A Prediction of Pediatric Cardiomyopathy Disease Associated Genes using Machine Learning Algorithms

K.Jayanthi, C. Mahesh

ABSTRACT--- Pediatric cardiomyopathy is considered as one of the heart diseases, which causes by abnormal disorder of the heart muscle. If pediatric cardiomyopathy remains untreated and unidentified at the early stages, it leads to heart failure. The global number of deaths and disability attributed to cardiomyopathy has steadily increased. Hence, machine learning approaches can solves the problem of identifying the critical problem by determining the pediatric cardiomyopathy disease associated genes from the collection of differentially expressed genes that are recognized by biological process of genes. The main objective of this study is to design a machine learning model which can predict the likelihood of pediatric cardiomyopathy in genes specified biological features with maximum of accuracy. Identified high throughput machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine were used in this experiment to determine the genes which can be derived from internal database repository having biological process of genes specified. Experiments are conducted on Gene Expression Omnibus (GEO) datasets which sourced from cardiogenomics.org and Biohunter tool. The performance of these machine learning algorithms is evaluated on various measures like Accuracy, Precision, Recall, F-Measure, and Receiver Operating Characteristics (ROC). From the obtained results shows that Random Forest provides high accuracy 84.4% when compared to other four machine learning algorithms.

1. INTRODUCTION

Cardio vascular disease is one of the primary causes of death in human life, and is abstract by environmental as well as genetic factors. Using the advances in various tools like genomic technologies, the possibility to diagnose and predict heart diseases. Several Researchers are conducting different kinds of experiments for diagnosing disease associated genes using classification Algorithms of machine learning techniques like Logistic Regression, Support vector machine, Decision tree etc., They proved that machine learning algorithms supports for diagnosing various types of diseases. In this work, Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest algorithms are used and evaluated on the biological dataset for the prediction of pediatric cardiomyopathy disease associated genes. Comparative Analysis of Performance metrics of all these algorithms is monitored with measures and results were obtained. Various kinds of disease predictions carried out in machine learning algorithms have been discussed. Using supervised machine learning techniques the performance of disease gene prediction were compared and proved that Random forest provided best results. D.le et al. proved in comparative study of classification based machine learning methods for novel disease gene prediction[1]. For the prediction of disease gene associations Ping Luo et al had proposed a new multimodal deep belief network method. It achieved an AUC of 0.969, also proved that with the help of Protein Protein Interconnection network can lead to predict disease gene associations with better accuracy [2]. Four machine learning algorithms such as Naive Bayes, Random Forest, multilayer perception and J48 were used to detect Demantia by Deepika Bansal et.al[3]. The comparative study proved that J48 provides best accuracy then other algorithms.To predict the prevalence of heart disease using Cleveland dataset Divyansh khanna et el [4] proved that logistic regression and SVM with linear kernel provides more accurate results and for performance analysis F1 Score and ROC were used. To find the minimum number of attributes required to enhance the precision of various supervised machine learning algorithms Dhomsekanchan and Mahalekishor [5]used Principal Component analysis for the prediction of diabetic disease. New convolution neural network were proposed by Min Chen et al. [6] for the prediction of chronic disease and they used latent factor to reconstruct the missing data. This new convolution algorithm provides 94.8% of accuracy. For the prediction of Alzhemier's disease Xiaoyan Huang et.al [7] used support vector machine and proved this algorithm reaches accuracy with 84.56 and ROC with 94%.

2. PROPOSED METHODOLOGY

The architecture of the proposed implemented work consists of four steps: Normalization of the Gene Expression Omnibus data, Biological and Statistical significance (Fold Change and P- Valve), differentially expressed genes identification using machine learning algorithms, and evaluation of results. The proposed methodology is described in Figure1.

Revised Manuscript Received on August 19, 2019.

K.Jayanthi, Research Scholar, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, TamilNadu, India.(E-Mail: 1jayanthi2contact@gmail.com)

C. Mahesh, Associate professor, Department of Information Technology Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, TamilNadu, India.(E-Mail: 2chimahesh@gmail.com)

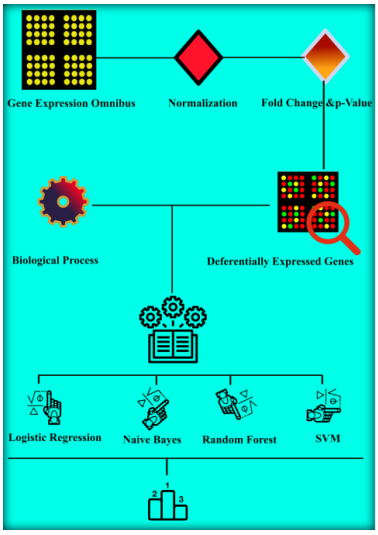

Figure: 1: Architecture of Proposed Methodology

3. METHODOLOGY

3.1 Data Collection

The dataset was taken from Gene Expression Omnibus (GEO) website. It is one of the repositories which provide the data publically available. Around 5700 genes identified with cardiomyopathy from that 1263 genes related to pediatric cardiomyopathy which gathered from different patients both typical and disease related genes.177 features which are identified as biological process that are purely related to cardiomyopathy genes. The class attributes: value is to determine whether the features supported to predict pediatric cardiomyopathy disease associated gene or not (Yes/No).

3.2 Normalization of Data

To identify the significant relationship between the gene expression patterns of pediatric cardiomyopathy log transformation were used. It also used to normalizing of data. For normalization logs along with geometric mean were used. Using these log transformation method distributions which are specified as highly skewed will be made less skewed. Further, normalization of genes which has less skewed are in the ranges {0 to 1}. And the highly skewed were specified in the range of {1 to 0}. To represent the data as more symmetric log transformation provides maximum support. It also leads parametric statistical test to provide both accuracy and relevancy at higher level of significance.

3.3 Biological Significance

It represents the measurement of how the quantity changes from the starting value to end value. Normally it's referred as fold change as given in equation 1. Here log2 fold change is obtained by taking the levels of expression between two conditions namely case and control[8].

$$log_2 F = log_2(CF)$$
$$= log_2\{of\ ration\ case\ and\ control\ data\}$$
$$= log_2 \frac{case}{control}$$

$$F = 2\left|log_2\ \frac{avg\ c}{avg\ N}\right| \quad \textbf{(1)}$$

3.4 Statistical Significance

Two set of conditions (case and control) are the different groups. To find out the significant difference between these two set of groups independent t-test can be used [8].While comparing the means of these conditions will represent how both are significantly different from one another. Let Case (C) and Control (T) represent the two groups to compare, (mc) and (mT) represent the means of groups C and T respectively. nC and nT represent the size of groups C and T. The t-test which determine the means are different as mentioned in equation 2. The formula is shown below.

$$t = \frac{m_c - m_t}{\sqrt{\frac{s2}{n_C} + \frac{s2}{n_T}}} \quad \textbf{2)}$$

where s^2 is the common variance of the two samples. It can be obtained by equation 3

$$s^2 = \frac{\sum(x-m_C)^2 + \sum(x-m_T)^2}{n_C + n_r - 2} \quad \textbf{(3)}$$

Once t- test value is calculated. Significance level of p-values which are less than 0.05 has choosen to determine the disease associated gene related to pediatric cardiomyopathy. In this paper, different machine learning algorithms that are utilized and experimentations such as Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machines used to determine the class whether the biological features are used to predict the cardiomyopathy disease associated genes.

3.5 Logistic Regression

Most popular predictive analysis algorithm is logistic regression. It is used to provide the relationship between one dependent variable and one or more independent variables,[4] and it is used to describe the data, some cases logistic regression interpretation of data is difficult. In this study, the biological features are independent variables. Using this, Prediction of dependent variables should be obtained as two set of classes whether the dependent variable is pediatric cardiomyopathy associated gene or not. Logistic regression is a non linear transformation of the linear regression analysis. It is derived with sigmoid function. It is called as shaped logistic distribution function. The main objective of logistic regression is to find the best model to fit our data. One of the standard classification techniques which is based on categorical variables. In this study we want to predict whatever the features we have used that are related to pediatric cardiomyopathy genes are not (Yes/No).Basically Logistic Regression deals with the probabilistic statistics values of the data [4]. Probabilities can be mapped from predicted values for that, we utilize the sigmoid function. The function maps any real value into another value somewhere in the range of 0 and 1. sigmoid function has used to outline the probabilities which is given in equation 4.

$$X(z) = \frac{1}{1+e^{-z}} \quad \textbf{(4)}$$

X(z) – Output that determines the values between 0 and 1.
z- Input to the function
e -Base of natural log.

3.6 Naive Bayes

Another classification technique is Naive Bayes which is defined with all features are independent values and not associated to each other. The status of one feature in a class does not affect any other feature. Basically, this algorithm works on conditional probability so it has been utilized for classification purpose. It supports in the need of data which has balancing problems and deals with missing data. This machine learning algorithm derived from Bayes Theorem. It is a posterior probability which is based on the training data. It uses Bayes theorem to find the pattern class. The assumption of the posterior probability is treating every feature as a class that follows conditionally independent.[5] So it makes the probabilistic classifier as very simpler. If this assumption of the naive bayes classifier is true. Then the results of this classifier will be used for making comparison over other classifier algorithms like neural network and classification trees. The main advantage of using naive bayes classifier is calculations are very simple so it can be used for large databases and results are calculated very fast with proper accuracy levels[9]. Since, it provides minimum error rate, so that we are using this classifier in their research we used naive bayes classifier. The maximum posterior probability will be

P (class: yes|no) | (B1, B2 .....Bd)
P (Class| Pattern)

Where B1, B2.....Bd are the features of biological process of a gene. We would like to predict the probability of how these features are associated to predict the disease related gene (X) .from the input samples (training samples) and labels(Y).Using this Bayes theorem.

P (predict disease associated) gene training samples) = P (training samples disease associated gene)*P (disease associated gene)/P (training samples).

Using Bayes theorem, If class yes can be as,

$$P(yes|B1, B2, \dots Bd) = \frac{P(yes) * P(B1, B2, \dots Bd|yes)}{P(B1, B2, \dots Bd)}$$

And if class no, then

$$P(no|B1, B2, \dots Bd) = \frac{P(no) * P(B1, B2, \dots Bd|no)}{P(B1, B2, \dots Bd)}$$

In this biological dataset there are 177 features and these features are used to predict pediatric cardiomyopathy disease associated gene or not. The naive Bayes classifier will predict this disease associated gene from the data which is used to test.

The prior probability of class yes as,

$$P(diseae\ associated\ gene = yes) = \frac{Total\ Number\ of\ yes}{No\ of\ biological\ features}$$

The prior probability of class no as,

$$P(diseae\ associated\ gene = no) = \frac{Total\ Number\ of\ no}{No\ of\ biological\ features}$$

For all the individual features will be analysed with the class label to find the class pattern (yes or no) that is pediatric cardiomyopathy disease associated gene or not. Finally naive Bayes classifier algorithm will obtain with the best hypothesis of given biological dataset.

3.7 Random Forest

Diversity of the data can be improved by machine learning classification algorithms. One of the Meta classifier is random forest that means, it consists of many trees (Individual learner). The random forest is the combination of multiple random trees they voting on a specific formation. The weight of the each vote is equal in the random forest algorithm.

Diversity of the data can be improved by machine learning classification algorithms. One of the Meta classifier is random forest that means, it consists of many trees (Individual learner). The random forest is the combination of multiple random trees they voting on a specific formation. The weight of the each vote is equal in the random forest algorithm. The process of random forest classifier algorithm shows in Figure 2.

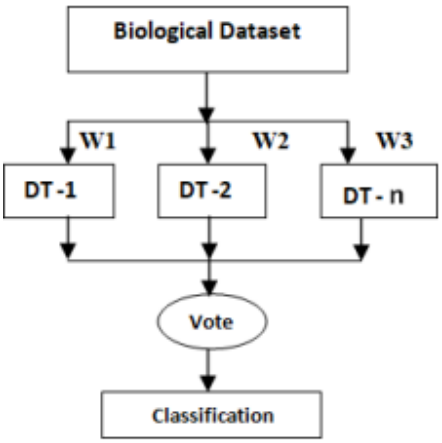

Figure 2: Random Forest Classifier.

1. A set of decision trees will be formed from training data using random selection.
2. Then the different vote aggregates from all the decision trees.
3. From that the final class will be decided by majority of votes from each of the decision tree which has been formed.

Random forest is most familiar classification technique which uses collection of decision trees. This algorithm randomly selects labels and features to generate multiple decision trees and then average will be calculated. That will predict as a result of this algorithm. Over fitting can be prevented by using this random forest algorithm.[10] Several possibilities of decision trees can be used to

determine which features provides more suitable to determine the labels which leads to conclude whether the inputs(features) predict the genes which is pediatric cardiomyopathy associated gene or not. (Yes/No).

3.8 Support Vector Machines (SVM)

One of the discriminative classifier which is defined by a separating hyperplane is known as SVM. It supports to separating the two classes deals with, from the biological dataset we have to predict whether the features of biological data is related to pediatric cardiomyopathy disease associated genes or not. In general, to separate two classes in x-y plane, one more dimension Z- axis is added to apply transformation when there is no line to separate the 2 classes. That Z- plane follows the equation 5.

$$W = x^2 + y^2 \quad (5)$$

SVM is one of the most accurate and robust method among various machine learning algorithms. Minimum number of examples is enough for training. In Binary classification learning task, the main objective of SVM is to capture the better classification function to differentiate the samples of the two classes in the training data. The measurement for the idea of the "best" classification capacity can be acknowledged geometrically. Since biological dataset is a linearly separable dataset, so that direct classification work relates to an isolating hyperplane s(x) that goes through the middle part of the two classes, isolating the two. When this capacity is resolved, new information occurrence xn can be ordered by just testing the indication of the capacity s(xn).

xn - positive class
if s(xn) > 0.
x - Number of features
s(x) – Function of x.
That classifies the two classes yes or no.
xn – testing data
s(xn) – function of testing data

There is more number of linear hyperplanes. SVM selects best function by maximizing the margin in between the two classes (yes/no).[11] This margin is termed as separation between the classes by the hyperplane. Using this SVM in our research will provide the geometric margin to predict the disease associated gene from the biological feature.

4. EXPERIMENTAL DESIGN AND RESULTS

4.1 Design

The experiment of this study done with dataset which deals 177 attributes that are related to biological process of various genes. Biological process is a process of a living organism. These biological processes are composed with various chemical reactions that provide chemical transformation. Metabolism is one the example for biological process. This process examined by different categories. Such as gene expression analysis, protein interaction with another protein. In this study biological processes of various genes can be analyzed and that will be used as an input to the machine learning algorithms. Determine is a label which is used to predict the disease associated genes related to cardiomyopathy.

4.2 Results

The pre-processed biological process dataset divided into two set of data. Training data and testing data. From the dataset 70% of data used as training and remaining 30% of data used for testing data. Training data can trained with various machine learning algorithms such as Logistic Regression, Naive Bayes, Random Forest, SVM. The results of these five algorithms were analysed. The testing data can used to for predicting the disease associated gene. While comparing these algorithms with performance measures like Accuracy, precision, recall, F- measure and ROC. The Random Forest provides best results with 84.4% accuracy. 83% precision, 96% recall, 89% F-measure and 0.9 Receiver operating characteristics.

5. PERFORMANCE EVALUATION

5.1 Metrics

The main objective of this paper is comparative analysis of machine learning algorithms with biological feature of genes. It can be experimented and performance evaluation of Machine Learning algorithms will used to determine the pediatric cardiomyopathy disease associated genes or not. The performance is measured in terms of accuracy, precision, recall, and F- measure, ROC in order to predict the cardiomyopathy disease associated genes. Performance measures are generally obtained by four measurements. Such as True positives, True Negatives, False positives and False Negatives. Based on significance of loss of our predicted model.

Accuracy: It is one of the important performance metrics which is used to determine out of all biological features how many features are correctly determines the class that is how many features are used to predict the pediatric cardiomyopathy disease associated gene.

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

Precision: It measures how many correct matches are found among all data sets. This metrics used to determine how many biological features are predicted as disease associated gene as correctly classified.

$$Precision = \frac{TP}{TP + FP}$$

Recall: It is the relation between correct matches and correspondences. It is the division of relevant instances that are retrieved as result.

$$Recall = \frac{TP}{TP + FN}$$

F-measure: F-Measure is the weighted average of precision and recall.

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

ROC (Receiver Operating Characteristics): ROC curves are used to compare the usefulness of tests.

5.2 Output table

The experimental results shown in the following Table1.

Table:1 Comparative Results of machine learning algorithms

Model	Accuracy (%)	Precision	Recall	F-Measure	Roc
Logistic Regression	71.2	0.71	0.95	0.81	0.68
Naive Bayes	66	0.69	0.86	0.77	0.6
Random Forest	84.4	0.83	0.96	0.89	0.9
Support Vector Machine	70.4	0.70	0.97	0.81	0.5

5.3 Analysis of Performance Measure

Comparision of machine learning algorithms for the prediction of pediatric cardiomyopathy disease associated gene with biological features performance measure shows in the following figure3.

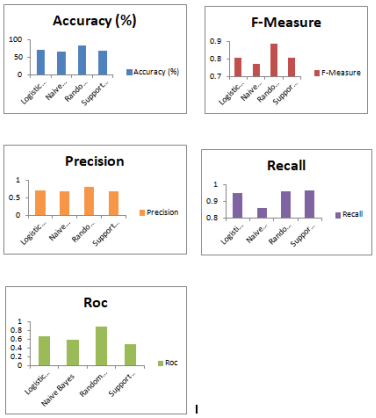

Figure:3 Comparision of Algorithms for disease associated gene prediction (Random Forest performs best)

5.4 Analysis of output

This paper deals with biological process of various genes can used to predict the pediatric cardiomyopathy disease associated gene or not. For better prediction we trained and tested with five machine learning algorithms LR, NB, Random Forest and SVM. From the obtained results Random forest classifier provides better results when comparing other four classifier algorithms in terms of accuracy, precision, recall, f-measure and ROC.

6. DISCUSSION

In this paper how features of biological process can used to predict the pediatric cardiomyopathy disease associated gene. In pre-processing internal database can used to filter the genes to determine the differentially expressed genes[12]. that are associated to cardiomyopathy. These data then can be used for prediction. Prediction analyzed with five machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest, and SVM. The comparative analysis done with five performance measures that concludes in all aspects Random Forest suggested to be best classifier to predict the cardiomyopathy disease associated genes.

7. CONCLUSION AND FUTURE WORK

Main focus of this paper is to enhance the prediction of pediatric cardiomyopathy disease associated genes in order to improve the accuracy of diagnosis From the literature studies understand that biological process of various genes can provide better solution to predict pediatric cardiomyopathy disease associated genes[13]. For prediction five machine leaning algorithms had used. From the comparative analysis of machine learning algorithms: Logistic Regression, Naive Bayes, Random Forest and SVM the performance in terms of accuracy, precision, recall, f-measure and ROC Random Forest is better as compared to other five machine learning algorithms. In this paper focused all 177 attributes which are related to predict the pediatric cardiomyopathy disease associated genes. In future the features can be analysed in order to minimize the number of features which are used to predict the disease associated gene. This can be extended with protein protein interaction[14] and pathways of genes features also used to provide the better prediction of pediatric cardiomyopathy disease associated genes.

REFERENCES

1. D. Le, N. X. Hoai, and Y. Kwon, “A Comparative Study of Classification-Based Machine Learning Methods for Novel Disease Gene Prediction A Comparative Study of Classification-Based Machine Learning Methods for Novel Disease Gene Prediction,” no. October 2014, 2015.
2. P. Luo, Y. Li, L.-P. Tian, and F.-X. Wu, “Enhancing the prediction of disease-gene associations with multimodal deep learning,” *Bioinformatics*, pp. 1–8, 2019.
3. D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, “ScienceDirect ScienceDirect Comparative Analysis of Various Machine Learning Algorithms for Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia Detecting Dementia,” *Procedia Comput. Sci.*, vol. 132, pp. 1497–1502, 2018.
4. D. Khanna, R. Sahu, V. Baths, and B. Deshpande, “Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease,” *Int. J. Mach. Learn. Comput.*, vol. 5, no. 5, pp. 414–419, 2015.
5. P. D. K. B, “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis,” pp. 5–10, 2016.
6. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” no. May, 2017.
7. X. Huang *et al.*, “Revealing Alzheimer ’ s disease genes spectrum in the whole-genome by machine learning,” pp. 1–8, 2018.
8. K. U. Maheswari and A. Valarmathi, “A Novel Feature Selection Algorithm for Coronary Artery Disease Prediction,” vol. 744, no. 4, pp. 735–744, 2018.

9. M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 09, no. 01, pp. 1–16, 2017.
10. M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," *Proc. 2nd Int. Conf. Comput. Methodol. Commun. ICCMC 2018*, no. Iccmc, pp. 997–1002, 2018.
11. S. Ekiz and P. Erdogmus, "Comparative study of heart disease classification," *2017 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2017*, pp. 1–4, 2017.
12. E. Neelima, "A comparative Study of Machine Learning Classifiers over Gene expressions towards Cardio Vascular Diseases Prediction," vol. 13, no. 3, pp. 403–424, 2017.
13. J. W. Rossano and M. J. O'Connor, "Sudden cardiac death in pediatric cardiomyopathy: The importance of well-designed population-based studies," *J. Am. Coll. Cardiol.*, vol. 65, no. 21, pp. 2311–2313, 2015.
14. X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and Validation of Disease Genes Using HeteSim Scores," vol. 14, no. 3, pp. 687–695, 2017.