Interpretable Machine Learning Fusion and Data Analytics Models for Anomaly Detection

Ahmed Abdelmonem[*1], **Nehal N. Mostafa** [2]

[1,2] Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt
Emails: aabdelmounem@zu.edu.eg; nihal.nabil@fci.zu.edu.eg

Abstract

Explainable artificial intelligence received great research attention in the past few years during the widespread of Black-Box techniques in sensitive fields such as medical care, self-driving cars, etc. Artificial intelligence needs explainable methods to discover model biases. Explainable artificial intelligence will lead to obtaining fairness and Transparency in the model. Making artificial intelligence models explainable and interpretable is challenging when implementing black-box models. Because of the inherent limitations of collecting data in its raw form, data fusion has become a popular method for dealing with such data and acquiring more trustworthy, helpful, and precise insights. Compared to other, more traditional-based data fusion methods, machine learning's capacity to automatically learn from experience with nonexplicit programming significantly improves fusion's computational and predictive power. This paper comprehensively studies the most explainable artificial intelligent methods based on anomaly detection. We proposed the required criteria of the transparency model to measure the data fusion analytics techniques. Also, define the different used evaluation metrics in explainable artificial intelligence. We provide some applications for explainable artificial intelligence. We provide a case study of anomaly detection with the fusion of machine learning. Finally, we discuss the key challenges and future directions in explainable artificial intelligence.

Keywords: artificial intelligence; Black-Box; machine learning; explainable artificial intelligence; Information fusion; intelligent methods; data fusion; anomaly detection;

1. Introduction

ML today appears in everyday technology use, which increases the contributions and studies in the research community toward Deep learning (DL) models and its significant performance and high results achievement in various fields such as medical diagnosis, economics, Network/IoT, and adversarial attacks. Most AI model is built as Black-box, White-Box, or grey-Box models. Where the Black-box model is the higher accuracy prediction, but these predictions cannot be interpretable based on non-linear functions. In contrast to White-box models that can be easy to interpret and simple in computation. In the grey-Box model, some partial knowledge can be known [1].

DL model is known as Black-Box, where predictions cannot be explained or interpreted. These increase the consideration of the model transparency and its ability to achieve unfairness. An example of the ML unfairness model is the COMPASS system that uses in most US justice community which determine more people who deserve parole. This system is based on giving a score to jailed people to determine if they are allowed parole or

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

not. Any error in the COMPASS score leads to extra years in jail. Many cases are denied parole for no clear reason until exploring the system found that it is biased toward race and age [2].

Another accident of ML model, during California, fires 2018 google daily air quality index that determines if it is safe to go outside or not indicates that it is fine to air quality and it is safe to go outside. Another biased ML model found that some hospital ML system recommendations for people who deserve extra medical help depend on race [3].

Utilizing Black-Box systems in multiple applications with the ambiguity on how it works and the inability to trace results and reasons for prediction will lead to hampering the trust of the user toward these systems. Obfuscated biases systems will harm some people and some groups, which will achieve unfairness and untrustiness in model transparency [4].

During appearing of these problems of the ML system of unfairness and algorithm biases results. The research community released that accuracy is not more important than making an ML system capable of explaining and interpreting its results. So many studies apply explainable methods to explain the results of their models in different fields. Explainable artificial intelligence (XAI) attracts many researchers to explain biases in many systems and models and in some ML methods. Also, building inherently interpretable models will lead to the use of an explainable white-box model instead of building a black-box model. Many fields integrated with XAI in the last years. So we conclude and review these studies in different domains and fields [5, 6].

Our study aims to investigate into aims to provide the main criteria that define the transparency model and the motivations behind building a transparency model. Also, we provide the most recent XAI methods with different categories and taxonomies supported with mathematical proof and drawbacks in each method. In an application such as anomaly detection, network/IoT, and adversarial attack, we introduce the most recent work in XAI in these applications.

The initial raw data are often unsure, imprecise, inconsistent, contradicting, and similar; data fusion is a method that fuses data to generate more consistent, useful, and correct information. Data fusion techniques come in a wide variety, each tailored to a certain set of use cases. Among the many applications of data, fusion is sensor networks, computer graphics, radars, target tracking, target rapid detection, penetration testing, scenario evaluation, and so on.

Bayesian fusion (e.g., Bayes fusion), evidentiary belief argumentation fusion (e.g., Vapnik theory), trying to arrange fusion, etc., are all classic approaches to data fusion. A new window of opportunity for progress in data fusion has opened up in recent years thanks to advancements in sensors, specialized tools, and numerous other data processing techniques. With its powerful computing and classification capabilities, machine learning is widely believed to be the key to dramatically enhancing the efficiency of data fusion methods.

As a crucial aspect of network security, anomalous traffic detection has risen to the forefront of academic and professional interest. In this context, "anomalous behavior detection" refers to the process of monitoring a system for any out-of-the-ordinary. Conventional detection strategies keep tabs on network activity in real time by combining data gathering, processing, and visualization. Existing systems for intrusion detection (IDSs) and IPSs are good complements to traffic monitoring, yet, there are issues of excessive false alarm rates or leak rates that prohibit them from being perfect.

The following paper structure is divided as follows. Section 2 provides the criteria that distinguish the transparency model and why Transparency is needed in the ML model. In section 3, we review different classifications of XAI with the most used method in XAI. In section 4, we provide the recent and few studies in anomaly detection, Network/IoT, and adversarial attacks. In Section 5, most challenges and future directions have been discussed. Section 6 provides the conclusions of this paper.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

2. Transparency in AI

Most ML systems are becoming Black-Box applications today, such as anomaly detection, elections, criminal justice, banking, medical decision-making, etc. Transparency of these secret models is essential as it found that these models are unfair in some situations, which may lead to trust issues in machine models to issues in justice fairness. In this section, we provide the important criteria that must exist in any transparent model and the reasons for using Transparency.

2.1 Transparency criteria

In this section, we aim to produce the important criteria of a transparent model:

Original/transformed data: The important point that differentiates whether the model is transparent or not is if the data is original or transformed. The transformed data is the process of converting the values, structure, or format of the original data. Usually, transformed data enhance the quality of data and protect it from null values, incomplete data, etc. also, and transformed sensitive data protect data during the adversarial attack that targets DL models. Model-sensitive value transformation guarantees the Transparency of the model, which raises the trustworthiness of the DL models.

Human-in-the-loop (HITL)/human-out-of-the-loop: HITL is known as a model that needs human interaction in both the training and testing phases. In this type of model, human cancel the ML decisions when it is wrong, which raise the learning of model confidence and accuracy of the decisions. In contrast, the HOTL is a fully automated AI system that is inherently controlled without the need for humans. HITL, in some cases, can allow for expert data manipulation or access only needed data which will lead to unfairness and untrustworthy. So, human judgment models do not achieve Transparency, unlike the HOTL model.

Knows about the model: Any known data about the model such as architecture, parameters, and inputs/outputs can cause a white-box adversarial attack. So, it is important for the model not to be known for all to achieve Transparency.

Model Security and privacy considerations: protecting data and model security associated with decision-making will enable Transparency to build trust. Adversarial attacks can easily control the DL model, which may control the decision of the model in sensitive cases such as in self-driving cars, which leads to insecure and untrust clients who use this service. So, to obtain model transparency must guarantee the security and privacy of sensitive data.

2.2 Reason of Transparency

Find "Clever Hans" decision based

The Clever Hans effect is referred to by scientists to be the risk of an unintentional hint of the required behavior by the questioner if experiments are not carefully designed. One ML example of this type of model, Lapuschkin et al. [7], found that PASCAL VOC solving methods use data correlations to achieve true classification. Also, the author shows the final fine-tuning stage in the DL model containing padding artifacts that have to influence the classifier output.

Boost confidence and Verifiability

Trust in AI applications is considered to be a significant increase in AI utilization in many critical areas of decision making such as medical diagnosis or criminal justice and self-driving cars. So, XAI helps to build trust between human-to-human interactions and between human-to-machine interactions, which can build confidence in future ML applications [4].

Transparency is essential for new ideas.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

Earlier, linear models were interpretable, and the extracted feature was understandable. The Recent DL methods are obtaining high accuracies without knowing the feature that the model is used. The necessity for strong interpretation is essentially more than predication itself which will lead to strong control. The non-linear models are needed without loss of interoperability [4].

To achieve LegislationLegislation

General Data Protection Regulation (GDPR) in Europe has discussed the justice aspects in the ML context. Also, emphasize the value of human-understandable for ML decisions. For example, if anyone has a loan refused by a bank using an AI system, has the right to know the reasons for the rejection. Although this considers a new challenge to LegislationLegislation, Transparency in the ML model will lead to achieving fairness of justices [4].

3. Explainable artificial intelligent methods (XAI)

Traditional methods or DL methods:

The traditional model can easily explain the model prediction. Also, it can deal with a small amount of data using standard hardware. Examples are SVM, Decision tree, and KNN. DL methods: difficult to interpret the predictions and deal with a large amount of data using GPU. For Example, CNN, RNN, and AE [8].

Discriminative methods or generative methods:

Discriminative methods can model a decision boundary between the classes, where it aims to estimate $P(A|B)$, such as SVM and CNN. On the other hand, the generative methods learn the actual distribution of each class which aims to estimate $P(A|B)$ to deduce $P(B|A)$ [8], such as naïve baye, DBN.

Global explanation or Local explanation:

The global aims to interpret the whole model individually from any input but cannot interpret a single instance. So local explanation aims to interpret only one prediction, which can deal with linear functions.

Post-hoc Vs. intrinsic:

The explanation may be during the prediction process, which is named "intrinsic," or needs post-processing generation to explain predictions that are named "post-Hoc ."The intrinsic methods aim to create white-box explainable models, and post-hoc methods aim to explain black-box models [5],[6].

Specific explanation or agnostic explanation:

Specific deals only with a certain type of algorithm. In contrast, model agnostic can be employed in any type of algorithm. Model-specific is can only explain models such as logistic regression, decision tree, and linear regression. But model agnostic has the flexibility to generate methods that can explain any type of model [5].

Data drove vs. Knowledge domain:

Data-driven domain refers to interpretations that compare the new data with only data inputs. On the other hand, the knowledge can be used to improve explanations and define why some features are more important than others.

Data type:

Explanation methods can be explained on many data types such as text, image, tabular data, graph, and video [5].

The aim of explanation can be to explain white-box models or intrinsic (specific) models, create inherently interpretable model (post-hoc) models, improve model fairness, or test prediction sensitivity [6].

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

Other explainable methods:

Instance-based aims to explain the actions of ML using a certain example in the dataset. These types of methods help for a better understanding of ML models using decision boundaries and properties of training data. This type of explainable method is divided into prototypes where the user is supplied with a sequence of examples that describe a class of the Black-box. And Counterfactuals are where the user is supplied with a sequence of examples same as the input query but with distinct class prediction.

Also, Neural network (NN) interpretations consternation on NN and DL explanation methods. These methods can divide to feature visualization, such as saliency maps which shows each image's quality. And distillation methods that aim to explain NN with a simpler model [9].

The global Surrogate model approximates the prediction in ML. It is also named model extraction. Such an example of the global surrogate is TREPAN [17] method which obtains a tree by querying NN and maximizing the gain ratio to estimate the network output. Also, an example of rule extraction is DeepRed [16], where the author explains the DL model based on decision tree inference.

Local interpretable model-agnostic explanations (LIME) [10] is the first local surrogate model based on a local approximation that mimics the local actions in the complex model by training part of the dataset related to a given example. The partial dataset is built by combining perturbation around the example. Using this dataset, the black-box model can approximate the alternative model as follows:

$$\xi(a) = arg \min_{z \in Z} L(f, z, \pi_a) + \Omega(z) \qquad (1)$$

Where f is the original model, L is the loss function which minimizes by model z. For instance, a. $\Omega(z)$ is applied as regularization for complexity reduction. The perturbation is weighted by π_a, which is defined as distance. The drawback of this method is finding the correct neighbor, which becomes a problem with tabular data. Also, depending on Gaussian distribution to define the example disregarding the correlation between features can cause improbable data points that can be used in local explanations. The anchor [11] method is introduced to solve issues that exist in the LIME method, which accurately defines local explanation areas using many rules instead of linear models. Based on LIME, Deep Argumentative eXplanations (DAX) [22] aims to explain CNN and FFNN on image and text data types. Zafar and Khan [23] provide a deterministic method of LIME to solve uncertainty problems by collecting data in hierarchal clusters and using KNN as a clustering method.

Another local data-driven method is Shapley Additive Explanations (SHAP) [12], based on optimal Shapley values. SHAP has two types of kernels. The first is KernalSHAP which assesses the Shapley values. And the other kernel is KernalTree which assesses tree-based models and consider them as global explanation based on a combination of Shapely values. The SHAP method obtains Shapley values based on game theory. The SHAP can be explained as follows:

$$E(\grave{a}) = \phi_0 + \sum_{i=1}^{M} \phi_i \grave{a}_i \qquad (2)$$

Where E is the explanation model, ϕ is the SHAP values, $\grave{a}$ nd ai is the feature vector of instance $\grave{a}$ .KernalSHAP is based on LIME to reduce the complexity of the sample and assess the SHAP values. Also, SHAP was applied more rapidly on Tree-based models. One disadvantage of this method is the slowness of KernalSHAP and neglect of the dependency on features which solves with TreeSHAP by introducing non-intuitive feature importance.

Partial Dependence Plot (PDP) [12] is an activation maximization method that defines the importance of features in the ML model. PDP indicates the marginal effect specific features have on the prediction. This method can visualize the non-linearity in the ML model. One drawback of this method is that part of the dependence feature is

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

not correlated with other features. Another Plot method is Individual Conditional Expectation (ICE) [13], which visualizes the prediction per one instance and shows the relation between prediction and instance changes. On the other hand, if the feature is strongly correlated, it is difficult to trust the initial dependency plot. So, the Accumulated Local Effects (ALE) [14] Plot can visualize as an average of the influence of features on the model prediction. ALE method is more complicated and less intuitive compared to other plots ICE and PDP.

Another perspective of model agnostic methods is permutation feature importance (PFI) [15] which computes the feature value and then measures the prediction error raising, which interrupts the correlation between the feature and the true outcome. Breiman [15] measures PFI based on random Forrest. Another study provides a model-agnostic version [40] where PFI is defined as:

$$PFI_T = \mathbb{E}(P(\hat{f}(\tilde{X}_T, X_M, Y)) - \mathbb{E}(P(\hat{f}(\tilde{X}), Y)) \qquad (3)$$

Where T is a subset that contains PFI features, $\tilde{X}_T$ is permuted reproduction of X_T, X_M is the feature matrix without T, and $P(\hat{f}, Y)$ is the performance measure of $\hat{f}$.

Another feature importance method, DALEX [18], is based on the method [41], which consists of decomposition for prediction. Each one is a local gradient for knowing the attribute importance for local XAI. And for Global XAI, many techniques are integrated into this method, such as calculating the performance, contribution of variables, residual diagnoses, and partial dependence plot. Another FI approach, the Neural additive model (NAM)[19], aims to integrate high-performance models Such as the DL model. This method can deal with the graph data and with additive DL models. Also, Contextual Importance and Utility (CIU) [20] depend on testing the ContextContext and determining the contribution of each feature in every ContextContext. This method utilizes two evaluation methods, Contextual Importance (CI), which approximates the overall significance of a feature in the present ContextContext, and Contextual Utility (CU) which gives an approximation for the satisfactory degree of present feature value from provided output class. Tree Space Prototype (TSP) [21] aims to obtain many tree prototypes in each class in the tree space to achieve a tree ensemble b explainable.

In instance-based, counterfactual methods aim to modify an example with the amount that can change its prediction. The first counterfactuals method is proposed by Wachter et al. [42], which minimizes the loss function using the following formula:

$$L(a, \grave{a}, \grave{b}, \lambda) = \lambda.(\hat{f}(\grave{a}) - \grave{b})^2 + d(a, \grave{a}) \qquad (4)$$

Where the quadratic distance between prediction and counterfactual $\grave{a}$ and desired output $\grave{b}$, and d is the distance between example a and counterfactual $\grave{a}$. Then obtain the Manhattan distance to be weighted by the inverse median absolute deviation (MAD)

$$d(a_i, \grave{a}) = \sum_{k \in F} \frac{|a_{i,k} - \grave{a}_k|}{MAD_k} \qquad (5)$$

where the feature set is F, and obtain MAD using the following formula:

$$MAD_k = median_{j \in P}(|A_{j,k} - median_{l \in p}(A_{l,k})|) \qquad (6)$$

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

Where p is the dataset, diverse Counterfactual Explanations (DICE) [24] methods aim to guarantee the feasibility of important ContextContext in CF and diversity among introduced CF. A further post-Hoc example aims to obtain CF using the shortest path distance of density weights metrics [26].CFX [25] produces CF for Bayesian classifiers, which indicates the reasons for the effects between variables.

Another instance-based method is Prototypes and Criticisms, where the prototype is the instance that represent all data and criticisms are the data that will not represent in the prototype. Most prototypes and criticism aim to select the optimum prototypes. As MMD-critic method [27] uses both prototypes and criticisms.

The NN interpretation considers the interpretation in NN and DL models. Such as Pixel Attribution (Saliency Maps) methods which provide each pixel contribution. One of these methods is Vanilla Gradient (Saliency Maps) [28] which is considered the main pixel attribution method. In this method, a gradient is obtained, and the higher gradient can cause modifications in predictions. Also, the deconvolution method [29] rebuild detected activation by unspooling, Relu, and Deconv in every CNN layer. One drawback of these methods is that they may mislead the explanation because of the gradient. So, layer-wise Relevance Propagation (LRP) [30] which based on Taylor decomposition for DL model interpretations.

Most previous methods are considered data-driven methods in the knowledge methods perspective, where exterior knowledge is applied to enhance the explainable model. One of the most common methods is Tcav [31] which provides the concept of activation vector (CAV) that explain the NN. Also, calculating the CA to obtain reasons. Another method also explains NN but in the NLP domain, which splits the input into parts to obtain reasons. Tcav depends on human existence to provide knowledge to model. Other knowledge-based methods example, Grad-CAM++ [32], aim to produce a holistic heatmap for input images. Another method [33] recommends songs for the user by integrating the history knowledge graph into the model. Then LSTM applies to give a score about the path that describes the user item and knowledge. Then combine all scores with exploring the user interests. Ma et al. [34] provide a knowledge method that combines recommendation and rule modules to explain the joint learning framework.

In inherently interpretable methods, Super-sparse Linear Integer Models (SLIM) [36] derived predictions depending on base math calculations. Another self-explainable method, the RB method [35], depends on the next hierarchical Bayesian probability optimization. Another method indicates the reasoning of the human expert in a certain field to create field interpretation [37]. DeepMiner [38] used interpretable CNN to drive the explanation of mmomogram classification. Also, in cancer detection, A recent XAI method interprets the NN model depending on mammography image [39].

3.1 Evaluation metric

In this section, we produce the metric that determines the effectiveness of the XAI model, which is considered to be guidance in XAI implementation.

Faithful and fidelity are considered accurate explanations for the model. Some studies use feature importance to calculate the probability of removed data. The important feature will indicate a high value [43, 44]. Also Tcav model has measured with faithful which indicates less value during noise images. Also, high fidelity on the surrogate model indicates that the model simulates the behavior in the black-box model, such as in [10], which provides a local surrogate model on an interpretable model and then measures recall of important features.

Robustness: refer to the ability of the model to protect its interpretation from small perturbation and adversarial attacks. Robustness is used to evaluate LIME and SHAP methods [44]. Also, some studies obtain robustness by measuring the difference in saliency map between original and perturbation inputs [45]. Also, some studies aim to reduce the bias model in the following three stages:

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

Pre-processing: methods in this stage obtain before training the model to remove bias. Such a method in [46] changes feature weights to eliminate bias in sensitive attributes.

In-processing: methods in this stage obtain during the model training process. Most studies aim to improve fairness constraints along with cost functions. Such as Adversarial Debiasing [47] which aims to improve the predictability of joint target variables while reducing the predictability of sensitive features using a GAN.

Post-processing: methods in this stage are applied after training. Unlike other stages, methods cannot perform any changes in input data or in the model, such as Equalized Odds [48], Which adopt the model thresholds to minimize the variations between the true positive rate and the false positive rate for each sensitive subgroup.

4. Application

4.1 XAI in Anomaly detection

Anomaly in ML means a data point in the space that does not follow the neighbor in the sample space, also named "outlier."So, anomaly detection (AD) is the process of a suspicious pattern that does not follow the norm. The AD is usually employed over unlabeled data. In practice, it is not important to detect outliers precisely, but it's essential to understand the reason for predictions. Anomalies explanations (AX) allow us to understand the reason for specific predictions from the AD model and also provide more safety to the model environment. Table 1 shows the summarized XAI methods in AD.

Table 1: Summarize XAI methods in AD.

Method	Type of method	Type
Based-on LRP	NN interpretation	Image
NEON	NN interpretations	Image
SFEs	Neuralization and sequential feature	Data points
X-PACS	Discriminative feature	Image
COIN	Discriminative feature	Image
FCDD	Perform FCNN architectures	Image
STAN	Space and time exploration	Audio-video
RP-tree	Discriminative feature	Text
Based on SHAP	Error reconstruction	Text
Based on SHAP and Local-DIFFI	Error reconstruction	Text

The 'neutralization means converting into a function like NN. In this ContextContext, Kauffmann et al. [49] study the influence of "Clever Hans" on AD. So, the author trained three models kernel density estimator KDE, Autoencoder (AE), and one class model on two datasets MINIST-C and MVTec ground truth and pixel-wise explanation. The results show that each model has its issues that lead to "Clever Hans" issues, so the author implements bagging techniques that combine the results of each individual model, which proves that it can reduce the "Clever Hans" AD. Another example of neuralization introduces Neuralization-Propagation (NEON), which was introduced to perform an explanation for the outlier. And apply LRP technique is built to predict the backward NN. In another example, the author introduces sequential feature explanations (SFEs) to explain the reason for considering anomalies points by indicating the importance of each point.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

In ContextContext of discriminative feature, Macha et al. introduce X-PACS for outlier explanations using clustering subspaces. Also, COIN explains outliers using classifiers knowledge that differentiates between normal and outlier points. Another study based on RP-tree explains the outliers by splitting tree regions.

Depending on Heatmap, Liznerski et al. explain one class to identify various types of energy consumption anomalies. And obtain Transparency by using an interpretation heatmap to detect the crucial area.

A study of the combination of data types in explanation, space-attention network (STAN) detecting Audio-video recognition, which introduces both space and time exploration.

Depending on Shapley value computing, Takeishi introduces a method to reconstruct error in Principle component analysis (PCA) which helps to explain the AD prediction based on PCA. Also, Brito et al. show similar behavior between SHAP and Local Depth-based Feature Importance for the Isolation Forest (Local-DIFFI).

4.2 XAI in Network/IoT

Cybersecurity attacks become a risk to the stability of communications and Network/IoT integrity. So, it is essential to understand the causes of any ML predictions that strengthen ML models. In this section, we provide an investigation on the most contribution in Network/IoT, which utilizes XAI methods. This investigation is summarized in Table 2.

Table 2: Summarize XAI methods in Network/IoT.

XAI Method	Type of attack	Dataset
LIME, SHAP, Boolean Decision Rules via Column Generation (BRCG), and Contrastive Explanation Method (CEM)	IDs	NSL-KDD
SHAP	IDs	NSL-KDD
SHAP	IDs	CSE-CIC-IDS2018, BoT-IoT, and ToN-IoT
LIME	Botnet	Selected dataset
Subspaces the features	Botnet	MAWI and ISP
Random Forrest	Botnet	UTC university dataset
Random Forrest, logistic regression, and NN	Malware	N/A

DFF depends on multiple XAI techniques to obtain measurable reasons that affect the attack prediction and its degree. Another method in IDs aims to differentiate between binary classifiers and multi classifiers depending on SHAP. Further contribution in IDs by Sarhan et al. aims to obtain the feature contribution depending on SHAP.

In Botnet detection, Guerra-Manzanares et al. analyze the impact of the selected feature on the accuracy and quality of the interpretation of data from the problem domain of nine IoT devices. Also, Araki et al. introduce subspaces of features to explain partial properties of objective hosts, combining with the cluster method to produce feature importance. Another study in Neris and Rbot detection aims to explain the model using Random Forrest.

In Malware detection, Saad et al. use the RB explanation to produce understandable features.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

4.3 Adversarial attacks

Table 3: Summarize XAI methods in adversarial attacks.

Local/Global	Specific/agnostic	Type
Both	Specific	Text
Both	Agnostic	Text
Both	Specific	Text
Both	Agnostic	Text
Both	Agnostic	Text
Both	Agnostic	Image
Both	Agnostic	Image
Both	Agnostic	Image
Both	Specific	Graph
Both	Specific	Image

In the text attack context, Cheng et al. consider attacks against the Seq2Seck model, which is widely used in text summarization and translation. Another study utilized the sememe substitution method and PSO to introduce an attack model. Furthermore, methods utilize the text perturbation method to perform attacks depending on replacing the BERT mask as an element in the original text with text fragments. Another example of NLP is utilizing the effect perturbation morphological word and taking only considerations that affect the element of the original word.

Another line of research for the image data type, A fast clipping approach that enforces the perturbation vector and computes the closest perturbation to the solution. Another method is gradient-based, stuck on the area near the boundary but shows its efficiency against gradient mask and is capable of optimizing many adversarial properties. Another line of methods that are not seeking optimal attack NATTACK found a good adversary in the exists in gradient area with neighbors around a certain point.

For graph data type, GNN-Meta-Attack is a method depending on meta-learning to solve time-training attacks and optimize the graph as hyperparameters. Table 3 shows the adversarial attacks.

A recent study explains prediction on CNN using the heatmap activation function in the last CNN layer.

4.4 Data Fusion

Data fusion aims to examine and explain pollution incidents to learn how they are connected. To do this, one must first identify anomalous graphs and then correlate such graphs with representative event subjects. The scope and significance of the event on the divergence of the pollution concentrations are also analyzed.

There are a total of 3629 photos used for train and 1725 used for testing in the MVTec Anomaly Detection dataset's 15 categories. There are no damaged pictures in the training set. You'll find faulty and perfect examples of each item in the test set.

There are 15 classifications; the first 5 represent distinct regular (carpeting, grid) or randomized (leather, tile, wood) patterns, while the next 10 represent various items. While some, like the bottles and metal nuts, are solid and unchanging in appearance, others, like the cables and rocks, are flexible and subject to change (hazelnut). Items like toothbrushes, capsules, and pills were collected in a somewhat aligned stance. In contrast, other things were put in front of the camera at arbitrary angles (e.g., metal nut, screw, and hazelnut). Surface flaws (e.g.,

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

scrapes, dents), structural flaws (e.g., warped portions of the item), and lack of components are only a few defects that may be found in test photos of weird samples. There are 73 distinct kinds of defects, with around five in each class.

By adjusting the total amount of nodes in the encoder and decoder, we were able to compare the detectability of LSTM-autoencoders with varying structures and draw conclusions about the impact of the system on LSTM-autoencoder efficiency. How many input layer nodes are represented by the decoder and decoding layers? The classification performance, model training duration, and variation in false alarm and leakage rates among networks were determined using a 5-fold cross-validation. The detection accuracy steadily increased as the number of nodes in the implied layer, both encoder, and decoder, grew, with little to no discernible variation in detection accuracies. However, the model's creation time ballooned as the number of nodes grew.

Choosing a proper threshold is unnecessary if you evaluate anomaly segmentation methods using cutoff point metrics, like considering the area beneath a curve. However, to use an algorithm effectively, a threshold value must be chosen to assess whether or not a certain component is flawed. Due to the scarcity of out-of-the-ordinary data points available during training, this is a difficult challenge to solve. We believe that estimating a threshold purely on anomaly-free data is better, even if a limited number of anomalous examples were available for estimation. This is due to the fact that it is not known whether the given samples represent the whole spectrum of anomalies, and the predicted threshold may have poor performance for other unknown kinds. To identify even slight deviations, we need to determine a point that differentiates the normal data distribution from the remainder of the information manifold.

Here, we look at three different approaches to threshold estimation for anomaly segmentation, all of which rely on testing on a collection of validation photos that have not been tampered with. We conducted studies to determine which criteria provide the best performance and how well each method translates from the validation to the test set.

Sustained Effort at Its Highest Level An ideal solution would label every pixel in the validation photos as normal. This may be done by setting the threshold equal to the highest value of any anomaly score in the validation data. Because even a single aberrant pixel with a high anomaly value might cause thresholds to underperform on the test set, this prediction is generally quite cautious in reality.

We rigorously compare and contrast several cutting-edge approaches to unsupervised anomaly classification on this dataset. Its ultimate purpose is to set a standard for procedures still to come. We compare and contrast the methods by discussing their performance on the dataset's objects and textures. We demonstrate that although each technique may identify specific kinds of abnormalities, none of the scenarios assessed really shines when applied to the whole dataset. For example, we demonstrate that strategies that use features learned by networks that have already been trained on a large Imagenet database provide the best results. It is clear that there is a lot of opportunity for development in deep learning-based unsupervised learning that is built from the start, such as multilayer autoencoders or generating stochastic networks. We evaluate how many performance parameters influence the final assessment score and compare several threshold estimation methods. We also detail how much memory and how long each assessed method takes to produce an assumption.

The train divide of the dataset is used to add 10,000 training sets to each dataset class. Textures are cropped uniformly across all training photos from randomly chosen patches. To create the effect of moving and rotating objects, we randomly transform the whole input picture and then scale it to the exact dimensions as the autoencoder's source. Mirroring is enhanced in areas where the item allows it.

Spot Anomalies Between Students and Instructors We employ 3 teacher network models trained on ImageNet to generate dense feature maps and evaluate Student-Teacher anomaly detection.

Spotlight Dictionary The suggested CNN feature dictionary is used; however, we have our implementation. To inspect textures, a GMM-based model was developed.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

Images are grayscaled and zoomed to a size of the input of 500 x 500 pixels, and information-related pyramids are built for training and assessment. A new GMM with a large correlation matrix is learned on each tier of the pyramid. To ensure uniformity throughout all pyramid tiers, we use a patch size of 9 by 9 pixels for all texture analyses. The evaluation model is shown in table 4.

Table 4: Indicators of performance that are not reliant on a threshold.

	Autoencoder	Texture	Feature Dictionary	Teacher	Validation
Receiver operator characteristic	0.651	0.721	0.910	0.941	0.812
Precision-Recall Curve	0.211	0.311	0.510	0.612	0.312
Pro Curve	0.723	0.711	0.899	0.921	0.835

Model of Variability The Variability Model is constructed by summing the means and standard deviations of all pixels across all training photos in the relevant dataset categories, regardless of scale. Aligned items in the photos are necessary for this to perform properly. In light of the fact that this is not still the situation, we developed a tailored alignment approach for our studies on the six main types of datasets. The shape-based comparison was used to align the bottle and steel nut, while template comparison with a normalised pass as the clustering method was used to align both grid and transistors.

Here is how you can get an anomaly map from a sample photograph. Every pixel in the anomalous map is assigned a value based on its distance from the training mean value, as determined by dividing the grey value of the associated test pixel by a fraction of the training standard deviation. In the case of multichannel pictures, we do this procedure independently for every signal, and then we generate a global anomaly map by taking the maximum of each show's map on a pixel-by-pixel basis.

Two layers, an encoder, and a decoder made up the LSTM-autoencoder architecture employed in the experiment. Nodes in the input neurons have the same sizes as the input eigenvalues, and half as many are used for the encoder and decoder layers. Figures 1 and 2 show the accuracy of the dataset.

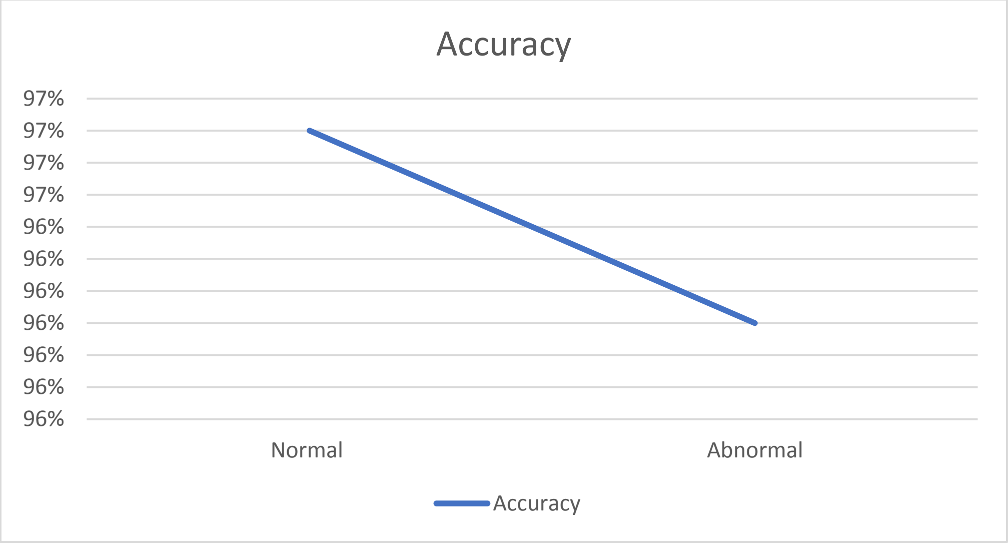

Figure 1: The threshold values for evaluation detection of the dataset.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

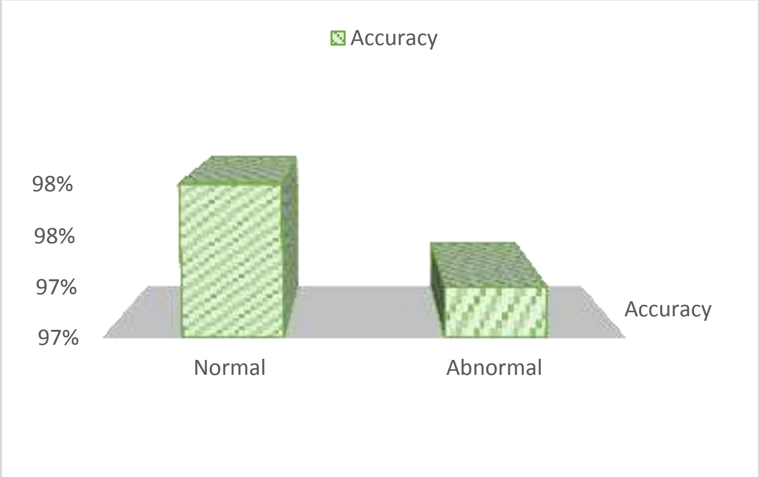

Figure 2: The statistics values for evaluation detection of the dataset.

5. Challenges and future work

We will discuss research directions for future work depending on the existing challenges of investigation in XAI. Having reviewed implemented methods in XAI and contributions in three fields (anomaly detection, Network/IoT, adversarial attacks).

Bias toward the Same type of data: it's noticed that many studies explore features of the image or tabular data, and just a few studies make combinations from audio and video.

Method limitations: methods such as LIME and SHAP can be easily manipulated or mistaken to access only the needed data. SHAP is a slow kernel. Also, LIME takes only samples from Gaussian distribution and ignores the correlation between features. Also, a very commonly used method, especially in image SM, confirms the bias. More efforts are needed to solve these issues or implement powerful methods.

Low feature explanation Some methods provide their explanation in terms of pixels, such as SM, but there is a need to build methods that give an explanation in higher explanation in terms of features or concepts. CA method that was reviewed in our work is to get the feature level explanation by determining the probability that a select feature has more contributions than other features. Further implementation using this method is needed.

Hard to implement inherently interpretable models Intrinsic White-box model is still difficult to implement, especially in NLP or image processing which may scarify with the performance during the race of DL models in knowledge transfer from one domain to another.

Performance against explanation still there is no exit for the model that can provide good DL model performance and good prediction explanation, especially since XAI methods can slow the learning process of the model.

Limited contribution in intrinsic models it is noted that there are not many studies implementing inherently interpretable methods, which makes this lack of comparisons and is employed in many fields.

Valuable knowledge in XAI knowledge methods still has a question on how to retrieve the valuable and important knowledge only.

Guarantee security and privacy. One of the criteria of model transparency that we define is that model must protect the user-sensitive data and be able to deal with security issues to provide trust in AI systems and obtain fairness. So, it's crucial study direction that is importantly needed to confirm the proper use of data.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

6. Conclusion

In this work, we introduce a comprehensive work on explainable artificial intelligent methods. Corresponding to our work, the transparency model must determine four criteria Original/transformed data, Human-in-the-loop (HITL)/human-out-of-the-loop, Knows about the model, Model Security, and privacy considerations. We show the effect of using explainable methods on machine and deep learning models. We introduce all categories of explainable artificial intelligent methods and provide advantages and disadvantages for each method. Evaluation metric was introduced, such as Faithful and fidelity, Robustness, and reduced bias in the model in Pre-processing phase, In-processing phase, and post-processing phase. Furthermore, we provide a review of some explainable artificial intelligent applications such as Anomaly detection, Network/IoT, and Adversarial attacks. Moreover, explainable artificial intelligent methods still face many limitations and problems that are concluded in the challenges section. In conclusion, although many deep learning-based intelligent multimedia forensics methods have achieved notable results, there is still much future research work in this area that must be discovered.

References

[1] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, *et al.*, "Privacy and Security Issues in Deep Learning: A Survey," *IEEE Access*, 2020.

[2] I. Northpointe, "Practitioner's Guide to COMPAS Core," 2015.

[3] M. MCGOUGH, "How bad is Sacramento's air, exactly? Google results appear at odds with reality, some say," in *Sacramento BEE*, ed, 2018.

[4] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, ed: Springer, 2019, pp. 5-22.

[5] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black-box models," *arXiv preprint arXiv:2102.13076*, 2021.

[6] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, p. 18, 2021.

[7] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever Hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, pp. 1-8, 2019.

[8] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning* vol. 1: MIT Press Cambridge, 2016.

[9] C. Molnar, *Interpretable machine learning*: Lulu. com, 2020.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135-1144.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[13] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, pp. 44-65, 2015.

[14] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 82, pp. 1059-1086, 2020.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[16] M. Sato and H. Tsukimoto, "Rule extraction from neural networks via decision tree induction," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, 2001, pp. 1870-1875.

[17] M. Craven and J. Shavlik, "Extracting Tree-Structured Representations of Trained Networks, Advances, Neural Information Processing Systems 8," 1996.

[18] P. Biecek and T. Burzykowski, *Explanatory model analysis: explore, explain, and examine predictive models*: CRC Press, 2021.

[19] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," *arXiv preprint arXiv:2004.13912*, 2020.

[20] S. Anjomshoae, T. Kampik, and K. Främling, "Py-CIU: A Python Library for Explaining Machine Learning Predictions Using Contextual Importance and Utility," in *IJCAI-PRICAI 2020 Workshop on Explainable Artificial Intelligence (XAI)*, 2020.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

[21] S. Tan, M. Soloviev, G. Hooker, and M. T. Wells, "Tree space prototypes: Another look at making tree ensembles interpretable," in *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, 2020, pp. 23-34.

[22] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, *et al.*, "Generating contrastive explanations with monotonic attribute functions," *arXiv preprint arXiv:1905.12698*, 2019.

[23] M. R. Zafar and N. M. Khan, "DLIME: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[24] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607-617.

[25] E. Albini, A. Rago, P. Baroni, and F. Toni, "Relation-based counterfactual explanations for Bayesian network classifiers," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI (2020, To Appear)*, 2020.

[26] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 344-350.

[27] B. Kim, O. Koyejo, and R. Khanna, "Examples are not enough, learn to criticize! Criticism for Interpretability," in *NIPS*, 2016, pp. 2280-2288.

[28] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, 2014, pp. 818-833.

[30] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, p. e0130140, 2015.

[31] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," *arXiv preprint arXiv:1606.04155*, 2016.

[32] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839-847.

[33] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T.-S. Chua, "Explainable reasoning over knowledge graphs for recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 5329-5336.

[34] W. Ma, M. Zhang, Y. Cao, W. Jin, C. Wang, Y. Liu, *et al.*, "Jointly learning explainable rules for recommendation with knowledge graph," in *The World Wide Web Conference*, 2019, pp. 1210-1221.

[35] H. Yang, C. Rudin, and M. Seltzer, "Scalable Bayesian rule lists," in *International Conference on Machine Learning*, 2017, pp. 3921-3930.

[36] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, pp. 349-391, 2016.

[37] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, *et al.*, "TED: Teaching AI to explain its decisions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 123-129.

[38] J. Wu, B. Zhou, D. Peck, S. Hsieh, V. Dialani, L. Mackey, *et al.*, "Deepminer: Discovering interpretable representations for mammogram classification and explanation," *arXiv preprint arXiv:1805.12323*, 2018.

[39] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, *et al.*, "IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography," *arXiv preprint arXiv:2103.12308*, 2021.

[40] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *Journal of Machine Learning Research*, vol. 20, pp. 1-81, 2019.

[41] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 589-600, 2008.

[42] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[43] M. Du, N. Liu, F. Yang, S. Ji, and X. Hu, "On the attribution of recurrent neural network predictions via additive decomposition," in *The World Wide Web Conference*, 2019, pp. 383-393.

[44] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021

[45] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, *et al.*, "The (un) reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ed: Springer, 2019, pp. 267-280.

[46] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems,* vol. 33, pp. 1-33, 2012.

[47] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335-340.

[48] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *arXiv preprint arXiv:1610.02413,* 2016.

[49] J. Kauffmann, L. Ruff, G. Montavon, and K.-R. Müller, "The Clever Hans effect in anomaly detection," *arXiv preprint arXiv:2006.10609*, 2020.

Doi: https://doi.org/10.54216/FPA.030104
Received: October 05, 2020 Revised: December 22, 2020 Accepted: March 21, 2021