International Journal of Computer Sciences and Engineering
Vol.**12**, Issue.**5**, pp.**59-67**, May **2024**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Article

Enhancing Interpretable Anomaly Detection: Depth-based Extended Isolation Forest Feature Importance (DEIFFI)

Rahul Singh[1*], Deepti Gupta[2]

[1,2]Computer Science and Engineering, UIET, Panjab University, Chandigarh, India

**Corresponding Author: kundlasrahulsingh@gmail.com*

Received: 02/Apr/2024; **Accepted**: 04/May/2024; **Published**: 31/May/2024. **DOI**: https://doi.org/10.26438/ijcse/v12i5.5967

Abstract: The research introduces a novel approach, Depth-based Extended Isolation Forest Feature Importance (DEIFFI), to enhance the interpretability of Extended Isolation Forest (EIF) algorithm in anomaly detection (AD). Anomaly detection is critical for identifying rare and significant deviations from norm in data. However, understanding the reasons behind classifying instances as anomalies poses a challenge. DEIFFI addresses this challenge by providing valuable insights, empowering users of EIF-based AD to conduct thorough root cause analysis. A noteworthy feature of DEIFFI is its capacity to improve interpretability without imposing heavy computational burdens. This is crucial for real world applications requiring efficient AD, particularly in situations demanding real-time decision-making. DEIFFI achieves remarkable results with low computational costs, making it an appealing option for practical implementations. With an accuracy of 0.914 and 0.942, precision of 0.607 and 0.64, recall of 0.773 and 0.96, and an F1 score of 0.68 and 0.768 on real and synthetic datasets, respectively. DEIFFI provides interpretable insights alongside competitive performance metrics, solidifying its suitability for real-time decision support. Importantly, DEIFFI contributes to AD by enhancing interpretability and assisting in unsupervised feature selection. This dual capability highlights practical utility of DEIFFI, improving EIF's capabilities and extending its applicability across diverse AD scenarios.

Keywords: Anomaly Detection, Explainable Artificial Intelligence, Extended Isolation Forest, Feature Selection, Interpretability, Outlier Detection.

1. Introduction

Anomaly detection (AD) in Wireless Sensor Networks (WSNs) [10] aims to identify deviations from normal behavior or patterns in the collected sensor data. Traditional rule-based approaches often struggle to handle the dynamic and complex nature of WSN data, making Machine Learning (ML)-based techniques an attractive solution. ML algorithms have the ability to learn from historical data patterns and detect anomalies based on learned patterns or statistical deviations. The effectiveness of AD techniques holds significant importance across a wide range of application domains, including WSNs [17], industrial cyber-physical systems [20], healthcare [16], driving systems [19], agriculture, environmental monitoring and biology [4].

The extensive applicability of AD algorithms originates from their capability to be trained and utilized in unsupervised environments. This trait is especially beneficial in situations where manually annotating data by human experts is both expensive and time intensive. In such cases, a human-centered design principle is crucial to minimize human efforts while ensuring effective AD.
In recent years, research has increasingly focused on using machine learning algorithms for anomaly detection tasks, particularly those involving supervised algorithms such as Random Forests, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN) [11]. However, obtaining labeled training data in WSNs can be challenging and may not always be feasible. Because of these considerations, traditional AD approaches, such as clustering-oriented algorithms [18] (e.g., K-means, DBSCAN, Hierarchical Clustering) continue to be useful in a wide range of applications. These approaches can identify clusters of similar instances and identify anomalies as data points that are not associated with any cluster. Density-based methods such as Local Outlier Factor (LOF) [18] and Gaussian Mixture Models (GMM) effectively detect anomalies based on the density distribution of data points and Angle-Based Outlier Detector [14] and Isolation Forest (IF) [15].

While AD algorithms have proved substantial value and efficacy, widespread application in companies and organizations with adequate infrastructures remains unrealized. The lack of confidence stems from the absence of labeled data points, which are necessary for proper testing and validation of the AD algorithms. This forces users to choose between blindly trusting the algorithm's conclusions and not using it at all, both of which have bad consequences. The second issue concerns getting extra knowledge regarding

the specific work at hand. Understanding the underlying causes of anomalies is critical since it can lead to actionable insights for troubleshooting and root cause analysis. In response to these obstacles, the concepts of eXplainable Artificial Intelligence (XAI) [8], become relevant. XAI endeavors to enhance the comprehensibility of black-box ML models for human understanding. By applying XAI principles, AD algorithms can be made more transparent and interpretable, allowing users to gain insights into how the algorithms arrive at their conclusions and enabling a better understanding of the underlying causes of anomalies.

This research work focuses on the interpretation of the EIF [9], one of the most popular and effective extension of the original IF approaches for AD. EIF builds upon the fundamental principles of the IF algorithm while incorporating additional features and improvements. The key features on which EIF operates are the isolation, extension level, randomness and path length. The proposed method offers a cost-effective and computationally efficient solution to address this problem effectively.

The authors review the available literature on AD and XAI in the subsections that follow. The following section presents a complete review of the research's primary contributions as well as the underlying motivations driving the proposed interpretability approaches. Section 2 of the paper is dedicated to presenting and analyzing the interpretability methods proposed in this research work. It provides a detailed description and in-depth analysis of these methods. Section 3 presents the experimental results obtained from applying the proposed interpretability methods. Section 4 summarizes the key findings and draws overall conclusions based on the research conducted.

In the following subsections, the authors examine the existing literature in the fields of AD and XAI. This section provides a comprehensive overview of the main contributions of this research and the underlying motivations that drive the proposed interpretability methods. Section 2 of the paper is dedicated to presenting and analysing the interpretability methods proposed in this research work. It provides a detailed description and in-depth analysis of these methods. Section 3 presents the experimental results obtained from applying the proposed interpretability methods. Section 4 summarizes the key findings and draws overall conclusions based on the research conducted.

2. Related Work

2.0.1 Anomaly Detection

As noted in Section 1, AD methods play a significant role in numerous applications, characterized by distance-based algorithms, density-based algorithms [13], approaches utilizing SVMs [22], and tree-based methods [15, 9]. This study aims to interpret the extended form of the original EIF [9]. The extensive usage and effectiveness of the EIF [9] algorithm has motivated the development of various adaptations and variants to address complex application scenarios and incorporate new methodological principles. The Functional Isolation Forest (FIF) [21] expands the application of the original Isolation Forest (IF) model from finite-dimensional observations to functional data. Various extensions of the IF algorithm have been proposed to address specific challenges and data types. For example, the k-means-based IF [12] integrates k-means clustering to establish division counts for each decision tree node. Additional adaptations, such as iForest ASD [5] and RS-Forest Isolation Mondrian Forest [23], have emerged for the analysis of streaming data. Amid the various IF model variations, this study opts to concentrate on the EIF algorithm in particular.

2.0.2 Interpretability

The main objective of Explainable AI (XAI) is to uncover the internal mechanisms of machine learning models, especially in regression and classification tasks. XAI prioritizes making algorithms, such as Deep Neural Networks (DNNs) and ensemble methods, more interpretable. These algorithm classes are known for their outstanding accuracy but are often difficult to understand from a human perspective [24]. This process involves extracting meaningful insights and explanations from complex models to understand how and why the model arrived at a specific outcome. It helps stakeholders, including domain experts, regulators, and end-users, to grasp the underlying factors, features, and patterns influencing the model's predictions. Considering the remarkable performance of DNNs in various challenging tasks such as time series forecasting, text classification, and image classification, it is understandable that a significant amount of research in the field of XAI has been dedicated to addressing the interpretability of DNNs. Interpretability of DNNs can be approached in two main ways. The first approach aims to provide explanations for the model's predictions or outputs, shedding light on the reasoning behind the decisions made by the model. This involves generating human-understandable explanations that can help users comprehend and trust the predictions made by the DNN. The second approach focuses on interpreting the internal representations learned by the DNN when processing the input data. This involves unraveling the complex internal structures and patterns captured by the DNN during its training process. By understanding the internal representations, researchers can gain in-sights into how the DNN processes and understands the data, leading to a deeper understanding of its decision-making process.

Although this research has largely focused on Random Forests (RFs) [2], it is crucial to note that many other studies in the literature focus on interpreting other ensemble approaches, like Gradient Boosting Decision Trees. Random Forests (RFs) are classification ơregression tree ensembles that use bagging to reduce prediction variability. RFs, as opposed to isolated Decision Trees, frequently provide improved accuracy at the sacrifice of interpretability. Many research efforts have been directed towards enhancing standard feature importance score methods in the context of RFs. As an instance, one study suggests augmenting the permutation importance metric through the introduction of a conditional permutation approach. A modified version of the Mean Decrease

Impurity (MDI) feature significance metric is presented in another research article to address the issue of bias in feature selection by MDI. These improvements aim to provide more reliable and meaningful feature importance scores within the RF framework. Moreover, recent research has also focused on detecting interactions between features, recognizing their importance in capturing complex relationships in data. These studies explore methods for identifying and quantifying feature interactions, which can provide valuable insights for understanding the underlying data patterns. In EIF, similar considerations can be applied to evaluate feature importance and explore interactions within the IF framework. It would be beneficial to investigate how these techniques can be adapted and incorporated into the EIF algorithm to enhance interpretability and provide deeper insights into the detected anomalies and their underlying causes.

Model-specific methods, which include the interpretability techniques previously discussed, are created especially for specific machine learning models. These techniques are very transparent, suggesting that they heavily depend on the ML model's internal structure. Conversely, flexible approaches have garnered significant attention because of their considerable mobility, which enables their application to a broad variety of machine learning models. These methods are known as *model-agnostic* methods. They offer the advantage of being applicable to different models without requiring specific adaptations for each model type. Conversely, *model-agnostic* techniques like accumulated local effects plots and partial dependency plots are examples of methods that are used to illustrate the overall behavior of the machine learning model. Even while *model-agnostic* methods seem portable, it's crucial to remember that choosing interpretability methods frequently comes after deciding on a particular model type. *Model-agnostic* methods become valuable in cases where there is a lack of *model-specific* alternatives for certain classes of ML models. Consequently, rather than being specialized to a single model, the value of *model-agnostic* approaches is in their capacity to offer interpretability across a variety of model types. It's crucial to recognize that *model-agnostic* techniques do have a few significant drawbacks, though:

- Limited exploitation of the model's inner structure: Since *model-agnostic* methods do not leverage the specific structure of the examined model, there may be concerns that the provided explanations are oversimplified and do not fully capture the genuine fundamental connection between the input and the output. This can raise doubts about the reliability and accuracy of the interpretability method.
- Manipulation of inputs and stability issues: Numerous *model-agnostic* techniques entail altering inputs and assessing the resultant impacts on corresponding predictions. This procedure demands care, as artificially generated input instances might not faithfully mirror the inherent data distribution. This can introduce instability and uncertainty into the interpretability process, potentially compromising the validity of the information conveyed by the method.
- Reliance on assumptions and methodological choices: *Model-agnostic* methods often require the adoption of restrictive assumptions or involve opaque methodological choices. For example, some methods assume independence between features or create perturbed input instances. This places a burden on the user to trust the interpretability method without fully comprehending its theoretical foundations. Consequently, the trust shifts from the model itself to the interpretability method, potentially raising concerns about the reliability and transparency of the method.

These limitations highlight the trade-offs involved in using *model-agnostic* methods for interpretability. While they offer flexibility and applicability across various models, users need to consider these shortcomings and critically evaluate the interpretability provided by such methods. In light of the increasing attention being paid to AD techniques that draw inspiration from the IF model and the shortcomings of *model-agnostic* interpretability approaches, this work presents novel *model-specific* interpretability strategies specifically designed to help comprehend the internal workings of the EIF.

2.2 Contributions

The EIF model, like the original IF, is highly valued and extensively utilized due to its excellent detection performance. The EIF often achieves remarkable results even when using default hyperparameter values, requiring minimal or no tuning. Additionally, it retains the computational efficiency that makes the IF model widely preferred in AD tasks.

However, similar to other ensemble learning methods, concerns and uncertainties regarding interpretability [7] may arise with the EIF: indeed, the EIF does not provide any information about the underlying logic behind its predictions. Additionally, it doesn't provide any guidance on what elements are most crucial to completing the AD assignment. To overcome the aforementioned challenges, the writers of this study effort propose the *model-specific* techniques.In particular, this research suggests:

- A variant of the DIFFI method offers Local Feature Importances (LFIs). The goal of LFIs is to interpret the distinct predictions generated by the EIF model during testing. An easy-to-use and effective solution for unsupervised feature selection in AD situations is the DEIFFI method.
- An appropriate proxy job for assessing interpretability methods in AD situations is unsupervised feature selection [6], which enables a useful assessment of these approaches without any prior information about the significance of the features.

All the contributions listed above adhere to the human-centered concept that guides this effort, which seeks to meet the user's needs to the greatest extent feasible. This guiding principle led us to prioritize certain characteristics, such as efficient computational times and simple hyperparameter

tuning procedures. It operates on simple computations using numbers that come naturally from the principles that underpin the EIF model. The method's details will be presented in the following sections, which distinguishes itself by avoiding any artificial manipulations of data points as well as the need for fitting interpretable local surrogate models. Unlike previous approaches, the DEIFFI method operates directly on the original model and data points, without the need of local approximations or data perturbations. This method provides an exact and fully transparent portrayal of the EIF's internal structure. Model-independent interpretability approaches, on the other hand, cannot achieve this degree of accuracy and transparency.

2.2 Motivations

By including the assessment procedure into the issue formalization process, the necessity for interpretable algorithms in the domain of AD is aligned with the inherent relationship between interpretability and incompleteness. This connection becomes particularly relevant in the context of AD, where the scarcity of labeled datasets limits the ability to test AD algorithms in unsupervised settings. To bridge this gap and overcome the potential reluctance towards adopting automated systems, it is essential to provide proxies that can assess the trustworthiness of these algorithms. These proxies serve as tools to interpret the inner workings of the model and evaluate whether it aligns with the expected behavior.

The alignment between the estimated feature importance scores and human prior knowledge is crucial in fostering users' confidence. When the importance scores align well with the existing understanding of the problem domain, users are more likely to rely on the EIF model's predictions and grant it greater autonomy, especially in non-critical scenarios. The goal of this effort is to close the gap between the EIF model's internal functioning and human understanding, allowing users to obtain insights about feature importance and, eventually, increase their faith in the EIF-based anomaly detection system. By addressing the interpretability needs specific to the EIF algorithm, DIFFI contributes to the broader objective of promoting the acceptance and adoption of the EIF model in various practical applications.

The impetus for DIFFI's *model-specificity* stems from a desire to identify the actual underlying logic controlling the EIF's behavior. This need presents a possible source of suspicion, as the surrogate model's fidelity to the original model may be questioned. By focusing on the specific characteristics and behavior of the EIF model, DIFFI aims to provide a more comprehensive and faithful interpretation that aligns with the EIF's inherent logic. This approach recognizes the limitations and challenges associated with using generic *model-agnostic* techniques in capturing the intricate workings of the EIF model.

DIFFI is a post-hoc method; This study aimed to provide global and local feature significance assessments computed a posteriori while retaining the performance of a proven and successful AD technique. Given the balance that exists between accuracy and interpretability [1], constructing a model that was innately interpretable would have entailed sacrificing some predictive capability. DIFFI's post-hoc method allows us to gain meaningful insights into the EIF's inner workings while not jeopardizing its robust AD capabilities. This research work therefore provides interpretable reasons for the EIF's forecasts while keeping its high accuracy in detecting abnormalities.

Ultimately, through the introduction of an interpretability approach, this research guarantees optimal adaptability. This empowers users to select the most fitting solution for their unique scenarios, considering factors like desired granularity or the available time for results analysis.

3. DEIFFI: Depth-based Extended Isolation Forest feature importance

The EIF [9] algorithm's underlying essential ideas has been briefly summarized in this section, along with the appropriate notation. The DEIFFI approach, which is especially suited for the interpretability [7] requirements of the EIF, is then further examined. This research work thoroughly reviews and evaluates each part of the DEIFFI approach, explaining its underlying assumptions.

This research work proposes a methodology called DEIFFI, which is intended to understand each prediction produced by the EIF model. This variation improves the model's interpretability at an instance level by enabling a more fine-grained knowledge of the variables driving particular predictions. This strategy has two advantages: it makes it easier to identify the most significant characteristics without prior knowledge of the problem, and it also serves as a *"proxy task"* for evaluating the correctness of the feature importance scores assigned. In relation to unsupervised Anomaly Detection, this paradigm for evaluating feature significance scores can be utilized as a functionally based evaluation technique. It also discusses and evaluates each component of the DEIFFI approach, clarifying its underlying assumptions.

This study enhances the EIF algorithm's interpretability and practical utility by providing these all-encompassing methodologies and methods, particularly in instances when prior knowledge about the problem is not immediately available. In the context of unsupervised AD, this paradigm to assess feature relevance scores can be employed as a functionally grounded evaluation technique.

3.1 Background: Extended isolation Forest

The EIF [9] is an enhanced version of the popular IF algorithm [15] that is widely used for AD tasks. The EIF algorithm retains the key principles of the IF algorithm, such as the use of random partitioning to isolate anomalies but incorporates additional improvements to enhance its performance and interpretability [7]. One of the key enhancements in the EIF algorithm is the consideration of the depth of each data point in the isolation process. By taking into account the depth at

which a data point is isolated, the EIF algorithm captures more nuanced information about the anomalies present in the data. This depth-based approach provides a finer granularity in quantifying the outlying behaviour of data points. Because branching is determined by how closely it resembles BST, this bias is the root cause. Because the branching points are parallel to one of the axes, bias is introduced. The general case requires a random slope at each branching point. It selects an arbitrary slope n for the branching cut and an intercept at random p over the feature and value. With boundaries derived from the sub-sample of data to be divided, the slope may be calculated using the N (0,1) Gaussian distribution and the intercept from the uniform distribution. The following are the branched conditions for data splitting at a certain point x:

$$(x - p) * n \leq 0 \quad (1)$$

This introduces a new generalization hyperparameter, *extensionLevel*. *extensionLevel* forces random items of n to be zero. The EIF algorithm's hyperparameter" *extensionLevel"* ranges from *0* to *(P-1),* where *P* is the dataset's number of dimensions or features. Setting the value to 0 means that all splits will be parallel to all axes, which is consistent with the behaviour of the original IF. A larger *extensionLevel* value specifies that the divides will be parallel to the specified number of axes. When *extensionLevel* is set to the maximum value of *P-1*, it indicates full extension. This means that the slope of the branching point is randomized for each split. This provides maximum flexibility in the EIF model. It is recommended to use a fully extended EIF. However, in cases where the range of minimum and maximum values differs significantly across features, a lower *extensionLevel* may be more appropriate. This allows the algorithm to adapt to the varying scales of the features and capture their importance accurately. The choice of *extensionLevel* in EIF is determined by the dataset's specific characteristics and the algorithm's desired behaviour in capturing feature importance and accommodating varied feature sizes.

The EIF algorithm has been formally described in the paragraphs that follow:
Algorithm-specific Parameters:

- *extensionlevel*: The value between *[0, P1],* where P is the overall feature count. IF behaviour is represented by the hyperparameter's default value of 0, which is its lowest value. The maximum, P1, denotes a complete expansion. The bias of a conventional IF decreases as the extension level is raised.
- *sample size*: The total number of observations that were randomly selected and utilised to train each EIF tree. Its default value is 256.
- *ntrees*: Specify the number of trees. This option defaults to 100.

During the training phase, the algorithm creates a tree by iteratively selecting a random dimension and comparing the value of each point to a randomly generated cutoff value for the selected dimension. According to the algorithm's criteria, the data points are then routed down the left or right branch. Anomaly scores are assigned by creating numerous trees and calculating the average depths of their branches. Any newly observed data point then follows these trained criteria to traverse down each tree. Equation 2 converts the overall depth of all the branches that the data point travels through into an anomaly score.

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2)$$

E(h(x)) is the average depth reached by a single data point (x) across all trees. The normalising factor, *c(n),* represents the average depth of an unsuccessful search in a Binary Search Tree (BST).

$$c(n) = 2H(n-1) - (2(n-1)/n) \quad (3)$$

H(i) is the harmonic number, which may be approximated using *ln(i) + 0:5772156649* (Euler's constant), and n is the number of points utilized in tree construction.

For more detail on the EIF algorithm and its properties, see [9]. Finally, it is worth mentioning that the EIF, as a tree-based ensemble model, has an inherent structure similar to the RF. However, random selections inside the EIF have a significantly larger influence. This is because, unlike RF, EIF selects properties associated with internal nodes at random rather than using established splitting criteria. Such a problem may appear overwhelming to academics seeking to improve the EIF's interpretability. However, the current study shows that a solution to this problem is indeed possible. *p*-dimensional vectors, where the *jth* component reflects the CFI (for inliers or outliers) of the *j*th feature.

Consider **P***ath* (**x**, *t*) as the trajectory from the root node to the relevant leaf node associated with data point **x** in tree *t*. This inquiry focuses on understanding the CFI update rule for inliers ($\boldsymbol{I}_{\mathrm{I}}$), with the extension to outliers ($\boldsymbol{I}_{\mathrm{O}}$) being a natural progression. Initially, the CFI for inliers, indicated as $\boldsymbol{I}_{\mathrm{I}}$, is initialized as a *p*-dimensional vector of zeros, denoted by **0***p*. The CFI for inliers is then adjusted additively. This update method iterates over all trees in the forest and, for each tree *t* and the subset of predicted inliers I*t*, conduct the following steps.

Finally, for the projected inlier $\boldsymbol{xI} \in \boldsymbol{It}$, the inner nodes in the path ***Path*** (***x****I*, *t*) are iterated. When the splitting attribute associated with a specific internal node *v* is marked as $\boldsymbol{f_j}$, the *j*th element of $\boldsymbol{I}_{\mathrm{I}}$ is updated by adding a unique value.

$$\Delta = \frac{1}{h_t(xI)} . \lambda I(v), \quad (4)$$

The expression $h_t(\boldsymbol{xI})$ denotes the depth of the leaf node linked to data point ***xI*** in tree *t*. This statement describes two variables that lead to DEIFFI. The update rule for $\boldsymbol{I}_O$ is identical to the one given above, but with a few modifications. Instead of iterating over *It*, use O*t*. In addition, $\lambda I(v)$ is substituted by $\lambda O(v)$ during the update.

3.2 DEIFFI

DEIFFI uses two basic hypotheses that guide the evaluation of feature significance:

- *Hypothesis I*: Significant features tend to separate outlier data points from shallower depths for trees constructed with the EIF model. In contrast, normal data points are moved to deeper levels of trees. This hypothesis is based on the intuition that anomalies have particular characteristics that lead to their isolation closer to the roots.
- *Hypothesis II*: Split tests of significant characteristics result in greater imbalance between abnormal data points compared to normal data points. This intuition indicates that salient features make a clearer distinction between abnormal and normal cases, thus aiding detection.

An Extended Isolation Forest (EIF) is trained by generating an ensemble of decision trees, each specialised in isolating anomalies. The basic idea underlying EIF is to establish isolation paths for data points, where the path length indicates the level of isolation inside a tree. While the exact training procedure consists of multiple parts:
For a dataset $D = \{x_1, x_2, \ldots, x_n\}$, where each x_i is a data point of d-dimensional feature vector: For each tree t (t = 1 to T):

- Randomly select a subsample D_t from the dataset D using bootstrap sampling.
- Choose a random feature subset F_t of size f, where f is the number of features to be considered for splitting at each node.
- Select the best split feature j_t and split value v_t based on a splitting criterion.
- Divide the data into left and right subsets: $D_l(x_i$ where $x_i[j_t] \leq v_t)$ and D_r $(x_i$ where $x_i[j_t] > v_t)$.
- If the stopping criteria are met, create a leaf node.
- Otherwise, continue recursively on D_l and D_r to create child nodes.

To understand individual predictions produced by the EIF, this research effort employs an approach similar to that described in [3], with variations due to the challenges in computing particular values in the local case (i.e., when assessing one sample at a time). Specifically:

- Cumulative Feature Importance (CFIs) are computed separately for inliers and outliers, represented as real-valued values. These quantities are subsequently appropriately normalized and merged to generate the ultimate feature importance scores. The CFIs are incrementally updated by utilizing data points in an additive manner, ensuring that the significance of features is accurately captured.
- The update mechanism is based on two metrics that reflect the two aforementioned intuitions: the depth of the leaf node where a given data point terminates (intuition I1) and the Induced Imbalance Coefficient (IIC) associated with a specific internal node (intuition I2).

Further information regarding CFI and GFI algorithm and its properties, can be referred in [3].

4. Experimental Results

This section discusses the outcomes of experiments using synthetic and real-world datasets. These experiments aim to evaluate the efficacy of the DEIFFI version, which generates feature importance ratings that are connected to individual predictions. Furthermore, this study examines the DEIFFI technique utilizing synthetic data to gain insights into the importance of each ingredient. All testing procedures were carried out on a standard consumer laptop with an Intel Core i5-8750H 2.20 GHz CPU and 8 GB RAM.

The DEIFFI approach was used to analyze individual predictions from the EIF [9] model. The DEIFFI approach's effectiveness has been evaluated using synthetic and real-world datasets. Both datasets provide prior knowledge of the most relevant aspects for the AD to address, which is critical for assessing DEIFFI's effectiveness. The experimental arrangement used in this work is based on a real-world scenario that is relevant to a wide range of applications. In this scenario, a trained EIF model is used in online settings, with the user looking for predictions and matching local feature importance scores for each processed particular data point.

4.1 Synthetic Dataset

This study initially examines the performance of the DEIFFI approach in controlled tests on synthetic data. The synthetic dataset was created by first working with two-dimensional data points, that were then enlarged by adding more noise features. Specifically, the generic data point xi is represented as a p-dimensional vector.

$$xi = [\rho \cos(\theta), \rho \sin(\theta), n1, \ldots, np - 2]\ T \quad (5)$$

The variables $nj \sim N(0, 1)$ *for* $j = 1, \ldots, p - 2$ denote white noise samples. The parameters p and θ are random variables with continuous uniform distributions. For regular points of data, the estimated distributions of those variables are organized as follows: $\theta \sim U(0, 2\pi)$; $\rho \sim U(0, 3)$. In contrast, anomalous data points have distributions as follows: $\theta \sim U(0, 2\pi)$ and $\rho \sim U(4, 30)$.

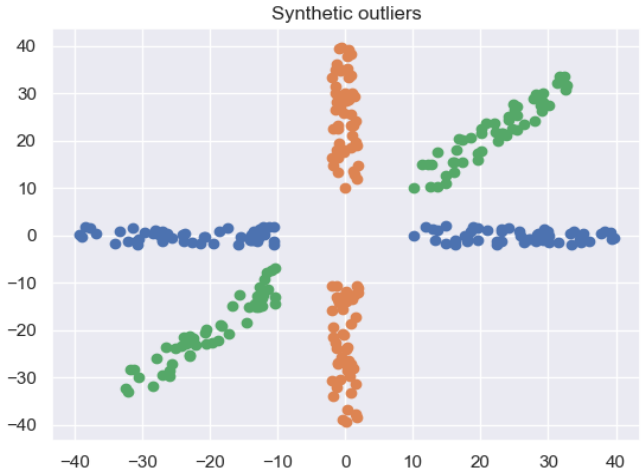

Figure 1. Synthetic outliers projected on the f1 − f2 plane.

For the experiments, a training set of 1000 data points with 6 dimensions was created, including four noise features. Approximately 10% of the data points were identified as abnormalities. Ten instances of the EIF algorithm were trained with 100 trees and a subsampling size of $\psi = 256$, which are standard values for the EIF hyperparameters [9]. The resulting models had an average F1-score of 0.768 on the training data, indicating excellent detection ability. This indicates that the informative features, notably *p cos(θ)* and *p sin(θ)*, are effectively used to address the problem at hand. The ranking of attributes based on these scores matches the prediction, with the coordinates of each point being selected as the two most essential characteristics.

However, removing IIC and focusing solely on projected outliers presents considerable issues. Without the IIC contribution, the distinction between relevant and noise features becomes less obvious, making it impossible to distinguish between them. When only predicted outliers are included, p cos(θ) fails to appear amongst the top two most significant features for three out of ten models, but p sin(θ) does in two out of ten models.

Ignoring the IIC contribution *(with λI (v) = 1)* minimizes the difference between the normalized relevance scores of informative and noisy features, which makes it more difficult to discriminate between them. Similar difficulties arise when only the impact of predicted outliers is examined *(i.e., when II/CI equals 1)*. In this situation, *p cos(θ)* cannot be found amongst the top two most important features for three (out of ten) models, while *p sin(θ)* does not appear for two (out of ten). The F1-score and accuracy on training data are 0.768 and 0.942, respectively. For the training stage, 300 extra ad-hoc anomalies were produced as shown in Figure 1 (projected onto the subspace of important features):

A total of 100 data points is positioned on the *x-axis* (blue points), the *y-axis* (orange points), and the bisector (green points). In the context of this Anomaly Detection (AD) job, it is determined that only feature *f1* is relevant for outliers on the *x-axis*, only feature *f2* is important for outliers on the *y-axis*, and both *f1* and *f2* are significant for outliers on the bisector. All other features, which function as white noise samples, are considered unimportant in all cases. After obtaining predictions for the created test outliers, this study uses the DEIFFI algorithm to calculate local feature importance scores and rankings. *Figure 2* shows color-coded features, with columns denoting the *y-axis*. DEIFFI finds both *f1* and *f2* to be significant in the third row, which is connected to points along the bisector, which is consistent with past findings. Interestingly, it is noted that, in contrast to the situation with outliers on the axes, the feature importance ratings supplied by DEIFFI for *f1* and *f2* are somewhat close in this case.

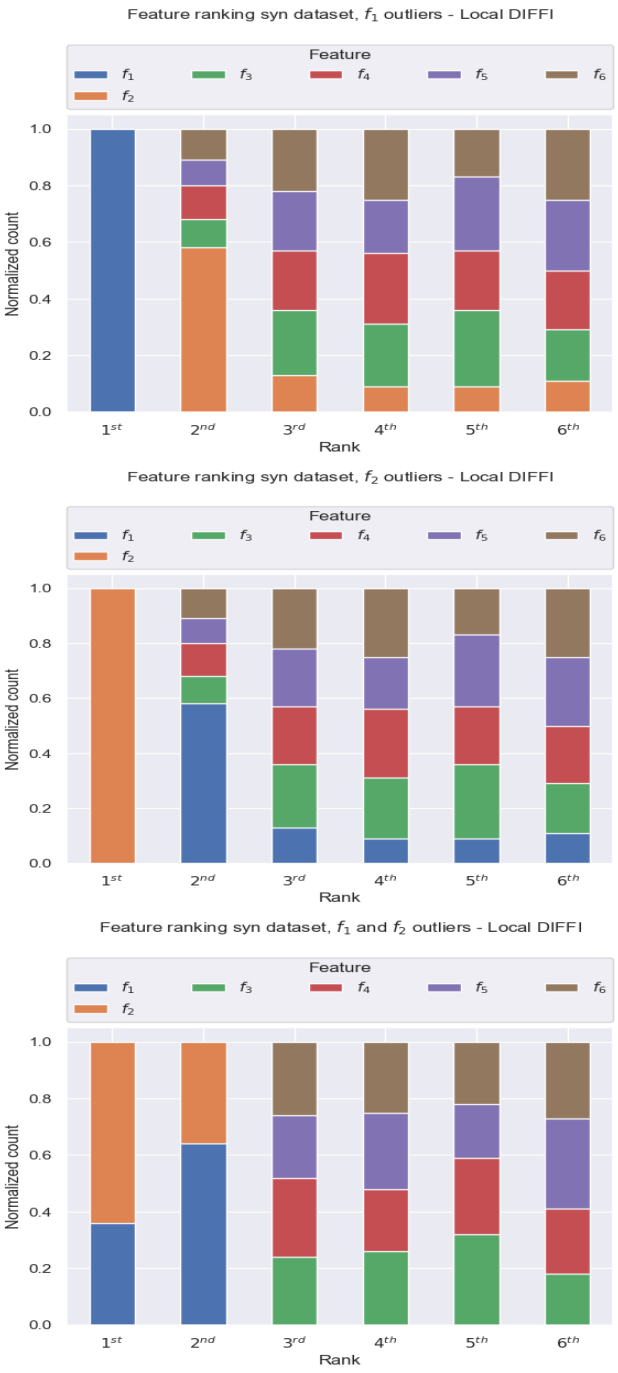

Figure 2 DEIFFI-based feature rankings for the synthetic dataset: outliers on the x-axis, y-axis, and bisector.

The height of each bar related to a feature shows the percentage of expected anomalies in which that feature has a particular rank. For instance, in the upper left plot of Figure, feature $f1$ consistently ranks first for all projected anomalies. Feature $f2$ has rank 2 for around 40. The first column (the most important feature according to the interpretability technique) in the first two rows (where correctly predicted outliers are taken into account) always matches the accurate feature, that is, *f1* for outliers on the *x-axis* and *f2* for outliers on the *y-axis*. DEIFFI finds both *f1* and *f2* to be significant in the third row, which is connected to points along the bisector, which is consistent with past

findings. Interestingly, it is noted that, in contrast to the situation with outliers on the axes, the feature importance ratings supplied by DEIFFI for *f1* and *f2* are somewhat close in this case.

4.2 Real-World Datasets

This study makes use of an altered version of the Glass Identification UCI dataset (referred to as [glass]), which was first created for multiclass classification tasks. There are 213 glass samples in the collection, and each one is represented by a 9-dimensional feature vector. One of these characteristics is the refractive index (RI), while the others stand for the concentrations of calcium (Ca), iron (Fe), sodium (Na), potassium (K), magnesium (Mg), silicon (Si), and barium (Ba). To create the class of regular data points, classes 1, 2, 3, and 4 (window glass) are combined in the completed tests. The three remaining classes ‘‘glass containers (class 5), glass dinnerware (class 6), and glass headlights (class 7)” are considered anomalous data points because they do not include window glass. The assessment is centred on evaluating the performance of DEIFFI using anticipated outliers from class 7, which serve as the test data points.

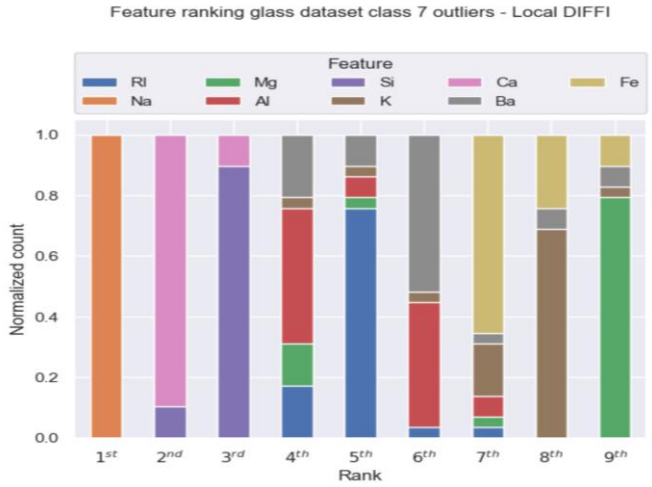

Figure 3 Based on DEIFFI scores, the glass dataset’s feature rankings are as follows: class 7 outliers (headlamps glass).

Drawing on previous research on headlamp glass, two key characteristics are considered:

The calcium concentration, which is used for the reflecting coating, and the sodium concentration, which is involved in the properties that make the glass heat resistant. These characteristics are expected to play a significant role in differentiating headlamp glass from window glass. A subsampling ratio of $\psi = 64$ and 100 trees were used to train an EIF instance.

Table 1. Results of Algorithm on Real-World Datasets

Evaluation Matrix	Original Method	Proposed Method
Accuracy	0.870	0.914
Precision	0.469	0.607
Recall	0.682	0.773
F1-Score	0.556	0.680

Table 1 displays the performance of the proposed approach with improved interpretability to quantify the feature importance for AD quantitatively using the EIF algorithm. The study of the training data yielded encouraging results: an F1-score of 0.68, accuracy of 0.914, precision of 0.607, and recall of 0.773 indicate that the model can correctly identify important patterns and forecast events. Figure 3 presents the findings from the analysis. In line with the accepted knowledge for the task, DEIFFI regularly identifies salt and calcium concentrations as the most significant features for the bulk of predicted abnormalities.

In summary, the statistical evaluation of DEIFFI on the EIF model demonstrated good accuracy and a high F1-Score. These findings highlight the DEIFFI methodology’s effectiveness in accurately detecting and classifying abnormalities, especially in the context of AD. By decreasing the number of false positives and false negatives, the model shows that it can support strong threat mitigation and progress the AD landscape.

6. Conclusion and Future Scope

This research work presents Depth-based Extended Isolation Forest Feature Importance (DEIFFI), an approach that enhances interpretability in the context of the EIF, which is a popular and effective AD algorithm. By providing insightful information, DEIFFI empowers end users of EIF-based AD solutions to gain deeper insights into the underlying process and facilitate root cause analysis. This promotes a better understanding of anomalies and enhances the overall interpretability of the EIF algorithm.

Surprisingly, DEIFFI attains these praiseworthy outcomes with significantly reduced computational costs, making it especially appealing for practical operational deployments, such as real-time situations. Furthermore, DEIFFI exhibits its ability to support unsupervised feature selection, which promotes the development of computationally efficient and maybe more accurate AD solutions. This demonstrates the concrete benefit of DEIFFI in enhancing the potential of the EIF algorithm and increasing its applicability in various AD settings.

Authors’ Contributions

Rahul Singh: oversaw the research framework's development and led the study design. conducted interviews, managed the analysis of quantitative data, gathered and evaluated qualitative data, and produced visual representations of the data.

Deepti Gupta: offered theoretical insights, carefully examined and rewrote the work for key intellectual elements and gave her approval before submitting the finished version.

Acknowledgements

This research has received no external funding.

References

[1] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp.**248–266, 2022.**

[2] G´erard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25, pp.**197–227, 2016.**

[3] Mattia Carletti, Matteo Terzi, and Gian Antonio Susto. Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. *Engineering Applications of Artificial Intelligence*, 119:105730, **2023.**

[4] Chengjie Chen, Hao Chen, Yi Zhang, Hannah R Thomas, Margaret H Frank, Yehua He, and Rui Xia. Tbtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular plant*, Vol.**13,** Issue.**8**, pp.**1194–1202, 2020.**

[5] Zhiguo Ding and Minrui Fei. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings* Vol.**46**, Issue.**20**, pp.**12–17, 2013.**

[6] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, **2017.**

[7] Timo Freiesleben, Gunnar König, Christoph Molnar, and Alvaro Tejero-Cantero. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv preprint arXiv:2206.05487, **2022.**

[8] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI magazine*, Vol.**40**, Issue.**2**, pp.**44–58, 2019.**

[9] Sahand Hariri, Matias Carrasco Kind, and Robert J. Brunner. Extended isolation forest. *IEEE Transactions on Knowledge and Data Engineering*, Vol.**33**, Issue.**4**, pp.**1479–1489, 2021.**

[10] Abderrahim BENI Hssane and Moulay Lahcen. Improved and balanced leach for heterogeneous wireless sensor networks. *IJCSE International Journal on Computer Science and Engineering*, Vol.**2**, Issue.**8**, pp.**2633–2640, 2010.**

[11] Vladislav Ishimtsev, Alexander Bernstein, Evgeny Burnaev, and Ivan Nazarov. Conformal *k*-nn anomaly detector for univariate data streams. In *Conformal and Probabilistic Prediction and Applications*, pages 213–227. PMLR, **2017.**

[12] Pawe-l Karczmarek, Adam Kiersztyn, Witold Pedrycz, and Ebru Al. K-means-based isolation forest. *Knowledge-based systems*, 195:105659, **2020.**

[13] Edwin M Knorr, Raymond T Ng, and Vladimir Tucakov. Distance-based outliers: algorithms and applications. *The VLDB Journal*, Vol.**8**, Issue.**3**, pp.**237–253**, **2000.**

[14] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.**444–452, 2008.**

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp.**413–422, 2008.**

[16] Lorenzo Meneghetti, Matteo Terzi, Simone Del Favero, Gian Antonio Susto, and Claudio Cobelli. Data-driven anomaly recognition for unsupervised model-free fault detection in artificial pancreas. *IEEE Transactions on Control Systems Technology*, Vol.**28**, Issue.**1**, pp.**33–47, 2020.**

[17] Hla Yin Min and Win Zaw. Performance evaluation of energy efficient cluster-based routing protocol in wireless sensor networks. *International Journal of Computer Science Engineering IJCSE*, Vol.**3**, Issue.**2**, pp.**71–76, 2014.**

[18] KM Archana Patel and Prateek Thakral. The best clustering algorithms in data mining. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pp.**2042–2046, 2016.**

[19] Andrew Pavlo, Gustavo Angulo, Joy Arulraj, Haibin Lin, Jiexi Lin, Lin Ma, Prashanth Menon, Todd C Mowry, Matthew Perron, Ian Quah, et al. Self-driving database management systems. In *CIDR*, Vol.**4**, pp.**1**, **2017.**

[20] Luca Puggini and Se'n McLoone. An enhanced variable selection and isolation forest-based methodology for anomaly detection with oes data. *Engineering Applications of Artificial Intelligence*, 67: pp.**126–135, 2018.**

[21] Guillaume Staerman, Pavlo Mozharovskyi, Stephan Cl´emen¸con, and Florence d'Alch´e Buc. Functional isolation forest. In *Asian Conference on Machine Learning*, pp.**332–347. 2019.**

[22] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov), pp.**45–66, 2001.**

[23] Ke Wu, Kun Zhang, Wei Fan, Andrea Edwards, and S Yu Philip. Rs-forest: A rapid density estimator for streaming anomaly detection. In *2014 IEEE international conference on data mining*, pp.**600–609, 2014.**

[24] Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. Dnn-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL international conference on advances in geographic information systems*, pp.**1–4, 2016.**

AUTHORS PROFILE

Rahul Singh earned his B.Tech. in Computer Science and Engineering from Atal Bihari Vajpayee Government Institute of Engineering and Technology, Pragatinagar, Shimla, Himachal Pradesh, India in 2020; M.E in Computer Science and Engineering for University Institute of Engineering and Technology, Panjab University, Chandigarh, India in 2023 specialising in Data Mining and Machine Learning Technologies. He worked as an intern in RAPS iTech, Chandigarh, India as a Machine Learning Engineer in year 2020.

Deepti Gupta Deepti Gupta received her BE in Computer Science and Engineering from University of Jammu, Jammu and Kashmir, India in 2006; MTECH in Computer Science and Engineering from National Institute of Technology, Jalandhar, Punjab, India in the year 2009 and PhD in Computer Science and Engineering from National Institute of Technology, Jalandhar, Punjab, India in the year 2015. She worked as Assistant Professor in the department of Computer Science and Engineering, National Institute of Technology, Delhi in the year 2014. She is currently working as Assistant Professor in the department of Computer Science and Engineering, University Institute of Engineering and Technology, Panjab University, Chandigarh, India. Her professional research activity lies in the field of wireless sensor networks and data mining. She has published 20 research papers in the International Journals/Conferences. She has supervised 06 M.E theses and is currently supervising 03 PhDs. She is Life Member of Advanced Computing & Communications Society, Indian Institute of Science, Bangalore, India, (L6233A1523472) and Indian Society for Technical Education (I.S.T.E.), New Delhi, India, (LM 110527).

