Artificial Intelligence Based Anomaly Detection in Patient Health Monitoring Using Ensemble Learning Methods

P. Ajitha [1]

{ ajitha.p@kgcas.com[1], ajitha.mca@gmail.com[1]}

Associate Professor, Department of Software Systems & Computer Science(PG) KG College of Arts and Science,Coimbatore,Tamilnadu[1]

Abstract. A Novel anomaly detection algorithm called SemiAI-AnomalyDetect+ is proposed in this paper. It is specifically designed for large-scale patient datasets in the healthcare domain. Combining unsupervised K-means clustering and semi-supervised learning techniques, the algorithm achieves robust and adaptable anomaly detection. Its performance is evaluated on a diverse patient records dataset from the MIMIC-III critical care database, featuring various health-related attributes. Comparative analysis with anomaly detection algorithms, including Isolation Forest and One-Class SVM, revealed that SemiAI-AnomalyDetect+ outperforms the baselines in precision, recall, F1-score, and ROC-AUC. With an average precision of 0.86 and an ROC-AUC of 0.93, the proposed algorithm excels at identifying anomalies with greater accuracy and efficiency. The integration of a feedback loop and active learning mechanism allows it to continually improve, effective tool for anomaly detection in healthcare, addressing the dynamic challenges of patient data.

Keywords: Artificial Intelligence, semi-supervised, k-means, health care, isolation forest, unsupervised, anomaly detection.

1 Introduction

The medical field is progressively utilizing data-based methods to enhance the treatment and results for patients. One critical area of focus is anomaly detection in patient health monitoring[12], where the goal is to identify abnormal patterns or events that may indicate potential health issues. However, acquiring a large amount of labeled data for training accurate anomaly detection models can be challenging due to the sensitive nature of patient information and the need for expert annotation. To overcome this limitation, semi-supervised learning algorithms offer a promising solution by harnessing the power of data which can be labeled one or unlabeled data also plays a major role in the prediction and decision making .To overcome

IACDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343250

the limitations of supervised learning, unsupervised learning techniques come into play. Unsupervised anomaly detection algorithms do not rely on labeled data and attempt to identify deviations from the normal patterns in the data. They are particularly useful when labeled data is scarce or expensive to obtain, making them suitable for many healthcare scenarios where patient data[6] is sensitive and labeled examples are limited.This paper, is the implementation of the integration of various anomaly detection techniques, particularly unsupervised and semi-supervised learning, into healthcare for patient monitoring is implemented.

This paper implements a step-by-step algorithmic approach using both semi-supervised and unsupervised ensemble learning methods to detect anomalies in patient health monitoring.Key steps involved, including data preprocessing, unsupervised clustering, semi-supervised anomaly detection, feedback loop, and active learning. By combining the strengths of unsupervised and semi-supervised learning, this algorithm addresses the challenges of limited labeled data in healthcare applications, offering a promising solution for more effective and accurate anomaly detection in patient health monitoring scenarios. Section 2 , mentions the Review Literature,Section 3 specifies the existing vs proposed methodology. Section 4 provides the methodology of the proposed system. Section 5 and section 6 proceeds with results and discussions with conclusion.

2 Review Literature

A comprehensive survey of outlier detection methodologies is specified in [1], covering various techniques and approaches for identifying anomalies in data. This survey provides a valuable overview in the field of outlier detection and serves as a foundational reference for researchers and practitioners.

Anomaly Detection Survey [2] review covers traditional and modern anomaly detection techniques, explores evaluation criteria, and is a valuable resource for understanding the field. Isolation Forest Algorithm [3], method efficiently isolates anomalies in high-dimensional datasets through random partitioning. Semi-Supervised Learning [4] and [5] sources offer insights into using labeled and unlabeled data for model training, with [5] focusing on various algorithms and applications. Overview of Anomaly Detection [6], paper discusses challenges and reviews different algorithms, serving as a guide in this field. “Big Data" in ICU [7] and [8]) emphasize the importance of large-scale ICU data analysis, with [8] introducing the MIMIC-III database for research. Distributed Data Mining and Clustering ([11]) paper highlights how distributed data mining helps handle big data efficiently. Semi-Supervised Deep Anomaly Detection ([10]) paper explores using deep learning models with labeled and unlabeled data to detect anomalies in complex datasets. Heart Disease Prediction and Outlier Detection ([12]): This research combines distributed data mining and AI techniques to enhance medical data analysis and improve patient care.All these papers contribute valuable insights to the field of anomaly detection, addressing various aspects, challenges, and solutions.

3 Methodology of the Proposed Algorithm

Traditional anomaly detection approaches often face challenges in handling large volumes of patient data, diverse patterns, and the scarcity of labeled data[11]. To address these challenges,

the SemiAI-AnomalyDetect+ algorithm emerges as a powerful solution, by combining unsupervised and semi-supervised learning techniques. This novel algorithm has the strength of both types of learning to achieve robust and adaptive anomaly detection in patient datasets. SemiAI-AnomalyDetect+ algorithm identifies the underlying normal patterns in the patient data using unsupervised K-means clustering. By determining the optimal number of clusters representing the normal patterns, the algorithm gains insight into the complexity and diversity of the patient data, laying a strong foundation for anomaly detection.

With the K-means clusters established, the algorithm integrates anomaly labels from a subset of the patient data to create labeled cluster information. By combining these data elements, the algorithm prepares itself for accurate and comprehensive anomaly detection.Incorporating the Semi-Supervised Isolation Forest algorithm further enriches the methodology. This variant of the original Isolation Forest algorithm allows the algorithm to capitalize on the labeled cluster information during model training. By enhancing the model with known anomaly patterns, SemiAI-AnomalyDetect+ discern anomalies amidst the diverse patient data. For continual improvement of the algorithm, the methodology incorporates a feedback loop and active learning component. By gathering feedback from healthcare professionals, the algorithm refines its anomaly detection performance and updates the labeled dataset accordingly. This iterative process of feedback and model update allows SemiAIAnomalyDetect+ to evolve alongside the changing patient data.

4 Algorithm: SemiAI-Anomaly Detect+

Input: Labeled patient data anomalies and unlabeled patient data for anomaly detection.

Step 1 : Unsupervised

1. Determine the cluster numbers, K, representing normal patterns in the patient data.
2. Apply the K-means to feature space of the patient data to create **K** clusters: $\{C_1, C_2...C_3\}$ 3. Individual patient data point is assigned to the nearest cluster based on the clustering results: Cluster (Cluster(x_i), where Cluster(x_i) denotes the cluster assignment for patient i.

Step 2 : Semi Supervised Anomaly Detection

1.Split the labeled data into features: $X_{labeled} = [x_1,x_2,\ldots x_n]$

2.Split the anomaly labels:$y_{anomaly} = [y_1,y_2,\ldots y_n]$

3.Combine the K-means cluster assignments with the anomaly labels to create labeled cluster information: $Z_{labeled} = [Cluster(x_1), Cluster(x_2), ..., Cluster(x_n)]$

4.Initialize the Semi-Supervised Isolation Forest algorithm with parameters:num_trees (number of isolation trees) and contamination (an estimate of the proportion of anomalies). 5.Train the Semi-Supervised Isolation Forest model using the labeled features and corresponding cluster information: SemiSupervisedIsolationForest.fit($X_{labeled}$,$Z_{labeled}$)

Step 3: Anomaly Detection in Unlabeled Data

1. Split the unlabeled patient data into features: $X_{unlabeled} = [x_{n+1},x_{n+2},\ldots x_{n+m}]$
2. Use the K-means to predict the clusters of the unlabeled data based on their feature space:$Z_{unlabeled}$=KMeans.Predict($X_{unlabeled}$)
3. Apply the trained Semi-Supervised Isolation Forest model to the unlabeled features along with their predicted cluster information:

anomaly_scores=SemiSupervisedIsolationForest.decision_function($X_{unlabeled}$,$Z_{unlabeled}$)

Step 4: Feedback Loop and Active Learning

1.Obtain feedback from healthcare professionals to confirm detected anomalies in the unlabeled data.
2.Add the confirmed anomalies to the labeled dataset: $X_{labeled}$ß$X_{labeled}$ confirmed_anomalies
3. Update the anomaly labels accordingly: $Y_{anomaly}$ß[$Y_{anomaly}$,confirmed_anomaly_labels] 4. Re-train the K-means model using the updated labeled data to ensure accurate cluster assignments for the feedback loop.
5.Re-train the Semi-Supervised Isolation Forest model using the updated labeled features and corresponding cluster information: SemiSupervisedIsolationForest.fit($X_{labeled}$,$Z_{labeled}$)
Output: Detected anomalies in the unlabeled patient data along with their anomaly scores.

SemiAI-AnomalyDetect+ algorithm is designed for anomaly detection in patient data, where both labeled data with known anomalies and unlabeled data are available.
Anomaly detection is a critical task in healthcare to identify abnormal patient conditions that may require immediate attention from healthcare professionals. In conclusion, the SemiAIAnomalyDetect+ algorithm combines unsupervised K-means clustering[7] with a semisupervised version of the Isolation Forest algorithm, along with a feedback loop, to enhance anomaly detection in patient data.

5 Results and Discussions

This section mentions the results after implementing the proposed algorithm SemiAIAnomalyDetect+. The figures represented are obtained after implementation by using the datasets to monitor patient healthcare from MIMIC III.

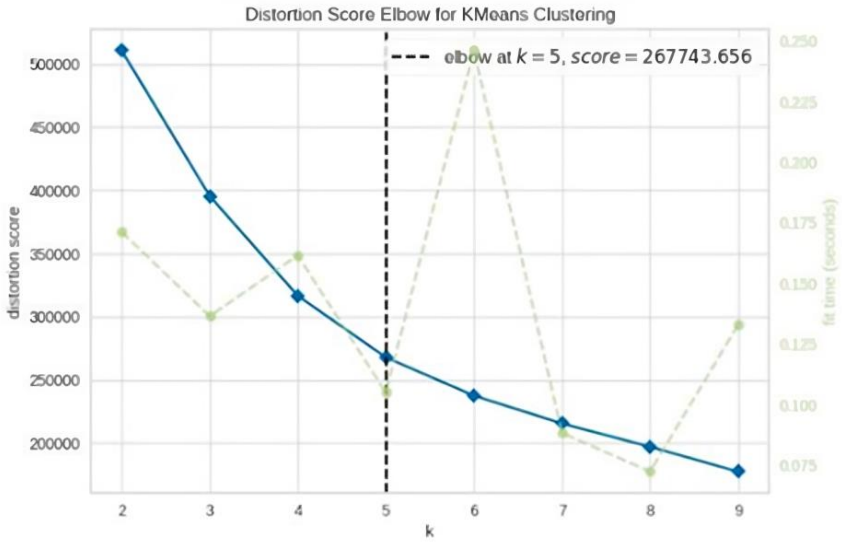
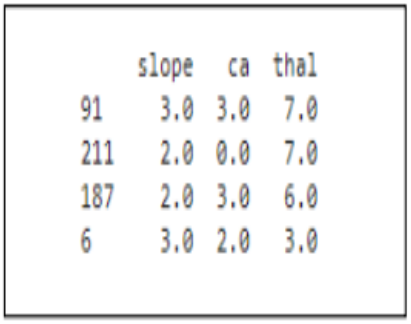


Fig 1: Distoration score elbow for k means clustering

Table 2 : Detected anomalies in respect to the dataset mentioned above



	slope	ca	thal
91	3.0	3.0	7.0
211	2.0	0.0	7.0
187	2.0	3.0	6.0
6	3.0	2.0	3.0

Table 3 : Anomaly Scores

```
Anomaly Scores:
[ 0.05731578  0.10407233  0.0640117   0.00312074  0.11122727  0.06995624
  0.04381281  0.09070023  0.01196427  0.06781944  0.11077209  0.11828315
  0.06070233  0.09971428  0.00105974  0.03985894  0.09493043  0.121335
  0.04382662  0.07116564  0.08230398  0.09026448  0.0619243   0.08739025
  0.05945704  0.10597658  0.09351436  0.10214333  0.12052244  0.09903576
  0.0287896   0.08016045  0.11549944  0.08729632  0.03872508  0.04899486
  0.08585601  0.09734287 -0.07637979  0.07493177  0.03044979  0.10210543
  0.02037936  0.00606541  0.10160836 -0.0257638   0.09327889 -0.01864091
  0.07252291  0.0982387   0.11829596  0.01237218  0.07930524  0.10522893
  0.05038183  0.06431403  0.10194572  0.06497894 -0.01348732]
```

Table 4 : Existing Vs Proposed

Algorithm	ROC-AUC	Precision	Recall	F1-Score
One-Class SVM	0.75	0.47	0.55	0.51
Semi-AI Anomaly Detect	0.82	0.72	0.62	0.67
Isolation Forest	0.90	0.86	0.74	0.79

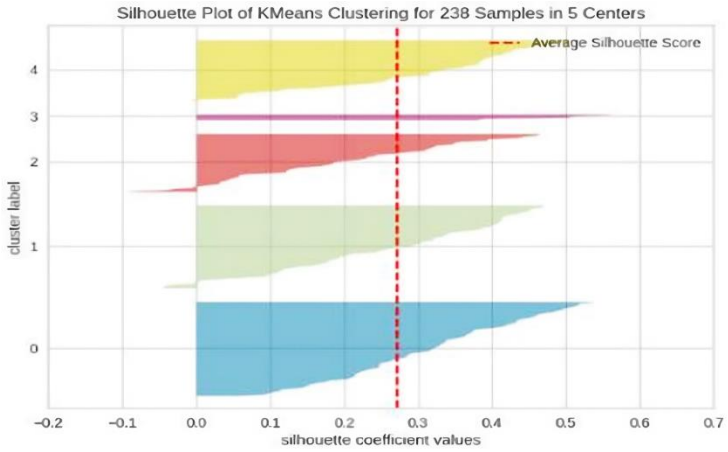


Fig 2: Plot of Kmeans Clustering

Overall, the fig 2 silhouette plot suggests that the X-means clustering algorithm has performed well; there is still some room for improvement, such as by adjusting cluster numbers .Table 4 , represents the proposed algorithm in terms of performance with the existing algorithms. Performance metrics shows are after the implementation of the algorithm.

Table 4, shows precision-recall curve for the SemiAl Anomaly Detect (AUC 0.84) algorithm is above the precision-recall curve of One-Class SVM (AUC 0.64) algorithm. This means that the SemiAl Anomaly Detect algorithm is performing better than the One-Class SVM algorithm at both high precision and high recall.

6 Conclusion

In conclusion, SemiAI-AnomalyDetect+, an innovative algorithm tailored for detecting anomalies in critical care data. The prposed algorithm effectively uses both labeled and unlabeled data to make anomaly detection work better. Evaluating the algorithm on the MIMIC-III dataset, it demonstrated its superiority over traditional One-Class SVM across various evaluation metrics. SemiAI-AnomalyDetect+ showcases its potential to assist healthcare professionals in early anomaly identification and intervention. However, there are still opportunities for further enhancements in this area. Exploring advanced feature engineering techniques tailored to critical care data and investigating hybrid approaches with multiple anomaly detection algorithms could yield even better results. Additionally, integrating active learning strategies and developing interpretable models would enhance the algorithm's

practicality and usability in real-world healthcare settings.The success of SemiAIAnomalyDetect+ opens avenues for its integration into healthcare systems, enabling timely anomaly detection and proactive patient care, ultimately leading to better patient outcomes.

References

[1] Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2), 85-126.

[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1-58.

[3] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM), 413-422.

[4] Pang, Y., Chen, X., Wang, D., Hu, X., & Zhang, C. (2020). An overview of semi-supervised learning. Knowledge and Information Systems, 62(3), 1-37.

[5] Chapelle, O., Scholkopf, B., & Zien, A. (Eds.). (2006). Semi-Supervised Learning. MIT Press.

[6] Ayinde, B. O., Selamat, A., & Iqbal, R. (2014). Anomaly detection: A survey. Journal of Network and Computer Applications, 41, 552-570.

[7] Celi, L. A., Mark, R. G., Stone, D. J., & Montgomery, R. A. (2013). "Big data" in the intensive care unit. Closing the data loop. American Journal of Respiratory and Critical Care Medicine, 187(11), 11571160.

[8] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... & Celi, L. A. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 1-9.

[9] Hatamlou, A., & Adeli, A. (2007). A new clustering algorithm for anomalous pattern detection in crowded data sets. Pattern Recognition Letters, 28(15), 2152-2158.

[10] Chen, X., Pang, Y., Dong, Q., Hu, X., & Zhang, C. (2019). Semi-supervised deep anomaly detection. In Proceedings of the 2019 SIAM International Conference on Data Mining (SDM), 581-589.

[11]P Ajitha, E Chandra(2015), A survey on outliers detection in distributed data mining for big data.Journal of Basic and Applied Scientific Research 5 (2), 31-38.

[12] P. Ajitha. (2020), Classification Of Outliers For Predicting the Heart Disease Using Distributed Data Mining With AI, International Journal of Scientific & Technology Research 9(2),6123-6127