International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I12P229

Volume 68 Issue 12, 182-186, December 2020

An Efficient Classification of Congenital Fetal Heart Disorder using Improved Random Forest Algorithm

K. Vimala[1*], Dr. D. Usha[2]

[1*]*Research Scholar, Department of Computer Science, Mother Teresa Women's University, Attuvampatti, Kodaikanal, Tamil Nadu, India*

[2]*Assistant Professor, Department of Computer Science, Mother Teresa Women's University, Attuvampatti, Kodaikanal, Tamil Nadu, India*

[1]*vimalakphd@gmail.com*

Abstract *— Congenital genetic disorders are one of the major complications in the medical application. Congenital disorders can be detected in the earlier stage, and patience could be diagnosed as soon as possible. This research work deals with the identification and detection of fetal heart congenital genetic disorders in humans. The gene dataset consists of the fetus from 20-weeks of conception. The dataset is pre-processed to check null criteria, and gene selection is performed using Principal component analysis, where the features are reduced for further processing. The classification was carried out using Machine Learning algorithms such as Improved Random forest classifier, Support Vector Machine, and Gradient Boosting algorithm. The performance of the random forest classification provided the best result of 87.85%.*

Keywords *— Congenital Heart Defects, Principal Component Analysis, Improved Random Forest Algorithm.*

I. INTRODUCTION

Congenital Heart Defects (CHD) results when the heart, or blood vessels near the heart, doesn't develop normally before birth. Identification of heart defects in the human fetus has a significant implication for the human's pregnancy, and it helps for delivery planning and detects the abnormalities in another organ. Gene expression analysis helps to find fetal heart anomalies before childbirth and formulate a proper procedure for immediate medical consultation, or surgical intervention is required. Prenatal recognition and detection of fetal cardiac anomalies are essential because congenital variance is the primary cause of toddler death and congenital heart disease leads to 35% of child death. Some of the heart defects that occur are Aortic Valve Stenosis, Atrial Septal Defect, Coarctation of the Aorta, Ventricular Septal Defect.

II. RELATED STUDIES

Fetal cardiac disorder and function are analyzed [1 – 2] and their abnormalities and treatment. The early embryonic development [3-6] and growth of the fetal heart, and the disease caused due to the disorder are discussed here. The metabolism and diabetes [7-9] of the embryo in the human fetal heart are diagnosed. Gene expression [10-13] plays a major role in identifying the genetic disease, and the work deals with the cardiac disorder that occurs in humans. The gene consists of [14] Deoxy Ribonucleic acid (DNA), a double-stranded coil that carries genetic information such as growth, development, reproduction, and functioning of different organs in the body. Congenital heart defects, blood pressure [15-17] in both infants and adults, are diagnosed early and proper treatment. [18] Here, the mass growth of human placental trophoblastic is observed, and the fetus's growth is observed from 6-8 weeks. The supervised and unsupervised algorithms are used to analyze the expressed data. Bayesian prediction analysis was used to relate the expression levels of the gene.

III. METHODOLOGY

The challenges that medical science faces today are high. The genetic-based disorder plays a major role in-case of the medical field. The misregulation or mutation in the gene makes the gene function different from the other gene pattern. It inherits certain disorder from their parents or through environmental changes. The gene expression analysis helps to find disorder at the initial stage and provides a pathway for the patients' diagnosis process. This research work detects the heart defects, even earlier at the gestation period itself.

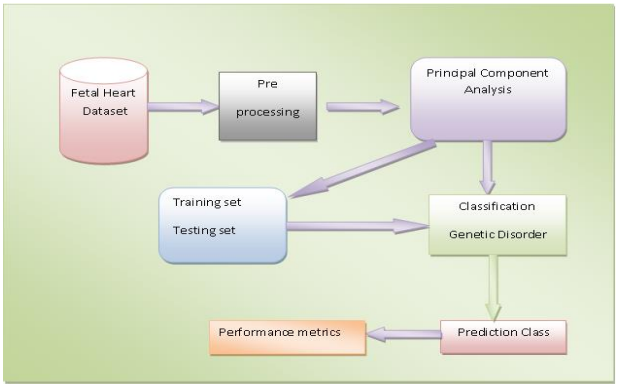

Fig. 1 Workflow

A. Database

The database chosen for the research work was the Fetal Heart dataset; this dataset provides the original data set, which proves to be the best for Gene Expression Analysis. Here, in this research work, the congenital fetal heart dataset was taken. The dataset consists of a combination of both normal and abnormal datasets. This dataset was found to have 16 samples and high in dimension. The data taken for the processing are in text format. The conception period of the fetus was 20-weeks of gestation period.

B. Data Pre-processing

The fetal congenital heart gene dataset consists of some noise in it. It is too removed, such that pure data are needed for research. The gene dataset consists of two null values in it, and it has been removed. The processing does not accept null criteria. The findings of missing values are evaluated, such that the unknown information has to be removed.

C. Feature Selection

The dataset consists of 45015 features. Since the genes set are large in volume, it is needed to reduce the dimension to make classification much easier. The feature selection is done using Principal Component Analysis.

a) Principal Component Analysis: The principal component analysis [19] is the best algorithm for feature selection in gene data, which provides the best feature as the resultant. The workflow of the principal component analysis is given as

- Standardization of matrix
- Covariance matrix
- Eigenvalues and Eigen Vectors

Algorithm

Step 1: Import the Congenital fetal heart gene dataset.

Step 2: Let D(X, Y) be that dataset matrix consisting of X rows and Y columns. The transformation matrix is given as T, R is the re-representation of the dataset.

$$TX = Y$$

Step 3: Find the standardization matrix by evaluating each fetal dataset's mean value and subtracting the mean value with each fetal data.

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(X - X')(Y - Y')}{n-1}$$

X, Y is the gene information
X' and Y' are the mean value the gene

Step 4: Calculate the covariance matrix

$$Cov = \begin{bmatrix} cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

$Cov(X,Y) = Cov(Y,X)$ is a symmetric matrix

Step 5: Calculate the Eigenvalues λ using covariance matrix $Cov$and identity matrix I.

$$E = |Cov - \quad \lambda I|$$

Step 6: Sort the Eigenvalues obtained from largest to smallest. Find Eigenvector V for the largest eigenvalue.

$$V = |Cov - \quad \lambda I| \begin{bmatrix} X \\ Y \end{bmatrix}$$

Step 7: Keep the largest k Eigenvector

Step 8: Convert the data into new space constructed by Eigenvectors

D. Classification

The classification was done using Machine Learning Algorithm. The Machine learning algorithms provide the best classification for the fetal heart gene dataset.

a) Support vector machine: The data extracted from the principal component analysis are given to the classifier. A support vector machine (SVM) is a supervised machine learning model [20]; here, the classification algorithms use a two-group classification problem; SVM builds on the linearly solvable concept because of high dimensional data. SVM works onset (xi, yi)

Training dataset of n points of the form

$$T = \{(x_1, y_1)(x_2, y_2) \ldots (x_n, y_n)\}$$

Where the value of y are either 1 or −1, each indicating the class to which the point x_i belongs

Hyperplanes are constructed as

$$w \,.\, x - b = 1$$

W is the normal vector; b is the plane. The hyperplane splits the boundary of one class, with label 1
and

$$w \,.\, x - b = -1$$

The hyperplane splits the boundary other class, as label -1
Soft margin

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - yi(wi \,.\, xi - b))\right] + \lambda||w||^2$$

Parameter λ provides the margin size, which ensures xi belongs to the correct class.

b) Gradient Boosting Algorithm: A gradient boosting algorithm [21] is also a machine learning algorithm that performs well for classification. This classifier is built mainly to enable many weak learners to perform as a strong learner. This algorithm is similar to Random Forest, but a

gradient boosting machine is an ensemble method that generates a number of decision trees sequentially. Each tree is grown using previously grown trees, unlike in bagging, where it creates multiple copies of original training data and fit separate decision tree on each.

Let S be the collection of data points

$$S = \{(x_1, y_1)(x_2, y_2) \dots (x_n, y_n)\}$$

It required to find start function *F* with a constant for the training set of data's

$$(x_{1-n}, y_{1-n})$$

The negative gradient is calculated as $gr(x)$. The fit for the base learner model is given as model $L(x, \theta S)$. Next, the gradient descent is found as pS

$$pS = argmin \sum_{i=0}^{n} \psi[yi, F\,S - 1(xi) + pL(xi, \theta S)]$$

The updating of gradient and the base-learner are given as,

$$F\,S = F\,S - 1 + pS\,L(x, \theta S)$$

The gradient boosting machine tries to approximate function *F* by minimizing the loss function $\psi[yi, F)]$

To avoid overfitting problem, the weight is added to each iteration, and random subsampling is used on the data where the base learner $L(x, \theta S)$is trained on.

c) Random Forest Classification: The classification of the Random improved forest [22-23] is based on the traditional method, where the prediction is unbalanced in the case of a medical database, so to overcome this issue, the algorithm is improved to provide high prediction accuracy and with best tolerance level for the abnormal dataset. These two phases in the Random Forest algorithm. First, the algorithm extracts the number of subsamples from the original samples with the bootstrap method and creates decision trees for every sample. Second, the algorithm proceeds with the classification of the decision trees and implements a voting process. The largest vote of the classification is chosen for the final result of the prediction

Improved Random Forest Algorithm

Step 1: The dataset is divided into the training set, validation set, and test set. The data are extracted randomly as an N+1 dataset from the original dataset. The bootstrap method is constructed for each data. Among the N+1 dataset, N sets are used as the training set and remaining as a validation set. The sample that has been not drawn are taken as the test dataset

Step 2: Construct the Random Forest classifier. The inputs of N training sets are used to build a random forest model. As final, the decision tree is constructed.

Let T be the collection of data points.

$$T = \{(x_1, y_1)(x_2, y_2) \dots (x_n, y_n)\}$$

Training Samples (N) are given as,

$$N = \begin{bmatrix} x_{A1} & \cdots & x_{Z1} \\ \vdots & \ddots & \vdots \\ x_{An} & \cdots & x_{zn} \end{bmatrix}$$

Step 3: Calculate the weighted value by evaluating the accuracy of each sub-classifier. The validation set is then given as input to the model, and the dataset is classified.

Step 4: Now, the test set is given as input, and performance is evaluated for the model.

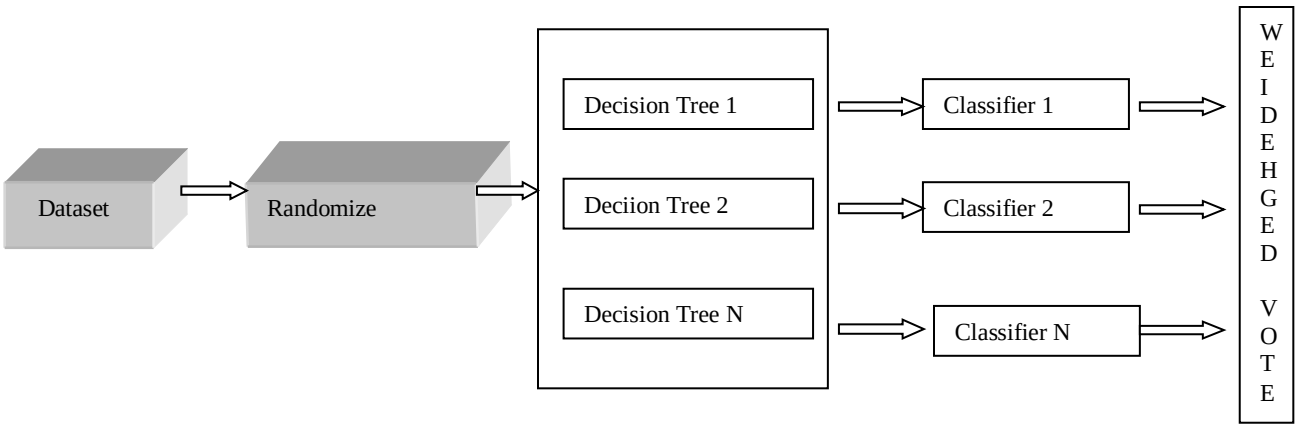

Fig. 2 Random Forest Algorithm

Step 5: The classification result of each sub-classifier K is evaluated, and the classification weight result of sub-classifier K is also found.

Algorithm

Input: Number of features $T(X_n Y_n)$
Output: Classification is normal and abnormal
Random Forest:
1.Check features
For each feature $T(X_n Y_n)$
Find the best split
2. best split
For each row in the dataset
X= all rows <= current X
Y = all columns > current Y
Calculate the score
If the score is best, then update it
3. Define the Gini Index (X_split, Y_split, N)

$$P(X) = \frac{X_i}{N_i}$$

$$P(Y) = \frac{Y_i}{N_i}$$

For each (i=1…n)
P = Probability of $\sum X_i \mathrel{+}= P^{\wedge 2}$
P = Probability of $\sum Y_i \mathrel{+}= P^{\wedge 2}$
Gini Index = $1 - \sum X_i$, $1 - \sum Y_i$
4. Prediction of labels

IV. PERFORMANCE

The fetal heart data set consists of 45015 features, and the total samples in the dataset are 16. The dataset consists of normal records and abnormal records. Principal component analysis reduces the dataset size, and efficient features are selected for further process. The counts of the features are reduced in fig 3. to 15870, such that it makes the classification process much easier. The accuracy of the classification algorithm is evaluated.

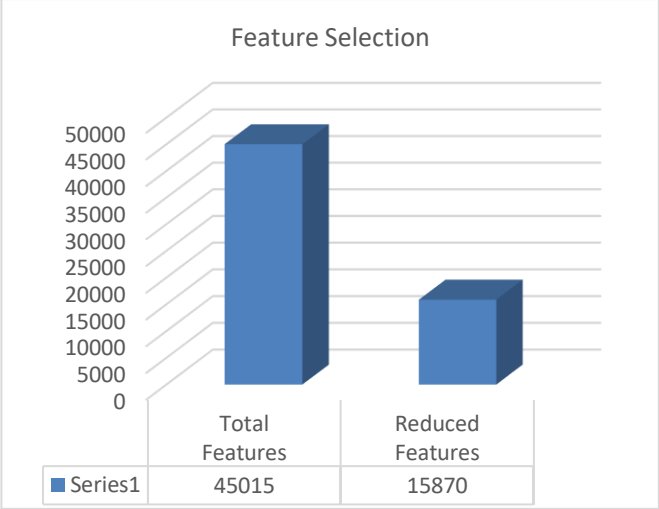

Fig. 3 Feature selection using PCA

Performance of Machine Learning Algorithm

In Machine Learning Classification, about of 12samples are given for the training set, and 4of extracted samples are forwarded for testing.SVM classifier is a fast and dependable classification algorithm, which performs well for gene dataset. Finally, the classifier provides a classification for the homo-sapiens fetal heart as normal and abnormal.

TABLE I
IMPROVED RANDOM FOREST

Forest (trees)	Testing Accuracy
50	85.32%
100	86.15%
200	87.85%

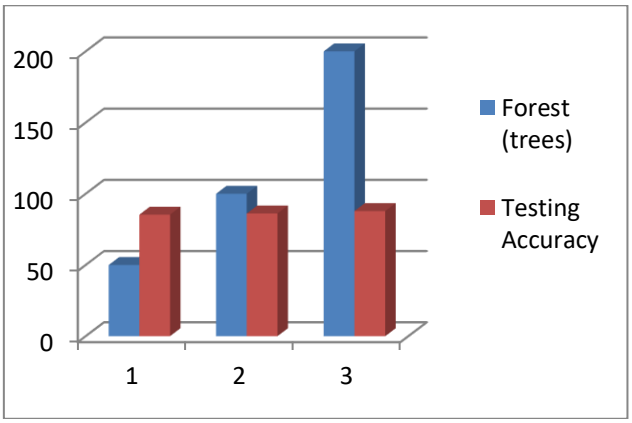

Fig. 4 Testing accuracy of Random forest

The accuracy level is 85.74%. Gradient Boosting Algorithm performs equally well as Random Forests for gene classification. This work required 13 iterations, along with 200 trees construction. The accuracy of the datasets is determined as 86.20%. The random forest provides the best result compared to the Support Vector Machine and gradient boosting. Table 1 the decision tree construction (50,100,200) and their accuracy are listed in Fig 4.

The performance of the Machine Learning algorithm is given in Table 2

TABLE II
ACCURACY OF MACHINE LEARNING ALGORITHM

Model - Algorithm	Accuracy
Support Vector Machine	85.74%
Gradient Boosting	86.20%
Improved Random Forest	87.85%

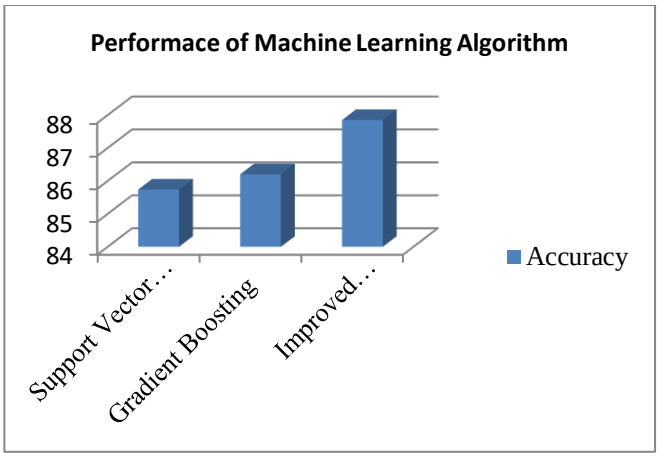

Fig. 5 Performance of Machine Learning Algorithm

V. CONCLUSIONS

Congenital Heart Defects are identified and detected efficiently using Principal Component Analysis and classified using Improved Random Forest Algorithm. This research work proves to be a pathway for analyzing many other congenital genetic disorders in human organs and animals also. Future work is based on increasing the algorithm's efficiency and detecting genetic disorders in Human Fetal Lungs. The accuracy of the algorithm can also increase by using deep learning concepts

VI. REFERENCES

[1] O. Valenti, S. Monte, and E. Giorgio, Fetal cardiac function during the first trimester of pregnancy, An International Journal of prenatal Diagnosis and fetal Maternal Medicine. 5(3) (2011) 59-62.

[2] H. Li, J. Wei, and Y. Ma, Prenatal diagnosis of congenital fetal heart abnormalities and clinical analysis, Journal of Zhejiang University Science B, 6(9) (2005) 903-906.

[3] C.M.J. Tan, The Transitional Heart: From Early Embryonic and Fetal Development to Neonatal Life, Fetal Diagnosis, and Therapy. 47(5) (2019) 373-386.

[4] J.M. Matinez, M. Comas, and A. Borrell, Abnormal first-trimester ductus venous blood flow: a marker of cardiac defects in fetuses with normal karyotype nuchal translucency Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology. 35(3) (2010) 267-272.

[5] E.C.M. Nelissen, A.P.A. Van Montfoort, and L.J.M. Smits, IVF culture medium affects human intrauterine growth as early as the second trimester of pregnancy, Human Reproduction, 28 (8), (2013) 2067-2074.

[6] S. Marchiano, A. Bertero, and C.E. Murry Learn from Your Elders: Developmental Biology Lesson to Guide Maturation of Stem Cell-Derived Cardiomyocytes. Pediatric Cardiology. 40(7) (2019) 1367-1387.

[7] J.I. Iruretagoyena, W. Davis and C. Bird, Metabolic gene profile in early human fetal heart development, Molecular Human Reproduction. 20(7) (2014) 690–700.

[8] R. Cerychova and G. Pavlinkova, HIF-1, Metabolism, and Diabetes in the Embryonic and Adult Heart, Frontiers in Endocrinology. 9 (2018) 460.

[9] J.M. Walejko, J.P. Koelmel, and T.J. Garrett, Multiomics approach reveals metabolic changes in the heart at birth, American Journal of Physiology. 315(6) (2018) 1212-1223.

[10] Z. Geng and J. Wang, Microarray Analysis of Differential Gene Expression Profile between Human Fetal and Adult Heart. Pediatric Cardiology. 38 (2017) 700 -706.

[11] D.S. Dizon- Townson, J. Lu and T.K. Morgan, Genetic expression by fetal chorionic villi during the first trimester of human gestation, American Journal of Obstetrics Gynecology. 183(3) (2000) 706-711.

[12] H. Li, S. Qin, and F. Xiao Predicting the first-trimester outcome of embryos with cardiac activity in women with recurrent spontaneous abortion, Journal of International Medical Research, 48(6) (2020).

[13] A.R. Singh, A. Sivadas, A. Sabharwal, S.K. Vellarikal, R. Jayarajan, A. Verma, S. Kapoor, A. Joshi, V. Scaria and S. Sivasubbu, Chamber Specific Gene Expression Landscape of the Zebrafish Heart, Plos One, 11, no. 1 0147823, 2016.

[14] T. Workalemahu, M. Ouidir, and D. Shrestha, Differential DNA Methylation in Placenta Associated with Maternal Blood Pressure During Pregnancy, Hypertension. 75(4) (2020) 1117 – 1124.

[15] B.A Firulli, R.M. George and J. Harkin, Hand1 loss-of-function within the embryonic reveals survivable congenital cardiac defects and adult heart failure. Cardiovascular Research, 116(3) (2020) 605-618.

[16] N. Velayutham and E.J. Agnew, Postnatal Cardiac Development and Regenerative Potential in Large Mammals. 40 (2019) 1345 – 1358.

[17] J. Binder, S. Carta, and J.S. Carvalho, Evidence for uteroplacental malperfusion in fetuses with a major congenital heart defect, Plos One. 15 (2) (2020) 0226741.

[18] M.A Khan, M Kar, S Mital, and S. Kumar, Small scale transcript expression profile of Human first trimester placental villi analyzed by a custom-tailored cDNA array, Indian Journal of Physiology and Pharmacology. 54(3) (2010) 235-254.

[19] K.Y. Yeung and W.L. Ruzzo, Principal component analysis for clustering gene expression data, Bioinformatics. 17(9) (2001) 763-771.

[20] S. Wang, J. Wang, and H. Chen, SVM-Based Tumor Classification with Gene Expression Data, Advanced Data Mining, and Applications. 2093 (2006) 864-870.

[21] O. Gonzalez-Recio and J.A. Jimenez-Montero, The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets, Journal of Dairy Science. 96(1) (2013) 614-624.

[22] R. Díaz-Uriarte and S.A. De Andres, Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 7(3) (2006)1471-2105.

[23] M. Ram, A. Najafi, and M.T. Shakeri, Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest, Iranian journal of Pathology.12(4) (2012) 339-347.