendations" that must be interpreted in the context of the overall forensic observation and investigation [15]. In addition, the evaluation apparatus must be verifiable, appropriate for the task it seeks to solve, and compatible with the other contextual analysis of the investigative model. Taking this into consideration, the methods for evaluating a DFAI techniques can be viewed in terms of two significant instruments: performance and forensic evaluation. Below, we discuss the significance and components of each of these instruments. These two instruments, in our opinion, are quite essential for a sound digital forensic process based on DFAI.

[2] Transparent digital forensics procedure that preserves the true context of the data for use in a legal proceeding.

2.1 Methods for Evaluating the Performance of DFAI Models

In a machine-driven system, evaluation produces value as a measure of the model's performance in accomplishing the task for which it was commissioned, which may be used to influence decision-making [10]. Depending on the problem the model attempt to solve, evaluation may be: a set of thresholds formulated as binary (i.e., 'yes' or 'no', or 0 or 1) or categorical (qualitative; one of a possible finite outcome) as the case maybe; discrete (enumeration of strength; e.g., range between 0 to 10); or continuous (e.g., probability distributions of real values between 0 and 1). Consequently, evaluating the performance of a DFAI model built to recognize specific faces in a CSEM is distinct from evaluating the performance of a model meant to classify faces as underage or adolescent. Similarly, distinct metrics are required for models that detect spam e-mails and those that attempt to infer intent from an e-mail content. The majority of DFAI tasks will fall into one of three categories: classification, regression, or clustering. The scientific methods used for evaluating the performance of these three categories are discussed below. It is worth mentioning, however, that these are standard metrics for ML tasks. Hence, we offer only a brief review of the methods, emphasizing the intersection and relevance of each metric to DFAI (including the weaknesses and strengths that make them appropriate or otherwise) where necessary. Therefore, readers are encouraged to consult additional publications on ML metrics for complete details.

2.1.1 Evaluating Classification Algorithms in DFAI

Classification models are predictive in nature, identifying the class to which a set of input samples belongs. Classification tasks are evaluated by comparing predicted class samples to ground-truth samples. In a vast majority of cases, classification model design will include both positive and negative examples. The former represent true samples obtained from data, whilst the latter are fictitious samples that do not exist in the real sense. A classification task is commonly modelled in ML as a binary representation that predicts a Bernoulli probability distribution [21] for each sample. Bernoulli

distributions are a type of discrete probability distribution in which events have binary outcomes such as 0 or 1. Therefore, the performance of a classification model is measured by its ability to correctly predict (assign a high probability value to) the class of positive samples and to assign a very low probability value to non-existent samples.

Prior to deploying a DFAI model, it is necessary to examine the characteristics of the investigation to determine whether the model is appropriate for that purpose. Practitioners are expected to be aware of the unique characteristics of learning algorithms and to use them appropriately. For instance, in a forensic investigation involving facial classification, two main techniques that can be applicable: verification and identification. Verification entails comparing an unknown face to a known face directly (One-vs-One) [22] and computing their similarity score. This can be adapted as a binary classification task, in which the system predicts whether or not two faces share a high degree of similarity, based on a predetermined threshold. On the other hand, identification involves One-vs-Rest [23] comparison, in which an unknown face is compared to the faces in a database of known persons. The Identification task is typically a "Multi-Class Classification" [24] problem, in which samples are classified into one of a set of known classes. Other classification models are: *Multi-label classification* [25] and *Imbalanced classification* [26].

Metrics such as accuracy, precision, recall, and F-Measure are all relevant depending on the investigation's characteristics. The measure of "accuracy" can be seen as the validity measure of a model. It is the ratio of the correctly classified samples to the total samples. Accuracy tells whether a model was correctly trained and how well it will function in general. However, caution should be exercised when using this information alone to reach a general conclusion in forensic investigation, as it provides little information about its application to the problem and performs poorly in circumstances of severe class imbalance. That is, if the dataset is asymmetric, e.g., if the proportion of false positives is not (or nearly) equal to the proportion of false negatives. Accuracy is calculated in terms of a confusion matrix while performing a binary classification task, such as predicting whether an e-mail is "spam" or "not-spam." The confusion matrix [27] [28] is applied to a set of test data, for which the true values are known. What a classifier seek to minimize is the number of *"False Positives"* and *"False Negatives."* A *true positive (tp)* is one in which the model accurately predicts the positive samples, while a *true negative (tn)* indicates the result of correctly predicted negative samples. Similarly, a *false positive (fp)* outcome occurs when the model incorrectly predicts positive samples, whereas a *false negative (fn)* outcome occurs when the model inaccurately predicts negative samples. Therefore, in terms of confusion matrix, an accuracy measure is represented as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

To ascertain the reliability of a DFAI model, **precision** metric [29] is critical. It provides additional assurance by posing the question: "how frequently is the model correct when it predicts a positive sample?" With precision, we affirm the classifier's ability not to label a negative sample as positive. Given that the outcome of a forensic investigation may be critical to the outcome of an inculpatory or exculpatory proceeding, the cost of a high rate of false positives may be detrimental.

Additionally, in situations where the cost of a false negative is potentially catastrophic, such as a facial recognition investigation to discover criminal materials via training examples. While the system is capable of identifying and classifying a large number of positive samples, it may be necessary to ascertain how many faces were correctly identified from the predicted samples. This is where **recall** [29] plays a critical role in DFAI. Recall is crucial for evaluating working hypotheses and can help in answering some potentially damning questions during court proceedings. Recall facilitates informed decisions on false negatives; for example, by highlighting crucial details that should not be overlooked.

To take advantage of both precision and recall's evaluative strength, the **F-Measure** (or F-Score) can be employed to measure the model's accuracy. It takes into consideration both false positives and negatives; with a low value indicating a good F-Measure. This has the potential to aid in the reduction of false assumptions during forensic investigations.

Another relevant metric for measuring a classifier's capacity to distinguish between classes is the **Area Under the Curve (AUC)** [30], which serves as a summary of the Receiver Operating Characteristic (ROC) curve [31]. The ROC curve is constructed by plotting the *tp* rate versus the *fp* rate at various threshold values. The AUC and **Average Precision (AP)** [32] are the quality measures used in link the performance of link prediction models, as well as the probability of a relationship between hypothetical variables.

There are instances when evaluating accuracy becomes preferable to F-measures; this is especially true when the cost of false positives and negatives is similar, meaning that the consequences are not negligible. If the situation is reversed, it is reasonable to evaluate the F-measure. However, some critical concerns about the F-measure's weaknesses are discussed in [33, 34]. Notable among them are its bias towards the majority class and its underlying

assumption that the actual and predicted distributions are identical. Additionally, caution should be exercised when evaluating performance on classified samples that involves the assignment of a threshold (as is the case in some logistic regression models). Increasing or decreasing the threshold value (in a classification model) has a major effect on the precision and recall results. In contrast to a model developed to optimize business decisions, it may be prudent to avoid including any threshold in DFAI—as it would be appropriate to have a realistic view of the analysis' outcome, unless there is certainty that doing so will not have a detrimental impact on the outcome. Nonetheless, accuracy is crucial; so, the threshold can be considered provided the trade-offs can be quantified and justified sufficiently.

2.1.2 Evaluating Regression Algorithms in DFAI

In contrast to classification models, which predict the classes of input samples, regression models predict an infinite number of possible (continuous; real-valued such as integer or floating point) outcomes. In DFAI, regression analysis can be utilized for two conceptually distinct purposes: forecasting and prediction; and inference of causal relationships between dependent (observed) and independent (predictors) variables. Before a regression analysis may be commissioned, the examiner must be convinced that the correlations present in the data possess the predictive power to infer a new context or that these correlations can induce a causal interpretation based on observational data [35, 36]. This is particularly important for forensic investigations. A significant factor that can improve the predictive capabilities of a regression model is when the input variables are arranged chronologically (according to event time), a notion referred to as time series forecasting. This is important for forensic tasks such as detecting deviations (anomalies), forecasting crime, predicting probable connections between data, and reconstructing events. Furthermore, while working with regression models, interpolation and extrapolation [37] are critical concepts to understand. Often, the former is preferable, as it involves the prediction of values within the range of data points in the dataset used to fit the model. The latter, on the other hand, depending on the task, might not be fully desirable for DFAI. Extrapolation is based on regression assumptions and requires predicting values outside the observed data range. Extrapolating over a range that is significantly larger than the actual data is risky and it is a sign of likely model failure.

A regression model's performance is measured as an error in prediction, i.e., how close the predictions were to the ground truth. To do this, the following error measures are frequently used: Mean Squared Error (MSE) [38–40], Root Mean Squared Error (RMSE) [41], Mean Absolute Error (MAE) [40], and Mean Absolute Percentage Error (MAPE) [42]. Although there are several other error metrics available; the choice of which is determined by the type of error being evaluated. We present a brief discussion about the above-mentioned metrics below.

MSE can be used to evaluate the quality of a predictor or an estimator. However, in DFAI, it better-off as a predictor since it can map arbitrary input to a sample of random variables. A MSE of zero indicates a perfectly accurate prediction, however this is rarely possible [43]. Unfortunately, other measures have been sometimes preferred to MSE due to its disproportionate weighting of outliers [44]. This occurs as a result of magnification of large errors than on small ones, due to each value being squared.

An extension of the MSE is the **RMSE**; which is always non-negative. A value of zero (0) is almost unrealistic; and if it does occur, it indicates that the model is trivial. RMSE is highly susceptible to outliers, as larger errors are significantly weighted. It may be prudent to establish a baseline RMSE for the working dataset in DFAI tasks by predicting the mean target value for the training dataset using a naive predictive model[3]. This can be accomplished by transforming or scaling the dataset's feature vectors between 0 and 1 (i.e., normalization).

In contrast to the previously stated error measures, which require squaring the differences, **MAE** changes are linear, intuitive, and interpretable; they simply represent the contribution of each error in proportion to the error's absolute value. MAE calculates the error difference between paired observations expressing the same event, i.e., it is scale-dependent; it uses the same scale as the data being measured[4] Moreover, it does not give greater or lesser weight to errors and hence provides a realistic view of the main prediction errors; thus, it is strongly recommended for DFAI. Additionally, it is a frequently used metric for forecasting error in time series analysis [45], which may be beneficial when examining an event reconstruction problems.

While **MAPE** appears to be well-suited for prediction, particularly when adequate data is available [46], caution should be exercised to prevent the 'one divided by zero' problem. Additionally, MAPE penalizes negative-valued errors significantly more than positive-valued errors; as a result, when utilized in a prediction task, it favours methods with extremely low forecasts, making it ineffective for evaluating tasks with large errors [46].

[3] A model in which the minimum possible effort or the less complex procedures are employed to make a prediction, often a random or constant prediction.

[4] See "Evaluating Forecast Accuracy." OTexts. Cited on Aug. 5, 2021. Available at https://otexts.com/fpp2/accuracy.html.

There are other error measures for regressors such as Max Error [47]; that calculates the maximum residual error and detect worst case errors [15], and R^2 (also known as R-Squared, Goodness of fit; Co-efficient of Determination) [48–50], which is the measure of variance proportion in the regressor.

Following the description of each of these error measurements for regression problems and their associated limitations in some cases, selecting which one is most appropriate for a specific forensic task can be somewhat puzzling. However, as demonstrated in [51], the RMSE is unreliable and unsuitable for determining the correctness of a time series analysis (such as temporal event reconstruction). Additionally, the study in [44, 52] stated that RMSE possessed "disturbing characteristics," rendering it ineffective as an error measure. MSE and all other squared errors were also deemed unsuitable for evaluation purposes (in the study). The work described in [53] somewhat challenged these conclusions by presenting arguments in support of RMSE. Nevertheless, MAE has been recommended in the majority of cases, which is understandable. As previously stated, the MAE metric is a consistent and compatible evaluation technique with DFAI; it is a more natural representation of the model's average error magnitude [52] that appropriately depicts the model's performance. The R^2 is another metric that deserves a role in DFAI. A recent comparison of regression analysis error measures is discussed in [54]. R^2 exhibit desirable features, including interpretability in terms of the data's information content and sufficient generality that span a relatively broad class of models [55]. Although a negative R^2 indicates a worse fit than the average line, this representation may be critical for determining how the learning model fits the dataset. Further on this, regardless of whether an examiner reports the R^2 score, or whether it helps to determine the performance of a regressor, it is a highly effective technique for evaluating the performance of a regression analysis and highly recommended for DFAI analysis.

2.1.3 Evaluating Clustering Algorithms in DFAI

Evaluating a clustering method can be challenging because it is mostly used in unsupervised learning [56, 57]; which means that no ground-truth labels are available. Clustering in a supervised (learning) [58] setting, on the other hand, can be evaluated using supervised learning metrics. One significant downside with unsupervised learning that fact-finders should be aware of is that applying clustering analysis to a dataset blindly would categorize the data into clusters (even if the data is random), as this is the algorithm's expected function. As a result, before deciding on a clustering approach, examiners must verify the non-random structure of the data. Three critical factors that should be considered in clustering are: (1) Clustering tendency; (2) Number of clusters, k; and (3) Clustering quality. We give a brief explanation of these factors below.

1. Clustering tendency: tests the spatial randomness of data by measuring the probability that a given dataset is generated by a uniform data distribution. If the data is sparsely random, clustering techniques may be meaningless. It is critical (especially in DFAI) for examiners to conduct this preliminary assessment, in part because it can assist reduce the amount of time required to analyze artifacts. A method for determining a dataset's cluster tendency is to utilize the Hopkins statistic [59], which is a type of sparse sampling test. The Hopkins statistic is used to test the null hypothesis (H_0) and the alternative hypothesis (H_a). the Hopkins statistic is close to 1 or $H > 0.5$, we can reject the null hypothesis and infer that there are significant clusters in the data.

2. Number of clusters: obtaining the ideal number, k, of clusters is critical in clustering analysis; while there is no definitive method for doing so, it can rely on the shape of the distribution, the size of the data set, and the examiner's preference. If k is set to a value that is too high, each data point has a chance of forming a cluster, whereas a value that is too low may result in inaccurate clusters. Additionally, the following approaches can help forensic examiners determine the cluster number:

- *Prior domain knowledge*—prior domain knowledge (based on experience on use case) can provide insight into the optimal number of clusters to choose.
- *Data driven approach*—employs mathematical methods to determine the correct value, such as rule of thumb method, elbow method [60, 61] and gap statistics [62].

3. Clustering quality: characterised by minimal intra-cluster distance and maximal inter-cluster distance.

To evaluate the performance of a clustering task, two validation statistics are key, namely: internal cluster validation and external cluster validation.

Internal cluster validation: evaluates a clustering structure's goodness without reference to external data. It frequently reflects the compactness, connectedness, and separation of the clusters. The **silhouette coefficient** (SC) [63, 64] and **Dunn index** (DI) [65] can be used to evaluate how well the algorithm performs in comparison to its internal clusters. By measuring the average distance between two observations, the SC determines how well they are clustered. SC has been applied in a variety of forensics-related clustering methodologies, including document forensics [173], image source identification [174, 175], and text forensics (e.g. authorship) [176, 177].

However, if computational cost is not an issue, the DI can be utilized. A practical application of DI in computer forensics is reported in [178], where it aids in the evaluation of ransomware sample similarity. There are further

indices (for example, the Davies-Bouldin index [66]); but, the silhouette and Dunn provide, in principle, the closest compatibility with DFAI in general, and specifically in terms of interpretability.

External cluster validation: compares and quantifies a cluster analysis' results against externally known benchmarks (e.g., externally provided gold standard labels). Such benchmarks are made up of a collection of pre-classified items, which are often created by (expert) humans. The evaluation approach quantifies the degree to which the clustering analysis result corresponds to predefined ground truth classes. To evaluate the performance of external cluster indices, the **Rand index** [67], the **Purity index** [68], the **F-measure** (with precision and recall; as indicated in the classification task), and the **Fowlkes-Mallows index** [69] can be utilized. As a matter of fact, it remains unclear how external cluster validation could improve DFAI. To elaborate on this fact, given the majority of digital artifacts from which evidence can be derived are sparse, unconventional, and previously unseen, having a ground truth label with which to compare may be impracticable. Moreover, given the majority of DF analysis are crime-specific (or relating to a particular case), the question is whether it is appropriate to compare crime-related data analysis to a general task ground truth labels. However, if gold standard, case-based labels are available, such as those for videos and photos in [70] or (though limited in scope and diversity) the "Computer Forensic Reference Dataset Portal CFReDS)[5]" or "Datasets for Cyber Forensics,[6]" then suitable comparisons can be established.

2.2 Forensic Evaluation

Upon the establishment of facts through a forensic investigation, decision-making follows, which is the adoption of a hypothesis as a conclusion [71]. While evaluation of forensic outcome is usually discussed in court contexts, review of forensic decisions is appropriate at all phases of the investigation [72]. It begins with evaluation of the individual hypothesis against all competing claims; the accuracy (including quantification of error rates) of the results obtained through automated tools used in the analysis; the extent to which experience and domain knowledge were helpful; and the ease with which the entire investigative process can be explained to a non-expert. Because automated systems are not self-contained and thus cannot take everything into account [15], it is possible that multiple DFAI approaches were used to find solutions to all competing hypotheses. As a result, forensic evaluation in this case will entail weighing the differing claims against the overall investigative problem. One way of determining this is to assign an evidential weight (strength of evidence) or "Likelihood Ratios" (LR) [73–75] to all contending claims. Although LR was originally created as a framework for evaluating forensic (science) evidence, the concept can be adopted to help make the DFAI's outcome more intelligible. Contrary to the factually deterministic requirements of evidence in a criminal or civil case, the majority of AI-based algorithms and their outputs are mostly probabilistic. However, forensic examiners do not pronounce judgments or issue final decisions; they rather provide expert testimony (or an opinion) or report of their findings to fact finders (attorneys, judges, etc.). Succinctly reporting forensic investigation findings remains a challenge [76], and while it may be comprehensible to state an opinion on a hypothesis and its alternatives as true (or false), such approach lacks the transparency and logical informativeness necessary to reach a verdict in a legal proceeding. Consequently, reporting DF findings in terms of weights or LRs enables the decision maker to assign the evidence an appropriate level of confidence [15]. LRs represent examiners' assessment of the relative probability of observed features under various hypotheses concerning a particular case. Furthermore, the European Network of Forensic Science Institutes (ENFSI) [75] recommends LR (simply in terms of numbers) even when examiners must make subjective decisions [75], because it makes the examiner's belief and inferential process explicit and transparent, facilitating the evaluation of strengths and weaknesses for those who rely on it [76]. While expressing subjective decision in terms of LRs has grown widespread in Europe, doubts have been raised in support of empirical data instead [73]. In other contexts, verbal expressions of LRs have been proposed; for example according to [73], consider an LR expression in the form: "*at least 1,000 times more likely*" and "*far more probable.*" The former is likely to receive scepticism regarding the basis for that figure, whereas the latter has a stronger possibility of acceptance [73].

Consequently, given the probabilistic (or stochastic) nature of the results of DFAI models, and the fact that these models have been empirically verified as accurate and well-suited for analytical purposes[7], as well as the inclusion of an "expert-in-the-middle[8]," it is still necessary to find the most appropriate way to report the results in the clearest and most understandable manner possible, albeit as recommendations. The recommended LR by the UK's Forensic Science Providers (AFSP) on "standard for the formulation of evaluative forensic science expert opinion" is available in [77].

[5] https://cfreds.nist.gov/

[6] https://datasets.fbreitinger.de/datasets/

[7] Via published studies, surveys, experiments, and peer review.

[8] Either by way of having human expert verify the results manually, or with a rule-based expert system.

Table 1 A proposed AI-adaptive C-Scale evaluation of strength of evidence for DFAI

C-Value	Accuracy Score (%)	False positive rate (%)	False negative rate (%)	Strength of support
C0	0–20	55–100	55–100	Erroneous (Incorrect)
C1	20–30	50–55	50–55	Extremely weak evidence
C2	30–40	40–50	40–50	Very weak evidence
C3	40–55	30–40	30–40	Weak evidence
C4	55–70	20–30	20–30	Strong evidence
C5	70–90	10–20	10–20	Very strong evidence
C6	90–100	0–10	0–10	Extremely strong evidence

However, in 2016, the US President's Council of Advisors on Science and Technology [78] recommended that forensic examiners reveal the error rates observed in black-box validation when reporting or testifying on forensic comparisons. Thus, error rates have become an intrinsic element of investigative outcome reporting, and with it, fact-finders have a greater logical and empirical understanding of the probative value of the examiner's conclusion [73]. It is not straightforward to express likelihood ratios in ways that are consistent with probabilistic distributions or error estimates (usually real values between 0 and 1). An approach was proposed in [79] which is based on the combination of prior probabilities and the likelihood ratio. However, when the conditional components of a hypothesis are transposed, evaluating its probability might be logically fallacious [72]. Probabilities are rarely acceptable in legal decisions, because an 80% probability is synonymous to the fact that one in five cases would be decided wrongly [80]. Given that probability is relative to certainty (or otherwise), we can align our DFAI evaluation intuition with the "Certainty Scale", or "Confidence Scale" (C-Scale) proposed in [72, 81, 82], which is reasonably appropriate for assigning strength of evidence to continuous values with respect to the hypothesis. As noted by [72]; *"...the strength of evidence does not exist in an abstract sense, and is not an inherent property of the evidence; it only exists when a forensic practitioner assigns value to the evidence in light of the hypothesis."* Therefore, in light of each working hypothesis resolved via DFAI, Table 1 represent a proposed C-Scale for expressing the strength of evidence that is compatible with DFAI analysis.

This is by no means a standard evaluation, but rather a tentative proposition that will need to be refined as research in this field progresses. Additionally, unlike the LR recommendation and the C-Scale proposals, which are based on hypothesis (or strength of hypothesis) about source identification during a forensic investigation, the DFAI C-scale evaluation method is fairly generic (for hypothesis and AI models) and applicable in a wide variety of situations, including strength of evidence. Furthermore, the *FP* and *FN* rating scales in Table 1 can be adjusted according to investigative tasks, as there are instances when a 50% to 60% false positive/negative rate would indicate "weak support".

As previously stated, human expert interpretation and evaluation are key components of DFAI in a partially automated setup because it is difficult to predetermine all of the reasoning required to do a forensic investigation work [15]. However, in a fully automated scenario, learning algorithms in conjunction with contextually structured expert systems can incorporate domain-specific knowledge-base rules. An expert system can also be built to evaluate every hypothesis at each modular level and make recommendations based on codified LRs.

3 Standardization in DFAI

The issue of standardization in digital forensics has persisted for several years; first because standard guidelines have been unable to keep up with the dynamic pace of technological sophistication, and second, because forensic stakeholders have been unable to agree on certain rules and standards, resulting in conflict of interest [83]. Additionally, the distinctiveness of investigation, the domain's diversity, and the existence of disparate legislative frameworks are all reasons cited as impediments to the standardization of the DF field [85, 86]. Nowadays, when it comes to standardization, the majority of what is available (in the form of guidelines) are check boxes; since the notion is that the more details, the better the standard [87]. Nonetheless, the "Forensic Science Regulator" in a 2016 guidance draft highlighted the validation of forensic methods as a standard, rather than the software tool [84]. This method validation entails a number of assessments, including the evaluation of data samples, which are relatively small in DF [88]. Standardization in DF (as well as DFAI) is a broad and intricate area of study, as every component of DF requires it. However, as part of the advancement of DFAI (for which further study is envisaged), we examine standardization within the context of forensic datasets and error rates.

3.1 DFAI Datasets

Datasets (or data samples) are a critical component of AI, as they define the validity of an AI model to a great extent. A dataset is a set of related, discrete elements that have varying meanings based on the context and are used in some type of experiment or analysis [89]. To evaluate or test novel approaches or to replicate existing procedures, similar data sets are required; for example, investigations on facial

recognition require human facial sample data. Similarly, an inquiry into message spamming necessitates the collection of e-mail samples. Datasets are often beneficial in the following ways, according to the National Institute of Standards and Technology (2019)[9]:

- *For training purposes:* dataset is generated for training purposes, i.e., simulation of case scenarios in order to train a model to learn the specifics of that environment, and to facilitate practitioner's training on case handling so that their ability to identify, examine, and interpret information can be assessed.
- *Tool validation:* wherein dataset is utilized to determine the completeness and correctness of a tool when it is deployed in a given scenario.
- *Familiarity with tool behavior:* for instance, a dataset collected from users' software interaction traces. As a result, such datasets are crucial for deciphering how certain software behaves on a device and for assisting in the interpretation of digital traces left by usage [86].

The process of creating a dataset is critical, even more so in the domain of DF, where each component must be verifiable, fit for purpose, and compliant with some set of standards. Therefore, the created dataset must be realistic and reliable [90]. This also entails having a high-quality, correctly labeled dataset that is identical to the real-world use case for testing and evaluation purposes, substantial enough for adequate learning, and is accessible to ensure reproducibility [89]. In the context of DFAI, there are a few considerations that must be made in order to conduct a forensically sound operation with respect to datasets.

Due to limited availability of datasets in DF, practitioners frequently overuse a single data corpus in developing several tools and methodologies, resulting in solutions gradually adapting to a dataset over time. For example, the Enron corpus has developed into a research treasure for a variety of forensic solutions, including e-mail classification [91–93], communication network analysis [19, 94], and other forensic linguistics works [95–97]. However, proving that a solution based on a single corpus is sufficiently generalizable to establish a conclusion in a forensic investigation will be difficult. Nevertheless, this is a widely recognized issue among stakeholders, and while it may be excusable in peer reviews, it is a major issue in the standardization of DF that requires immediate resolution. Similarly, while a workable DF dataset is constantly being sought, it is worth emphasizing that using a (single) dataset to assess the validity of a tool or method may not appropriately represent the general case scenario.

Datasets are created as a "mock-up" of a specific scenario, representing the activities/events that occur within an environment; supposedly within a specified time period. Each use case is time-dependent; as such, the continued relevance of a particular use case (from a previous period) in a future period may be debatable. This is particularly true in the domain of DF. For instance, given the advancements in computer network architecture, it may be illogical to use a dataset of network traffic from the 1990s to model an intrusion detection system today. This is also a point made in [98]. Similarly, it may seem counter-intuitive to argue that a model trained on images retrieved from an older (e.g., 2000) CCTV footage or camera is helpful for identifying objects in a contemporary crime scene image - technology has improved. However, in an ideal circumstance and for a robust model, updating the dataset with a collection of new features compatible with recent realities, rather than completely discarding the old dataset, should be viable.

Criminal cases such as hate speech [99] may involve local nuances [101], and while global dimension may not be impossible [100], investigations should take into account regional differences. For instance, in a typical forensic linguistics investigation [95–97] (e.g., cyberbullying [102]), a language corpus plays a vital role. However, native speakers' use of language (for example, English) may differ greatly from those of non-native speakers. Language, in usage and writing, varies across borders. An AI model trained to identify instances of bullying using a message corpus derived from British databases may not be completely representative of the same use case in Anglophone Africa – some English phrases are offensive to native speakers but inconsequential to non-natives. As such, a DFAI training dataset should accurately represent the use case (in terms of geographical location and dimensionality) for which application is intended.

Lastly, the demand for synthetically generated datasets is increasing in the DF domain, and rightly so. The issues of privacy, unavailability, and non-sharing policy continue to be a barrier to getting forensically viable datasets for the purpose of training, testing, and validating forensic tools. Synthetic data, first introduced in [103, 104], is described as an artificially generated data that contains statistical properties of the original data. While synthetic data can be extremely beneficial for research and education, the question is whether any novel technique can be tested on fictitious data [105], and particularly for DF; whether a perfect simulation of a crime event can be achieved. Nonetheless, several research (not related to DF) have demonstrated the usefulness of synthetic data in comparison to actual data [106, 107], in which a model was trained on synthetic data and tested on real data. The results indicated that the accuracy

[9] National Institute of Standards and Technology, 2019. The CFReDS Project. Available at https://www.cfreds.nist.gov/. (Accessed 20 June 2021).

of a variety of ML approaches were slightly decreased and varied when a synthetic dataset was used. Synthetic data can be used to augment or enhance an existing dataset, as well as to adjust for data imbalances created by an event's rarity. In DFAI, modeling with synthetic data is sometimes useful, but not always. Synthetic data generation requires a purpose-built dataset that may be too narrow for general-purpose solutions; demonstrating the results' applicability to real-world crime data may be difficult. This point is highlighted in [108], while some other challenges are emphasized in [109]. Furthermore, synthetic datasets are randomised, which means that the data do not follow a regular pattern. We foresee an extended challenge if the dataset is used to train an unsupervised neural network model – the model may learn non-interpretable patterns. While it is natural to assume that random data is less biased, there is no means to verify this claim. Thus, while synthetic datasets may be advantageous for solving specific ML problems, their usage in DFAI should be carefully considered.

3.2 DFAI Error Rates

As critical as accuracy is in determining the correctness of an evidence mining process, so also is the error rate. The error rate not only indicates the probability that a particular result is correct, or the strength of a technique, but also its limitations. According to the Scientific Working Group on Digital Evidence (SWGDE) [110], the term "error" does not allude to a mistake or blunder, but rather to the inevitable uncertainty inherent in scientific measurements. Numerous factors can influence these uncertainties, including algorithmic flaws, statistical probability, physical measurements, and human error [110]. One of the criteria for validating scientific methods under Daubert standard[10] is the assessment of error rate. Indeed, some of the other requirements (in the Daubert standard) are heavily weighted around error rate. For example, the Daubert standard requires the validation (or test) of a theory or methodology. The question is how can we validate a hypothesis and its alternatives, or a method, without determining the rate of uncertainty? Additionally, peer-review publishing of the method(s) used in forensic examination of digital artifacts is critical. Peer-review enables scientific validation of the technique and quantification of methodological uncertainties. This demonstrates the importance of publishing error rates for forensic methods alongside accuracy values. Thus, in contrast to conventional approaches to AI/ML methods that place a premium on accuracy (or precision), we propose that the results of DFAI algorithm include additional information regarding the method's errors and uncertainties. That way, the method's limitations are known in advance, allowing for an assessment of whether the method's outcomes are sufficiently (and scientifically) valid as evidence.

In alignment with the guidelines offered in [110], the uncertainty associated with any DFAI technique can be assessed in two ways: random and systematic [111]. Random uncertainties are related with the technique's algorithm and are commonly associated with measurements, whereas systematic uncertainties are typically associated with implementation and it occur in tools. DF tools represents implementation of a technique, and their functionality varies according to the task they seek to resolve. It is not uncommon for software to possibly contain intrinsic bugs [112], which are caused by logical flaws or incorrect instructions. For instance, an erroneous string search algorithm can cause a tool to report certain critical evidence incompletely. In this case, the tool will extract some relevant strings but will likely under-report their extent. Due to the fact that these flaws are not random, the tool frequently produces the same output when given the same input, which may be inadvertently deceptive to an examiner. Consequently, additional error mitigation methods may be required to detect and fix the error.

Due to the probabilistic nature of DFAI algorithms (the outcome of which may be random), the error rates are expressed in terms of false positive and false negative rates (which we discussed earlier). Depending on the percentages of these errors, and as long as adequate confidence in the algorithm's optimality exists, the error rates may only indicate the technique's limitations, not its true efficiency. It is critical to report and publish error rates for techniques in the DF domain [113], and this should be especially true for DFAI. This increases the technique's transparency and ensures that, in the event of method replication, the intended outcome is known. Additionally, disclosing error rates provides prospective researchers with a baseline understanding of the components that function efficiently, where improvements are anticipated, as well as prevent potential biases in interpretation. Mitigating this error may not be straightforward scientifically, as it is dependent on a variety of factors; however, algorithm optimization, sufficient datasets, accurate labelling (in supervised settings), and strong domain knowledge (for proper interpretations) are some of the ways to achieve a fairly reasonable success. Additional mitigating strategies for systematic errors include training, written procedures, documentation, peer-review, and testing [110].

[10] https://www.law.cornell.edu/wex/daubert_standard.

4 Methods for Optimizing DFAI Techniques

Developing an AI/ML model involves initializing and optimizing weight parameters via an optimization method until the objective function[11] tend towards a minimum value, or the accuracy approaches a maximum value [114]. In addition to learning in predictive models, optimization is necessary at several stages of the process, and it includes selecting: (1) the model's hyper-parameters (HPs) [115]; (2) the transformation techniques to apply to the model prior to modelling; and (3) the modelling pipeline to apply. This section will not explore the depth of optimization in AI, but will instead describe hyper-parameter optimization (HPO) [116] as a component in DFAI models.

Two parameters are critical in ML models: (1) the model parameters, which can be initialized and updated during the learning process; and (2) the HPs, which cannot be estimated directly from data learning and must be set prior to training a ML model – because they define the model's architecture [117]. Understanding which HP is required for a given task is critical in a variety of scenarios, ranging from experimental design to automated optimization processes. The traditional method, which is still used in research but requires knowledge of the ML algorithm's HP configurations, entails manually tuning the HP until the desired result is achieved [11]. This is ineffective in some cases, particularly for complex models with non-linear HP interactions [118]. Numerous circumstances may necessitate the use of HPO techniques [119]; we highlight few of them below, specifically focusing on DFAI tasks.

1. Conducting a digital forensic investigation requires an inordinate amount of time, and minimizing this time has been a primary focus of research in this domain for years. Similarly, machine-driven techniques can be time consuming, depending on the size of the dataset or the number of HPs. Applying AI techniques on already complicated forensic investigations almost always adds complexity. HPO can significantly reduce the amount of human effort required to tune these HPs, hence considerably shortening the entire forensic analysis time.
2. We have already highlighted the importance of performance in the context of DFAI procedures. ML methods require a range of HP settings to obtain optimal performance on a variety of datasets and problems. Numerous HPO techniques exist to assist in optimizing the performance of AI-based models by searching over different optimization spaces in quest of the global optimum for a given problem.
3. As previously stated, reproducibility is a necessary condition for a standard DF technique. HPO can assist in a variety of ways in achieving this goal. When evaluating the efficacy of several AI algorithms on a certain analysis, for example, adopting the same HP settings across all models establishes a fair comparison process. This can also be used to determine the optimal algorithm for a particular problem. Reporting these HP configurations can be advantageous in the event of DFAI model replication.

As with conventional AI models, when developing a DFAI model with HPO in mind, the process will include the following: an estimator (a classifier or regressor) with its objective function, a search (configuration) space, an optimization method for identifying suitable HP combinations, and an evaluation function for comparing the performance of various HP configurations [118]. A typical HP configuration can be continuous (e.g., multiple learning rate values), discrete (e.g., the number of clusters, k), binary (e.g., whether to use early stopping or not), or categorical (type of optimizer), all of which can be combined to produce an optimized model. Because the majority of ML algorithms have well-defined open-source frameworks (such as scikit learn[12]) that can assist in solving problems by tuning (changing values) some already pre-set HPs, we will focus on HPOs related to DL models because they require self/auto-tuning of un-set parameters. HP in DL are set and tuned according to the complexity of the dataset and the task, and they are proportional to the number of hidden layers and neurons in each layer [120]. The initial parameter setting for a DL model is to specify the loss function (binary cross-entropy [121], multi-class cross-entropy [122], or squared error loss) appropriate for the problem type. Then comes the type of activation function (e.g., ReLU [123], sigmoid,[13] etc.) that describes how the weighted sum of the input is transformed into the output. Finally, the optimizer type is specified, which may be stochastic gradient descent (SGD) [124], Adaptive Moment Estimation (Adam) [125], or root mean square propagation (RMSprop) [126]. In what follows, we describe several optimization techniques that can be vital to the optimization of a DFAI model.

[11] or loss function: is a function that maps an event or values of one or more variables onto a real number intuitively representing some cost associated with the event.

[12] https://scikit-learn.org/stable/index.html.

[13] See https://deepai.org/machine-learning-glossary-and-terms/sigmoid-function.

4.1 Methods for Hyper-Parameter Optimization in DFAI

A. Trial and error method

This method involves tuning parameters manually. It entails testing a large number of HP values based on experience, guesswork, or analysis of prior results. The approach is to improve parameter guesses iteratively until a satisfying result is obtained. This approach may be impractical for a variety of issues, particularly those involving DF analysis, that could involve large number of HP or complex models [118]. However, this technique can improve interpretability by allowing for the assessment of the model's various working parts as the parameters are tuned.

B. Grid search (GS) This is a frequently used tech nique for exploring the HP configuration space [127]. It does a parallel search of the configuration space and is suitable within a limited search space; otherwise, it may suffer from the "curse of dimensionality" [129]

When DF examiner has sufficient knowledge about the (finite) set of HP to specify [95] for the search space, GS is preferable. Because computational intensity is one of GS's drawbacks [128], its usage in DFAI is mostly focused on comparing the performances of many ML algorithms [169] in order to identify which one achieves the best performance on a certain forensic task. The authors in [130] described a botnet detection method using GS optimization techniques.

C. Randon search (RS)

RS was proposed in [131] as a way to circumvent GS's limitations. Unlike GS, however, RS randomly selects a predefined number of candidate samples between a specified upper and lower bound and trains them until the budget is exhausted or the target accuracy is reached. It does this by allocating resources to best-performing regions with parallelization [132].

Due to the simplicity with which RS parallelizes, it is an ideal choice for DFAI tasks involving convolutional networks (CNN) [133], such as multimedia forensics (e.g., sound and video), image forensics, and so on, in which (low-dimensional) features are mapped from one layer to the next. This method can be time and memory-intensive. To optimize the process, a batching strategy [135] is implemented that takes advantage of the batch size and learning rate to reduce training time without compromising performance. In this case, RS may be useful in terms of determining the ideal range of values for these parameters [134], as just the search space must be specified. Additionally, RS's use in optimizing multimedia forensics analysis suggests that it may be key for recurrent neural networks (RNN) [136], although RS has the disadvantage of not taking past results into account during evaluation [118]. As a result, using RS in recursive tasks such as event reconstruction in DFAI may result in less-than-optimal outcomes.

D. Gradient descent (GD)

The gradient descent [137] optimization computes the gradient of variables in order to determine the most promising path to the optimum. Gradient-based optimization techniques converge faster to the local minimum than the previously described techniques, but they are only applicable to continuous HP, such as the learning rate in NN [138], as other types of HP (e.g., categorical) lack gradient direction. The application of GD in DFAI approaches is almost ubiquitous, as it is used in virtually all DL models. It is one of the most straightforward optimization architectures to understand and interpret. However, the findings published in [172] proved the occurrence of "Catastrophic Forgetting" when gradient descent is used, particularly for reproduction. That is, when trained on a new task, ML models may forget what they learned on a previous task with only gradient descent. A combination with dropout [172] is recommended, however.

E. Bayesian Optimization (BO)

BO [139, 140] is an iterative algorithm that calculates future evaluation points based on the prior results. It is a typical model for all sorts of global optimization, with the goal of becoming less incorrect with more data [141]. BO identifies optimal HP combinations faster and it is applicable regardless of whether the objective function is stochastic, discrete, continuous, convex, or non-convex. Gaussian process (GP) [142], Sequential Model-based algorithm configuration (SMAC) [143], and Tree Parzen Estimator (TPE) [144] are an examples of common BO algorithms. BO is especially useful in tools like the Waikato Environment for Knowledge Analysis (WEKA) [145], an open-source tool with collections of ML and data processing algorithms. Numerous DF analyses methods [146–148] have been proposed or conducted using WEKA—leveraging its robust data mining capabilities and the possibility to choose from, or compare a variety of extensible, base learning algorithms for a specific forensic task. Selecting the right algorithm and HPs for optimal performance and accuracy in a WEKA-based DFAI analysis might be challenging. In this case, the excellent properties of BO can aid in choosing the optimal ML method and HP settings that minimizes analytical errors.

The works presented in [149] and [150] demonstrates how BO can be used (more precisely, with SMAC and TPE) as meta-learning to guide the choice of ML algorithms and HPO settings that outperform conventional selections on a classification task.

F. Multi-fidelity optimization (MFO)

MFO techniques are frequently used to overcome the time constraints limitations imposed by other HPO due to huge configuration space and datasets. MFO evaluates practical applications by combining low and high-fidelity measures [151]. In low-fidelity, a relatively small subset is evaluated at a low cost and with poor generalization performance; while in high-fidelity, a larger subset is

examined at a higher cost and with improved generalization performance [152].

MFO techniques include "bandit-based" [153] methods that allocates computational resources to the "best-arm" (most promising) HP settings. Successive halving (SHA) and Hyperband (HB) are the two most often used bandit-based algorithms [152, 154].

The application of MFO techniques to DFAI can be exemplified with transfer learning (TL) [155], which is the process by which previously stored knowledge is used to solve different but related problems. TL has been deployed in a variety of DFAI methods [156, 157], most notably on image forensics and detection problems using labeled samples. Thus, low or high fidelity optimization can be helpful for determining the optimal solution depending on the size of the stored knowledge (dataset), the investigative problem, and available computational resources. [158] describes an example of work on detecting (signature-based and unknown) malware-infected domains based on HTTPS traffic, using TL and optimized with Hyperband optimization. Additionally, a state-of-the-art HPO technique called Bayesian Optimization Hyperband (BOHB) [159], which combines BO and HB to maximize the benefits of both, is gaining attention, and it will be interesting to see how DF research employs this promising technique in the future.

G. Metaheuristic algorithms

Metaheuristic algorithms are a popular type of optimization technique that are primarily inspired by biological evolution and genetic mutations. They are capable of resolving problems that are not continuous, non-convex, or non-smooth [118]. Population-based optimization algorithms (POAs) [160] are an excellent example of metaheuristic algorithms since they update and evaluate each generation within a population until the global optimum is found. The two most frequently utilized types of POA are genetic algorithms (GA) [161] and particle swarm optimization (PSO) [162]. PSO, specifically, is an evolutionary algorithm that functions by allowing a group of particles (swarm) to traverse the search space in a semi-random fashion [116], while simultaneously discovering the optimal solution through information sharing across swarms.

Network forensics with DL is an ideal use for PSOs, as training such models can be time-consuming since it requires identifying complex patterns from large amounts of data. To detect network intrusion or attack, iterative reverse-engineered codes on parser and network traffic logs are required; this can be challenging for humans [163]. The work described in [163] shows the efficacy of PSO as a useful instrument to minimize/maximize an objective function, and to determine the optimal HPs (such as epochs, learning rate, and batch size) that contribute to the deep forensic model's AUC accuracy and the reduction in false alarm rate.

4.2 General Discussion on HPO in DFAI

It is worth emphasizing that the techniques discussed here are by no means exhaustive in terms of definition, components, and applicability. These few are chosen for their popularity and as a means of briefly discussing optimization techniques in the context of DFAI models. As such, in depth discussions about HPOs are available in [114, 118]. In general, depending on the size of the data, the complexity of the model (e.g., the number of hidden layers in a neural network (NN) [164–166] or the number of neighbours in a k–Nearest Neighbors (KNN) [167, 168]), and the available computational resources, an HP configuration may lengthen the time required to complete a task. Further along this line, in most cases, only a few HP have a substantial effect on the model's performance in ML methods [118]. As such, having many HP configurations exponentially increases the complexity of the search space. However, with DL, HPO techniques will require significant resources, particularly when dealing with large datasets. Considering all of these complexities, especially in the context of DFAI, where timeliness, transparency, and interpretability are critical, a well-chosen HPO technique should aid in rapid convergence and avoid random results. However, given that DF analysis are case-specific, often distinctive, with interpretability as a fundamental requirement, decomposing complexity should be a priority. Thus, unless forensic investigators have sufficient computing resources and a working knowledge of the parameter settings for various HPO techniques, they may choose to consider the default HP settings in major open-source ML libraries, or make use of a simple linear model with reduced complexity, where necessary. In case of a self-defined DNN model, basic HP settings and early stopping techniques can be considered. Finally, to summarize the various HPO algorithms mentioned thus far, table 2 compares these HPO algorithms and their respective strengths and drawbacks, as adapted from [118] but extended with additional inputs.

5 Conclusion and Future Works

In this paper, we addressed common misunderstandings about "AI Forensics" and "Digital Forensics AI" (DFAI). We presented the notion of AI Forensics as specified in the literature, while also providing a conceptual description of "Digital Forensics AI" as a generic term referring to all the components and instruments used in the application of AI in digital forensics. As a result, we examined techniques and methods for evaluating the effectiveness of classification and regression algorithms, as well as algorithms based on clustering that are employed in digital forensics

Table 2 The comparison of HPO techniques (n denote the number of HP values and k is the number of HP)

HPO Technique	Strengths	Drawbacks	Time Complexity
GS			$O(n^k)$
	* Simple	* Effective with categorical HP	
	* Independent (Parallelization)	* Time consuming	
	* Exhaustive use of the search space	* HP grows exponentially	
		* Possible overfitting	
RS			$O(n)$
	* Effective parallelization	* Less-effective with conditional HP	
	* Improvement over GS	* Ignores previous result during evaluation	
	* Better with low-dimensional data	* Potential for variance since it is random	
	* Reduce overfitting		
	* No HP tunning except for specifying search space		
GD			$O(n^k)$
	* Fast convergence speed for continuous HP such as learning rate	* Support only continuous HP	
		* Detects only a local optimum	
BO(BO-GP, SMAC, BO-TPE)			$0(n^k)$ (BO-GP), $O(nlogn)$ (SMAC, BO-TPE)
	* Fast convergence speed for continuous HP	* Poor parallelization capacity	
	* Effective with all types of HP (in SMAC and BO-TPE cases)	* Slow convergence with dimension > 1000	
	* Computes mean and variance	* Specification of prior is difficult	
HP			$0(nlogn)$
	* Better parallelization	* Less-effective with conditional HP	
		* Subset with small budget required	
BO-HP			$0(nlogn)$
	* Effective with all types of HP	* Subset with small budget required	
	* Better parallelization		
GA			$0(n^2)$
	* No initialization	* Poor parallelization capacity	
	* Effective with all types of HP	* Computational complexity	
	* Produces multiple optimal solutions		
	* Possible global optimal solution		
	* Large solution space		
	* Support multiple objective function		
PSO			$0(nlogn)$
	* Better parallelization	* Initialization required	
	* Effective with all types of HP	* Weak local optimum search space	
	* Efficient global search algorithm		
	* Insensitive to caling of design variables		

investigation. We focused on indicators that should not be disregarded while evaluating a predictive model's correctness. Additionally, we examined forensic (decision) evaluation and proposed an AI-adaptive confidence scale reporting system that takes into account the error rates associated with false positives and negatives in a forensic output. We laid great emphasis on the datasets and error rates of AI-based programs used in digital forensics when it comes to standardization.

Finally, we conducted a comparative review of the key optimization techniques used in machine learning models, focusing on their application (and suitability) for digital forensics. We summarized these techniques and their various strengths and drawbacks, as well as their corresponding time complexities. Additionally, we presented our opinion

on the usage of hyper-parameter optimization in AI-based DF analysis under discussion section.

As this is an attempt to formalize the concept of DFAI with all its prospective components, future work will strive to expand standardization beyond the two areas addressed thus far: datasets and error rates. Furthermore, the idea of expanding the methods for evaluating DFAI techniques to include comparative analysis of the various methods in practical settings appears to be a promising development for the domain, and it will be fascinating to see how it evolves in the future. Additionally, the explainability/interpretability and understandability of AI models employed in forensic investigation (and, more widely, in general) remains a concern. This is also a critical instrument of DFAI for which resources can be expanded; hence, our future work will look to broaden the research focus in this direction.

References

1. Hand DJ (2007) Principles of data mining. Drug Saf 30(7):621–622
2. Turin AM (1950) Computing machinery and intelligence. Mind 59(236):433
3. McCarthy J (2004) What is artificial intelligence? Available online at: https://kewd.pw/what_is_artificial_intelligence.pdf
4. Pasquale F (2015) The black box society: the secret algorithms that control money and information. Cambridge: Harvard Univ. Press, 320. ISBN 978-0674368279
5. Palmiotto F (2021) The Black Box on Trial: The Impact of Algorithmic Opacity on Fair Trial Rights in Criminal Proceedings. In Algorithmic Governance and Governance of Algorithms 49-70
6. Carbonell JG et al (1983) An overview of machine learning. Mach Learn 1:3–23
7. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspective, and prospects. Science 349(6245):255–260
8. Doowon J (2020) Artificial intelligence secuurity threat, crime, and forensics: taxonomy and open issues. IEEE Access 8:184560–184574
9. Behzadan V, Baggili IM (2020) Founding the domain of AI forensics. In SafeAI@ AAAI
10. Pollitt M, Casey E, Jaquet-Chiffelle D, Gladyshev P (2018) A framework for harmonizing forensic science practices and digital/multimedia evidence. In Organization of Scientific Area Committees for Forensic Science (OSAC)
11. Sunde N, Dror I (2019) Cognitive and Human Factors in Digital Forensics: Problems, Challenges, and the way Forward. Digit Investig 29:101–108
12. Ask K, Granhag PA (2005) Motivational Sources of Confirmation Bias in Criminal Investigations: The Need for Cognitive Closure. J Investig Psychol Offender Profil 2:43–63
13. Lau T, Biedermann A (2020) Assessing AI output in legal decision-making with nearest Neighbors. Penn State Law Rev 24(3). https://elibrary.law.psu.edu/pslr/vol124/iss3/1/
14. Graham J, Jones S, Booth G, Champod C, Evett I (2006) The Nature of Forensic Science Opinion - a Possible Framework to Guide Thinking and Practice in Investigation and in Court Proceedings. Sci Justice 46(1):33–44
15. Bollé T, Casey E, Jacquet M (2020) The role of evaluations in reaching decisions using automated systems supporting forensic analysis. Forensic Sci Int Digit Investig 34:301016
16. Islam M et al (2019) Toward detection of child exploitation material: a forensic approach. Deep Learn Appl Cyber Secur 221–246
17. Steel C et al (2022) Technical behaviour of child sexual exploitation material offenders. J Dig Forensic Secur Law 17(1):2
18. Anda F, David L, Kanta A, et al (2019) Improving the accuracy of automated facial age estimation to aid CSEM investigations. Dig Investig 28(S142)
19. Solanke A, Chen X, Ramiréz-Cruz Y (2021) Pattern recognition and reconstruction: detecting malicious deletions in textual communications. In: Proceedings of the IEEE Intl. Conf. on Big Data 2021:2574–2582
20. European Network of Forensic Science Institutes (ENFSI) (2015). Best Practice manual for the Forensic Examination of Digital Technology. ENFSI-BPM_FIT-01 (Version 01). Available online at: https://enfsi.eu/wp-content/uploads/2016/09/1._forensic_examination_of_digital_technology_0.pdf
21. Dai B, Ding S, Wahba G (2013) Multivariate Bernoulli distribution. Bernoulli 19(4):1465–1483
22. Gidudu A, Hulley G, Tshilidzi M (2007) Image classification using SVMs: one-against-one vs one-against-all. Asian Conf. on Remote Sensing 2007
23. Hong J, Cho S (2008) A probabilistic multi-class strategy of one-vs-rest support vector machines for cancer classification. Neurocomputing 71(16–18):3275–3281
24. Wu T, Lin C, Weng R (2004) Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res 5:975–1005
25. Tsoumakas G, Katakis I (2007) Multi-label classification: an overview. Int J Data Warehouse Mining (IJDWM) 3(3):1–13
26. Tang Y, Zhang Y, Chawla N, Krasser S (2008) SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern Part B (Cybern) 39(1):281–288
27. Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some english consonants. J Acoust Soc Am 27(2):617
28. Townsend JT (1971) Theoretical analysis of an alphabetic confusion matrix. Percept Psychophys 9(1):40–50
29. Zhu M (2004) Recall. University of Waterloo, Precision and Average Precision. Dept. of Statistics & Actuarial Science
30. Bradley A (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
31. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36
32. Zhang E, Zhang Y (2009) Average precision. Encyclopedia of Database Systems

33. Powers DM (2014) What the f-measure doesn't measure. Technical report, Beijing University of Technology, China & Flinders University, Australia Tech. Report
34. Hand D, Christen P (2018) A note on using the F-measure for evaluating record linkgae algorithms. Stat Comput 28(2):539–547
35. Cook D, Weisberg S (1982) Criticism and Influence Analysis in Regression. Sociol Methodol 13:313-361. Jstor
36. Freedman D (2009) Statistical Models: Theory and Practice (2nd Ed.). Cambridge University Press. ISBN: 9780521743853
37. Khoury R, Harder DW (2016) Interpolation, regression, and extrapolation. In: Numerical methods and modelling for engineering 77–113
38. Toro-Vizcarrondo C, Wallace T (1968) A test of the mean square error criterion for restrictions in linear regression. J Am Stat Assoc 63(322):558–572
39. Allen D (1971) Mean square error of prediction as a criterion for selecting variables. Technometrics 13(3):469–475
40. Sammut C, Webb G (2010) Mean absolute error. Encyclopedia of Machine Learning, 652
41. Nevitt J, Hancock G (2000) Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. J Exp Educ 68(3):251–268
42. De Myttenaere A, Golden B, Le Grand B, Rossi F (2016) Mean absolute percentage error for regression models. Neurocomputing 192:38–48
43. Lehmann E, Casella G (1998) Theory of point estimation (2nd ed.). New York: Springer. ISBN 978-0-387-98502-2
44. Willmott C, Matsuura K, Robeson S (2009) Ambiguities inherent in sums-of-squares-based error statistics. Atmos Environ 43(3):749–752
45. Hyndman R, Koehler A (2006) Another Look at Measures of Forecast Accuracy. Int J Forecast 22(4):679–688
46. Ren L, Glasure Y (2009) Applicability of the revised mean absolute percentage errors (mape) approach to some popular normal and non-normal independent time series. Int Adv Econ Res 15(4):409–420
47. Garofalakis M, Kumar A (2004) Deterministic wavelet thresholding for maximum-error metrics. In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems pp. 166–176
48. Wright S (1921) Correlation and causation
49. Barrett G (2000) The coefficient of determination: understanding r squared and R squared. Math Teach 93(3):230–234
50. Di Bucchianico A (2008) Coefficient of determination (R2). Encyclopedia of Statistics in Quality and Reliability 1
51. Armstrong J, Collopy F (1992) Error measures for generalizing about forecasting methods: Empirical comparisons. Int J Forecast 8(1):69–80
52. Willmott C, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root man square error (RMSE) in assessing average model performance. Clim Res 30(1):79–82
53. Chai T, Draxler R (2014) Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguments against Avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250
54. Chicco D, Warrens M, Jurman G (2021) The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci 7:e623
55. Cameron C, Windmeijer F (1997) A R-Squared Measure of Goodness of Fit for some Common Nonlinear Regression Models. J Econom 77:329–342
56. Barlow HB (1989) Unsupervised learning. Neural Comput 1(3):295–311
57. Hinton G, Sejnowski TJ (Eds.) (1999) Unsupervised learning: foundations of neural computation. MIT Press
58. Caruana R, Niculecu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In proceedings of the 23rd international conference on Machine learning 161–168
59. Hopkins B, Skellam J (1954) A new method for determining the type of distribution of plant individuals. Ann Bot 18(2):213–227
60. Ng A (2012) Clustering with the k-means algorithm. Mach Learn
61. Kodinariya T, Makwana P (2013) Review on determining number of Cluster in K-Means Clustering. Int J 1(6):90–95
62. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Stat Methodol) 63(2):411–423
63. Rousseeuw P (1987) Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis Computational and Applied Mathematics. 20:53–65
64. Aranganayagi S, Thangavel K (2007) Clustering categorical data using silhouette coefficient as a relocating measure. In: International conference on computational intelligence and multimedia applications (ICCIMA 2007) 2:13–17
65. Dunn J (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern 3(3):32–57
66. Davies D, Bouldin D (1979) A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell PAMI-1 (2):224-227
67. Rand W (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66(336):846–850
68. Manning C, Raghavan P, Schütze H (2008) Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5
69. Fowlkes E, Mallows C (1983) A method for comparing two hierarchical clusterings. J Am Stat Assoc 78(383):553–569
70. Ferreira S, Antunes M, Correia M (2021) A dataset of photos and videos for digital forensics analysis using machine learning processing. Data 6(8):87
71. Lau T, Biedermann A (2020) Assessing AI output in legal decision-making with nearest Neighbors. Penn State Law Rev. 24(3). https://elibrary.law.psu.edu/pslr/vol124/iss3/1/
72. Casey E (2020) Standardization of forming and expressing preliminary evaluative opinion on digital evidence. Forensic Sci Int Digit Investig 32:200888
73. Berger C, Buckleton J, Chmapod C, Evett I, Jackson G (2011) Evidence evaluation: a response to the court of appeal judgement in R v T. Sci Justice 51(2):43–9
74. Kerkhoff W, Stoel R, Mattijssen E, Hermsen R (2013) The likelihood ratio approach in cartridge case and bullet comparison. AFTE J 45(3):284–289
75. European Network of Forensic Science Institutes (ENFSI) (2015) Guideline for evaluative reporting in forensic science: strengthening the evaluation of forensic results across Europe. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf
76. Thompson W (2017) How should forensic scientists present source conclusions. Seton Hall Law Rev. 48:773
77. Association of Forensic Science Providers (AFSP) (2009) Standards for the formulation of evaluative forensic science expert opinion. Sci Justice 49(3):161–4
78. President's Council of Advisors on Science And Technology (PCAST) (2016). Forensic science in crimi-nal courts: ensuring scientific validity of feature-comparison methods. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf
79. Morrison G (2011) Measuring the validity and reliability of forensic likelihood-ratio systems. Sci Justice 51:91–98
80. Atkinson K, Bench-Capon T, Bollegala D (2020) Explanation in AI and Law: Past, Present and Future. Artif Intell 289:103387
81. Casey E (2002) Error, uncertainty and loss in digital evidence. Int J Digit Evidence 1(2)

82. Casey E (2011) Digital evidence and computer crime 3rd Edition. Forensic Science, computers, and the Internet. Academic Press. eBook ISBN: 9780080921488
83. Bennett D (2012) The challenges facing computer forensics investigators in obtaining information from mobile devices for use in criminal investigations. Information Security Journal: A Global Per-spective v21(3):159–168
84. Forensic Science Regulator (2016) Draft guidance: digital forensics method validation. Crown Pros-ecution Service
85. Palmer G (2001) A road map for digital forensic research. in Proceedings of the 1st Digital Forensic Research Workshop, Utica, NY. 27-30
86. Reith M, Carr C, Gunsch G (2002) An examination of digital forensic models. Int J Digit Evid 1(3):1–12
87. Sommer P (2018) Accrediting digital forensics: what are the choices? Digit Investig 25:116–120
88. Arshad H, Aman B, Abiodun O (2018) Digital forensics: review of issues in scientific validation of digital evidence. J Inf Process Syst 14(2)
89. Grajeda C, Breitinger F, Baggili I (2017) Availability of datasets for digital forensics: and what is missing. Digit Investig 22:S94–S105
90. Gobel T, Schafer T, Hachenberger J, Turr J, Harald B (2020) A Novel approach for generating synthetic datasets for digital forensics. Adv Digit Forensic XVI, IFIP ACT 589:73–9
91. Miyamoto D, Hazeyama H, Kadobayashi Y (2008) Detecting methods of virus email based on mail header and encoding anomaly. In: Advances in Neuro-Information Processing
92. Guo H, Jin B, Qian W (2013) Analysis of Email Header for Forensics Purpose. In Proceedings of the International Conference on Communication Systems and Network technologies
93. Morovati K, Kadam S (2019) Detection of phishing emails with email forensic analysis and machine learning techniques. International Journal of Cyber-Security and Digital Forensics (IJCSDF) 8(2):98-107, 2019
94. Diesner J, Carley KM (2005) Exploration of communication networks from the Enron email corpus. SIAM Intl. Conf. on Data Mining, Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA, pp 3–14
95. Farkhund I, Rachid H, Benjamin F, Mourad D (2008) A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digit Investig 5:42–51
96. Bogawar P, Bhoyar K (2016) A novel approach for the identification of writing traits on email database. In 2016 1st India International Conference on Information Processing (IICIP) 1-6
97. Emad A et al (2019) Simplified features for email authorship identification. Int J Secure Network 8(2):72–81
98. McHugh J (2001) Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Trans Inf Syst Secur 3(4):262–294
99. Han EH (2006) Hate crimes and hate speech. Geol J Gender L 7:679
100. Nikoloska S, Liman X (2019) Criminal investigation of hate speech through a computer system in the Republic of North Macedonia. Facult Secur 63
101. Agustina A et al (2020) Light and shadow in hate-speech cases: a forensic linguistics. Advances in Social Science, Education and Humanities Research
102. Riadi I, Widiandana P (2019) Mobile Forensics for Cyberbullying Detection using Term Frequency-Inverse Document Frequency (TF-IDF) 5:68–76
103. Rubin D (1993) Statistical disclosure limitation. J Off Stat 9(2):461–468
104. Little R (1993) Statistical analysis of masked data. J Off Stat 9(2):407
105. Baggili I, Breitinger F (2015) Data sources for advancing cyber forensics: what the social world has to offer. in Proceedings of the 2015 AAAI Spring Symposium Series, Palo Alto, CA
106. Heyburn R et al (2018) Machine learning using synthetic and real data: similarity of evaluation metrics for different healthcare datasets and for different algorithms. In: Proceedings of the 13th International FLINS Conference, pp 1281–1291
107. Rankin D et al (2020) Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inf 8(7):e18910
108. Yannikos Y et al (2014) Data corpora for digital forensics education and research. In: IFIP International conference on digital forensics, pp 309–325
109. Horsman G, Lyle J (2021) Dataset construction challenges for digital forensics. Forensic Sci Int Digit Investig 38:301264
110. Scientific Working Group on Digital Evidence (SWGDE) (2018). Establishing confidence in digital and multimedia evidence forensics results by error mitigation analysis (Version 2.0)
111. Hughes I, Hase T (2010) Measurements and their uncertainties: a practical guide to modern error analysis. OUP Oxford
112. Walker IR (2011) Reliability in scientific research: improving the dependability of measurements, calculations, equipment, and software. Cambridge University Press, Cambridge
113. Henry F et al (2003) The impact of daubert on forensic science. Pepp Law Rev 31:323
114. Sun S, Cao Z, Zhu H, Zhao J (2020) A survey of optimization methods from a machine learning perspective. IEEE Trans Cybern 50(8):3668–3681
115. Probst P et al (2019) Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 20(1):1934–1965
116. Steinholtz O (2018) A comparative study of black-box optimization algorithms for tuning of hyper-parameters in deep neural networks. M.S. thesis, Dept. Elect. Eng., Luleå Univ. Technology
117. Kuhn M, Kjell J (2013) Applied predictive modelling. Springer, ISBN: 9781461468493
118. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415:295–316
119. Hutter F, Kotthoff L, Vanschoren J (eds) (2019) Automatic Machine Learning: Methods, Systems. Springer, Challenges. ISBN 9783030053185
120. Koutsoukas A, Monaghan K, Li X, Huan J (2017) Deep-learning: investigating deep neural net-works hyper-parameters and comparison of performance to shallow methods for modelling bioactivity data. J Cheminf 9(42):1–13
121. Ramos D et al (2018) Deconstructing cross-entropy for probabilistic binary classifiers. Entropy 20(3):208
122. Aly M (2005) Survey on multiclass classification methods. Neural Netw 19(2):1–9
123. Nair V, Hinton GE (2010) Rectified Linear Units Improve Restricted Boltzman Machines. ICML pp 807–814
124. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press. ISBN: 9780262035613
125. Kingma DP, Ba LJ (2015) Adam: a method for stochastic optimization. ICLR
126. Tieleman T, Hinton G (2012) Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4(2):26-31
127. Injadat M et al (2020) Systematic ensemble model selection ap-proach for educational data mining. Knowl-Based Syst 200:105992

128. Lorenzo P et al (2017) Particle swarm optimization for hy-per-parameter selection in deep neural networks. Proceeding of the Genetic and Evolutionary Com-putation Conference 481-488
129. Bach F (2017) Breaking the curse of dimensionality with convex neural networks. The Journal of Machine Learning Research 18(1):629–681
130. Gonzalez-Cuautle D et al (2019) An Efficient Botnet Detection Methodology using Hyper-parameter Optimization Through Grid-Search Techniques. IWBF
131. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(1):281–305
132. Krivulin N, Dennis G, Charles H (2005) Parallel implementation of a random search procedure: an experimental study. In 5th WSEAS International Conference on Simulation, Modeling and Optimiza-tion (SMO'05)
133. Albawi S et al (2017) Understanding of a convolutional neural network. Intl. Conf. on Engineering and Technology (ICET) 1-6
134. Ari N, Heru S (2020) Hyper-Parameter Tuning based on Random Search for DenseNet Optimization. Intl. Conf. on Inf. Tech., Computer, & Elect. Eng. (ICITACEE)
135. Pavlo R (2017) Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets. Info. Tech. & Mgt. Science 20(1):20–24
136. Mikolov T et al (2010) Recurrent neural network based language model. Interspeech 2(3):1045–1048
137. Bengio Y (2000) Gradient-based optimization of hyperparameters. Neural Comput 12(8):1889–1900
138. Maclaurin D, Duvenaud D, Adams R (2015) Gradient-based Hyperparameter Optimization through Reversible Learning. Intl. Conf. on Machine Learning 37:2113–2122
139. Jones D, Schonlau M, Welch W (1998) Efficient global optimization of expensive black-box functions. J Global Optim 13(4):455–492
140. Snoek J, Larochelle H, Adams R (2012) Practical Bayesian optimization of machine learning algorithms. Neural Information Processing Systems 2:2951–2959
141. Koehrsen W (2018) Comparison of activation functions for deep neural networks. https: //towardsdatascience.com/bayes-rule-applied-75965e4482ff
142. Seeger M (2004) Gaussian processes for machine learning. International Journal Neural System 14(2):69–106
143. Hutter F, Hoos H, Leyton-Brown K (2011) Sequential model-based optimization for general algorithm configuration. Proc. LION 5:507–523
144. Bergstra J et al (2011) Algorithms for hyper-parameter optimization. Proceedings of the Neural Information Processing Systems 2546-2554
145. Hall M et al (2009) The WEKA data mining software: an update. ACM SIGKDD Exploration Newsletter 11(1):10–18
146. Bhatt VH (2010) A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application. IACSIT 2(3):313–319
147. Nirkhi SM et al (2012) Data Mining: A Prospective Approach for Digital Forensics. IJDKP 2(6):41–48
148. Maheswari UK, Bushra NS (2021) Machine learning forensics to gauge the likelihood of fraud in emails. Intl. Conf. on Comm. & Elect. Systems, IEEE
149. Thornton C et al (2013) Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. ACM SIGKDD 847-855
150. Kunang YN et al (2020) Improving Classification Attacks in IOT Intrusion Detection System using Bayesian Hyperparameter Optimization. ISRITI 146-151. IEEE
151. Zhang S et al (2016) A new optimal sampling rule for multi-fidelity optimization via ordinal transformation. IEEE International Conference om Automation Science and Engineering 670-674
152. Jamieson K, Talwalkar A (2015) Non-stochastic best arm identification and hyperparameter optimization. In Artificial Intelligence and Statistics 240-248
153. Karnin Z et al (2013) (2013) Almost optimal exploration in multi-armed bandits. Int. Conf. Mach. Learn. ICML 28:2275–2283
154. Li L et al (2017) Hyperband: a novel bandit-based approach to hyper-parameter optimization. J Mach Learn Res 18(1):1–52
155. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
156. Zhan et al (2017) Image forensics based on transfer learning and convolutional neural network. ACM Workshop on Information Hiding and Multimedia Security 165-170
157. Al Banna MH et al (2019) Camera model identification using deep CNN and transfer learning approach. ICREST (pp. 626-630). IEEE
158. Prasse P et al (2019) Joint detection of malicious domains and infected clients. Mach Learn 108(8):1353–1368
159. Falkner S, Klein A, Hutter F (2018) BOHB: robust and efficient hyperparameter optimization at scale. Int. Conf. Mach. Learn. ICML 2018(4):2323–2341
160. Eggensperger K et al (2013) Towards an empirical foundation for assessing Bayesian optimization of hyperparameters. NIPS Workshop on Bayesian Optimization in Theory and Practice Work 1-5
161. Shapiro J (2001) Genetic Algorithms in Machine Learning, Paliouras G. et al. (Eds.): ACAI'99. LNAI (2049):146-168
162. Shi Y, Eberhart R (1998) Parameter Selection in Particle Swarm Optimization. Evolutionary Programming VII 591-600
163. Koroniotis N et al (2020) A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. Futur Gener Comput Syst 110:91–106
164. Jure Z (1994) Introduction to artificial neural network (ANN) methods: what they are and how to use them. Acta Chim Slov 41:327–327
165. Wang SC (2003) Artificial neural network. interdisciplinary computing in java programming. Springer, Boston, MA, pp 81–100
166. Dongare AD et al (2012) Introduction to Artificial Neural Network. International Journal of Engineering and Innovative Technology (IJEIT) 2(1):189–194
167. Fix E, Hodges JL (1951) Discriminatory Analysis. Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas, Nonparametric Discrimination
168. Altman NS (1992) An introduction to kernel and nearest-neigbor nonparametric regression. Am Stat 46(3):175–185
169. Rami MA, Mohammed A (2019) A comparison of machine learning techniques for file system forensic analysis. Journal of Information Security and Application 46:53–61
170. LeCun Y et al (2015) Deep Learning. In. Nature 521(7553):436–444
171. Goodfellow IJ et al (2014) An empirical investigation of catastrophic forgetting in gradient-based neural networks. In proceedings ofthe ICLR
172. Dahl GE et al (2013) Improving deep neural networks for LVCSR using rectified linear units and dropout. In 2013 IEEE intl. conf. on acoustics, speech and signal processing 8609-8613
173. Nassif LF, Hruschka ER (2013) Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. IEEE Trans Inf Forensics Secur 8(1):46–54
174. Villalba LJ et al (2015) Smartphone image clustering. Expert Syst Appl 42(4):1927–1940

175. Li C, Lin X (2017) A fast source-oriented image clustering method for digital forensics. EURASIP Journal on Image and Video Processing 2017(1):1–16
176. Almaslukh B (2019) Forensic Analysis using Text Clustering in the Age of Large Volume Data: A Review. Intl. Journal of Advanced Computer Science and Application 10(6):71-77
177. Layton R et al (2012) Evaluating authorship distance methods using the positive Silhouette coefficient. Nat Lang Eng 9(4):517–535
178. Naik N et al (2019) A Ransomware Detection Method Using Fuzzy Hashing for Mitigating the Risk of Occlusion of Information Systems. Itl. Symposium on Systems Engineering (ISSE) 1-6