INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique in IoT Healthcare Data

S Subha[1]*, **J G R Sathiaseelan**[2]

1 Research Scholar, Department of Computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Trichy-17, Tamil Nadu, India
2 Associate Professor, Department of Computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Trichy-17, Tamil Nadu, India

OPEN ACCESS

Received: 04-07-2023
Accepted: 02-01-2024
Published: 23-01-2024

Citation: Subha S, Sathiaseelan JGR (2024) Combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique in IoT Healthcare Data. Indian Journal of Science and Technology 17(5): 386-396. https://doi.org/10.17485/IJST/v17i5.1645

* **Corresponding author**.

subhaebenezerraja@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This study addresses the concept drift issue in anomaly detection for IoT systems. The objective is to develop a novel approach that effectively handles the dynamic nature of IoT data. **Methods:** The proposed COMCADSET (Combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique) addresses the concept drift challenge. It adapts to evolving data distributions, detects anomalies in IoT healthcare data, mitigates class distribution imbalances through under-sampling, and enhances performance with ensemble techniques. The approach involves four phases: multi-class anomaly spotting, one-class anomaly isolation, concept-drift-free dataset creation, and robust anomaly detection using ensembles. Evaluation utilizes the ”Heart Failure Prediction” dataset from Kaggle, with comprehensive experiments and three classification algorithms. COMCADSET's innovation merges one-class and multi-class anomaly detection, under-sampling, and ensemble classification. It's compared against gold standards for classification accuracy, concept drift management, and anomaly detection performance. **Findings:** Conduct comprehensive experiments using a concept drift dataset and three classification algorithms to evaluate the efficacy of the COMCADSET technique. The experimental result shows the proposed COMCADSET technique attains an impressive 98.401% accuracy, decisively enhancing classification accuracy by adeptly addressing concept drift and identifying anomalies in IoT data. Early detection of abnormal behaviour prevents more significant issues and potential security vulnerabilities in IoT systems. **Novelty:** The novelty of the COMCADSET technique lies in its ability to address the concept drift issue and improve anomaly detection accuracy in IoT systems. By integrating one-class and multi-class anomaly detection, under-sampling, and ensemble techniques, the proposed approach provides a robust solution for handling the dynamic nature of IoT data.

Keywords: Anomaly Detection; Concept Drift; Ensemble Classification; Internet of Things; UnderSampling

1 Introduction

The Internet of Things (IoT) has revolutionized daily life by connecting devices, allowing remote control, and streamlining activities[(1)]. However, IoT faces challenges in dealing with anomalies in sensed data, which are deviations caused by sensor issues, unforeseen events, or malicious attacks (Figure 1). Detecting these anomalies is crucial for real-time identification and proactive intervention using techniques like statistics, machine learning, and deep learning[(2)]. Concept drift-based anomaly detection is vital in the dynamic IoT landscape due to evolving data distributions over time[(3)]. For instance, equipment behaviour changes in IoT-driven manufacturing make static anomaly detection models ineffective. Concept drift-based methods adapt to changing data distributions, ensuring adaptability across diverse contexts. Recent studies on IoT anomaly detection often neglect the concept drift challenge. For example, Savic et al., Abu-Alhaija et al., Yang et al., Wu et al., Ullah et al., Pathak et al., and Bhatti et al. propose various approaches but do not specifically address concept drift[(4–10)]. Although concept drift-based anomaly detection has gained attention for maintaining accuracy in evolving data distributions, existing methodologies have limitations[(11–14)]. Ding et al. use a Transformer for time series anomaly detection, but it goes beyond concept drift[(11)]. MS AR et al. focus on healthcare data and sudden concept drift detection but lack concept drift adaptation[(12)]. Gemaque et al. provide an overview of unsupervised drift detection methods without a dedicated concept drift-focused strategy for anomaly detection[(13)]. Sarnovsky and Kolarik concentrate on dynamic class-weighted ensembles, a part of the solution but not a holistic approach to concept drift[(14)]. This study introduces COMCADSET (A combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique) to address concept drift-based IoT anomaly detection challenges. COMCADSET combines one-class and multi-class anomaly detection, concept drift management, and ensemble classification. This technique aims to overcome existing method limitations through comprehensive experiments, contributing to a more robust IoT anomaly detection solution. The subsequent sections detail the COMCADSET approach (Section 2), present experimental results (Section 3), and conclude with final remarks and future research recommendations (Section 4).

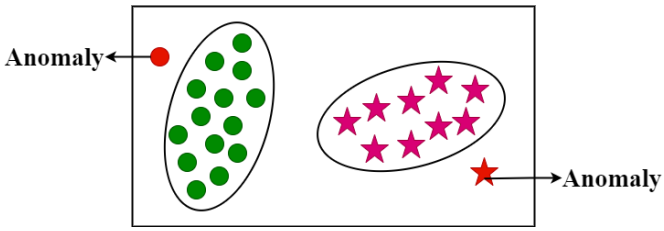

Fig 1. Example of anomalies

2 Methodology

This section delves into the intricacies of the proposed methodology, COMCADSET (Combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique), addressing the concept drift challenge in IoT anomaly detection and introducing innovative fusion techniques. The ”Heart Failure Prediction” dataset from Kaggle forms the foundation for the evaluation, tailored

explicitly to IoT healthcare data. The methodology operates within the dataset's parameters to ensure real-world relevance. The evaluation employs three distinct classification algorithms, namely KNN, RF, and SVM, facilitating a comprehensive assessment of COMCADSET's anomaly detection performance. The dataset is segmented into training and testing subsets, ensuring a rigorous evaluation of the approach.

COMCADSET's innovation lies in the convergence of one-class and multi-class anomaly detection, under-sampling to balance class distributions, and ensemble classification for enhanced accuracy. These strategies collectively distinguish the approach, enabling it to excel in identifying evolving data patterns and adapting to concept drift. In contrast to prevailing approaches, which often overlook concept drift intricacies, COMCADSET fills this gap. The approach surpasses conventional accuracy, precision, recall, and f-measure models by addressing evolving data distributions. This distinctiveness, coupled with its innovation, positions COMCADSET at the forefront of IoT anomaly detection methodologies. The proposed COMCADSET technique consists of four phases: multi-class anomaly detection, one-class anomaly detection, training dataset generation using concept drift removal and ensemble classification-based anomaly detection.

The COMCADSET technique randomly extracts 50 % of samples from the given input dataset for training dataset generation and takes the remaining samples as testing datasets. After that, the COMCADSET technique applies the combination of multi-class and one-class anomaly detection techniques for data labelling.

Data labelling is allocating labels to data to be discovered and analyzed. It is frequently utilized in machine learning and artificial intelligence to train models to recognize patterns and generate predictions. For instance, data labelling for image recognition might entail assigning labels to images like ”dog,” ”cat,” ”vehicle,” etc. It enhances the model's accuracy by assisting it in comprehending what it is looking at.

Training datasets require data labelling to provide the model with the correct data to learn from. The model wouldn't know a given input's correct output or outcome without labelled data. To assess the model's effectiveness and make enhancements, labelled data also enables testing and evaluation. Labelled data could also optimize the model for certain activities or applications. The COMCADSET technique generates two labels. They are normal and abnormal. These two data labels indicate whether each data instance in the extracted 50% samples is typical or an anomaly. The COMCADSET technique uses multi-class and one-class anomaly detection techniques to detect this.

The multi-class anomaly detection technique spots out-of-the-ordinary or anomalous data points in a dataset with numerous categories or classes. It aids in locating outliers or anomalies among various groups or classes of data and could be used to find patterns or behaviours that differ from the usual. Multi-class anomaly detection, for instance, might be utilized to find consumers with abnormally high buying habits or a large number of returns in a dataset, including customer data. It is a technique for locating data points deviating from the set's expectations.

Another method for locating unexpected or anomalous observations in a dataset that only includes instances of a single class is called one-class anomaly detection. When there are very few instances of the unusual class available for training, people typically use it. The model trains on typical data before applying it to find observations that vary considerably from the others. People consider these various observations as anomalies. It could help find anomalous patterns in time series data, detect fraud, and locate equipment faults. Figure 2 shows the example of multi-class and one-class anomaly detection.

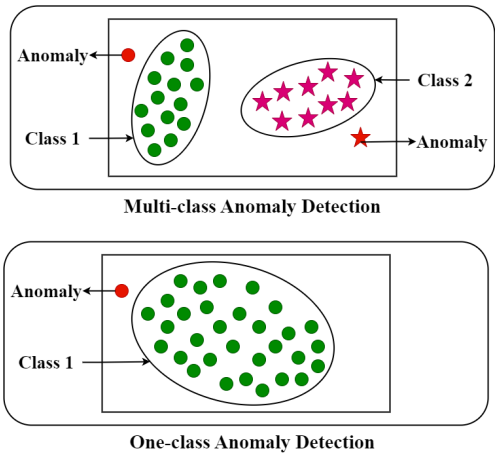

Fig 2. Examples of multi-class and one-class anomaly detection

The multi-class anomaly detection algorithm consists of four steps. The algorithm takes dataset D (extracted 50% samples) in the first step. The dataset D is assumed to have P classes and Q features. In the second step, for each class p in D, the algorithm computes its anomaly score of the q-th feature, denoted as ASpq. Equation (1) explains the calculation of ASpq.

$$AS_{pq} = \sqrt{\frac{\sum_{i=1}^{n} (X_{ipq})^2}{n-2}} * 2 \tag{1}$$

Here, n represents the number of instances in D, X_{ipq} Represents the data point of the i-th row at the q-th feature in the p-th class, and AS_{pq} represents the anomaly score of the q-th feature in the p-th class. After the anomaly score computation of each feature in each class, the proposed multi-class anomaly detection technique checks whether a data point is an anomaly based on Equation (2).

$$A_{ipq} = X_{ipq} - AS_{pq} \tag{2}$$

Here A_{ipq} represents the anomaly score of X_{ipq} Data point. If A_{ipq} is greater than 0, the data point X_{ipq} is an anomaly, otherwise not. After calculating the anomaly score for each data point in each instance, the proposed multi-class anomaly detection technique assigns an anomaly label to the instance if any data point in an instance is anomalous. Otherwise, it sets the normal label. This process is called stage 1 data labelling.

After multi-class anomaly detection, the proposed COMCADSET technique executes one-class anomaly detection to further label each normal instance as normal or anomaly. Performing one-class anomaly detection on normal instances is necessary because the multi-class anomaly detection step might misclassify some anomalies as normal. By applying one-class anomaly detection, these misclassified instances can be correctly labelled as anomalies, which is essential for downstream applications that require accurate classification of anomaly instances. Therefore, the COMCADSET technique executes one-class anomaly detection to mark each normal instance as normal or anomaly. Like multi-class anomaly detection, One-class anomaly detection involves working with dataset D (extracted 50% samples) and has P classes and Q features. The first step is to compute the centroid of D using the median. It is an essential step in one-class anomaly detection as it helps identify normal instances in the dataset. Table 1 shows an example of generating centroids using the median.

Table 1. Centroid computation using median

Instance_Id	Attribute_1	Attribute_2	Attribute_3	Attribute_4
1	15	22	31	10
2	19	34	11	11
3	34	15	9	8
4	13	65	16	6
Centroid	17	28	13.5	9

After centroid computation, the proposed one-class anomaly detection technique computes Mahalanobis distance (MD) from the centroid to each instance. Mahalanobis distance is a measure of similarity between two data instances. Frequently, it is used in multivariate statistical analytics while considering the covariance of the data. Moreover, it serves as a measure for determining the distance between a point and a distribution. Simply put, it's a method for calculating the distance between a point and the centre of distribution while considering the data's spread across all dimensions. Equation (3) shows the Mahalanobis distance between two vectors, P and Q.

$$MD = 1 - \sqrt{(P-Q)^T S^{-1} (P-Q)} \tag{3}$$

Here, P represents the centroid, Q represents each instance, and S represents the covariance matrix of P and Q. After calculating the MD of each instance, the proposed one-class anomaly detection technique assigns an anomaly label to an instance if its MD value exceeds the cutoff point value. Otherwise, it sets the normal label. Here, the cutoff point value is 0.95. This process is called stage 2 data labelling.

The proposed COMCADSET technique compares the stage 1 and 2 data labels for final data labelling. If either of these two stages confirms the data label as an anomaly, it marks the instance as an anomaly. Otherwise, it proves it as normal. Figure 3 shows the final data labelling generation process.

After the final data labelling generation, the proposed COMCADSET technique detects concept drift. There are numerous techniques to discover concept drift in a dataset. Typical techniques include:

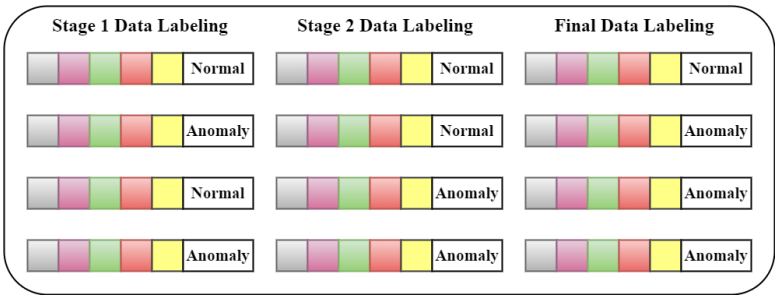

Fig 3. Final data labelling generation process

• **Tracking the effectiveness of a model trained on the dataset:** If the model's effectiveness degrades over time, it can signify that the dataset's notion has drifted.

• **Visualizing the data:** The distribution or patterns could vary over time, suggesting concept drift. It is revealed by plotting the data across time.

• **Using statistical tests:** Comparing statistics of the present data to prior data or a control group can reveal changes that may signify concept drift.

• **Using change detection algorithms:** Particular algorithms like Page-Hinkley, ADWIN, and DDM, designed to identify variations in data streams, can be used to detect concept drift.

• **Monitoring the data distribution:** The data's distribution may vary over time, signifying a change in the fundamental concept.

The proposed COMCADSET technique detects concept drift using visualizing the data method. It envisions the final data label (normal, anomaly) and detects which type of concept drift occurred. There are five concept drifts in literature: sudden, gradual, incremental, recurring and blips concept drift. Sudden concept drift refers to a significant and rapid change in the underlying data distribution in a machine-learning model. "Gradual concept drift" describes a slow, gradual shift in data distribution. Incremental concept drift describes a small-scale, over-time variation in data distribution. The term "recurring concept drift" describes periodic changes in data distribution. Blips concept drift refers to a temporary variation in the distribution of data that swiftly returns to its previous condition. Figure 4 shows these five types of concept drift.

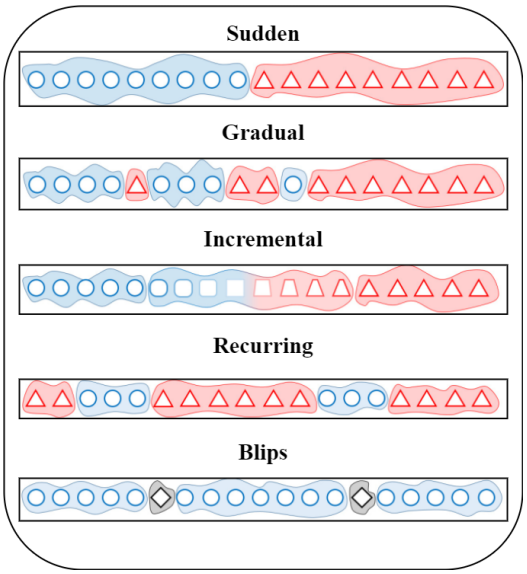

Fig 4. Five types of concept drift

If there is concept drift in the training dataset, it will reduce the prediction accuracy of the classification algorithms. So, solving concept drift is very important. An excellent way to deal with concept drift is under-sampling. Machine learning uses

the under-sampling technique to balance the distribution of instances in a dataset. Getting an equal distribution of instances entails lowering the number of instances from the dataset's over represented class (or classes). It can enhance the effectiveness of classification systems and assist in preventing bias. One way to resolve concept drift based on under-sampling is to periodically retrain the model on a subset of the data that reflects the current concept. Randomly selecting samples from the over-represented class (or classes) in the dataset, equal to the number of samples in the under-represented class, can create this subset.

Additionally, ensemble approaches, combining multiple models trained on different data subsets, can reduce concept drift. Therefore, the proposed COMCADSET technique solves concept drift based on under-sampling and ensemble classification-based anomaly detection. The proposed COMCADSET technique takes the dataset obtained after under-sampling as the training dataset.

After training dataset generation, the proposed COMCADSET technique applies MVC-based anomaly detection using KNN, RF and SVM classification algorithms. MVC is a technique that combines the predictions of various classifiers to find anomalies in data (as ensemble classification). The same training dataset is used in MVC to train KNN, RF, and SVM classifiers. Each classifier predicts whether a particular data instance in the testing dataset is normal or an anomaly. Using the majority vote of the classifiers' predictions determines the final prediction. By limiting the errors that could happen when employing a single classifier, MVC could help boost the effectiveness of anomaly detection. Figure 5 shows the architecture of the proposed MVC-based anomaly detection.

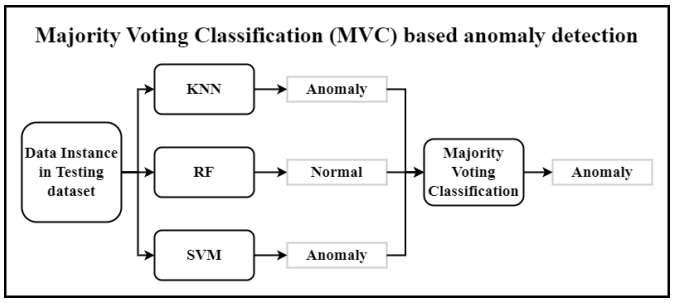

Fig 5. The architecture of the proposed MVC-based anomaly detection

This architecture consists of three main components:

1. **KNN Classifier:** This non-parametric approach classifies new cases based on a similarity metric after storing all existing cases (e.g. distance functions).
2. **RF Classifier:** Multiple decision trees are combined in this ensemble method to increase the model's accuracy.
3. **SVM Classifier:** This linear model determines the ideal hyperplane to partition various data classes.

Using the same dataset to train all three classifiers, they would then be applied to classify novel data. The three classifiers would vote in order of preference, selecting the majority class as the final result. The testing data passes through the three classifiers in this architecture, choosing the class with the majority among the three, which determines the final output. The Majority Voting Classifier (MVC) employs anomaly detection, determining the output class with the highest probability through the majority vote of the classifiers. Algorithm 1 shows the proposed COMCADSET technique.

Algorithm 1: Combination of One-Class and Multi-Class Anomaly Detection Using Under-Sampling and Ensemble Technique (COMCADSET)

Input : Dataset D of IoT sensor data
Output : Labelled dataset D_labelled with normal and anomaly labels.
Step 1 : Randomly split D into D_train and D_test with 50% each
Step 2 : Apply multi-class anomaly detection on D_train to label each instance in D_train as normal or anomaly.
For each class p in D_train, do
For each feature q in D_train, do
Compute anomaly score ASpq of feature q in class p using Eq. (1)
End for
End for
For each instance in D_train, do
Compute anomaly score A of each data point in an instance using Eq. (2)
If A > 0 then

Label instance as an anomaly
Else
Label instance as normal
End if
End for

Step 3 : Apply one-class anomaly detection on D_train to further label each normal instance in D_train as normal or anomaly.

Compute centroid C of D_train using median
For each instance in D_train, do
Compute Mahalanobis distance MD from centroid C to instance using Eq. (3)
If MD > threshold, then
Label instance as an anomaly
Else
Label instance as normal
End if
End for

Step 4 : For each instance i in D_train:

a. If the label for i from the multi-class anomaly detection stage is ”anomaly” OR the label for i from the one-class anomaly detection stage is ”anomaly”, then label i as ”anomaly.”

b. Otherwise, label i as ”normal.”

Step 5 : Detect concept drift using visualizing the data method

```
drift_type = visualize_data(D_train, D_test)
```

Step 6 : Resolve concept drift using under-sampling and ensemble classification-based anomaly detection

```
if drift_type is not None:
  D_train = undersample(D_train)
  D_train = select_samples(D_train)
  clf1 = KNN(D_train)
  clf2 = RF(D_train)
  clf3 = SVM(D_train)
  For each instance i in D_test:
    y_pred1 = clf1.predict(i)
    y_pred2 = clf2.predict(i)
    y_pred3 = clf3.predict(i)
    y_pred = majority_voting(y_pred1, y_pred2, y_pred3)
  End For
```

Step 7 : # Perform MVC-based anomaly detection using KNN, RF and SVM classification algorithms

```
def majority_voting(y_pred1, y_pred2, y_pred3):
  vote = []
  if y_pred1 == 'anomaly':
    vote.append(1)
  Else:
    vote.append(0)
  if y_pred2 == 'anomaly':
    vote.append(1)
  Else:
    vote.append(0)
  if y_pred3 == 'anomaly':
    vote.append(1)
  Else:
    vote.append(0)
  majority_vote = ''
  if sum(vote) > 1:
    majority_vote = 'anomaly'
```

```
    Else:
      majority_vote = 'normal'
  return majority_vote
```

3 Experimental results and discussions

This section explains the effectiveness of the proposed COMCADSET technique. Assessing the performance of the proposed model involves using real-time data, with data on Heart Failure Prediction gathered from Kaggle. This part assesses the performance using the datasets before and after anomaly removal and oversampling. The following evaluation metrics, Accuracy, Recall, F-Measure, and Precision, are used to assess the proposed COMCADSET technique.

$$Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Recall(Rec) = \frac{TP}{TP+FN}$$

$$F-Measure\,(F1) = 2 * \frac{Pre * Rec}{Pre+Rec}$$

$$Precision(Pre) = \frac{TP}{TP+FP}$$

3.1 Comparison of Anomaly Removal Using COMCADSET Technique on Heart Failure Dataset: Before and After:

To evaluate the effectiveness of the novel COMCADSET technique, we conduct a comprehensive assessment across the K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) classification algorithms. This examination delves into the performance improvements brought about by the COMCADSET technique when applied to the Heart Failure Prediction dataset.

Table 2 presents a detailed depiction of the outcomes obtained using the COMCADSET technique on the Heart Failure Prediction dataset. By employing the KNN, RF, and SVM classification algorithms, this evaluation scrutinizes the technique's impact on anomaly removal within the dataset. The results of this analysis capture the dataset's condition both before and after applying the COMCADSET technique.

This comparative assessment provides insights into the technique's efficacy in identifying and eliminating anomalies, subsequently enhancing the dataset's quality. The outcomes recorded in Table 2 testify to the COMCADSET technique's ability to improve the accuracy and reliability of anomaly detection, particularly in heart failure prediction.

Table 2. Evaluation metrics for heart failure prediction dataset before and after anomaly removal

Metrics	KNN		RF		SVM	
	Before Using COMCADSET	After Using COMCADSET	Before Using COMCADSET	After Using COMCADSET	Before Using COMCADSET	After Using COMCADSET
Precision	95.33	96.977	93.87	98.337	89.32	97.066
Recall	95.95	96.892	93.94	98.211	89.38	97.331
F-Measure	95.64	96.935	93.9	98.274	89.35	97.198
Accuracy	95.77	97.469	93.58	98.401	89.44	97.468

The evaluation outcomes underscore a notable improvement in accuracy upon implementing the COMCADSET technique to address anomaly and concept drift challenges within the dataset. The accuracy metric demonstrates a discernible increase after applying the COMCADSET technique.

Of the three classification algorithms under scrutiny, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF), it is evident that the Random Forest algorithm attains the highest accuracy value following the removal

of anomalous data from the dataset. This observation indicates the technique's proficiency in refining the dataset by detecting and eliminating anomalies. The higher accuracy achieved by RF accentuates its suitability for applications where precision is paramount.

This analysis provides a comprehensive overview of the impact of the COMCADSET technique on accuracy enhancement, reaffirming its efficacy in anomaly detection and concept drift management.

3.2 Comparison of proposed COMCADSET technique with existing technique:

This section also presents the results and subsequent discussion of the comparative analysis between the proposed COMCADSET technique and the existing iF_Ensemble technique[10] for heart failure prediction using IoT healthcare data. The iF_Ensemble technique is an ideal comparator for the COMCADSET technique due to its focus on anomaly detection in the IoT. It employs unsupervised and supervised methods to spot outliers in Wi-Fi signal strengths for location determination. Combining different classifiers, the technique's ensemble approach enhances accuracy by identifying and rectifying data anomalies. It makes iF_Ensemble a strong benchmark for assessing COMCADSET's performance addressing concept drift and anomaly detection in IoT healthcare data. Key anomaly detection metrics, including Precision, Recall, F-measure, and Accuracy, are used to assess the outcomes across different classification algorithms.

Table 3 comprehensively compares the performance between the COMCADSET and iF_Ensemble techniques across various metrics and classification algorithms. The evaluation considers K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) as the classification algorithms.

Table 3. Comparison of proposed COMCADSET technique with existing iF_Ensemble technique for heart failure prediction dataset before and after anomaly removal

Metrics	KNN		RF		SVM	
	iF_Ensemble	COMCADSET	iF_Ensemble	COMCADSET	iF_Ensemble	COMCADSET
Precision	96.89	96.977	94.79	98.337	96.18	97.066
Recall	96.86	96.892	94.57	98.211	94.57	97.331
F-Measure	96.85	96.935	93.73	98.274	95.64	97.198
Accuracy	96.71	97.469	94.6	98.401	96	97.468

Furthermore, Figure 6 shows the pictorial diagram comparing the proposed COMCADSET technique with the existing iF_Ensemble technique for the heart failure prediction dataset before and after anomaly removal.

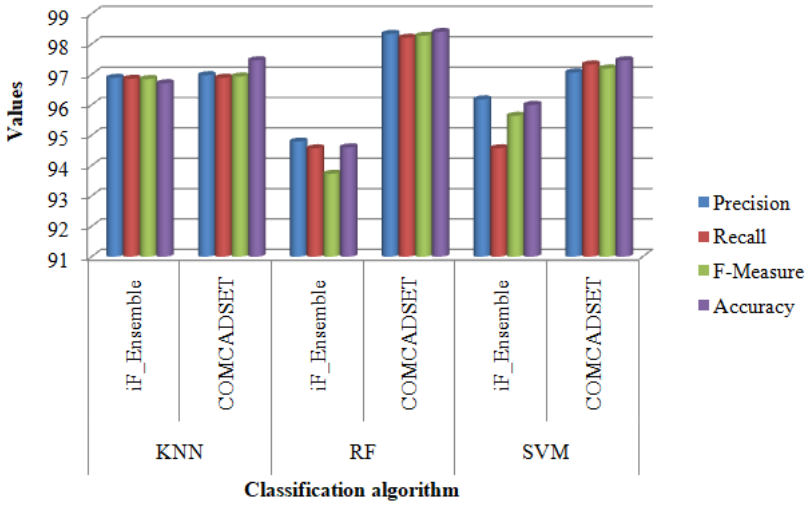

Fig 6. Comparison of proposed COMCADSET technique with existing iF_Ensemble technique for heart failure prediction dataset before and after anomaly removal

Table 3 and Figure 6 compare the performance of the proposed COMCADSET technique with the existing iF_Ensemble technique in the context of heart failure prediction using IoT healthcare data. The metrics evaluated include Precision, Recall, F-Measure, and Accuracy for three classification algorithms: K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM).

COMCADSET outperforms iF_Ensemble in precision for all three classification algorithms (KNN, RF, SVM). This means that COMCADSET has a higher ability to correctly identify anomalies while minimizing false positives compared to iF_Ensemble. COMCADSET also exhibits higher recall across all algorithms. This implies that it identifies more true anomalies while reducing the number of false negatives, which are anomalies mistakenly classified as normal data. The F-Measure for COMCADSET is consistently higher than that of iF_Ensemble, indicating a better balance between precision and recall. This implies that COMCADSET achieves a better trade-off between minimizing false positives and false negatives. COMCADSET achieves higher accuracy than iF_Ensemble, indicating better overall predictive performance. This means that COMCADSET's predictions are more accurate in classifying both normal and anomalous data points.

Its unique approach attributes to COMCADSET's superior performance. COMCADSET's merging of one-class and multi-class anomaly detection, along with under-sampling and ensemble techniques, enhances its ability to adapt to concept drift and accurately identify anomalies in dynamic IoT data. This approach addresses the challenge of evolving data distributions, ensuring improved accuracy in detecting abnormal behaviour early.

Overall, the comparison table underscores the advantages of the COMCADSET technique over the iF_Ensemble technique in anomaly detection metrics. The proposed approach's innovation and comprehensive strategy contribute to its superior performance, making it a promising solution for addressing the concept drift issue in IoT systems.

3.3 Novelty of the COMCADSET Technique:

This section delves into the innovative aspects of the COMCADSET technique and its significance in the context of IoT anomaly detection.

The primary novelty of the COMCADSET technique lies in its holistic approach to tackling the concept drift challenge in IoT anomaly detection. Unlike traditional methods, COMCADSET merges one-class and multi-class anomaly detection techniques, effectively accommodating evolving data distributions. This amalgamation enhances the ability to identify anomalies accurately, even in dynamic environments. Moreover, the technique incorporates under-sampling strategies to mitigate class distribution imbalances, fostering balanced and unbiased anomaly detection. This innovative combination sets COMCADSET apart as a comprehensive solution that caters to the intricacies of IoT data.

Furthermore, the ensemble classification technique utilized in COMCADSET, combining the predictive strengths of multiple classifiers, contributes to enhanced anomaly detection precision. By amalgamating the outputs of K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM), the technique capitalizes on diverse perspectives, resulting in more accurate anomaly identification. This approach surpasses standalone algorithms and showcases the technique's proficiency in detecting anomalies in IoT healthcare data.

The significance of this novelty becomes evident through the comparative analysis with the iF_Ensemble technique. While both techniques address anomaly detection, COMCADSET's comprehensive integration of methods distinguishes it as a robust solution. The improved Precision, Recall, F-measure, and Accuracy metrics across multiple classification algorithms demonstrate COMCADSET's effectiveness in early anomaly identification. This innovative approach can potentially revolutionize anomaly detection in the evolving landscape of IoT systems.

Overall, the novelty of the COMCADSET technique is rooted in its holistic integration of one-class and multi-class anomaly detection, under-sampling, and ensemble classification. This comprehensive strategy presents a paradigm shift in addressing the concept drift challenge and accurately identifying anomalies within dynamic IoT healthcare data. Through rigorous evaluation and comparison, COMCADSET emerges as a promising advancement in anomaly detection, catering to the evolving demands of IoT systems.

4 Conclusion

The conceptual realm of Internet of Things (IoT) systems witnessed a substantial leap forward with the advent of the proposed concept drift-based anomaly detection approach, embodied by the innovative COMCADSET technique. This paradigm shift in anomaly detection marks a promising stride towards precise and effective identification of abnormal behaviour in IoT systems. The distinctiveness of the COMCADSET technique resides in its comprehensive strategy that amalgamates multi-class anomaly detection, one-class anomaly detection, concept drift removal, and ensemble classification-based anomaly detection techniques. This holistic integration empowers the technique to scrutinize IoT data from multiple perspectives, accurately

pinpointing anomalies. By embracing the ever-evolving nature of IoT systems, this approach fortifies stability and security, especially amidst the escalating complexity and proliferation of IoT devices. Throughout this study, the performance evaluation unfolds across three classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF). RF emerges as the front runner, yielding superior accuracy values after eliminating anomalies and concept drift from the dataset. This pivotal finding highlights RF's potential to expedite anomaly detection and underscores the prowess of the COMCADSET technique. The novel contributions of this research emphasize its strengths while also delineating its areas for improvement. The robustness of the COMCADSET technique in addressing concept drift and enhancing anomaly detection precision solidifies its position as a groundbreaking advancement. However, avenues for further exploration and refinement remain. The technique's capacity to excel in diverse IoT contexts and its potential to catalyze the evolution of anomaly detection strategies warrant continued exploration. Looking ahead, recommendations for future research echo the call to explore deeper into the uncharted territories of IoT anomaly detection. A comprehensive understanding of the limitations and prospects of the COMCADSET technique catalyzes devising strategies that transcend the existing benchmarks. In essence, the COMCADSET technique addresses the demands of contemporary IoT systems and extends a beckoning invitation to researchers to elevate the efficacy and innovation of anomaly detection strategies. This conclusion resonates as an encouragement to action for further research, kindling a drive for creativity and brilliance that will influence the development of IoT systems globally.

References

1) Jiang J, Liu F, Liu Y, Tang Q, Wang B, Zhong G, et al. A dynamic ensemble algorithm for anomaly detection in IoT imbalanced data streams. *Computer Communications*. 2022;194:250–257. Available from: https://doi.org/10.1016/j.comcom.2022.07.034.
2) Chatterjee A, Ahmed BS. IoT anomaly detection methods and applications: A survey. *Internet of Things*. 2022;19:1–17. Available from: https://doi.org/10.1016/j.iot.2022.100568.
3) Togbe MU, Chabchoub Y, Boly A, Barry M, Chiky R, Bahri M. Anomalies Detection Using Isolation in Concept-Drifting Data Streams. *Computers*. 2021;10(1):1–21. Available from: https://doi.org/10.3390/computers10010013.
4) Šabić E, Keeley D, Henderson B, Nannemann S. Healthcare and anomaly detection: using machine learning to predict anomalies in heart rate data. *AI & SOCIETY*. 2021;36(1):149–158. Available from: https://doi.org/10.1007/s00146-020-00985-1.
5) Abu-Alhaija M, Turab NM. Automated Learning of ECG Streaming Data Through Machine Learning Internet of Things. *Intelligent Automation & Soft Computing*. 2022;32(1):45–53. Available from: https://doi.org/10.32604/iasc.2022.021426.
6) Yang K, Kpotufe S, Feamster N. An efficient one-class SVM for anomaly detection in the Internet of Things. 2021. Available from: https://arxiv.org/pdf/2104.11146.pdf.
7) Wu Y, Dai HNN, Tang H. Graph Neural Networks for Anomaly Detection in Industrial Internet of Things. *IEEE Internet of Things Journal*. 2022;9(12):9214–9231. Available from: https://doi.org/10.1109/JIOT.2021.3094295.
8) Ullah I, Mahmoud QH. Design and Development of a Deep Learning-Based Model for Anomaly Detection in IoT Networks. *IEEE Access*. 2021;9:103906–103926. Available from: https://doi.org/10.1109/ACCESS.2021.3094024.
9) Pathak AK, Saguna S, Mitra K, Ahlund C. Anomaly Detection using Machine Learning to Discover Sensor Tampering in IoT Systems. In: ICC 2021 - IEEE International Conference on Communications. IEEE. 2021;p. 1–6. Available from: https://doi.org/10.1109/ICC42927.2021.9500825.
10) Bhatti MA, Riaz R, Rizvi SS, Shokat S, Riaz F, Kwon SJ. Outlier detection in indoor localization and Internet of Things (IoT) using machine learning. *Journal of Communications and Networks*. 2020;22(3):236–243. Available from: https://doi.org/10.1109/JCN.2020.000018.
11) Ding C, Zhao J, Sun S. Concept Drift Adaptation for Time Series Anomaly Detection via Transformer. *Neural Processing Letters*. 2023;55:2081–2101. Available from: https://doi.org/10.1007/s11063-022-11015-0.
12) Razak MSA, Nirmala CR, Aljohani M, Sreenivasa BR. A novel technique for detecting sudden concept drift in healthcare data using multi-linear artificial intelligence techniques. *Frontiers in Artificial Intelligence*. 2022;5:1–14. Available from: https://doi.org/10.3389/frai.2022.950659.
13) Gemaque RN, Costa AFJ, Giusti R, dos Santos and EM. An overview of unsupervised drift detection methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020;10(6):1–18. Available from: https://doi.org/10.1002/widm.1381.
14) Sarnovsky M, Kolarik M. Classification of the drifting data streams using heterogeneous diversified dynamic class-weighted ensemble. *PeerJ Computer Science*. 2021;7(2):1–31. Available from: https://doi.org/10.7717/peerj-cs.459.