IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 June 2024, accepted 18 July 2024, date of publication 22 July 2024, date of current version 31 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3432174

RESEARCH ARTICLE

Isolation Forest With Exclusion of Attributes Based on Shapley Index

ALBERT RACHWAŁ, PAWEŁ KARCZMAREK, ALICJA RACHWAŁ, AND RAFAŁ STĘGIERSKI
Department of Computational Intelligence, Lublin University of Technology, 20-618 Lublin, Poland
Corresponding author: Albert Rachwał (a.rachwal@pollub.pl)

ABSTRACT Recognizing anomalies is an extremely important process in data analysis, aimed at identifying patterns in data that deviate from known norms or typical standards. These anomalies are often indicative of significant, and sometimes critical, issues such as fraud, network intrusions, and system failures. Traditional anomaly detection algorithms primarily focus on the attributes of individual observations within a dataset, typically establishing a 'normal' profile and flagging deviations from this profile as anomalies. This paper introduces an innovative enhancement to the Isolation Forest algorithm, a renowned method for anomaly detection known for its effectiveness and efficiency, especially in large datasets. The Isolation Forest algorithm operates by randomly partitioning the data space and constructing a binary tree, where the oddity score of a data point is ascertained based on its separation from the extremity to the base of the structure, enabling the autonomous detection of outliers in a completely unsupervised manner. The methodology presented in the paper is based on repeatedly building Isolation Forest models on datasets from which individual attributes are excluded. In our research, we used the SHAP (SHapley Additive exPlanations) method which comes from game theory and is used to determine the impact of individual features on the result of the model. When training the Isolation Forest on the full dataset, the SHAP method is used to obtain the coefficients of influence of model attributes on the prediction result. Both negative and positive influences are considered significant when counting the anomaly score. On the foundations of the results from all sub-models, a weighted average is calculated, to which weights are calculated based on the SHAP model. The comparative analysis of evaluation metrics revealed a substantial enhancement attributed to the implemented methodology. The metrics used for evaluation have shown improvement in most cases from 3.5 to 6 percent point. One of the metrics have shown an improvement of 12 percent. Obtained results demonstrate that this integrated approach not only enhances the prediction accuracy of the Isolation Forest algorithm but also offers a more interpretable understanding of the data. This advancement in anomaly detection methodology promises significant implications for various fields where quick and accurate detection of outliers is paramount.

INDEX TERMS Anomaly detection, outliers detection, isolation forest, shapley index, attribution exclusion.

I. INTRODUCTION

With technological advancement, society seeks to anticipate situations that diminish the quality of life. Early disease diagnosis and the foresight of potential electronic equipment malfunctions can prevent detrimental health and life risks [1]. To address such kind of challenges, computer science researchers develop anomaly detection algorithms to identify outlier observations. These algorithms usually process unlabeled data, termed as unsupervised anomaly detection. Two primary assumptions guide anomaly detection: Anomalies are rare in the dataset, and their features significantly diverge from normal instances.

Several methods exist for anomaly detection, including density-based [2], [3], [4], [5], cluster-based [6], Bayesian-network [7], and neural network techniques [8], [9], [10], [11]. The isolation forest method, a popular topic of discussion, offers diverse applications, ranging from spam

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

filters [12], social network spam detection, [13], and faulty detection [15] to various applications in medicine [16], [17].

Isolation Forest (IF) [18], [19], is an anomaly detection algorithm, utilize binary trees for detecting anomalies. Unlike traditional outlier detection methods, which focus on measuring distances between points or belonging to particular areas of a dataset based on the density of observations, Isolation Forest isolates outlier elements by randomly selecting features and dividing the values until the individual observations are isolated. The IF method randomly selects an attribute and then chooses a random division value from the range between the smallest and largest instances of that selected feature. The set space is consistently segmented into subsets based on randomized partition values. These steps are repeated until a single observation is isolated. The number of splits needed to isolate an observation indicates its anomaly; fewer splits indicate an outlier observation.

This algorithm boasts linear time complexity and minimal memory requirements, accommodating large datasets. Its rise in popularity stems from its unsupervised nature, capable of handling unlabeled data sets. The algorithm divides the data space with lines parallel to the standard basis, assigning higher anomaly scores to data points requiring fewer splits for isolation. The Isolation Forest algorithm has been implemented across a multitude of industrial sectors, affirming its versatility and effectiveness. Its applications extend to anomaly detection in deepwater drilling data [20], fault identification in wind turbines [21], solar power plants [22], [23] and nuclear power facilities [24]. The flexibility of the Isolation Forest is further illustrated by its adoption in blockchain technology, particularly in the detection of cyber attacks in real-time blockchain transactions [25]. It is also employed to pinpoint anomalies in wireless sensor networks [26] and to undertake the expansive task of detecting anomalous behavior and patterns [27], [28], [29]. In the medical field alone, there are some examples of applications like predicting chronic kidney disease [30], [31], [32], [33].

To compare our propose, we have considered popular anomaly detection techniques such as Isolation Forest, HDBSCAN, neural networks, and autoencoders. They offer diverse approaches and, depending on the dataset, yield good results. However, sometimes it is beneficial to look at the anomaly detection problem from a broader perspective so that a single iteration of the algorithm on the data becomes part of a more comprehensive methodology. We sought efficient descriptive methods that could be used to gain insight into how anomaly detection models operate, and consequently, which components of the analyzed data are particularly important to the model's outcome, while others have no impact or a negative impact on the result. Some known descriptive methods do not have direct application in anomaly detection task but only in multi-class classification. However, a method that have proven to work excellently with tree-structured algorithms is SHAP (SHapley Additive exPlanations). SHAP is a method for explaining models based on game theory [34]. It uses Shapley values - a concept in game theory for assessing the 'value' of a participant's contribution in a multiplayer game - to ascertain the significance of features in predictive models. SHAP values operate under the premise that every feature entered into the model contributes to the prediction to some extent. SHAP calculates the importance of each feature by considering every possible combination of features and measuring the effect of adding or removing a feature on the prediction score. The above approach gives intuitive as well as clear explanations of the model. In the quest for interpretable machine learning models, the SHAP model [35] emerges. It assigns dataset attribute values, reflecting their influence on the final prediction during the model's learning phase. SHAP is now widely used in medicine [36], [37] and computer science, including anomaly detection [38].

In the realm of anomaly detection, the SHAP model collaborates with machine learning models [39], [40]. This collaboration aids in understanding feature impacts on model outputs and occasionally guides maintenance engineers on which factors deserve keen focus to prevent future issues [41], [42]. Some studies have even utilized the tree SHAP explanatory model, which has also been employed in this work to elucidate the Isolation Forest model [43].

Although Isolation Forest currently has a plurality of applications, it is being intensively studied for possible improvements. One avenue of development may be to note that IF focuses on isolating individual observations of a dataset without paying attention to attribute properties. Several ways to enhance the effectiveness of the isolation forest by modifying the tree structure's construction method have already been developed [46], [48], but to the best of our knowledge none of them involved interfering with the features of the observations. Some Isolation Forest development methods try to modify the way the tree is created, such as using trees based on the $k - means$ algorithm [44] or minimal spanning tree algorithm [45] instead of binary trees. There are also innovative works focusing on extending the isolation forest technique with methods based on fuzzy logic and granular computing [47].

Many classic anomaly detection methods, not based on neural networks, give good results and high efficiency when used in modern industrial problems, but they still give room for development by creating larger, more complex anomaly detection systems that utilize aggregation techniques to use data from different models to maximize the results. Therefore, descriptive analysis methods such as SHAP presented in this work can be an excellent tool that, in cooperation with existing techniques, can serve as feedback, providing "positive reinforcement" by maximizing the impact of positive features and minimizing the impact of negative features.

The motivation for this work is to develop the anomaly detection technique using binary trees, known as Isolation

Forest, by combining it with descriptive model techniques, exemplified by the SHAP model. The aim of the study is to demonstrate that the feedback provided to the algorithm by the descriptive model during the learning phase can be utilized in a way that takes into account the impact of data on the result, enhancing the final prediction and leading to greater accuracy and improved algorithm performance. Additionally, other goals that guided the creation of this work include developing new expansions of the Isolation Forest method characterized by higher metrics for result evaluation, and gaining a deeper understanding of dataset properties to advance methods that reveal significant features within the studied field. In this work, the SHAP method is employed to assess the impact of dataset characteristics on the Isolation Forest (IF) model's final output during its training phase. The IF model is trained for each scenario where a single attribute is omitted from the dataset. Following this, the anomaly scores obtained from the individual models are averaged based on the weights derived through SHAP. This approach, where the prediction is performed on the averaged vector, consistently yields better results compared to the standard baseline IF method.

It is worth noting that we assume that all these methods that modify the way of building a forest can be adapted for use with the descriptive method based on SHAP presented in this article, because SHAP approaches the examined Isolation Forest model as a black box from which it obtains information about the importance of individual attributes on prediction results using SHAP.

The structure of the article is as follows: Section II delves into the theory of the Isolation Forest and the SHAP INDEX operation. Section III then presents a pioneering approach that merges SHAP and Isolation Forest to enhance anomaly detection. Subsequently, Section IV elucidates the experimental results and draws conclusions. The final section, Conclusions, wraps up the findings, highlighting potential future trajectories and the novel applications of the proposed approach.

II. THEORETICAL FOUNDATIONS OF SHAP AND ISOLATION FOREST

In this section, we will delve more deeply into the scientific theory underpinning the Isolation Forest algorithm as well as its association with Shapley values. The isolation forest algorithm used for anomaly detection consists of two stages. In the first stage, a set of binary trees is constructed based on the data. In the second stage, operations are performed to read the properties of individual trees, specifically the path lengths from the root to the leaf. Based on this information, the anomaly score is determined. The SHAP model is an approach derived from game theory, which serves to determine the influence of players on the outcome of a game. It is based on Shapley values. In the context of machine learning, it can be directly applied by treating the model's attributes as players in the game and the model's output as the game's result. Additionally, the TreeExplainer model, which is utilized for computing Shapley values, will be described.

A. THEORETICAL BACKGROUND OF ISOLATION FOREST

Consider a node T in an isolation tree. Such a node can be categorized as either an external node, devoid of children, or an internal node characterized by a singular condition and precisely two child nodes, denoted as (T_l, T_r). The defining condition of an internal node involves an attribute q and a corresponding split value p, orchestrating a division of data points into T_l and T_r based on the criterion $q < p$.

In the construction of an isolation tree from a dataset $X = \{x_1, \ldots, x_n\}$ comprising n instances, a recursive partitioning strategy is employed. This involves random selection of an attribute q and a split value p, which continues until the emergence of one of the following scenarios:

(i) attainment of a predefined tree height,
(ii) reduction of X to a singular element,
(iii) uniformity in the attribute values within X.

An isolation tree exemplifies a proper binary tree structure, each node of which either has no children or exactly two. Under the assumption of distinct instances, each instance becomes isolated in an external node upon full maturation of the isolation tree. Consequently, the count of external nodes equals n, while that of internal nodes is $n - 1$, yielding a total node count of $2n - 1$. This structure ensures a linear memory requirement relative to n.

The primary objective in anomaly detection is to generate a ranking reflective of anormality. An effective approach entails ordering data points based on their path lengths or anomaly scores, with top-ranked points signifying anomalies. The definitions of path length and anomaly score are as follows. The path length $h(x)$ of a data point x is determined by the quantity of edges traversed in an isolation tree from the root to an external node. The challenge in anomaly score formulation lies in the variance between the maximum potential height of the isolation tree, which scales with n, and the average height, which scales with $\log n$. Consequently, normalizing $h(x)$ with either of these metrics presents limitations in either bounding or comparability.

Given the structural parallel between Isolation Trees and Binary Search Trees (BST), the estimation of average $h(x)$ at the termination at external nodes mirrors the unsuccessful search scenario in a BST. Adopting analytical techniques from BST, the average path length in an isolation tree for a dataset of n instances is estimated as [50]:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{1}$$

where $H(i)$ represents the harmonic number, approximated by $\ln(i) + 0.5772156649$ (Euler's constant).

Utilizing $c(n)$ as the average benchmark for $h(x)$, the anomaly score for an instance x is delineated as:

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

where $E(h(x))$ indicates the mean path length across multiple isolation trees. In this context:

- As $E(h(x))$ approaches $c(n)$, s gravitates towards 0.5.
- When $E(h(x))$ is minimal, s trends towards 1.
- Conversely, as $E(h(x))$ nears $n-1$, s approaches 0.

Employing this anomaly score, s, facilitates the following assessments:

(a) Instances yielding s values close to 1 are unequivocally anomalous.
(b) Instances with s significantly below 0.5 can be safely classified as normal.
(c) A uniform distribution of s around 0.5 across all instances indicates an absence of pronounced anomalies within the sample.

In the context of an Isolation Forest, which integrates multiple isolation trees:

(a) Anomalies are identified as data points exhibiting shorter path lengths.
(b) The ensemble of trees serves as a collective of 'experts', each targeting distinct anomalies.

B. SHAPLEY ADDITIVE EXPLANATIONS THEORY

Originating from game theory, the Shapley value method is a classic technique that equitably allocates the total payoff of a cooperative game among its participants [49]. Formally conceptualized, a cooperative game comprises a set of players, $\mathcal{M} = \{1, \ldots, M\}$, collectively referred to as *the grand coalition*. The game is characterized by a set function, $v : 2^{\mathcal{M}} \to \mathbb{R}$, in which $v(S)$ represents the payoff attributable to any coalition $S \subseteq \mathcal{M}$, with the assumption that $v(\emptyset) = 0$.

The computation of Shapley value for player i, denoted as $\Phi_i(v)$, involves a weighted mean of the player's incremental contributions across all conceivable coalitions:

$$\Phi_i(v) = \frac{1}{M} \sum_{S \subseteq \mathcal{M} - \{i\}} \binom{M-1}{|S|}^{-1} \left(v(S \cup \{i\}) - v(S)\right). \quad (3)$$

Notably, Shapley method is distinguished as the sole approach that satisfies four axiomatic principles: Dummy, Symmetry, Efficiency, and Linearity. This unique confluence of axioms renders it a robust and equitable metric for assessing contributions.

C. SHAP IN MACHINE LEARNING

Within the domain of machine learning, SHAP (SHapley Additive exPlanations) values play a pivotal role in evaluating the significance of each feature within a predictive model. Specifically, for a feature denoted as j, the corresponding SHAP value Φ_j quantifies the relative influence of the j-th feature on the prediction outcome of a specific instance, compared against the average prediction across the dataset [51].

The computation of SHAP values includes an exhaustive analysis of the model's predictive response across every feasible permutation of feature combinations. As the feature set expands, this computational task escalates in complexity, often exponentially. In scenarios where models incorporate a substantial number of features, the precise calculation of SHAP values becomes impractical. Consequently, approximation methodologies are frequently adopted. [52] propose an approximation technique employing Monte-Carlo sampling. Assume that x is the data point, z is a randomly selected data point from the dataset, M is the total number of samples, and j is the feature index for which we are computing the SHAP value. The vector x^m_{+j} represents the instance where feature j is taken from x and the remaining features are taken from z. Conversely, x^m_{-j} is similar to x^m_{+j} but includes the feature j from the sampled point x^m_j. using the notation introduced above, the formula then has the following form:

$$\hat{\Phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}(x^m_{+j}) - \hat{f}(x^m_{-j})\right), \quad (4)$$

where $\hat{f}(x^m_{+j})$ signifies the prediction model's output when certain feature values are substituted with those from a randomly selected data point z, with the exception of feature j. The vector x^m_{-j} closely resembles x^m_{+j}, however, it additionally incorporates the value x^m_j derived from the sampled point. In essence, SHAP values, which are extrapolated from Shapley values in game theory, offer an intuitive and transparent framework for dissecting the influence of individual features within a machine learning model.

The SHAP TreeExplainer [53] is an algorithmic framework designed to offer interpretable explanations for predictions made by machine learning models, particularly those based on tree structures such as decision trees, random forests, and gradient boosting. It employs a polynomial-time algorithm to compute SHAP values efficiently, bypassing the need for the exponential-time computations previously required.

Fundamentally, the TreeExplainer works by tracing the decision paths in a tree, assessing the contribution of each feature to the final prediction by evaluating the change in prediction probability conditioned on the feature's presence or absence.

This computational model operates in $O(TLD^2)$ time, where T is the number of trees, L is the maximum number of leaves in any tree, D is the depth of the tree. This signifies a substantial reduction in complexity from the exponential time that exact computation of SHAP values would require, making it feasible to interpret even large ensemble models.

III. PROPOSED METHODOLOGY

The innovative methodology proposed in this study is grounded in the premise that when constructing a model for anomaly detection in a dataset, certain attributes hold greater potential than others to yield accurate predictions. This potential is quantifiably measured using the values of the assessment metrics applied for the final prediction evaluation through the model. This is confirmed by the fact that attributes to which higher weight values have been assigned reflect a stronger impact on the model's outcome, as evidenced by an increase in the values of the used

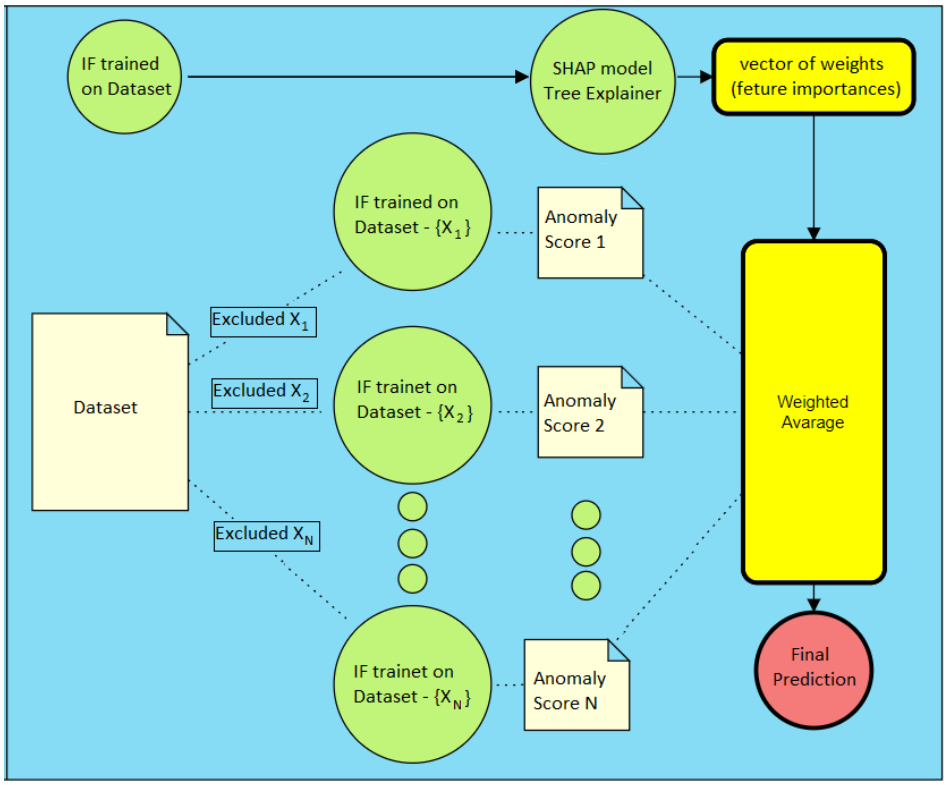

FIGURE 1. A graph presenting the methodology used in a simplified manner.

metrics indicating higher accuracy in the detected anomalies. Crucial to enhancing the algorithm's performance is a deep understanding of the processes inherent in a given field being studied. Such comprehension not only facilitates a better grasp of the subject area but also sheds light on the influence of various characteristics on the associated processes.

By focusing on attributes with higher weights, researchers can delve deeper into the nuances of these features, exploring their interrelations and the mechanisms by which they exert their influence. This nuanced understanding is vital for refining the model, as it allows for the adjustment of the algorithm to better capture the complexities of the data. As a result, the model becomes more adept at not only identifying anomalies but also at providing insights into why certain data points are classified as such.

A graph illustrating the methodology proposed in this work is presented in the Figure 1. In each dataset under examination, the SHAP method is employed to derive a vector that encapsulates the influence coefficients, quantitatively expressing how significantly each attribute affects the operational efficiency of the model. Initially, for the main analysis, an isolation forest model is specifically trained on a substantial portion, amounting to 80% of the training dataset. This model's purpose is to identify and quantify anomalies within the data. Following this, the trained Isolation Forest model is directly subjected to the SHAP framework's Tree-Explainer, employing Tree SHAP to elucidate the model's decision-making processes. For the purpose of this study, the Isolation Forest model provided by the scikit-learn library is employed, operating with a contamination factor set at 10%. In our experiments, it has been shown that the algorithm performs slightly better when focusing solely on the strength of a feature's impact, rather than its positive or negative direction. Consequently, the results were transformed into absolute values and appropriately scaled so that the number assigned to each feature reflects its proportion, i.e., the share of each element in the total sum, which can later be directly translated into probabilities.

Once trained on the isolation forest model, the Tree-Explainer is capable of calculating SHAP values for new data. Therefore, we provide it with the 20% test dataset to generate SHAP values for this subset. This process results in the determination of a weight vector, which fundamentally represents the impact of each attribute on the model's decision-making process, with a particular focus on the test data. This methodical approach ensures a comprehensive and thorough understanding of the model's behavior, particularly highlighting how each feature influences the identification of anomalies in new, unseen data. The next step involves repetitively training the isolation forest model on a modified dataset, where each iteration excludes a different attribute. Subsequently, the decision functions that each model yields for all observations in the set are recorded. With all the partial anomaly scores collected, a weighted average is

Algorithm 1 Methodology for Anomaly Detection Using Isolation Forest and SHAP

Data preparation:
for each column X_i of the data set D **do**
    Save X_i as a separate file/data frame

Building the Isolation Forest model:
Model_IF ← TrainIsolationForest(D)

Using the SHAP model:
SHAP_Weights ← SHAP(Model_IF)

Detecting anomalies:
for each column X_i from D **do**
    $D\_minus\_Xi \leftarrow D - \{X_i\}$
    Model_IF ← TrainIsolationForest(D_minus_Xi)
    Anomaly_Scores[i] ← Model_IF.Predict(D_minus_Xi)

Aggregating results:
Final_Scores ← WeightedAverage(Anomaly_Scores, SHAP_Weights)

Final prediction:
Predictions ← DetermineAnomalies(Final_Scores)

Evaluating results:
Evaluate(Predictions, metrics=["AUC","Accuracy", "Balanaced accuracy", "F1", "PRAUC", "Precision", "Recall])

calculated using the SHAP values. This process aggregates the decision functions of models trained on the modified datasets, resulting in a consolidated value. Based on this aggregated value, the final prediction is conducted. The methodology described above is also presented using pseudo code in Alghoritm 1. n our algorithm, we scaled the obtained anomaly scores to be within the range [-1; 1]. We then set a threshold value at 0 to separate normal observations from anomalies, where achieving a threshold of 0 is still treated as a normal value, despite being on the boundary.

Instead of Tree Explainer, the KernelExplainer, another component of the SHAP framework can be used. It requires the decision function of the model and the training set records as inputs rather than the constructed model itself. While KernelExplainer is a more universal model, offering broad applicability across various algorithms, for tree-based structures like Isolation Forests, the TreeExplainer is more apt due to its specialized design that ensures enhanced interpretability tailored to the intrinsic workings of tree ensembles.

IV. EXPERIMENTAL RESULTS

This chapter presents the results of the performed experiments.

A. DESCRIPTION OF USED DATASET

In this work, we used data sets popular in the field of data science, available publicly on the Internet, which are specially prepared for the task of detecting anomalies. Each of these sets has been marked for anomalous observations by experts, adding labels indicating "0" for normal observations and "1" for outstanding observations. These datasets include Breastw, Cardio, Cover, Ionosphere, Letter, Mammography, MNIST, Mulcross, Musk, Optdigits, Pima, Satellite, Satimage-2, Seismic-Bumps, Thyroid, Vertebral, Vowels, WBC, Wine, and Ecoli, totaling 20 distinct datasets. Details about each of these datasets are described in the Table 1.

TABLE 1. Details of the datasets used in experiments.

Dataset	Records	Attributes	Anomalies	Anomalies [%]
Breastw	683	9	239	34.99
Cardio	1831	21	176	9.61
Cover	286048	10	2747	0.96
Ecoli	336	7	9	2.68
Ionosphere	351	33	126	35.90
Letter	1600	32	100	6.25
Mammography	11183	6	260	2.32
Mnist	7603	100	700	9.21
Mulcross	262144	4	26214	10.00
Musk	3062	166	97	3.17
Optdigits	5216	64	150	2.88
Pima	768	8	268	34.90
Satellite	6435	36	2036	31.64
Satimage-2	5803	36	71	1.22
Seismic-bumps	2584	18	170	6.58
Thyroid	3772	6	93	2.47
Vertebral	240	6	30	12.50
Vowels	1456	12	50	3.43
Wbc	378	30	21	5.56
Wine	129	13	10	7.75

The selection of 20 datasets is significant, as this number provides a robust basis for evaluating the real-world efficacy of the model. Such a diverse and substantial collection of datasets ensures that the observed improvements in the model's performance are not confined to a narrow set of data characteristics but are broadly applicable across various types of data. This extensive range of datasets enhances the reliability of the conclusion that the new methodologies applied to the classic isolation forest algorithm contribute to a statistically significant improvement in anomaly detection.

In these datasets, every observation is labeled to indicate whether it is an anomaly or not. However, for the purpose of training the algorithm in this experiment, these labels were removed to simulate an unsupervised learning scenario. This approach aligns with real-world situations where anomaly labels are often unavailable, and the algorithm must learn to identify anomalies without this guidance. This methodology reflects a common practice in the field, where the primary goal is to allow the algorithm to independently discover patterns and irregularities that signify anomalous behavior. By employing datasets commonly used in the scientific community, this study adheres to the standards of reproducibility and comparability, enabling other researchers to validate and compare the results with their own findings.

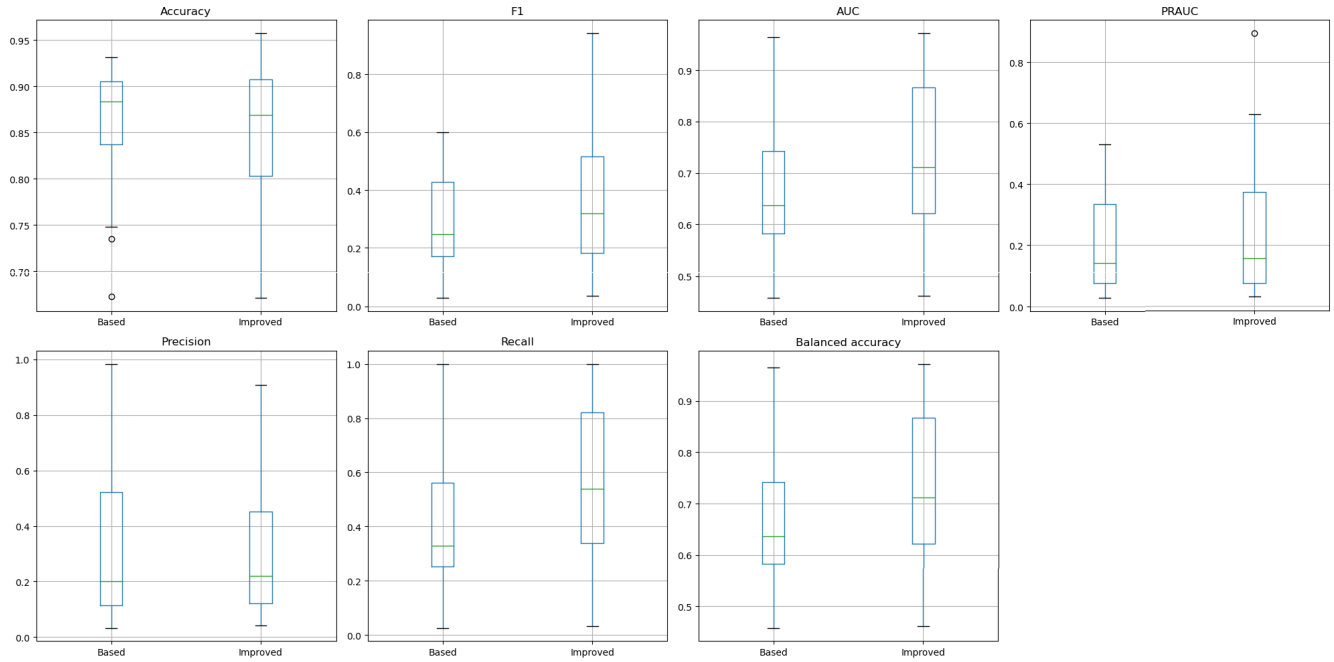

FIGURE 2. Metrics comparison using boxplots.

TABLE 2. Comparison of metrics used in base and improved model.

Metric	Base IF	Improved IF
AUC	0.682	**0.733**
Accuracy	0.856	**0.859**
Balanced accuracy	0.682	**0.733**
F1	0.283	**0.343**
PRAUC	0.215	**0.249**
Precision	**0.336**	0.315
Recall	0.436	**0.565**

B. METHODS COMPARISON

Table 2 presents a comparative analysis of seven performance metrics across 20 datasets, providing mean scores for both the base and improved models. A systematic examination of the data reveals several key findings. The improved model demonstrates a superior mean performance in five out of seven metrics when compared to the base model. This is particularly evident in the AUC (Area Under the Curve) and balanced accuracy metrics, where the improved model exhibits a mean increase of approximately 5.1 percentage points.

A notable increase in the F1 Score by 6 percentage points for the improved model suggests a more balanced performance in terms of precision and recall. This implies that the improved model is better at managing the trade-off between the false positives and false negatives.

While the improved model shows a slight decrease in precision by 2.1 percentage points, it concurrently presents a significant improvement in recall by approximately 12.9 percentage points. This suggests that the improved model is more sensitive and able to identify a higher number of true positive cases, even at the expense of a slight increase in false positives. Table 3 compares all 7 metrics used across 20 datasets. The values in the table represent the differences in average scores that the improved model achieves compared to the base model. Positive values signify superior performance by the improved model, whereas negative values suggest the base model fares better. The final two entries consolidate these results into average and median values. Notably, the median value is consistently positive, and the mean value also remains positive with the exception of the precision metric. Table 4 itemizes the four principal metrics and delineates the corresponding values for both the baseline and the enhanced models. The concluding records present the mean and median of these metrics. The data exhibited in the tables corroborate the premise that the adopted approach yielded significant improvements in metric assessments. Notably, metrics such as AUC and balanced accuracy show increases exceeding 5 percentage points for the improved model, and the F1 metric even reaches an enhancement of 6 points, while the Recall metric registers a substantial rise of over 12 points. Comparing the results presented in table 3 with the data specification shown in table 1, it can be concluded that the algorithm more often achieves higher improvement in metrics in situations when there is a large percentage of anomalies in the set. On the other hand, let's look at the F1 metric. For example, out of 20 sets, its values decreased in only 4 cases, which had anomalies percentage levels of 0.96, 2.68, 6.25, and 2.32, which are relatively small values. Of course, this is not the only factor influencing the result, but there is some correlation.

TABLE 3. Difference in average metrics values for the improved and base models expressed in percentage points.

Dataset	AUC	Accuracy	Balanced acc.	F1	PRAUC	Precision	Recall
Breastw	32.18	20.97	32.18	50.09	36.54	-7.68	69.51
Cardio	-0.32	0.29	-0.32	0.23	0.20	1.49	-1.06
Cover	6.38	-4.53	6.38	-0.56	0.30	-0.44	17.5
Ecoli	-0.15	-0.30	-0.15	-0.74	-0.46	-0.59	0
Ionosphere	12.89	7.12	12.89	26.85	10.58	-16.29	33.33
Letter	-0.15	3.08	-0.15	-0.59	-0.01	1.4	-3.85
Mammography	3.34	-3.26	3.34	-1.83	-0.07	-1.72	10.26
Mnist	7.22	-7.36	7.22	2.92	1.77	-5.69	25.1
Mulcross	14.98	-5.08	14.98	0.51	2.92	-16.63	40.06
Musk	0.76	1.41	0.76	5.75	5.18	5.17	0.06
Optdigits	8.52	-8.88	8.52	3.92	0.95	1.79	26.97
Pima	1.03	-0.13	1.03	5.01	0.52	-3.35	4.85
Satellite	5.09	2.14	5.09	11.52	3.82	-10.28	13.11
Satimage-2	0.33	0.65	0.33	1.31	0.81	0.83	0
Seismic-bumps	2.65	-2.71	2.65	1.97	0.74	-0.34	8.82
Thyroid	-0.83	0.42	-0.83	0.81	0.27	0.77	-2.15
Vertebral	0.38	0.17	0.38	0.74	-0.07	0.83	0.67
Vowels	0.56	0.65	0.56	1.07	0.29	0.92	0.46
Wbc	0.56	0.65	0.56	1.07	0.29	0.92	0.46
Wine	7.02	0.70	7.02	9.83	3.75	7.05	14.50
Average	5.122	0.30	5.122	5.994	3.416	-2.092	12.93
Median	1,84	0,355	1,84	1,19	0,63	0,215	6,835

TABLE 4. Detailed results of the metrics that achieved the highest average improvement.

Dataset	AUC		PRAUC		F1		Recall	
	Base	Improved	Base	Improved	Base	Improved	Base	Improved
Breastw	0.640726	0.962513	0.529955	0.895323	0.440728	0.941650	0.283975	0.979079
Cardio	0.730323	0.727161	0.304174	0.306168	0.506253	0.508571	0.516307	0.505682
Cover	0.700805	0.764562	0.029361	0.032372	0.087224	0.081625	0.497754	0.672734
Ecoli	0.847604	0.846075	0.166083	0.161508	0.325581	0.318182	0.777778	0.777778
Ionosphere	0.632698	0.761587	0.524371	0.630123	0.422360	0.690909	0.269841	0.603175
Letter	0.525867	0.524333	0.067354	0.067260	0.114231	0.108374	0.148500	0.110000
Mammography	0.729005	0.762380	0.080335	0.079608	0.206426	0.188091	0.547423	0.650000
Mnist	0.622905	0.695154	0.159580	0.177271	0.309774	0.338954	0.323271	0.574286
Mulcross	0.777436	0.927286	0.403876	0.433110	0.599376	0.604434	0.599387	1.000000
Musk	0.964267	0.971838	0.315594	0.367424	0.479901	0.537396	0.999381	1.000000
Optdigits	0.518976	0.604143	0.030821	0.040334	0.061134	0.100329	0.136933	0.406667
Pima	0.557687	0.567940	0.394806	0.399995	0.272464	0.322581	0.175373	0.223881
Satellite	0.645221	0.696128	0.503843	0.542000	0.453731	0.568921	0.298625	0.429764
Satimage-2	0.941254	0.944569	0.115760	0.123837	0.211656	0.224756	0.971831	0.971831
Seismic-bumps	0.578583	0.605095	0.089599	0.096967	0.195804	0.215501	0.247059	0.335294
Thyroid	0.944730	0.936424	0.231210	0.233920	0.382166	0.390244	0.967742	0.946237
Vertebral	0.458095	0.461905	0.122889	0.122222	0.029630	0.037037	0.026667	0.033333
Vowels	0.621743	0.627326	0.061915	0.064855	0.171122	0.181818	0.335400	0.340000
Wbc	0.621743	0.627326	0.061915	0.064855	0.171122	0.181818	0.335400	0.340000
Wine	0.583592	0.653782	0.115675	0.153178	0.221739	0.320000	0.255000	0.400000
Average	0.682163	0.733376	0.215456	0.249616	0.283121	0.343060	0.435682	0.564987
Median	0.636712	0.711645	0.141234	0.157343	0.247101	0.319091	0.329336	0.539984

Table 5 contains the execution times of the various components of the approach proposed in this work. These include: SHAP computation, Isolation Forest model computation, results aggregation, and total times. The calculations were performed using Python 3.11.7 on a machine with an AMD Ryzen 7 4800H processor.

C. GRAPHICAL PRESENTATION OF THE OBTAINED RESULTS

The boxplot visualization of performance metrics presented on Figure 2 yields several insights into the comparative efficacy of the base and improved models. A key observation is that the median values, depicted by the horizontal lines within the boxes, are generally higher for the improved model across the metrics of F1 score, AUC, PRAUC, recall, and balanced accuracy. This denotes a central tendency towards improved performance, suggesting that methodological enhancements may have positively influenced these metrics. For the accuracy and precision metrics, the boxplots indicate no substantial difference in median values between the base and improved models. The similarity in median performance implies that the improvements introduced in the improved model did not significantly affect these specific metrics.

Another point of interest is the variability in results, which can be inferred from the range of the 'whiskers' and the

TABLE 5. Measurements of the processing time of a single iteration of the method in seconds.

Dataset	Time of calculation in seconds			
	SHAP	IF	Aggregation	Summary
Breastw	0.7629	1.4411	0.0304	2,2344
Cardio	0.8630	3.1714	0.0752	4,1096
Cover	63.8487	31.4479	15.0924	110,389
Ecoli	0.5254	1.0956	0.0170	1,638
Ionosphere	0.6877	5.3143	0.0380	6,04
Letter	0.9658	5.6682	0.0730	6,707
Mammography	1.9309	1.6635	0.2410	3,8354
Mnist	3.0665	26.6616	0.5060	30,2341
Mulcross	36.9407	11.7860	7.4775	56,2042
Musk	1.5641	33.0244	0.4000	34,9885
Optdigits	2.3283	3.6342	0.2603	6,2228
Pima	0.6350	1.2553	0.0280	1,9183
Satellite	2.2404	7.7378	0.2340	10,2122
Satimage-2	1.9534	7.4128	0.2140	9,5802
Seismic-bumps	0.8807	3.0102	0.0820	3,9729
Thyroid	0.8974	1.0889	0.0880	2,0743
Vertebral	0.4450	0.8666	0.0130	1,3246
Vowels	0.8407	1.8592	0.0470	2,7469
Wbc	0.4610	4.2898	0.0370	4,7878
Wine	0.3540	1.8056	0.0080	2,1676

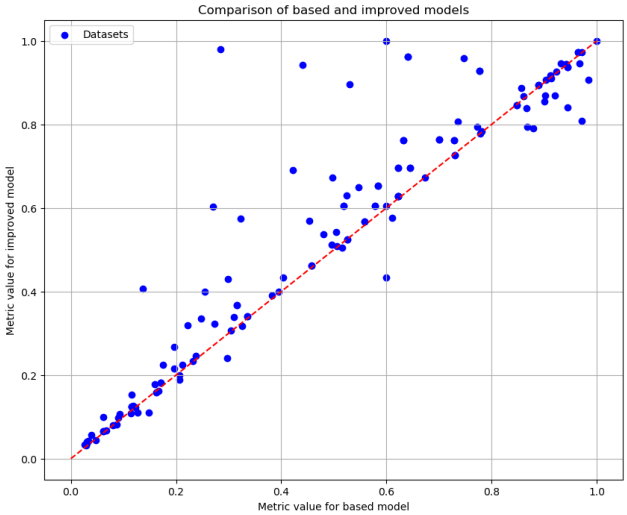

FIGURE 3. Comparison of various metrics for each dataset.

width of the boxes. The improved model exhibits a potentially reduced variability in metrics such as AUC and PRAUC, as indicated by the narrower boxes compared to those of the base model. This suggests a consistency in performance, which is a desirable characteristic, indicating a robustness that may translate to more reliable predictions in practical applications.

To summarize, the boxplots provide robust evidence for the superiority of the improved model in terms of F1 score, AUC, PRAUC, recall, and balanced accuracy. However, in terms of accuracy and precision, the models perform comparably. The presence of outliers in the improved model's PRAUC results, along with reduced variability in certain metrics, highlights areas where the improved model distinguishes itself from its predecessor. These results emphasize the need for ongoing model enhancement and point to future research avenues to better exploit the noted advancements.

Figure 3 juxtaposes the outcomes of all metrics for the baseline and enhanced models. The depicted red diagonal line signifies the threshold of equivalence, wherein a given metric yields identical results for both methodologies. Proximity of data points to the y-axis indicates superior performance of the enhanced model relative to the baseline, while a closer alignment with the x-axis suggests the baseline model outperforms the enhanced. It is observable that a predominant number of points are situated above the diagonal, denoting a trend towards the efficacy of the proposed enhancement.

Figures 4 to 10 present bar charts that compare the average value of each metric for each of the 20 datasets, juxtaposing them for the baseline and the improved models.

In Figure 4, the accuracy metric appears to maintain a similar level across most cases, with minor deviations in either direction. Figure 5 illustrates that the AUC metric shows deterioration in only two instances for the improved model, remains unchanged in two, while in the remaining 16, the improved model exhibits enhancements.

Figure 6 demonstrates that, while most metric comparisons are at a similar level, there are several instances where the improvements brought by the enhanced model are both significant and noticeable. Figure 7 displays the balanced accuracy metric, whose results are very close to those of the previously presented AUC metric.

Figure 8 showcases comparisons of the PRAUC values, which generally assume relatively low values compared to other metrics. For the majority of datasets, PRAUC indicates an improvement in the enhanced method.

Figure 9 presents the Precision metric, which uniquely favors the original isolation forest. Precision defines the ratio of observations correctly identified as anomalies to the total number of observations labeled as anomalous. Hence, it can be inferred that the improved model, while augmenting the values of other metrics, does so partly because it tends to classify observations as anomalous more frequently, which ultimately reduces its precision.

Finally, Figure 10 describes the Recall metric, which is directly related to precision. Recall determines the ratio between the detected anomalies and all anomalies present in the dataset. A substantial increase in this metric, coupled with a decrease in precision, indicates that although the model classifies more observations as anomalous, a portion of them are correctly identified, leading to the detection of nearly all the anomalies hidden within the dataset.

D. STATISTICAL TESTS

The statistical analysis in this study is conducted using the Wilcoxon signed-rank test, a non-parametric test used to compare two related samples. This test is particularly useful when the data do not conform to a normal distribution, which is a common scenario in real-world datasets, also occurring in data analyzed in this study.

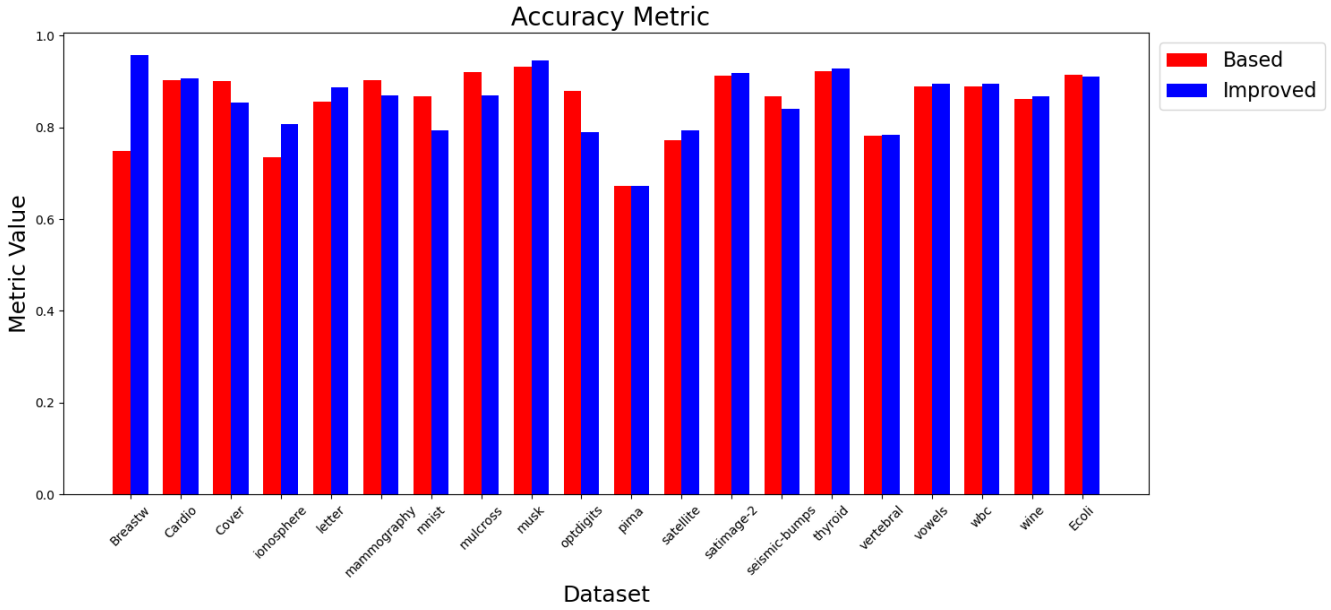

FIGURE 4. Comparison of the accuracy metric value for the classic IF model and the improved model.

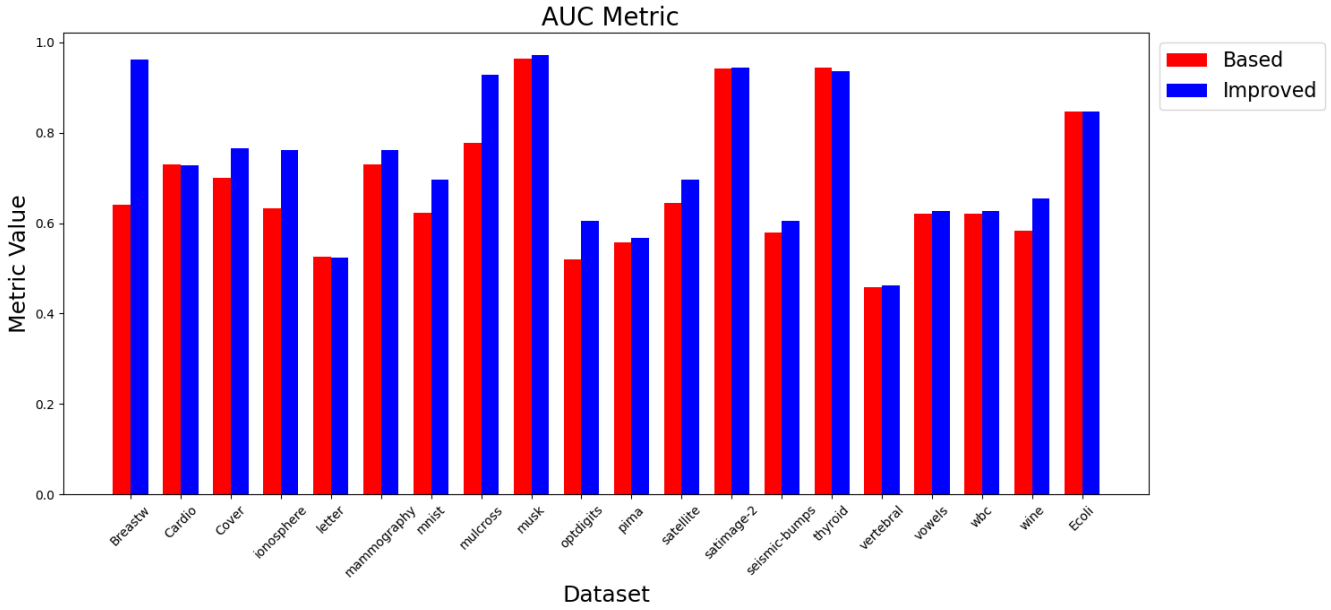

FIGURE 5. Comparison of the AUC metric value for the classic IF model and the improved model.

TABLE 6. Wilcoxon test results for the seven metrics used.

Metrics	p-value
Accuracy	0.4636
F1	0.0006
AUC	0.0001
PRAUC	0.0001
Precision	0.6762
Recall	0.0015
Balanced accuracy	0.0001

The Wilcoxon test operates under the null hypothesis that the median differences between pairs of observations in the two samples are zero. The alternative hypothesis, depending on the direction of the test, can suggest that one median is greater than or less than the other. The analysis conducted adheres to a statistical significance threshold of 0.05 for the p-value. In our analysis, the Wilcoxon test is applied to compare metrics of two models: base and improved. The test results for various performance metrics like accuracy, F1 score, AUC, PRAUC, precision, recall, and balanced accuracy are as follows:

- Accuracy: With a p-value of 0.4636, the test does not provide sufficient evidence to reject the null hypothesis, suggesting no significant difference in Accuracy.

TABLE 7. Comparison of the metrics for various datasets (Part 1).

Dataset	Metric	Mann-Whitney p-value	Kruskal-Wallis p-value	Mean Improved IF	Mean Base IF	Better Method
breastw	accuracy	<0.001	<0.001	0.96	0.75	Improved IF
breastw	f1	<0.001	<0.001	0.94	0.44	Improved IF
breastw	precision	<0.001	<0.001	0.91	0.98	Base IF
breastw	recall	<0.001	<0.001	0.98	0.28	Improved IF
breastw	balanced accuracy	<0.001	<0.001	0.96	0.64	Improved IF
breastw	auc	<0.001	<0.001	0.96	0.64	Improved IF
breastw	prauc	<0.001	<0.001	0.89	0.53	Improved IF
cardio	accuracy	<0.001	<0.001	0.91	0.90	Improved IF
cardio	f1	0.8622	0.8613	0.51	0.51	Improved IF
cardio	precision	<0.001	<0.001	0.51	0.50	Improved IF
cardio	recall	<0.001	<0.001	0.50	0.52	Base IF
cardio	balanced accuracy	0.0282	0.0281	0.73	0.73	Base IF
cardio	auc	0.0282	0.0281	0.73	0.73	Base IF
cardio	prauc	0.6582	0.6574	0.31	0.30	Improved IF
cover	accuracy	<0.001	<0.001	0.85	0.90	Base IF
cover	f1	<0.001	<0.001	0.08	0.09	Base IF
cover	precision	<0.001	<0.001	0.04	0.05	Base IF
cover	recall	<0.001	<0.001	0.67	0.50	Improved IF
cover	balanced accuracy	<0.001	<0.001	0.76	0.70	Improved IF
cover	auc	<0.001	<0.001	0.76	0.70	Improved IF
cover	prauc	0.0090	0.0089	0.03	0.03	Improved IF
ionosphere	accuracy	<0.001	<0.001	0.81	0.73	Improved IF
ionosphere	f1	<0.001	<0.001	0.69	0.42	Improved IF
ionosphere	precision	<0.001	<0.001	0.81	0.97	Base IF
ionosphere	recall	<0.001	<0.001	0.60	0.27	Improved IF
ionosphere	balanced accuracy	<0.001	<0.001	0.76	0.63	Improved IF
ionosphere	auc	<0.001	<0.001	0.76	0.63	Improved IF
ionosphere	prauc	<0.001	<0.001	0.63	0.52	Improved IF
letter	accuracy	<0.001	<0.001	0.89	0.86	Improved IF
letter	f1	<0.001	<0.001	0.11	0.12	Base IF
letter	precision	<0.001	<0.001	0.10	0.10	Improved IF
letter	recall	<0.001	<0.001	0.11	0.15	Base IF
letter	balanced accuracy	0.0031	0.0031	0.52	0.53	Base IF
letter	auc	0.0029	0.0029	0.52	0.53	Base IF
letter	prauc	0.1614	0.1610	0.07	0.07	Base IF
mammography	accuracy	<0.001	<0.001	0.87	0.90	Base IF
mammography	f1	<0.001	<0.001	0.19	0.21	Base IF
mammography	precision	<0.001	<0.001	0.11	0.13	Base IF
mammography	recall	<0.001	<0.001	0.65	0.55	Improved IF
mammography	balanced accuracy	<0.001	<0.001	0.76	0.73	Improved IF
mammography	auc	<0.001	<0.001	0.76	0.73	Improved IF
mammography	prauc	0.0422	0.0420	0.08	0.08	Base IF
mnist	accuracy	<0.001	<0.001	0.79	0.87	Base IF
mnist	f1	<0.001	<0.001	0.34	0.31	Improved IF
mnist	precision	<0.001	<0.001	0.24	0.30	Base IF
mnist	recall	<0.001	<0.001	0.58	0.32	Improved IF
mnist	balanced accuracy	<0.001	<0.001	0.70	0.62	Improved IF
mnist	auc	<0.001	<0.001	0.70	0.62	Improved IF
mnist	prauc	<0.001	<0.001	0.18	0.16	Improved IF
mulcross	accuracy	<0.001	<0.001	0.87	0.92	Base IF
mulcross	f1	0.5340	0.5332	0.60	0.60	Improved IF
mulcross	precision	<0.001	<0.001	0.43	0.60	Base IF
mulcross	recall	0	0	0	0	identical values
mulcross	balanced accuracy	<0.001	<0.001	0.93	0.78	Improved IF
mulcross	auc	<0.001	<0.001	0.93	0.78	Improved IF
mulcross	prauc	<0.001	<0.001	0.43	0.41	Improved IF
musk	accuracy	<0.001	<0.001	0.95	0.93	Improved IF
musk	f1	<0.001	<0.001	0.54	0.48	Improved IF
musk	precision	<0.001	<0.001	0.37	0.32	Improved IF
musk	recall	0	0	0	0	identical values
musk	balanced accuracy	<0.001	<0.001	0.97	0.96	Improved IF
musk	auc	<0.001	<0.001	0.97	0.96	Improved IF
musk	prauc	<0.001	<0.001	0.37	0.32	Improved IF
optdigits	accuracy	<0.001	<0.001	0.79	0.88	Base IF
optdigits	f1	<0.001	<0.001	0.10	0.06	Improved IF
optdigits	precision	<0.001	<0.001	0.06	0.04	Improved IF
optdigits	recall	<0.001	<0.001	0.40	0.14	Improved IF
optdigits	balanced accuracy	<0.001	<0.001	0.60	0.52	Improved IF
optdigits	auc	<0.001	<0.001	0.60	0.52	Improved IF
optdigits	prauc	<0.001	<0.001	0.04	0.03	Improved IF

TABLE 8. Comparison of the metrics for various datasets (Part 2).

Dataset	Metric	Mann-Whitney p-value	Kruskal-Wallis p-value	Mean Improved IF	Mean Base IF	Better Method
pima	accuracy	<0.001	<0.001	0.67	0.67	Improved IF
pima	f1	<0.001	<0.001	0.32	0.26	Improved IF
pima	precision	0.0232	0.0231	0.58	0.57	Improved IF
pima	recall	<0.001	<0.001	0.22	0.16	Improved IF
pima	balanced accuracy	<0.001	<0.001	0.57	0.55	Improved IF
pima	auc	<0.001	<0.001	0.57	0.55	Improved IF
pima	prauc	<0.001	<0.001	0.40	0.39	Improved IF
satellite	accuracy	<0.001	<0.001	0.79	0.77	Improved IF
satellite	f1	<0.001	<0.001	0.57	0.45	Improved IF
satellite	precision	<0.001	<0.001	0.84	0.94	Base IF
satellite	recall	<0.001	<0.001	0.43	0.30	Improved IF
satellite	balanced accuracy	<0.001	<0.001	0.70	0.64	Improved IF
satellite	auc	<0.001	<0.001	0.70	0.64	Improved IF
satellite	prauc	<0.001	<0.001	0.54	0.50	Improved IF
satimage-2	accuracy	<0.001	<0.001	0.92	0.91	Improved IF
satimage-2	f1	<0.001	<0.001	0.22	0.21	Improved IF
satimage-2	precision	<0.001	<0.001	0.13	0.12	Improved IF
satimage-2	recall	<0.001	<0.001	0.97	0.98	Base IF
satimage-2	balanced accuracy	<0.001	<0.001	0.95	0.94	Improved IF
satimage-2	auc	<0.001	<0.001	0.95	0.94	Improved IF
satimage-2	prauc	<0.001	<0.001	0.12	0.12	Improved IF
seismic-bumps	accuracy	<0.001	<0.001	0.84	0.87	Base IF
seismic-bumps	f1	<0.001	<0.001	0.21	0.20	Improved IF
seismic-bumps	precision	<0.001	<0.001	0.16	0.16	Base IF
seismic-bumps	recall	<0.001	<0.001	0.33	0.25	Improved IF
seismic-bumps	balanced accuracy	<0.001	<0.001	0.60	0.58	Improved IF
seismic-bumps	auc	<0.001	<0.001	0.60	0.58	Improved IF
seismic-bumps	prauc	<0.001	<0.001	0.10	0.09	Improved IF
thyroid	accuracy	<0.001	<0.001	0.93	0.92	Improved IF
thyroid	f1	<0.001	<0.001	0.39	0.37	Improved IF
thyroid	precision	<0.001	<0.001	0.24	0.23	Improved IF
thyroid	recall	0.5359	0.5351	0.95	0.95	Improved IF
thyroid	balanced accuracy	<0.001	<0.001	0.94	0.93	Improved IF
thyroid	auc	<0.001	<0.001	0.94	0.93	Improved IF
thyroid	prauc	<0.001	<0.001	0.23	0.22	Improved IF
vertebral	accuracy	0.3220	0.3213	0.78	0.78	Base IF
vertebral	f1	0.3532	0.3525	0.04	0.03	Improved IF
vertebral	precision	0.3532	0.3525	0.04	0.04	Improved IF
vertebral	recall	0.0081	0.0080	0.03	0.03	Improved IF
vertebral	balanced accuracy	0.3636	0.3629	0.46	0.46	Improved IF
vertebral	auc	0.3636	0.3629	0.46	0.46	Improved IF
vertebral	prauc	<0.001	<0.001	0.12	0.12	Base IF
vowels	accuracy	<0.001	<0.001	0.89	0.89	Improved IF
vowels	f1	0.0353	0.0352	0.18	0.17	Improved IF
vowels	precision	0.0034	0.0034	0.12	0.12	Improved IF
vowels	recall	0.0203	0.0203	0.33	0.34	Base IF
vowels	balanced accuracy	0.8399	0.8390	0.62	0.62	Base IF
vowels	auc	0.8399	0.8390	0.62	0.62	Base IF
vowels	prauc	0.3503	0.3497	0.06	0.06	Improved IF
wbc	accuracy	<0.001	<0.001	0.93	0.92	Improved IF
wbc	f1	<0.001	<0.001	0.51	0.50	Improved IF
wbc	precision	<0.001	<0.001	0.42	0.39	Improved IF
wbc	recall	<0.001	<0.001	0.67	0.70	Base IF
wbc	balanced accuracy	<0.001	<0.001	0.81	0.82	Base IF
wbc	auc	<0.001	<0.001	0.81	0.82	Base IF
wbc	prauc	<0.001	<0.001	0.30	0.29	Improved IF
wine	accuracy	<0.001	<0.001	0.87	0.86	Improved IF
wine	f1	<0.001	<0.001	0.31	0.22	Improved IF
wine	precision	<0.001	<0.001	0.26	0.19	Improved IF
wine	recall	<0.001	<0.001	0.40	0.25	Improved IF
wine	balanced accuracy	<0.001	<0.001	0.65	0.58	Improved IF
wine	auc	<0.001	<0.001	0.65	0.58	Improved IF
wine	prauc	<0.001	<0.001	0.15	0.11	Improved IF

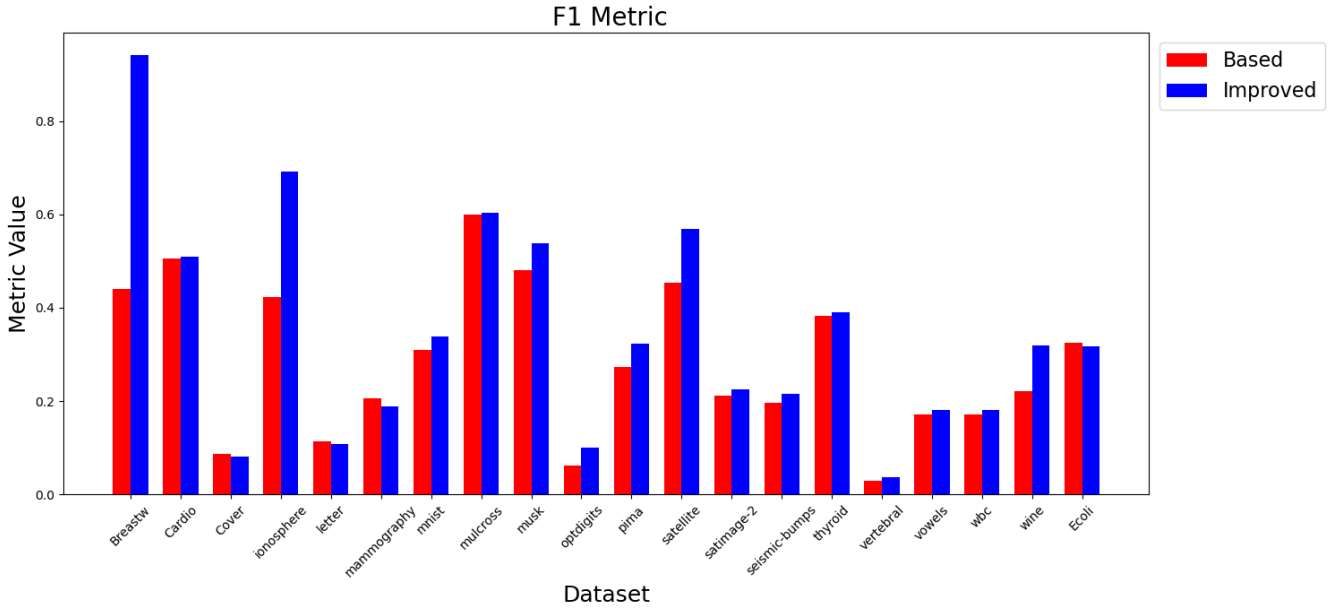

FIGURE 6. Comparison of the F1 metric value for the classic IF model and the improved model.

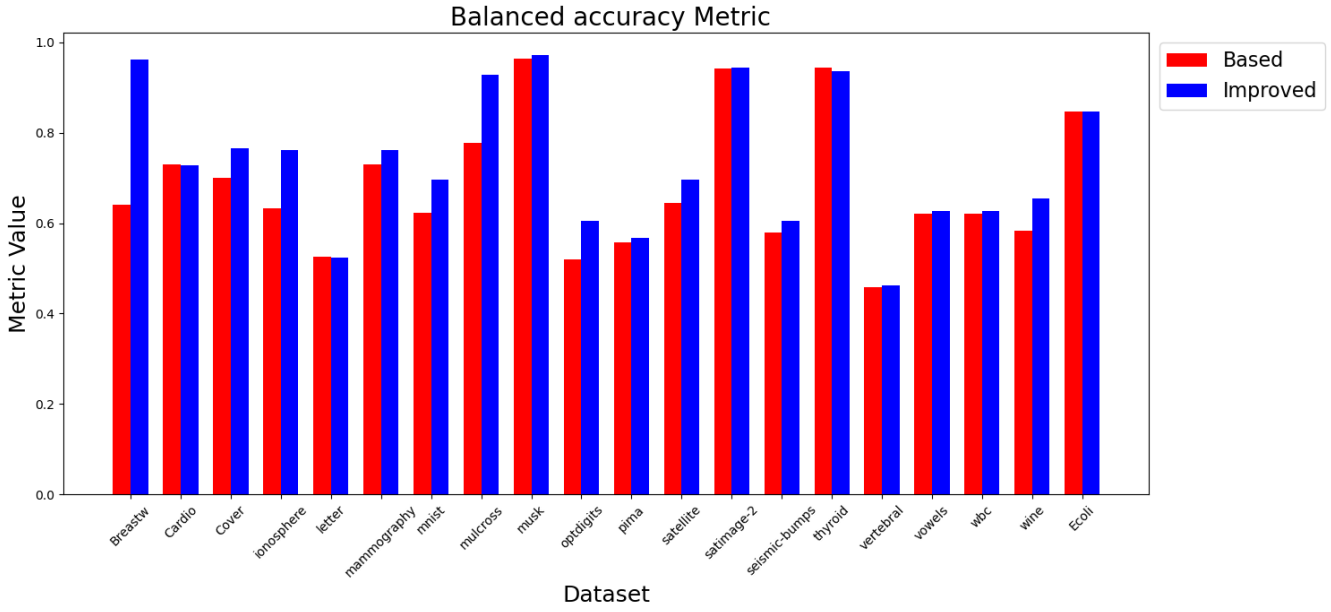

FIGURE 7. Comparison of the balanced accuracy metric value for the classic IF model and the improved model.

- F1 score: The p-value of 0.0006 indicates a significant difference, favoring the improved model.
- AUC and PRAUC: Both have very low p-values (0.0001), strongly suggesting the improved model's superiority in these aspects.
- Precision: Similar to Accuracy, the higher p-value (0.6762) implies no significant difference.
- Recall and balanced accuracy: Both metrics show significant differences with low p-values, indicating better performance in the improved model.

These results allow us to conclude that, except for Accuracy and Precision, the improved model generally outperforms the base model in the other metrics. The application of the Wilcoxon test in this context provides robust evidence for the effectiveness of the improved model in specific areas.

Additional statistical tests conducted on data collected during the experiment include the Mann-Whitney U test and the Kruskal-Wallis test. The results of these statistical tests are presented in tables 7 and 8. For 20 datasets and 7 metrics,

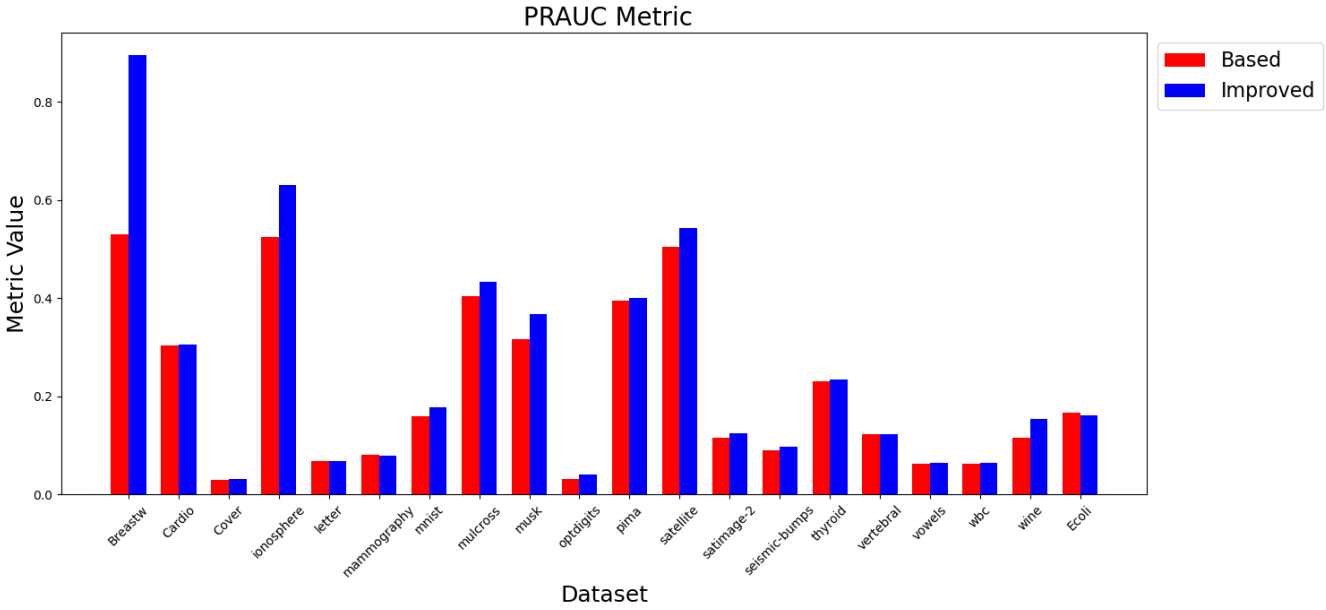

FIGURE 8. Comparison of the PRAUC metric value for the classic IF model and the improved model.

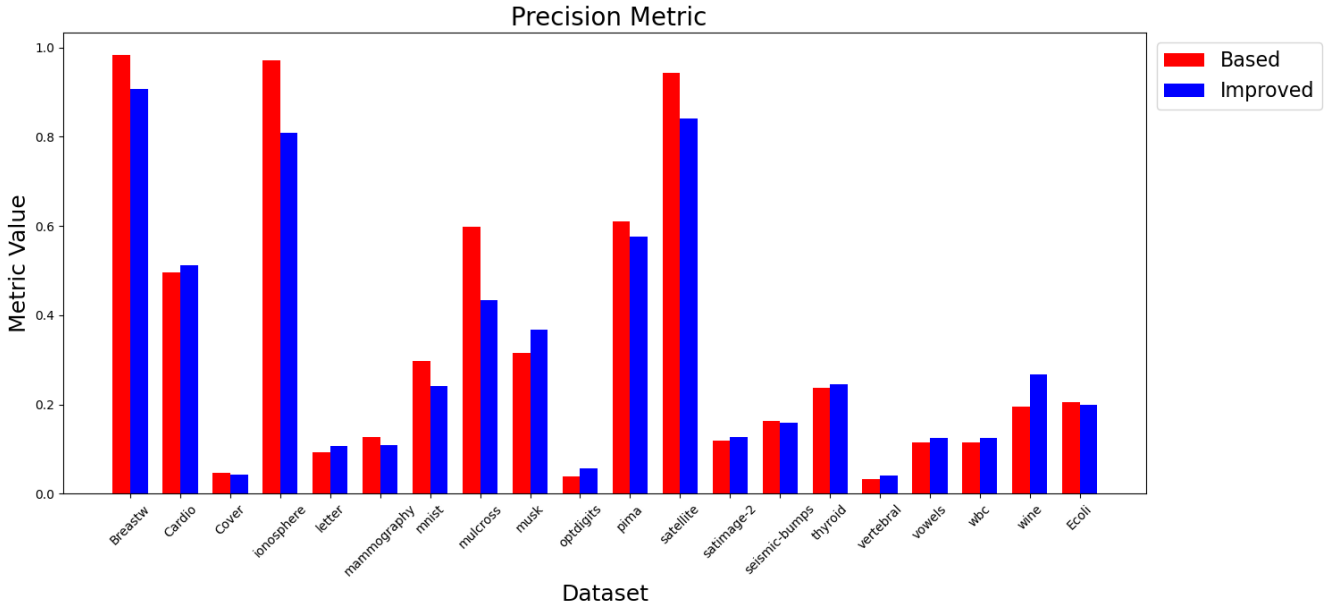

FIGURE 9. Comparison of the precision metric value for the classic IF model and the improved model.

results are compared across 140 records. It is apparent that in 88 out of 140 cases, the p-value indicates a statistically significant improvement favoring the method proposed in this study, while in 30 cases, the baseline method yielded statistically higher metrics. In the table 9 it can be seen that the most significant improvements are in the metrics: AUC, balanced accuracy, F1, and PRAUC, which are statistically better in 15 or more cases. Accuracy, precision, and recall are slightly worse, with improvements observed in only 11-12 cases out of 20.

V. CONCLUSION AND FUTURE WORK

This paper proposes a novel improvement of the Isolation Forest model. In this enhanced method, the Isolation Forest algorithm is repeatedly applied to the dataset, each time excluding a different attribute. By assigning SHAP-derived weights to the anomaly detection results from each version of the dataset, we calculate a weighted average for the anomaly coefficients of individual observations. This calculation is tailored such that attributes deemed less significant by SHAP are assigned lower weights, whereas attributes identified as

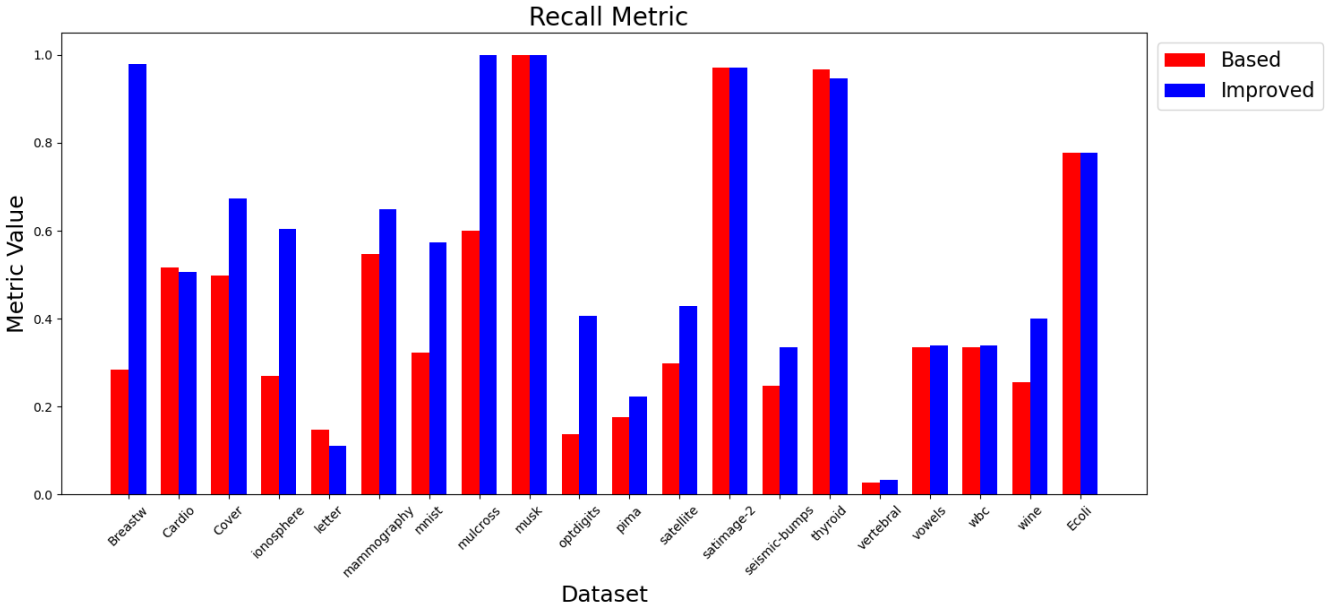

FIGURE 10. Comparison of the recall metric value for the classic IF model and the improved model.

TABLE 9. The number of statistically significant improvements in a given metric for the enhanced method compared to the baseline.

Metric	Enhanced Better Count
accuracy	12
auc	15
balanced accuracy	15
f1	16
prauc	16
precision	11
recall	12

more influential receive higher weights in the final average. This approach, therefore, not only enhances the robustness of the Isolation Forest in identifying anomalies but also offers a nuanced view of the relative importance of different attributes in the dataset, leading to more insightful and interpretable anomaly detection.

In the experiments conducted, 20 datasets are examined, and 7 metrics are used for evaluation. The results of the experiments show that all metrics achieve higher values for the presented method. Except for the metrics ''Accuracy'' and ''Precision'', all other metrics give statistically significant improvement, which is tested by the Wilcoxon test. For the evaluation, the Mann-Whitney and Kruskal-Wallis tests are also used, both of which assess the superiority of values for two independent samples. The tests revealed that for 20 datasets and 7 examined metrics, significant improvement is observed in 88 out of 140 instances. Specifically, metrics such as F1 and PRAUC indicated statistically significant improvement in 16 out of 20 cases, while AUC and balanced accuracy showed improvement in 15 out of 20 cases. This suggests that the method described in this study is highly effective at enhancing these metrics compared to the baseline method.

The remaining metrics also indicated statistically significant improvement in most cases, although not as frequently as the previously mentioned metrics. Accuracy and recall showed improvement in 12 cases each, while precision showed improvement in 11 cases. The experiments demonstrated that the majority of metrics exhibited significant improvements, such as the F1 metric, which increased its value by 6 percentage points. The proposed methodology also identified a substantially larger proportion of anomalies present in the datasets, as evidenced by the Recall metric, which exhibited an average increase of over 12 percentage points.

A potential direction for exploration could be the examination of an expanded spectrum of aggregation operations within the algorithm, including the incorporation of fuzzy operators such as the Choquet integral and the OWA operator. In addition to experimenting with various explainable models to interpret the significance of attributes in final prediction outcomes, there is also potential in aggregating results from multiple explainability models to achieve a more comprehensive understanding. Such inquiries may also lead to the creation of a custom explainability tool, providing customized and in-depth interpretations of the model's decision-making processes. This multifaceted approach could significantly enhance the interpretability and effectiveness of anomaly detection systems.

REFERENCES

[1] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[2] M. Çelik, F. Dadaser-Çelik, and A. S. Dokuz, "Anomaly detection in temperature data using DBSCAN algorithm," in *Proc. Int. Symp. Innov. Intell. Syst. Appl.*, Jun. 2011, pp. 91–95.

[3] T. M. Thang and J. Kim, "The anomaly detection by using DBSCAN clustering with multiple parameters," in *Proc. Int. Conf. Inf. Sci. Appl.*, Apr. 2011, pp. 1–5.
[4] C. H. Jin, H. J. Na, M. Piao, G. Pok, and K. H. Ryu, "A novel DBSCAN-based defect pattern detection and classification framework for wafer bin map," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 3, pp. 286–292, Aug. 2019.
[5] L. McInnes and J. Healy, "Accelerated hierarchical density based clustering," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 33–42.
[6] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Proc. 4th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2013, pp. 1–7.
[7] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," *Soft Comput.*, vol. 20, no. 1, pp. 343–357, Nov. 2014.
[8] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 733–742.
[9] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. MLSDA 2nd Workshop Mach. Learn. Sensory Data Anal.*, Dec. 2014.
[10] B. Kiran, D. Thomas, and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *J. Imag.*, vol. 4, no. 2, p. 36, Feb. 2018.
[11] A. Taylor, S. Leblanc, and N. Japkowicz, "Anomaly detection in automobile control network data with long short-term memory networks," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2016, pp. 130–139.
[12] C. Laorden, X. Ugarte-Pedrero, I. Santos, B. Sanz, J. Nieves, and P. G. Bringas, "Study on the effectiveness of anomaly detection for spam filtering," *Inf. Sci.*, vol. 277, pp. 421–444, Sep. 2014.
[13] K. Anand, J. Kumar, and K. Anand, "Anomaly detection in online social network: A survey," in *Proc. Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Mar. 2017, pp. 456–459.
[14] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014, doi: 10.1016/j.ins.2013.11.016.
[15] J. Miao, Y. Liu, Q. Yin, B. Ju, G. Zhang, and H. Wang, "A novel soft fault detection and diagnosis method for a DC/DC buck converter based on contrastive learning," *IEEE Trans. Power Electron.*, vol. 39, no. 1, pp. 1501–1513, Jan. 2024.
[16] V. E. Papageorgiou, P. Dogoulis, and D.-P. Papageorgiou, "A convolutional neural network of low complexity for tumor anomaly detection," in *Proc. 8th Int. Congr. Inf. Commun. Technol.*, Sep. 2023, pp. 973–983.
[17] T. Shi, H. Jiang, M. Wang, Z. Diao, G. Zhang, and Y.-D. Yao, "Metabolic anomaly appearance aware U-Net for automatic lymphoma segmentation in whole-body PET/CT scans," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 5, pp. 2465–2476, May 2023.
[18] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, Pisa, Italy, 2008, pp. 413–422.
[19] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, Mar. 2012.
[20] R. Zhao, Z. Yin, and Y. Li, "Time series data processing algorithm in deep water drilling," in *Proc. 7th IEEE Int. Conf. Netw. Intell. Digit. Content (IC-NIDC)*, Nov. 2021, pp. 459–463.
[21] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, "A data-driven design for fault detection of wind turbines using random forests and XGboost," *IEEE Access*, vol. 6, pp. 21020–21031, 2018.
[22] M. Ibrahim, A. Alsheikh, F. Awaysheh, and M. Alshehri, "Machine learning schemes for anomaly detection in solar power plants," *Energies*, vol. 15, no. 3, p. 1082, Feb. 2022.
[23] M. K. Boutahir, Y. Farhaoui, and M. Azrour, "Towards an effective anomaly detection in solar power plants using the AE-LSTM-GA approach," in *Proc. Int. Conf. Artif. Intell. Smart Environ.*, 2023, pp. 794–799.
[24] E. Gursel, B. Reddy, A. Khojandi, M. Madadi, J. B. Coble, V. Agarwal, V. Yadav, and R. L. Boring, "Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance," *Nucl. Eng. Technol.*, vol. 55, no. 2, pp. 603–622, Feb. 2023.
[25] Z. Feng, Y. Li, and X. Ma, "Blockchain-oriented approach for detecting cyber-attack transactions," *Financial Innov.*, vol. 9, no. 1, May 2023.
[26] X. Yang, Y. Chen, X. Qian, T. Li, and X. Lv, "BCEAD: A blockchain-empowered ensemble anomaly detection for wireless sensor network via isolation forest," *Secur. Commun. Netw.*, vol. 2021, pp. 1–10, Nov. 2021.
[27] X. Liu, F. Jiang, and R. Zhang, "A new social user anomaly behavior detection system based on blockchain and smart contract," in *Proc. IEEE Int. Conf. Netw., Sens. Control (ICNSC)*, Oct. 30, 2020, pp. 1–5.
[28] E. A. Refaee and S. Shamsudheen, "A computing system that integrates deep learning and the Internet of Things for effective disease diagnosis in smart health care systems," *J. Supercomput.*, vol. 78, no. 7, pp. 9285–9306, Jan. 2022.
[29] R. Jin, B. Wei, Y. Luo, T. Ren, and R. Wu, "Blockchain-based data collection with efficient anomaly detection for estimating battery state-of-health," *IEEE Sensors J.*, vol. 21, no. 12, pp. 13455–13465, Jun. 15, 2021.
[30] L. Antony, S. Azam, E. Ignatious, R. Quadir, A. R. Beeravolu, M. Jonkman, and F. De Boer, "A comprehensive unsupervised framework for chronic kidney disease prediction," *IEEE Access*, vol. 9, pp. 126481–126501, 2021.
[31] R. F. Mansour, A. E. Amraoui, I. Nouaouri, V. G. Díaz, D. Gupta, and S. Kumar, "Artificial intelligence and Internet of Things enabled disease diagnosis model for smart healthcare systems," *IEEE Access*, vol. 9, pp. 45137–45146, 2021.
[32] R. Tr, U. K. Lilhore, P. M. S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian J. Comput. Sci.*, pp. 132–148, Mar. 2022.
[33] L. Meneghetti, M. Terzi, S. Del Favero, G. A. Susto, and C. Cobelli, "Data-driven anomaly recognition for unsupervised model-free fault detection in artificial pancreas," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 1, pp. 33–47, Jan. 2020.
[34] L. S. Shapley, "A value for N-person games," in *Contributions to the Theory of Games*, 1953, pp. 307–317.
[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[36] P.-Y. Tseng, Y.-T. Chen, C.-H. Wang, K.-M. Chiu, Y.-S. Peng, S.-P. Hsu, K.-L. Chen, C.-Y. Yang, and O. K.-S. Lee, "Prediction of the development of acute kidney injury following cardiac surgery by machine learning," *Crit. Care*, vol. 24, no. 1, Jul. 2020.
[37] D. Jiang, Z. Wu, C.-Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu, and T. Hou, "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models," *J. Cheminformatics*, vol. 13, no. 1, Feb. 2021.
[38] H. Choi, D. Kim, J. Kim, J. Kim, and P. Kang, "Explainable anomaly detection framework for predictive maintenance in manufacturing systems," *Appl. Soft Comput.*, vol. 125, Aug. 2022, Art. no. 109147.
[39] F. Carrera, V. Dentamaro, S. Galantucci, A. Iannacone, D. Impedovo, and G. Pirlo, "Combining unsupervised approaches for near real-time network traffic anomaly detection," *Appl. Sci.*, vol. 12, no. 3, p. 1759, Feb. 2022.
[40] D. H. Kim and S. J. Hong, "Use of plasma information in machine-learning-based fault detection and classification for advanced equipment control," *IEEE Trans. Semicond. Manuf.*, vol. 34, no. 3, pp. 408–419, Aug. 2021.
[41] L. C. Brito, G. A. Susto, J. N. Brito, and M. A. V. Duarte, "An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery," *Mech. Syst. Signal Process.*, vol. 163, Jan. 2022, Art. no. 108105.
[42] E. Christoforou, K. Blom, Q. Gao, M. Börü, and T. Cataltepe, "MRI condition monitoring with explainable AI and feature selection," in *Proc. 30th Signal Process. Commun. Appl. Conf. (SIU)*, May 2022, pp. 1–4.
[43] X. Liu and C. Aldrich, "Explaining anomalies in coal proximity and coal processing data with Shapley and tree-based models," *Fuel*, vol. 335, Mar. 2023, Art. no. 126891.
[44] P. Karczmarek, A. Kiersztyn, W. Pedrycz, and E. Al, "K-Means-based isolation forest," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105659.
[45] L. Galka, P. Karczmarek, and M. Tokovarov, "Isolation forest based on minimal spanning tree," *IEEE Access*, vol. 10, pp. 74175–74186, 2022.
[46] Ł. Gałka, P. Karczmarek, and M. Tokovarov, "Effective enhancement of isolation forest method based on minimal spanning tree clustering," *Inf. Sci.*, vol. 628, pp. 320–338, May 2023.
[47] P. Karczmarek, A. Kiersztyn, and W. Pedrycz, "Fuzzy set-based isolation forest," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2020, pp. 1–6.

[48] P.-F. Marteau, "Random partitioning forest for point-wise and collective anomaly detection—Application to network intrusion detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2157–2172, 2021.
[49] L. Merrick and A. Taly, "The explanation game: Explaining machine learning models using Shapley values," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*, 2019.
[50] B. R. Preiss, *Data Structures and Algorithms With Object-Oriented Design Patterns in Java*. New York, NY, USA: Wiley, 2000.
[51] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2023.
[52] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Aug. 2013.
[53] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*.

ALBERT RACHWAŁ received the master's degree in computer science with a specialization in internet applications from Lublin University of Technology, in 2022. Currently, he is a Research Assistant with the Department of Computational Intelligence, Lublin University of Technology. His scientific work primarily involves the topics of anomaly detection and outlier observation. Engaged in software development, mainly web applications and decision-making support systems.

PAWEŁ KARCZMAREK received the Ph.D. degree in mathematics from the University of Gda'nsk, Poland, in 2010, and the Habilitation (D.Sc.) degree in computer science from the Systems Research Institute of Polish Academy of Sciences, in 2019. He is currently a Professor with Lublin University of Technology, Poland, and the Head of the Department of Computational Intelligence. He is the author of over 90 research articles and one monograph. His current research interests include computational intelligence, anomaly and deep fake detection, multicriteria decision-making theory, and pattern recognition.

ALICJA RACHWAŁ received the M.Sc. (Eng.) degree in mathematics from Lublin University of Technology, in 2022. She is currently a Research Assistant with the Department of Computational Intelligence, Lublin University of Technology. Her current research interests include fuzzy operator aggregation and mulitcriteria decision-making theory. In her scientific work, she studies new modifications of aggregation operators based on OWA and the Choquet integral.

RAFAŁ STĘGIERSKI received the M.Sc. degree from Maria Curie-Skłodowska University, Lublin, and the Ph.D. degree from the Silesian University of Technology, specializing in digital image processing with applications in computer facial reconstruction based on anthropological data. His research involved developing algorithms for the analysis of medical imaging data. Over the years, he has dedicated his focus to the application of artificial intelligence methods, placing particular emphasis on deep learning in robotics, including human–robot interaction (HRI) and image processing and analysis, including anomaly detection. He is currently an Academic Ambassador with the Nvidia Deep Learning Institute.