A Hybrid Method: Hierarchical Agglomerative Clustering Algorithm with Classification Techniques for Effective Heart Disease Prediction

Farha Akhter Munmun[1], Sumi Khatun[2]

[1,2]*Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh*

***Abstract*: Prediction of heart disease is challenging because countless data are collected for clinical data analysis, but all this information is not equally important for making the right decisions. We have proposed a hybrid method: Hierarchical Agglomerative Clustering algorithm combined with conventional classification techniques such as K-Nearest Neighbors (K-NN), Decision Tree (J48), and Naïve Bayes which aims to reduce the prediction time by clustering the patients having almost similar symptoms of heart failure. This approach minimizes the forecasting time based on clusters of patients instead of individual patients. Moreover, a comparison between the classification techniques and our approach is depicted based on precision, recall, F1 score, accuracy, and prediction time. The accuracies of the classifiers (K-NN-66.67%, J48-83.33%, and Naïve Bayes83.33%) of our system have slightly decreased compared with the conventional methods (K-NNN-69.128%, J48-83.8926%, and Naïve Bayes-87.248%) but the prediction time was significantly low (K-NNN-230ms, J48-203ms, and Naïve Bayes-195ms).**

***Keywords*: heart disease, feature selection, hybrid method, agglomerative clustering, classification**

I. INTRODUCTION

Heart disease is a common name for various types of diseases and disorders affecting the heart and blood vessels directly. Symptoms can vary depending on the type of heart disease. Most hospitals nowadays use systems for managing patient data [3] which generate enormous amounts of data taking the form of images, text, and numbers. Besides, most of the time patients suffering from multiple diseases provide unnecessary symptoms. But all of this information is hardly used to make the right decisions for any specific kind of disease. So, it becomes challenging to turn these data into efficient and useful information for making intelligent clinical decisions. Data mining is an excellent solution for solving this type of real-life problem. Different data mining techniques with efficient algorithms can solve the problem of extracting hidden knowledge from large databases. Different tools for data mining carry out data analytics for discovering secret patterns.

The main objective of this work is to propose a method that can reduce the prediction time for heart disease. All the algorithms used for heart disease prediction take a significant amount of time. Our proposed methodology tries to solve this problem by grouping the patients with almost similar symptoms and then applying data mining techniques to the groups for predicting whether they have the risk of heart disease or not. Our proposed methodology reduces the prognosis time by clustering the patients having similitude symptoms of heart disease utilizing the Hierarchical Agglomerative Clustering algorithm and then different classification techniques K-NN, Decision Tree (J48), and Naïve Bayes have been applied to the clusters for predicting the heart disease risk as a group instead of a single patient.

The rest of the paper is organized as follows: In Section II, related existing works. In Section III, the proposed methodology is explained along with the proposed architecture. Section IV presents the implementation details and result analysis. Finally, we have concluded the paper in Section V.

II. LITERATURE REVIEW

A large number of works directly related to our work has been proposed to generate high accuracy for heart disease prediction. K. Srinivas et al. [4] proposed different classification-based techniques for efficient decision support systems. Tanagra was used as a tool for statistical data analysis. A dataset of a total of 3000 entities was used. A comparison was shown among these algorithms and as per their result, Naïve Bayes performed best (52.33%).

N. Bhatla et al. [5] compared the performance of Naïve Bayes, Decision Tree, and Neural Network using 15 and 13 attributes. A total of 909 records from the Cleveland dataset were used. Their perceptions uncovered that Neural systems with a dataset of 15 qualities have performed much way better than all other information mining techniques.

A. k. Pandey et al. [6] developed a model for heart disease classification using the J48 Decision Tree. The model was developed based on a dataset with 14 attributes against pruned, pruned with minimized error, and unpruned pruning approach. Their experimental result showed that the approach for pruning the J48 with a minimized error performed better than the other two approaches.

A framework based on back-propagation was displayed by N. Al-milli et al. [7] for the forecast of heart disease. 13 attributes were used for their prediction system. The execution of their proposed approach was great compared with other existing methods.

A. Khemphila et al. [8] compared the performance of different classifiers for predicting heart disease risk. The analysis was done with 303 records. They included decision tree, Logistic Regression (LR), and ANNs for comparing the performance. ANNs gave the highest accuracy with the least error rate.

K. M. Almustafa et al. [9] compared the accuracy of different classifiers using 14 attributes. They included Naïve Bayes, K-NN, J48, Support Vector Machine (SVM), JRip, Adaboost, Decision Table (DT), and Stochastic Gradient Decent (SGD) to evaluate the performance.

S. Joshi et al. [10] performed a heart disease prediction system using three classification techniques such as K-NN, Decision Tree, and Naïve Bayes. They also used a dataset with

303 entities and 14 attributes from the UCI repository. They conducted their experiments in WEKA and among the three techniques, K-NN showed the best performance.

S. M. M. Hasan et al. [11] remove unwanted features from the dataset using the info gain feature selection technique. Distinctive classification methods were utilized on the heart disease dataset for superior prediction.

S. Mohan et al. [12] proposed a novel strategy that points to finding significant features by applying machine learning strategies coming about progressing the precision within the prediction of heart disease. They created an improved execution level with an accuracy of 88.7% through a Hybrid of Random Forest with a Linear Model (HRFLM).

S. Bashir et al. [15] used different feature selection techniques to improve the accuracy of heart disease prediction. Various classification methods such as Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree, Random Forest, and Naïve Bayes have been applied individually to the UCI dataset to compare with the result of previous research.

M. A. Khan [16] proposed an IoT-based framework for improving the accuracy while evaluating heart disease. A Modified Deep Convolutional Neural Network (MDCNN) was used and compared with the existing deep neural networks and logistic regression. The experimental result illustrated that the proposed model (MDCNN) performs better.

III. PROPOSED METHODOLOGY

Fig. 1 describes our proposed methodology. As shown, our method works by following the steps: i) Data Collection ii) Data Preprocessing iii) Feature Selection iv) Clustering the Data, and v) Prediction

A. Data Collection

Among the four databases (Hungarian, Cleveland, Longbeach-VA, and Switzerland) of UCI machine learning repository[1], we have used the Cleveland database for our experiment. This Cleveland database contains 303 patients' information and 76 attributes. All of these attributes allow three types of values: Real, Integer, and Categorical.

B. Data preprocessing

The dataset we have selected had missing attribute values for some of the patients. Those types of patient information have been deleted from the dataset and the rest of the data have been used. Algorithm 1 presents the steps which we have followed for data preprocessing. After implementing this algorithm, 6 records have been deleted, and the rest 297 records have been used.

Algorithm 1 Remove Empty valued data

```
1: Start
2: Input: Workbook
3: for index, row do
4: for cell in row do
5: if cell.value == NULL then
6: Workbook.delete row(index)
7: End
```

C. Feature Selection

The most common problem in the field of data mining is to predict whether all the features are equally important and relevant to determine the output or not. Some of the most popular feature selection techniques used for heart disease datasets: Chi-Squared Test, Principal Component Analysis (PCA), Symmetrical Uncertainty (SU), and ReliefF.

1) *Principal Component Analysis (PCA):* is a dimensionality-reduction approach for eliminating the dimensionality of a huge dataset, which transforms the dataset into a reduced version that contains the most appropriate information related to the dataset.
2) *Chi-Square Test:* is a statistical test for determining the relevance of an input with output for prediction purposes. The Chi-Square notation is given in (1)

$$X_i^2 = -\sum(A_i - E_i)^2/E_i \quad (1)$$

Where, A = Actual frequency, E = Expected frequency, $\sum$ = Summation, and X^2= Value of Chi-Square; $0 < i \leq n$: n is the number of patients in the dataset.

Among all the features of our dataset, some were unnecessary, unexplained, and repeated. These features were not included in our analysis. Garate et al. [1] performed three´ different tests for feature selection: In the first experiment, the Chi-Square technique was applied for obtaining a reduced set of features. The second one used a subset of reduced datasets applying the chi-square and then applied PCA. The last test was done by directly using the raw data. After analyzing all the experiments, they have chosen 14 features and Table I reveals the full description of those features. They implemented their classification model using all these reduced datasets and among all the experiments, the second test (combination of Chi-square and PCA) provides the best performance.

D. Clustering

The dataset is clustered by using the Hierarchical Agglomerative Clustering algorithm. This is a bottom-up

approach for hierarchical clustering that starts to create small clusters and then combines the small clusters to create large clusters using Hierarchical Agglomerative Clustering Algorithm [13]. We have implemented hierarchical clustering using Scikit-Learn (sklearn. cluster. Agglomerative Clustering). We have used the number of clusters, n = 10 for our implementation. After clustering the data classifiers are applied to the 10 clusters to estimate the time and measure the performance.

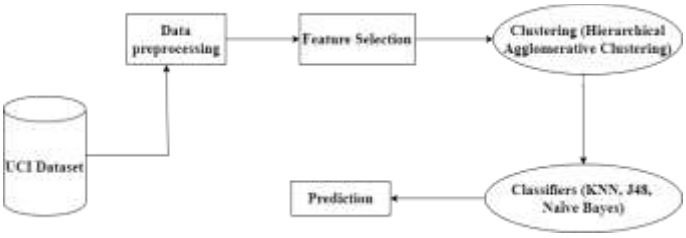

Fig. 1. Architecture of the proposed methodology

E. Classification Models

1) K-NN: K-NN is a type of supervised learning technique that classifies data using the Euclidean distance function. In the K-NN classification technique, test data are used as input, and class membership is an output. It calculates the Euclidean distance between each test data for all training data. Equation (2) is used for measuring the distance:

$$D = \sqrt{\sum_{i=1}^{n}(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_i - y_i)^2} \quad (2)$$

2) J48: J48 is a WEKA implementation of the Iterative Dichotomiser 3 (ID3) algorithm [2]. Based on the training data J48 generates a decision tree that is used to classify the test data. The tree construction is very simple and follows the divide and conquer method. One of the attributes is chosen first and this attribute is used as the root node. Each branch of the tree denotes a test, and each leaf node denotes a class label.

3) Naïve Bayes: Naïve Bayes algorithm calculates the probability of occurring one event using Bayes Theorem assuming the interdependency

among the attributes. Equation (3) calculates the probability of event C in the occurrence of event E.

$$P(C|E) = (P(E|C).\ P(C))/P(E) \quad (3)$$

Where, P(C)= is the probability of occurring event C, P(E)= is the probability of occurring event E, and P(E/C) = probability of event occurrence of event C.

IV. IMPLEMENTATION AND RESULT ANALYSIS

A. Experimental setup

Three algorithms for classification (K-NN, Decision Tree (J48) and Naïve Bayes) were used for performing our experiment. We have used Python (3.9) for data preprocessing and implementing our algorithms. WEKA (Waikato Environment for Knowledge Analysis) was also used for analytical purposes.

B. Train the dataset

After performing the processing, the dataset contains 298 patients. The training data contains 50% of the records and the remaining 50% records were used for testing purposes. Also, for measuring the performance of our proposed method, a total of the first four clusters (40%) were used as training data and the remaining six clusters (60%) were used for testing.

C. Test the dataset for proposed model

For conducting our testing, we have used three classification algorithms. We performed out testing considering:

- *Case 1*: We have implemented the classification algorithms for individual patient and measure the performance for each algorithm.
- *Case 2*: The same process was done for clusters to measure the performance of each classification model.

D. Performance evaluation

To measure the performance of our proposed method we have selected precision, recall, F1 score, and accuracy as performance metrics [14]. The outcome of these four metrics was generated by a confusion matrix: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) and this confusion matrix is shown in Table II.

Table I Description of Selected 14 Attributes

SLN	Attribute	Description
1	age	age in years
2	sex	1 = male; 0 = female
3	cp	chest pain type(1: typical angina, 2:atypical, 3: non-anginal, 4:asymptomatic)
4	trestbps	resting blood pressure (in mm Hg)
5	chol	serum cholesterol in mg/dl
6	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7	restecg	resting electrocardiographic results(0: normal, 1: having ST-T wave abnormality , 2:left ventricular hypertrophy)
8	thalach	maximum heart rate achieved
9	exang	exercise induced angina (1 = yes; 0 = no)
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	slope of the peak exercise ST segment(1:up, 2:flat,3:down)
12	ca	Number of major vessels(0-3)
13	thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	num	diagnosis of heart disease (0: < 50% diameter narrowing, 1: > 50% diameter narrowing)

Precision attempts to find the proportion of correctly classified positive instances (TP) and the total instances which are classified as positive (TP+FP) in the experiment. Recall, contrarily, calculates the percentage of actual positive instances

that are identified correctly. F1 score provides a single number that combines the performance of both precision and recall. We have calculated the above four performance measurements for both the conventional system and our proposed system. Time taken for both systems has also compared which was generated by WEKA. Performance measurement for the conventional system has been shown in Table III and our proposed method in Table IV.

E. Result comparison with the existing method

The accuracy, precision, and recall of the proposed method have been compared with the conventional method. Fig. 2 compares the accuracy between conventional and proposed method graphically. Fig. 3 and Fig. 4 provides information on precision and recall for these two methods respectively. Fig. 5 represents the comparison between the time taken for prediction.

The above four bar graphs illustrate the comparison among conventional machine learning algorithms (K-NN, Naïve Bayes, and J48) and our developed approach in terms of accuracy, precision, recall, and time (in milliseconds). Overall, with regard to the former three, our method reveals slight lower parentage, however, the latter: time- our system is the lowest among the algorithms because we have used clustering on the whole data set.

As a result, the data set is reduced in size by groups rather than considering all the rows of patients in the data set. Hence, our system requires less time for predicting heart disease.

Table Ii Confusion Matrix

Confusion Matrix	Classified as affected by heart disease	Classified as not affected by heart disease
Actually affected by heart disease	TP	FN
Actually not affected by heart disease	FP	TN

Table Iii. Confusion Matrix Obtained For Conventional Method

Classifier	Confusion Matrix		Precision	Recall	F1 score	Accuracy	Time required for prediction (ms)
K-NN	TP=63	FN=23	0.7325	0.7325	0.7325	69.128%	900
	FP=23	TN=40					
Naïve Bayes	TP=89	FN=10	0.9081	0.8989	0.9035	87.248%	586
	FP=9	TN=41					
J48	TP=78	FN=12	0.8667	0.8667	0.8667	83.8926%	694
	FP=12	TN=47					

Table Iv. Confusion Matrix Obtained For Proposed Method

Classifier	Confusion Matrix		Precision	Recall	F1 score	Accuracy	Time required for prediction (ms)
K-NN	TP=3	FN=1	0.75	0.75	0.75	66.67%	230
	FP=1	TN=1					
Naïve Bayes	TP=5	FN=0	0.83	1.0	0.91	83.33%	195
	FP=1	TN=0					
J48	TP=5	FN=0	0.83	1.0	0.91	83.33%	203
	FP=1	TN=0					

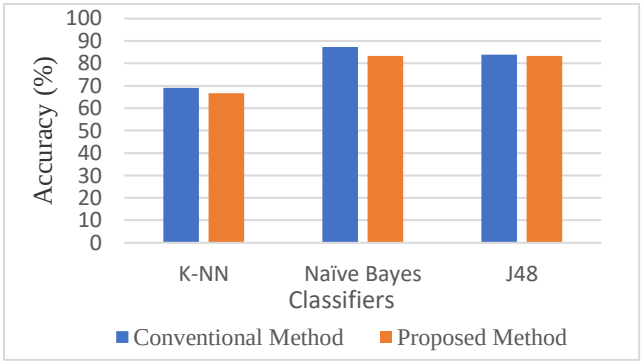

Fig. 2. Comparison of Accuracy of Conventional Method and Proposed Method

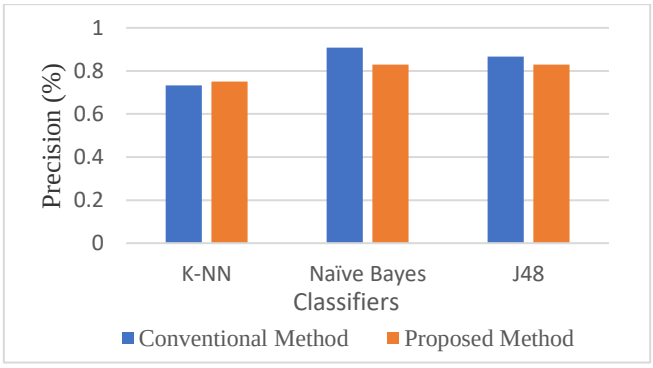

Fig. 3. Comparison of Precision of Conventional and Proposed Method

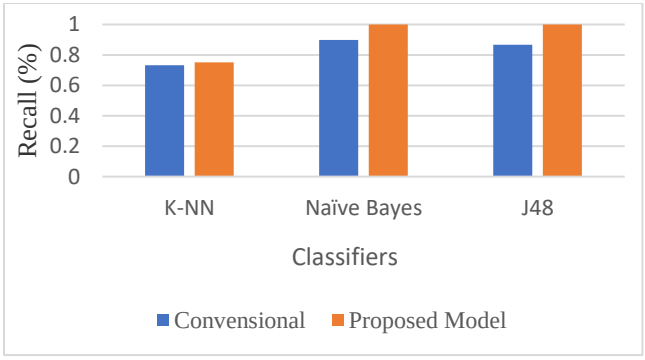

Fig. 4. Comparison of Recall of Conventional Method and Proposed Method

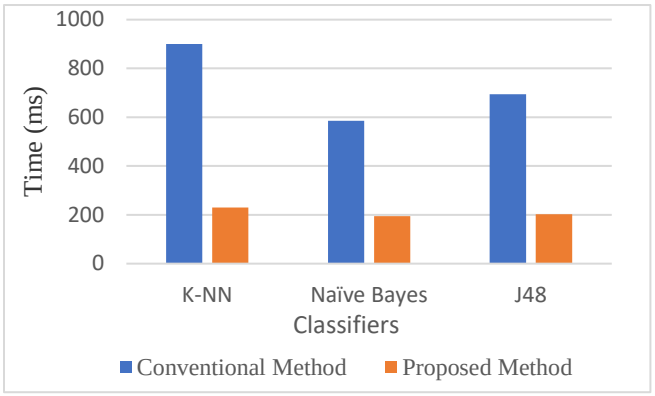

Fig. 5. Comparison of Time of Conventional Method and Proposed Method

V. CONCLUSION

Prediction of heart disease seems to be very tough and challenging for clinical data. Classification algorithms are used for effective heart disease prediction and some of these types of systems have been discussed in the literature review. Although these systems perform well in terms of accuracy the prediction time is significantly high. Minimizing the prediction time will reduce the chance of anomalies in long term heart disease. In this paper, a hybrid method is proposed based on a combination of Hierarchical Agglomerative Clustering algorithm with conventional classification techniques to perform this task. The prediction time of our proposed method is satisfactory in comparison with existing methods, but the accuracy is slightly less. In future, testing can be performed using real data from various hospitals for improved accuracy with less prediction time.

REFERENCES

[1] A. K. Garate-Escamila, A. H. El Hassani and E. Andr´es, “Classification Models for Heart Disease Prediction Using Feature Selection and PCA,” Informatics in Medicine Unlocked (IMU 2020), Amsterdam, The Netherlands, Elsevier, vol. 19, p 100330, 2020.

[2] K. Thenmozhi, P. Deepika, “Heart disease prediction using classification with different decision tree techniques,” IJERGS, vol. 2, no. 6, pp. 6-11, 2014.

[3] M. K. Obenshain, MAT, “Application of Data Mining Techniques to Healthcare Data,” The University of Chicago Press on behalf of The Society for Healthcare Epidemiology of America, vol. 25, no. 8, pp. 690-695, 2004.

[4] K. Srinivas, B. K. Rani, A. Govrdhan, “Applications of data mining techniques in healthcare and prediction of heart attacks,” International Journal on Computer Science and Engineering (IJCSE), vol. 2, no. 02, pp. 250-255, 2010.

[5] N. Bhatla, K. Jyoti, “An analysis of heart disease prediction using different data mining techniques,” International Journal of Engineering, vol. 1, no. 8, pp 1-4, 2012.

[6] A. K. Pandey, P. Pandey, K. L. Jaiswal, A. K. Sen, “A heart disease prediction model using decision tree,” Journal of Computer Engineering (IOSR-JCE), vol. 12, no. 6, pp. 83-86, 2013.

[7] N. Al-Milli, “Backpropagation neural network for prediction of heart disease,” Journal of theoretical and applied information Technology, vol. 56, no. 1, pp. 131-135, 2013.

[8] A. Khemphila, V. Boonjing, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, “A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease,” IEEE symposium on computers and communications (ISCC), pp. 204-207, 2010.

[9] K. M. Almustafa, “Prediction of heart disease and classifiers’ sensitivity analysis,” BMC bioinformatics, vol. 21, no. 1, pp. 1-18, 2020.

[10] S. Joshi, M.K. Nair, “Prediction of heart disease using classification based data mining techniques,” Computational Intelligence in Data Mining, Springer, vol. 2, pp. 503-511, 2015.

[11] S.M.M. Hasan, M.A. Mamun, M.P. Uddin, M.A. Hossain, “Comparative analysis of classification approaches for heart disease prediction,” International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), pp. 1-4, 2018.

[12] S. Mohan, C. Thirumalai, G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, pp. 81542-81554, 2019.

[13] D. Mullner, “Modern hierarchical, agglomerative clustering algorithms,”” Cornell University, Mar. 12, 2011, Accessed on Mar. 02, 2021. [Online]. Available: https://arxiv.org/abs/1109.2378

[14] Accuracy, Precision, Recall F1 Score: Interpretation of Performance Measures, Accessed on Mar. 02, 2021. [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-scoreinterpretation-of-performance-measures/

[15] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, K. Bashir, “Improving heart disease prediction using feature selection approaches,” 16th international bhurban conference on applied sciences and technology (IBCAST), pp. 619-623, 2019.

[16] M. A. Khan, “An IoT framework for heart disease prediction based on MDCNN classifier,” IEEE Access, vol. 8, pp. 34717-34727, 2020.