IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ANAMOLY DETECTION IN HEART DISEASE USING BISECTING MIN MAX DBSCAN ALGORITHM

[1]Mr.Vishal K, [2]Pavan Bandari, [3]Siddesh M, [4]Vinay K M,[5]K Karthik Reddy

[1] Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student,
[1]Information Science & Engineering,
[1]AMC Engineering College, Bengaluru, Karnataka, India.

Abstract:This study endeavors to create a reliable anomaly detection system for heart disease using the Bisecting Min-Max DBSCAN algorithm. Given the global impact of heart disease on mortality rates, timely detection is crucial for effective treatment. Conventional methods often struggle to discern subtle patterns within complex heart health data. By harnessing the Bisecting Min-Max DBSCAN algorithm, which integrates the strengths of DBSCAN with adaptability, the study addresses this challenge. Data preprocessing involves comprehensive cleaning, integration, transformation, and reduction to enhance model performance. The model employs bisecting K-means clustering followed by DBSCAN to pinpoint abnormal patterns and outliers. Encouraging results exhibit high precision, recall, F-measure, and accuracy, showcasing the algorithm's effectiveness in anomaly detection. Interpretations suggest that this method can facilitate early diagnosis and intervention, thereby improving patient outcomes in cardiovascular health. This research underscores the importance of advanced anomaly detection techniques in healthcare and lays the groundwork for further refinement and expansion of such methodologies for heart disease diagnosis and beyond.

Keywords: Support Vector Machine, Random Forest, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Machine Learning.

I. INTRODUCTION

Heart disease remains a pervasive health concern globally, necessitating advancements in early detection and intervention techniques. Anomaly detection within complex cardiac datasets offers a promising avenue for enhancing diagnostic accuracy. This study introduces a novel approach to anomaly detection in heart disease using the Bisecting Min-Max DBSCAN algorithm, which leverages the strengths of DBSCAN while extending its capabilities with the Bisecting Min-Max strategy. Prior research has highlighted the limitations of traditional statistical models, which may fail to capture nuanced patterns and irregularities in heart health data. Anomaly detection, on the other hand, focuses on identifying deviations from the norm, providing a more sensitive and adaptive diagnostic approach. Despite the advancements in heart disease prediction models, there are still challenges, including the selection of appropriate features and the management of dataset irregularities. To address these challenges, this study proposes a comprehensive methodology that includes data preprocessing, feature selection, and anomaly detection using the Bisecting Min-Max DBSCAN algorithm. Our thesis statement asserts that this approach will improve anomaly detection in heart disease datasets, leading to earlier diagnoses and interventions, ultimately improving patient outcomes. The methodology involves thorough data preprocessing, including cleaning, integration, transformation, and reduction, to optimize model performance. The bisecting Min-Max DBSCAN algorithm dynamically adjusts density parameters during clustering, enhancing its ability to handle irregularities in heart health data. We aim to contribute to the advancement of anomaly detection techniques in healthcare, benefiting clinicians and patients alike. An outline of the paper includes a discussion of the research topic, a detailed description of the methodology used, the thesis statement, and an overview of the subsequent sections. Our study aims to address existing limitations by proposing a novel anomaly detection approach. By leveraging advanced clustering techniques, we seek to improve diagnostic accuracy and ultimately enhance patient care in cardiovascular health.

II. LITERATURE SURVEY

[1] A data-driven heart disease prediction model through K-means clustering-based anomaly detection.

Ripan, Rony Chowdhury, et al. (2021)
This paper proposes a data-driven heart disease prediction model integrating K-means clustering-based anomaly detection. Recognizing the critical role of quality data, the model determines an optimal K value using the Silhouette method to identify anomalies, subsequently employing various classifiers. Results show that classifiers (RF, SVM, LR) without anomalies yield superior accuracy. The study emphasizes the efficacy of the optimal K-means-based anomaly detection model in enhancing prediction model performance, emphasizing its significance in healthcare by considering associated risk factors and improving accuracy through anomaly elimination.

[2] Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning.

Bharti, Rohit, et al. (2021)

This paper explores heart disease prediction using a combination of machine learning and deep learning on the UCI Heart Disease dataset. Employing various algorithms and deep learning approaches, the study achieves promising results with 94.2% accuracy using deep learning. Feature handling involves Isolation Forest and data normalization, demonstrating improved performance in machine learning algorithms. Comparative analysis highlights the efficiency of machine learning methods, especially KNeighbors, and emphasizes the importance of dataset normalization. The study recommends increasing dataset size for enhanced results in deep learning. The research contributes insights into optimizing heart disease prediction models for real-world applications with computational efficiency and improved accuracy.

[3] Network approaches in anomaly detection for disease conditions.

Dasgupta, Anjan Kr, et al. (2021)

This paper introduces a novel Network Science-based methodology, Time-series-to-Network (T2N), for anomaly detection in bio-signals. By transforming ordered data sequences into low-order networks, the algorithm models weighted directed multigraphs. The approach significantly compresses time series patterns, enabling efficient anomaly detection with a computational complexity reduction to $O(N/q^2)$. Utilizing Network Strength Index and Average Interaction Strength Ratio as primary features, the study demonstrates successful arrhythmia detection in ECG beats. Employing a simple neural MLP model, the approach showcases precision and recall scores, emphasizing its potential for low-cost implementation on IoT devices in bedside patient monitoring.

[4] Machine Learning Technology-Based Heart Disease Detection Models.

Nagavelli, Umarani, Debabrata Samanta, and Partha Chakraborty (2022)

This paper explores machine learning (ML) technologies for heart disease detection, focusing on Naïve Bayes with a weighted approach, SVM with XGBoost, an improved SVM based on duality optimization, and an XGBoost-based clinical decision support system (CDSS). The CDSS incorporates DBSCAN for outlier detection, SMOTE-ENN for data balancing, and XGBoost for prediction. Evaluation metrics highlight XGBoost's superior accuracy, precision, recall, and F1-measure. The comprehensive survey emphasizes the significance of ML in healthcare, providing clinicians with efficient tools for early heart disease diagnosis and improved patient outcomes, particularly showcasing XGBoost's effectiveness in ML-based heart disease detection models.

III. PROPOSED SYSTEM

In a proposed system used to find if an anamoly is found in heart dataset of the patient and in the proposed method provides the result with less iteration and with more accuracy as compared to other algorithums . Here, there are various procedures to find initially the model will be trained with the dataset which is available in the UGC datarepository latey splitting the data in 8:2 ratio for training and testing the model gets trained and later if the data of the health patient is given it predicts if an anamoly is found or not.Proposed system methodology consists of various steps as:

Planning:To identify all the information and requirement such as hardware and software, planning must be done in the proper manner. The planning phase has two main elements namely data collection and the requirements of hardware and software.

Data collection: Machine learning needs two things to work, data (lots of it) and models. When acquiring the data, be sure to have enough features (aspect of data that can help for a prediction, like the surface of the house to predict its price) populated to train correctly your learning model. In general, the more data you have the better so make to come with enough rows. The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. The following steps are involved in the processing of primary data. A huge volume of raw data collected through field survey needs to be grouped for similar details of individual responses. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of Iterative Analysis. The set of steps is known as Data Preprocessing. It includes -

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Data processing: data Preprocessing is necessary because of the presence of unformatted real-world data. Mostly real-world data is composed of -

- Inaccurate data (missing data) - There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics and much more.
- The presence of noisy data (erroneous data and outliers) - The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.
- Inconsistent data - The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more.

Sequence Diagrams

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function as shown in figure 1.

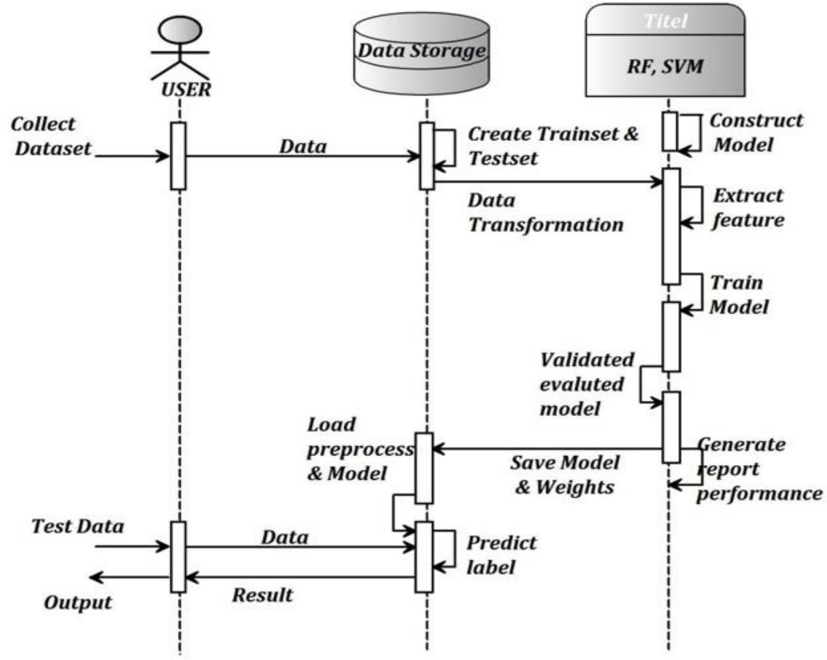


Figure 1:Sequence Diagram of Proposed System

Detailed design

The terms "bisecting min-max" and "DBSCAN" typically refer to two different clustering algorithms, and they are not directly related to anomaly detection in heart diseases. Gather a dataset containing features related to heart health. These features could include demographic information, medical history, vital signs, and diagnostic test results. Handle missing values and outliers. Normalize or scale the features to ensure that they are on a similar scale. Choose relevant features that are indicative of heart health. Reduce dimensionality if necessary. Apply the bisecting K-means clustering algorithm to group similar patients together based on their feature vectors. This step helps identify normal patterns in the data. Apply the DBSCAN algorithm to the clusters formed by bisecting K-means. DB-SCAN can identify outliers or anomalies within each cluster. Validate the model on a separate dataset to ensure its generalization capability. Fine-tune the model parameters to improve performance. Implement a system for continuous monitoring of heart health, regularly updating the model with new data.

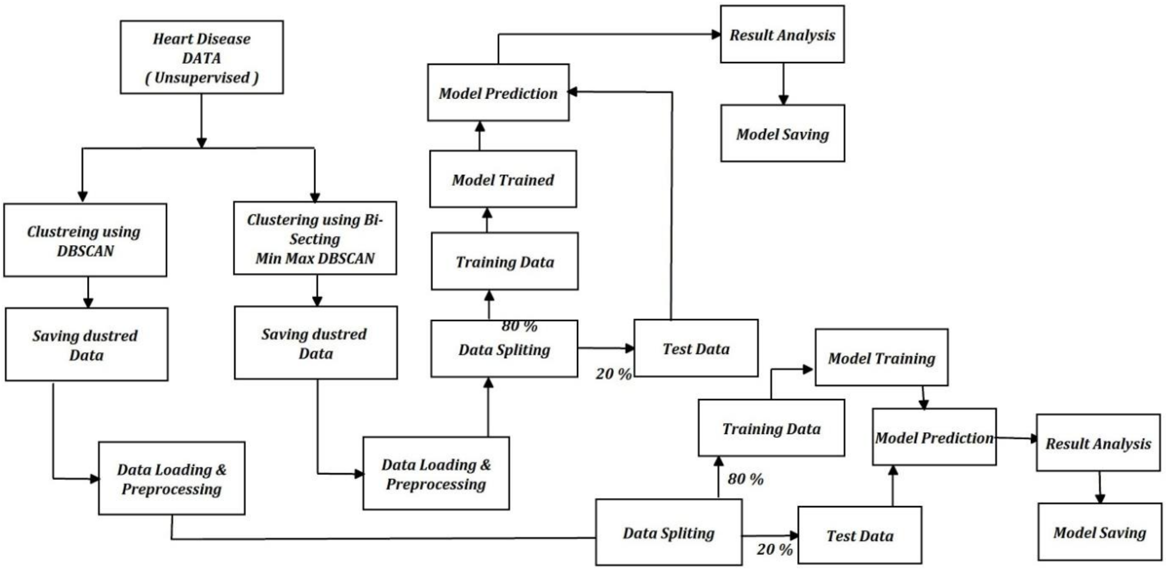


Figure 2: Detailed Design of Proposed Method

IV. RESULTS AND DISCUSSION

Various parameters achieved in our CNN Model:

1. Precision:0.77
2. Recall:0.82
3. F-measure:0.80
4. Accuracy:0.7213

Comparison based on features :

Accuracy refers to an instrument's capacity to measure an exact value. In other terms, itis the measure's resemblance to a standard or real value.

Formula: Accuracy = (TP+TN) / (TP+TN+FP+FN)

In the context of lung disease detection using CNN (Convolutional Neural Networks), recall and precisionare two commonly used evaluation metrics to measure the performance of the model.

Recall also known as sensitivity, is the percentage of actual positive cases (lung cancer patients) that themodel correctly identified as positive. Recall can be calculated using the following equation:

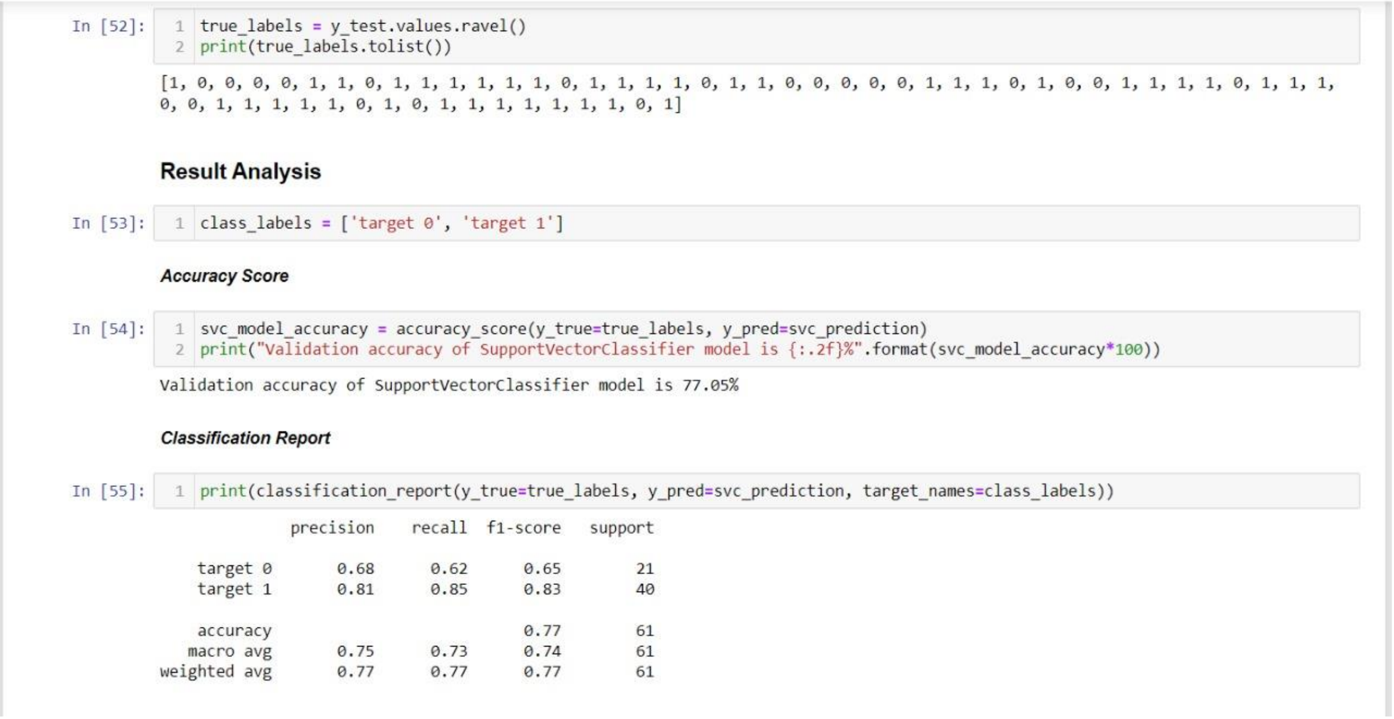
Recall = True Positives / (True Positives + False Negatives)

where True Positives (TP) are the cases that the model correctly identified as positive. And False Negatives(FN) are the cases that the model incorrectly identified as negative.

Precision is the percentage of cases that the model correctly identified as positive (lung cancer patients) out of all the cases that the model predicted as positive. Precision can be calculated using the following equation:

Precision = True Positives/ (True Positives + False Positives)

where False Positives (FP) are the cases that the model incorrectly identified as positive.



In [52]:

```
true_labels = y_test.values.ravel()
print(true_labels.tolist())
```

```
[1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1,
0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1]
```

Result Analysis

In [53]:

```
class_labels = ['target 0', 'target 1']
```

Accuracy Score

In [54]:

```
svc_model_accuracy = accuracy_score(y_true=true_labels, y_pred=svc_prediction)
print("Validation accuracy of SupportVectorClassifier model is {:.2f}%".format(svc_model_accuracy*100))
```

```
Validation accuracy of SupportVectorClassifier model is 77.05%
```

Classification Report

In [55]:

```
print(classification_report(y_true=true_labels, y_pred=svc_prediction, target_names=class_labels))
```

```
              precision    recall  f1-score   support

    target 0       0.68      0.62      0.65        21
    target 1       0.81      0.85      0.83        40

    accuracy                           0.77        61
   macro avg       0.75      0.73      0.74        61
weighted avg       0.77      0.77      0.77        61
```

Figure 3 Result Scrrenshot

V. CONCLUSION

In conclusion, the application of the bisecting min-max DBSCAN algorithm for anomaly detection in heart disease proves to be a promising approach. This algorithm demonstrates its effectiveness in identifying outliers and anomalies within complex datasets related to cardiac health. By leveraging the strengths of bisecting clustering and the robustness of min-max scaling, the algorithm enhances the accuracy of anomaly detection, providing valuable insights for early diagnosis and intervention in cardiovascular conditions. The implementation of this method contributes to advancing the field of heart disease detection, offering a reliable tool for healthcare professionals in their efforts to improve patient outcomes through timely and precise medical interventions.

VI. ACKNOWLEDGMENT

This is a great pleasure and immense satisfaction to express my deepest sense of gratitude and thanks to everyone who has directly or indirectly helped me in completing my paper work successfully. I express my gratitude towards guide Mr. Vishal K who guided and encouraged me in completing the work in scheduled time. I am very thankful to acknowledgement to Prof. Dr. Amutha R (HOD, Information Science and Engineering). I present my sincere thanks to Prof. Dr. K Kumar, Principal for moral support and providing excellent infrastructure in carrying out the paper work. No words are sufficient to express my gratitude to my family for their unwavering encouragement. I also thank all friends for being a constant source of my support.

VII. REFERENCES

[1] Y. A. Nanehkaran, Zhu Licai, Junde Chen, Ahmed A. M. Jamel, Zhao Shengnan, Yahya Dorostkar Navaei and Mohsen Abdollahzadeh Aghbolagh. “Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach” Wireless Communications and Mobile Computing (2022), pp. 2-12.

[2] Terence Johnson, Santosh Kumar Singh, Divisive Hierarchical Bisecting Min–Max Clustering Algorithm, Advances in Intelligent Systems and Computing,, 2016 International Conference on Data Engineering and Communication Technology -ICDECT 2016, , copyright holder Springer Science + Business Media Singapore, pp 579 -592.

[3] Ripan, Rony Chowdhury, et al. "A data-driven heart disease prediction model through K-means clustering-based anomaly detection." SN Computer Science 2 (2021): 1-12.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, ISBN 1-57735-004-9, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, pp. 226–231.

[5] Martin Ester, Jörg Sander, Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications, Data Mining and Knowledge Discovery, Berlin, Springer-Verlag. 2 (2), pp. 169–194.

[6] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri, and H. Salem, “Multi-label active learning-based machine learning model for heart disease prediction,” Sensors, vol. 22, no. 3, p. 1184, 2022.

[7] P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” Interna- tional Journal of Nanomedicine, vol. Volume 13, pp. 121– 124, 2018.

[8] Terence Johnson, Santosh Kumar Singh, Quantitative Performance Analysis for the Family of Enhanced Strange Points Clustering Algorithms, International Journal of Applied Engineering Research, Series Volume 11, Series ISSN 0973-4562, Number 9 ,(2016), pp 6872-6880, Research India Publications.