A New Cygnus Optimization Algorithm for Prediction Of Cardio Vascular Disease

K. Shyamala, T. Marikani

Abstract: ***Machine learning is an emerging field in the present day due to a massive improvement in the size of data. However, with the current studies, the prologue of artificial intelligence and medical sciences, help in averting any such category of disease. With the intention to take good decision in health care, data mining methodology and technology plays a foremost role to transmit the enormous data into valuable information. Emerging as a primary source of fatality in the early 20th century and peaking in prevalence from 1980s, Cardio-Vascular Disease (CVD) remains a major global threat. Early prediction of CVD is most vital for sensible preclusion and treatment. This paper is to investigate the work on the attribute selection approach and propose an improved Cygnus Optimization Algorithm by carrying a random search through the whole attributes, which was obtained from UCI machine learning repository. The proposed method shows 97% of accuracy, which is better than early.***

Keywords: Machine learning, Data mining techniques, Naïve Bayes classifier, Intelligent data analysis, Cygnus Optimization.

I. INTRODUCTION

Lifestyle diseases have been a source of distress in the recent times. Cardio-Vascular Disease (CVD) ruins at the peak in this hierarchy. CVD deals with the heart and allied vascular conditions at immense. According to the American Heart Association (AHA) statistics, 83% of fatality occurs in patients' ≥65 years of age [7]. Cardio-Vascular Disease includes stroke, hypertensive heart disease, cardiomyopathy, congenital heart disease, endocarditic and few more. These may be caused by high blood pressure, smoking, high blood cholesterol, poor diet, diabetes, lack of exercise, obesity. Blockage of the coronary arteries is one of the most universal causes of heart disease [3]. Heart attack is the stumbling block of the arteries and vessels which supply oxygen and nutrient-rich blood to heart. Moreover, it is known as coronary heart disease and is the foremost source of fatality in humans [2]. This occurs when the coronary arteries turn out to be blocked or congested. Commonly this leads to an irregular heartbeat called an arrhythmia which lead a rigorous decline in the pumping function of the heart. If the blockage is failed to treat within few hours, then it lead the affected muscle to expire. In paper [11], prediction of heart disease using supervised learning algorithm had been proposed. In this work the dataset are extracted from Cleveland heart disease dataset, where it contain 303 instances, which contains 72 attributes in the dataset. Among them 14 attributes are selected using data dependency concept to do the research activity. In paper [13], in-order to reduce the size of the attribute for research purpose, a Novel Dyno-Quick Reduct Algorithm is used. This algorithm uses the feature selection method to reduce the attribute size. Finally, in the previous paper [14] Modified Multinomial Naive Bayes Algorithm has been proposed where the accuracy percentage was 74.8%. The proposed algorithm helps to improve accuracy percentage in prediction of heart disease.

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author

K. Shyamala*, Associate Professor, P.G & Research Dept. of Computer Science, Dr. Ambedkar Government Arts College(Autonomous), Chennai, Tamilnadu, India

T. Marikani, Assistant Professor, Dept. of Computer Science, Sree Muthukumaraswamy College, Affiliated to University of Madras, Chennai, Tamilnadu, India

II. RELATED WORK

Dhruvi Ragesh Parikh et.al. [1], proposed the work on Meta-Classifier Bayesian Multinomial. The author provided a proficient technique for predicting the probability of Cardiac-Vascular Disease. They proved that, for the disease prediction grouping of heterogeneous classifiers gave better accuracy compare to individual base classifiers. In order to provide real time summarized statistical report, they use web scraping and text summarization techniques. Harini D K et al. [2], implemented structured and unstructured data to predict the risk of disease in healthcare industry. Authors have undergone three methods in this work. In the first phase, the authors used latent factor model to reconstruct the missing data from the medical record. In the second phase, by using the statistical tool they determined the Cardio-Vascular Diseases. In the third phase, for unstructured text data they selected the features automatically using CNN algorithm. Finally, for disease prediction the author proposed novel CNN-based Multimodal Disease Risk Prediction (CNN-MDRP) algorithms which contain both structured and unstructured data. The proposed algorithm predicted the accuracy of 94.8% than that of other prediction algorithm. Chaithra N et al. [3], surveyed and compared three different classification techniques for heart disease prediction. The authors used various classification techniques like J48 Decision Tree, Naive Bayes and Neural Network on the prediction of Cardio-Vascular Disease. The analysis shows that Neural Network performed improved outcome in predicting the heart disease with 97.91% of accuracy. This analysis helps the junior cardiologists and echo technicians to monitor the patients who have a high probability of having the disease and relocate those patients to senior cardiologists for further analysis.Mrutyunjaya Panda et al.

[4], performed a work on Discriminative Multinomial Naïve Bayes with various filtering analysis in order to build a network intrusion detection system**.** The proposed work that combines discriminative parameter learning using Naïve Bayes classifier with principal component analysis as a filtering approach. The author used the Data Mining Naive Bayes classifier as base classifier and binary supervised filtering approach with the intention of constructing a well-organized network intrusion detection system. They proposed a novel filtered meta classifier approach which is faster and accurate in comparison to other existing approaches.Frantisek Babic et al. [5], presented work on heart disease which describes a range of conditions affecting our heart. The authors used various statistical and data mining methods to recognize different medical data sets, to produce the result. The selected methods are Decision Trees, Naive Bayes and Support Vector Machine. Authors focused on two directions: a predictive analysis and descriptive analysis. The predictive analysis based on Decision Trees, Naive Bayes, NN and Support Vector Machine while the descriptive analysis based on association and decision rules.In paper [11], in our earlier research, supervised learning algorithms are used to analysis the heart disease and also experimented that how these algorithms performed better results by applying in orange tool. The research work of this paper concluded that the Naive Bayes algorithm produces 81.7% of accuracy, whereas classification tree produce 95.4% of accuracy and Random forest produce 96.3% of prediction. Further the research work focused on attribute reduction method to reduce the attribute in-order to make more accuracy for the prediction

III. SUPERVISED LEARNING ALGORITHM

Supervised machine learning is the exploration for algorithms that motive from externally supplied instances to yield overall hypotheses, which then make predictions about future instances. A common method for comparing supervised machine learning algorithms is to perform statistical comparisons of the accuracies of trained classifiers on specific datasets [8]. Supervised classification is one of the tasks most recurrently conceded out by machine learning intelligent systems. Accordingly, a large amount of techniques have been developed based on artificial intelligence, Perceptron-based techniques and statistics method like Naïve Bayes algorithm, Bayesian network and Multinomial Naïve bayes.

IV. IMPLEMENTATION

Classification method using Cygnus Optimization algorithm proposed in this work. Naïve Bayes algorithm is one of the classification algorithm, where this algorithm is analysised and compared with the proposed work. Naïve Bayes Algorithm consists of various methods like Multinomial Naïve Bayes, Binomial Naïve Bayes, Gaussian Naïve Bayes. Naive Bayes classifiers are a family of simple probabilistic classifiers by applying bayes theorem with strong independence assumptions between the features [6]. It is the simplest and the fastest probabilistic classifier especially for the training phase [7].

A. NAIVE BAYES ALGORITHM

Bayesian networks are the most well-known representative of statistical learning algorithms. The foremost improvement of the Naive Bayes classifier is, it take short computational time for training the data. It is very useful to do probabilistic prediction [10]. It can handle both continuous and discrete value. Naïve bayes algorithm can be used for both binary and multi-class classification problems. The set of factors and their possible values are shown in Table 1. Figure 1 shows Naïve Bayes computation for the heart disease problem.

Table 1. Factors and values for Heart disease problem

Factors	Values
Chest pain	**{H, M, L }**
Chol	**{V}**
Fbs	**{T,F}**
Restecg	**{N, ABN}**
Exang	**T,F**
Thal	**{F,N,R}**

L=Low, H= High, M= Medium, F= Fixed, N= Normal, R = Reversable, T= True, F= False, N= Normal, ABN =Abnormal, V=Value.

B. CYGNUS OPTIMIZATION

Cygnus optimization algorithm follows the principle of the bird swan. The unique feature of the swan is to extract the milk alone, from a mixture of milk and water, leaving aside the water. In the same way, the unique attributes are selected from the Cleveland dataset by using the proposed algorithm. The algorithm works as, how the swan put its neck down into the deep water, brings up food for itself from inside the water by means of same approach, from the collected dataset the irrelevant data are filtered by using the filtering technique, then by using the feature subset selection method the random values are removed from the dataset. The following fig 1 clearly depicts the workflow of a New Cygnus Optimization algorithm.

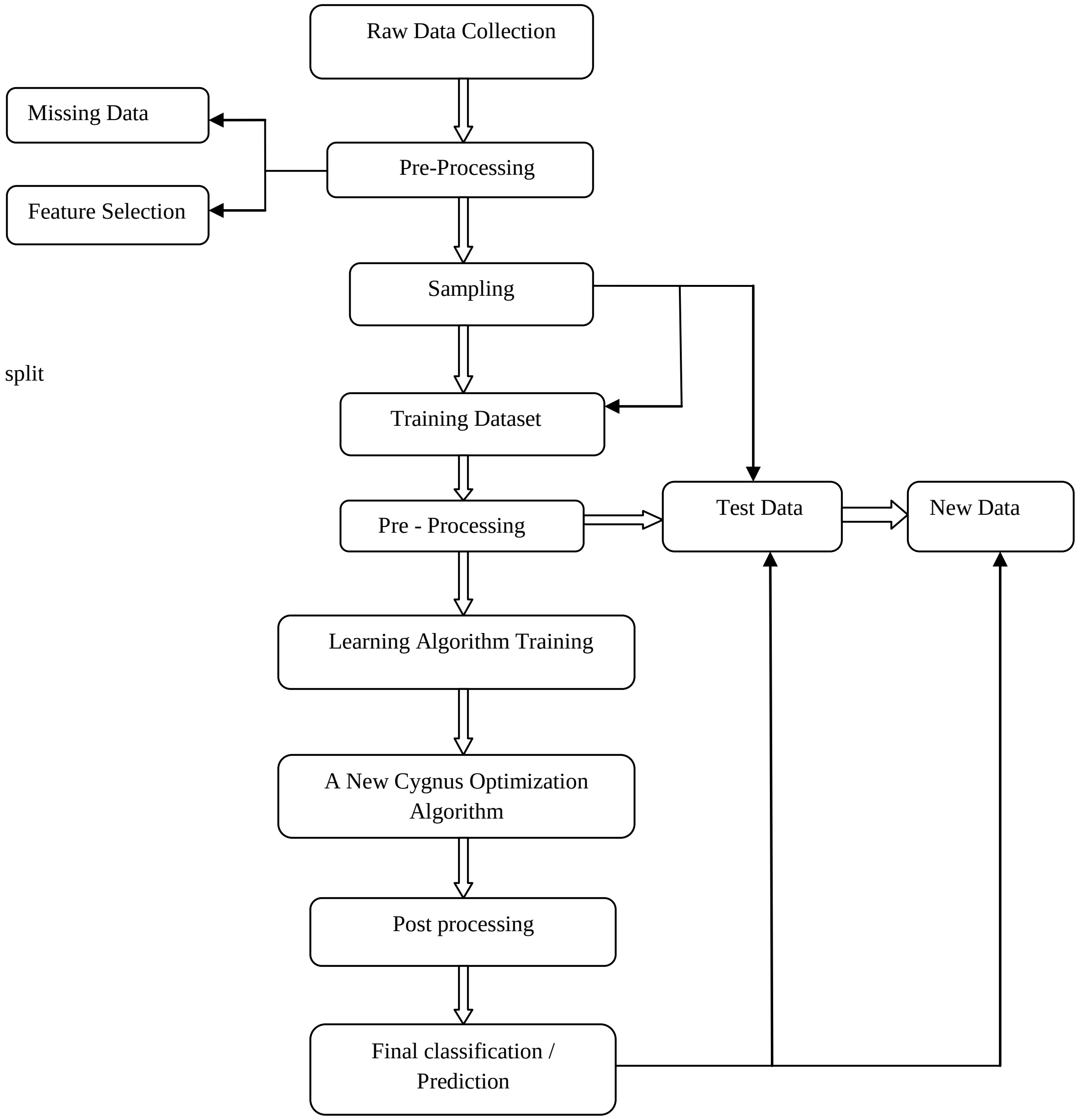

Fig B 1.Work flow of New Cygnus Optimization Algorithm

C. ALGORITHM FOR CYGNUS OPTIMIZATION In this segment, the research work focus on the methodology to elucidates the performance of Cygnus Optimization Algorithm classifier for prediction of heart disease. Algorithm 1.for proposed work is given below:

Algorithm 1: Cygnus Optimization Algorithm
Input: Dataset, count_docs, attributes, sum_prob, predict _prob, pos_pred
Output: Accuracy prediction percentage

```
Begin
    A= Total data, C=count_probabilities, Col_prob= column probability, pos_pred= positive predict, Cal=calculate_probability, R=Row, cmp=compare, s=sum_prob, T=total number of terms, p=probabilities
    new data← Removed Random values(data) //Random values are removed from the dataset and moved to new data
        for each c € A
          repeat
              Count the col_prob(A,Cal)
          until element found in the set
        for each s € C
          repeat
            do R←count for the total prob and check if(C > 0.3) //count total prob and check if count_prob is greater than 3
          until the total row counted
        for each p € R
        repeat
          do condprob[p][R]←[totalprob / 2] //Divide the total obtained probabilities by 2 and assume to condprob
```

Retrieval Number: L3686081219/2019©BEIESP
DOI: 10.35940/ijitee.L3686.1081219
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
15.         until conditional probability obtained
16.       for each $cmp \leftarrow P$
17.         repeat
18.           do pos_pred[p][R] $\leftarrow$maxpred/2
19.         until pos_pred count
20.     Return pos_pred
21. End
```

The above algorithm conspicuously predicts the percentage of heart disease among the patients which was collected from machine learning repository [15]. The algorithm commence from getting the input values as total dataset, column probability, row probability, count probabilities, positive predict values, sum probabilities and total number terms. In line number 3, the random values are removed from the dataset using filtering technique and remaining values are stored in new location. From step number 4 to 7, column probabilities are calculated from list of attributes and counted until the search element found in the dataset. From step 8 to 11, sum probabilities are calculated and do the row count for the total probabilities and check if condition is greater than 0.3 until the total row counted. From step 12 to 15 calculate the total prediction probabilities and divide the total probabilities by 2, do the same operation until conditional probability obtained and finally in the step 16 to 20 compare the probabilities and find the positive predict row by dividing the maximum prediction value by 2. Finally, the algorithm counts the positive predict data and predicts the positive value of occurrence of heart disease among the patients.

V. RESULT ANALYSIS

A. NAÏVE BAYES CLASSIFIER

Naïve Bayes algorithm used to predict the accuracy of heart disease among the patients which is selected from Cleveland dataset. The dataset has been implemented in Orange tool with Python software. The following table 2 shows the prediction percentage of heart disease.

Table 2. Naive Bayes Prediction Percentage

Classification Method	Scoring Method	Accuracy %
Naïve Bayes Algorithm	AUC	81.80%
	CA	81.70%
	Precision	84%
	Recall	81.30%
	F1	82.60%

B. CYGNUS OPTIMIZATION ALGORITHM

The proposed work Cygnus Optimization algorithm select the attributes which most positive discrimination to predict the disease among the patients. It selects the most relevant six attributes from the dataset among the 72 attributes by using the feature selection concept from the rough set theory. Feature selection is also recognized as attribute selection method or variable selection method. It is the method of selecting automatically the attributes of our data which is appropriate to the predictive modelling problem [11]. The following table 3 shows the result of Cygnus Optimisation algorithm. In the proposed method, the accuracy of Naive Bayes classifier is improved by using the Cygnus Optimization Algorithm. The data collected from Cleveland dataset UCI repository, in which the dataset that is being used in the method contains six features including one class label and 230 instances with no missing values. Here the positive phenomenon of heart disease is predicted on the basis of cholesterol, chest pain type, fasting blood sugar, resting ECG, exercise angina, smoke.

Table 3. Cygnus Optimization Algorithm with Prediction Percentage

Proposed Method	Scoring Method	Accuracy %
Cygnus optimization algorithm	Precision	97.6%

Where, AUC is Area under curve. CA is classification Accuracy. The precision is the ratio of;

Precision = tp/(tp + fp)

Where, tp is the number of true positives and fp the number of false positives. Where, precision is computed as accuracy percentage of the algorithm.The recall is the ratio of;

Recall = tp / (tp +fn)

Where, tp is the number of true positives and fn the number of false negatives. F1 score, also known as balanced F-score or F-measure. The F1 score can be interpreted as a weighted average of the precision and recall. The formula for the F1 score is:

F1 = 2 * (precision * recall) / (precision + recall)

C. COMPARISON ANALYSIS

Based on analysis of the results for heart disease prediction from the table 2 and 3, the classification with Naïve Bayes and Cygnus Optimization Algorithm summarized in table 4. Naïve Bayes Algorithm produces accuracy percentage of 84% (i.e) the positive predicted value (precision). While in the proposed Cygnus Optimization Algorithm, the accuracy percentage was increased to 97%.In the same way, various supervised algorithms prediction percentage are calculated. The predictive model with Random forest classification is 89%, SVM predict the accuracy of 85.6% and classification tree produce 94%.The research concludes that, the proposed research work improves the accuracy of prediction percentage which helps the doctor to advice the patient for further diagnosis process. The following table 4 shows the prediction performance of various supervised learning algorithm.

Table 4. Prediction performance of various supervised algorithm

Classification methods	Prediction %
Naïve Bayes	84%
Random Forest	89%
SVM	85.6%
Classification Tree	94%
Cygnus Optimization	97%

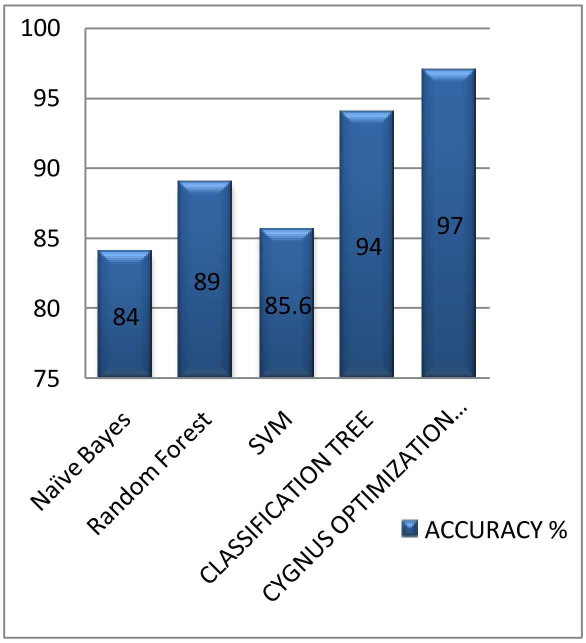

Figure 2. Comparison result of various supervised algorithm

The following chart clearly depicts the evaluation result of various classification algorithms. Fig 2 helps to find out the comparison result (i.e) accuracy percentage of different learning algorithms.

VI. CONCLUSION

It was concluded that the proposed work is effective and efficient to improve the accuracy of the classification algorithm using the Cygnus Optimization for feature subset selection which achieves better classification performance. The goal to predict the heart disease among the patients with minimum attributes was successfully achieved by developing a new adaptive Cygnus Optimization Algorithm. From evaluation results, it is analyzed that COA could automatically evolve a feature subset selection with a less number of features and increase classification performance than using all the features of a dataset. In future, Cygnus optimization feature subset selection can be implemented on decision tree classification to obtain better outcome.

REFERENCE

1. Dhruvi Ragesh Parikh, Yavnika Rajendra Bhagat and Nutan Ramesh Ghanwat “Prediction of Probability of Chronic Diseases and Providing Relative Real Time Statist ical Report using data mining and machine learning techniques”, International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2278 – 7798, Vol.5(4), pp. 1009-1014, April 2016.
2. Harini D K and Natesh M , “Prediction of probability of disease based on symptoms using machine learning algorithm”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, Vol.5(5), pp. 392-395, May 2018.
3. Chaithra N and Madhu B, ”Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques”, Journal of Cardiovascular Diseases and Diagnosis, ISSN: 2329-9517, Vol.6(6), pp. 1-4, 2018.
4. Mrutyunjaya Panda, Ajith Abraham and Manas Ranjan Patra, “Discriminative Multinomial Naïve Bayes for Network Intrusion Detection”, In proceedings of Sixth International Conference on Information Assurance and Security, pp.23-25, Aug.2010.
5. František Babič, Jaroslav Olejár, Zuzana Vantová and JánParalič, “Predictive and Descriptive Analysis for Heart Disease Diagnosis”, Proceedings of the Federated Conference on Computer Science and Information Systems, FEDCSIS, ISSN 2300-5963, Vol. 11, pp. 155–163, 2017.
6. Santhi.P and Dr.V.Murali Bhaskaran,”Performance of Classification Algorithms in Heart Disease Data”, International Journal of Advanced Research and computer science, ISSN No. 0976 -5697, Vol.2(3), pp.63-70, June 2011.
7. Uma N Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization”, Biomedical Research, Vol 29(12), pp.2646-2649, 2018.
8. Kotsiantis.S.B.,”Supervised Machine Learning: A Review of Classification Techniques”, Informatica, Vol. 31, pp. 249-268. 2007.
9. Madhu.G, Rajinikanth.T.V and Govardhan.A, “Feature Selection Algorithm with Discretization PSO Search Methods for Continuous Attributes”, International Journal of Computer Science and Information Technologies (IJCSIT), ISSN No. 0975 – 9646, Vol. 5 (2), pp. 1398-1402, 2014.
10. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
11. Marikani.T and Dr.K.Shyamala, “Prediction of Heart Disease using Supervised Learning Algorithms” International Journal of Computer Applications (IJCA), ISSN No.0975 – 8887, Volume 165(5), pp.41-45, May 2017.
12. Sfenrianto, Indah Purnamasari and Rizal Broer Bahaweres, “Naive Bayes classifier algorithm and Particle Swarm Optimization for classification of cross selling (Case study: PT TELKOM Jakarta)”, In proceeding of 4th International Conference on Cyber and IT Service Management, pp. 1-4, April 2016.
13. Marikani.T and Dr.K.Shyamala, “A Novel Dyno-Quick Reduct Algorithm for heart disease prediction using supervised learning algorithm”, Accepted for publication in ICCVBIC Springer Proceedings, International conference on computational vision and Bio-inspired computing(ICCVBIC), Nov (2018).
14. Marikani.T and Dr.K.Shyamala, ,”Modified Multinomial Naïve Bayes Algorithm For Heart Disease Prediction”, Accepted for publication in ICICV Springer Proceedings, International conference on Intelligence Communication Technologies and Virtual Mobile Network (ICICV), Feb(2019).
15. https://archive.ics.uci.edu.