Hindawi
Journal of Sensors
Volume 2021, Article ID 1055770, 10 pages
https://doi.org/10.1155/2021/1055770

Research Article

Early Detection of Pediatric Cardiomyopathy Disease Using Window Based Correlation Method from Gene Micro Array Data

K. Jayanthi[1], **C. Mahesh**[1], **A. Arthi**[2], **K. T. Rajendran**[3], **B. Vijayalakshmi**[4] and **N. R. Shanker**[5]

[1]*Veltech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology, Chennai, India*
[2]*R.M.K Engineering College, Chennai, India*
[3]*Saveetha School of Engineering, Courtallam, India*
[4]*Sri Parasakthi College for Women, Courtallam, India*
[5]*Aalim Muhammed Salegh College of Engineering, Chennai, India*

Correspondence should be addressed to K. Jayanthi; jayanthi2contact@gmail.com

Received 10 May 2021; Revised 7 July 2021; Accepted 11 August 2021; Published 25 August 2021

Academic Editor: Bruno da Silva

Disease prediction through gene is a challenging task. Researchers have proposed algorithms to identify disease from genes. Traditional algorithms prioritize through annotation and combines the structures in biological process or molecular functions and compared with annotations of known disease genes for classification. Pediatric Cardiomyopathy is a disease due to disorder in heart muscle and identification at early stage is a challenging problem. In this paper, the above problem solves through Window Based Correlation (WBC). In WBC, Global data is reduced to spatial data using block reduction technique. After Data reduction, strong relationship analysis between the genes is identified through RMSE values between the genes. This RMSE values helps to detect the pediatric cardiomyopathy at early stage using Window based correlation method. From the results, ablation study proves an accuracy of prediction is about 85%.

1. Introduction

In human body, DNA Structure is similar in all cells and they are dissimilar in sequence, when affected by diseases. DNA consists of gene which generates a code of sequence for proteins. Genes are expressed through proteins. Proteins are specified by encoding Genes and different proteins are produced during cell regeneration. The production of protein is affected through any biological process change, which arises due to disease, stress, food and ambient changes. The proteins are produced through process of molecular biology.

Transcription of a gene from DNA into temporary molecule is called as RNA. Furthermore, the translation of the gene is represented as cellular components which builds a protein using the RNA. The DNA and RNA have similar property where each has a chain of chemicals known as bases. The bases are termed as Adenine, Cytosine, Guanine and Thymine and generally represented as A, C, G and T. Four bases are common for DNA and RNA. Thymine RA has Uracil referred as U. Genes are building blocks of inheritance and genes are passed from one to other generation. Genes contains DNA holds information of protein synthesis.

Protein performs building block in cells. If irregularity occurs other above process results in genetic disorder. However, mutation change in DNA content of cell will change genes. Changes in gene mutation cause's irregularities in making a protein. The irregular protein never performs well and leads to genetic disorder.

Disease prediction through genes is a difficult challenge. Researchers proposed algorithms [1–3] to identify the disease from genes. Traditional algorithms prioritize through annotation and combine structures in biological process or

Table 1: Literature Survey of Gene Prediction for Pediatric Cardiomyopathy.

Ref/ year	Problem	Dataset	Methodology	Advantages/disadvantages
[7]/ 2019	Image classification and tumer subtype prediction	Multi-omics dataset	Bayesian optimization.	Advantages: Feature selection considers biological context, more effective than existing reduction algorithms. Disadvantages: Integrating two molecular layers and multi-omics dataset is never used over fitting. Limited with same sample group.
[1]/ 2019	Cancer gene prediction	Acute myeloid leukemia (AML) cancer-gene dataset	Support vector machines (one-class classification)	Advantages: Retrieve validated negative data. Disadvantages: More genes need to be and no real negative samples.
[2]/ 2019	Tumer prediction	Online Mendelian inheritance in man (OMIM) database	Multimodal DBN (dgMDL)	Advantages: Disease gene pair and lung cancer disease related genes used to predict new disease genes. Disadvantages: Gene is not included data, and its associations never correctly predicted.
[8]/ 2020	Novel disease	Online Gene Essentiality & Database of essential genes	Synthetic minority over-sampling technique.	Advantages: Algorithms is base on intrinsic, extrinsic gene and protein enabled systems for prediction. Predictions made from protein sequences.
[9]/ 2019	Novel disease	DisGeNET	Deep neural network (DNN) along with SVM, NB, random Forest	Advantages: Large-scale prediction from infectious diseases Disadvantages: Model limited with smaller dataset.
[10]/ 2019	Disease mutation	Benign disease-associated mutations dataset	Extracted sequential and spatial features using deep learning model with convolutional layer with sixteen convolution filters evaluate performance of the MCCNN (multichannel convolutional neural network) model	Advantages: Deep learning incorporates both spatial and sequential features used for prediction of disease associated with mutation genes.
[11]/ 2019	Parkinson disease gene	Clinvar dataset used for Parkinson disease gene.	Node2vec and auto encoder along with support vector machine.	Advantages: Vector representation of gene extracted. Dimensions reduced by autoencoder. Predictions done with SVM classifier.
[12]/ 2020	Cancer disease	Comparative Toxicogenomics database and online Mendelian inheritance in man database.	CNN for disease-associated genes identification.	Advantages: Relationship between disease and symptom. Map the gene-disease association.
[13]/ 2010	HIV-1 and human proteins	NIAID database	Semi-supervised multi-task framework. Classification, ranking and embedding	Advantages: Labelled examples trained with multi-layer perception network. Prediction task based on classification, ranking and embedding improves accuracy.
[14]/ 2010	Human diseases	Online Mendelian inheritance in man (OMIM)	Network-based approach called PRINCE (PRIoritizatioN and complex elucidation) propagation-based algorithm	Advantages: Similarity measure and protein-protein interactions identified.

Table 1: Continued.

Ref/ year	Problem	Dataset	Methodology	Advantages/disadvantages
[15]/ 2017	Human genes	HumanNet	Probability-based collaborative filtering predict pathogenic human genes.	Advantages: Gene – Gene similarities with gene disease similarities are identified and dimension reduced using probability based filtering.
[16]/ 2019	Human disease genes	Online Mendelian inheritance in man and BioGPS	Graph convolution networks (GCN)	Advantages: Using GCN features easily identified by their similarity graph. Gene and disease embedding defined factors from matrix factorization. Loss functions optimized by Adam method.
[17]/ 2018	Gene protein interaction	Simons Foundation autism research initiative gene database.	Naive Bayes, support vector machine, decision tree and random forest used for classification.	Advantages: Functional similarity matrix constructed between the genes.
[1]/ 2019	Cancer disease	NCBI GEO dataset	Boosted tree regression considers gene expression.	Advantages: OCSVM method with linear and RBF kernel classify and predict novel disease genes using gene expression.
[18]/ 2020	Novel disease genes	Online Mendelian inheritance in man and BioGPS	Ensemble algorithm predict disease through genes.	Advantages: Prediction of disease through genes with models trained by centrality features extracted.

molecular functions and compared with annotations of known disease genes for classification.

The annotated based approaches [3] are finite and never captures indirect association among genes. Common behavior or function of genes are never utilized for disease identification. Ontology - based disease gene similarity network method is applied for prioritize genes [4] compared with annotation-based genes [3] and identified the interaction between cellular, molecular, proteins and microarray data. Genes based pediatric cardiomyopathy disease identification is proposed in this paper.

Pediatric Cardiomyopathy disease is a disorder in heart muscle. In this world, the deaths and disability of cardiomyopathy increased rapidly due to genetic disorder. In pediatric cardiomyopathy Registry, 1.1 to 1.5 per 100,000 children under the age of 18 is diagnosed with cardiomyopathy. Approximately, 50% of patients suddenly died in childhood or at the time of cardiac transplantation due pediatric cardiomyopathies.

Children with cardiomyopathies of approximately 17.5 million patients die which is the 31% of all global deaths. The death rate may extend 23.6 million by the year 2030. In India, 1.7 to 2.0 million people is affected by coronary illness. In Kerala, heart disease is about (187 - >350 passing/100,000 man/year). The South Indian area one fifth of total causality of India. In 2030, cardiomyopathies difficulties will increase to 17.9 million in India as per prediction from medical board of India. Global Society of Heart and Lung Transplant in Prague (2012) found that almost 25% of patients had heart transplant due to cardiomyopathies and 4% of patients infected due transplant. The Individuals with cardiomyopathies never show any signs or symptoms

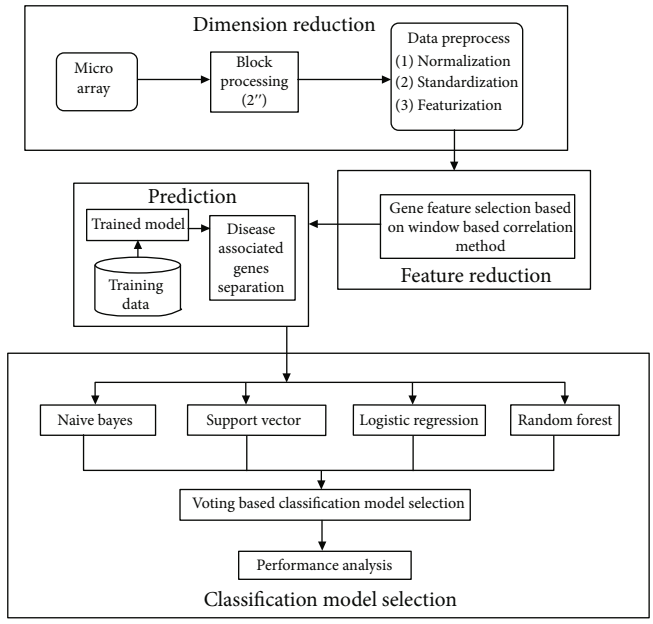

Figure 1: Overview of Proposed Window Based Correlation Method.

treatment cardiomyopathies signs and symptoms depend on age. Diagnosing pediatric cardiomyopathy is very important and should be performed precisely. However, it is a challenging task for Researchers.

1.1. Problem Statement

Step 1: Create a list consisting of initial condition (BEST < - initial condition) and create a close list and leave empty.
Step 2: Calculate S = arg max e(x)
Step 3: Omit S from initial list and added to close list.
Step 4: If e(s) > = e(BEST), BEST < — S
Step 5: Every child t from S which does not belong to initial open list or end list, estimate and add to initial open list.
Step 6: If BEST changes in last set, repeat step 2
Step 7: Back BEST

Algorithm 1: Window Based Correlation Algorithm.

(i) Global data-based prediction of gene never provides the relationship between individual gene parameters. For example, gene data such as SNHG9, ACTG1, EXTL3 needs relational analysis for their protein transport, cellular catabolic process, organic substance catabolic process, protein – containing complex assembly, Programmed cell death, apoptotic process, cellular protein localization, cellular macromolecule localization, positive regulation of molecular function, and response to endoplasmic reticulum stress

(ii) Global Analysis of pediatric cardiomyopathy disease can detect the features of prognosis and characteristics of disease

(iii) The Global data analysis has more time complexity and accuracy of prediction is less due to size of the data and structure of the data

1.2. Contributions. Microarray data analysis of genes are used for diagnosis of various diseases such as heart, diabetes, tumor and cancer. Microarray datasets are bigger in size and difficult to predict. Hence, requires appropriate statistical method with high accuracy in prediction.

Micro array data describes the expression levels of hundreds of thousands of genes in cells, which are correlated with the corresponding protein, allowing to understand the cellular processes involved in biological processes in a better way. There are many genes in Microarray data analysis, to locate the most relevant genes, dimension of genes is reduced using 2n window processing and correlation method is used for identifying the relation between proteins and related diseases for early detection of pediatric cardiomyopathy is proposed in this paper.

The proposed method is called as Window-based correlation method (WBC) and compared the efficiency of proposed algorithm with traditional algorithms [5, 6] . However, the proposed algorithm identifies the correlation between proteins and then the prediction is performed. Whereas traditional methods are applied to global data for prediction, so less accuracy and proposed method WBC is based on spatial data of proteins related to disease are used for earlier prediction of pediatric cardiomyopathy.

(i) In WBC, Global data is reduced to spatial data using block reduction technique (2^n) reduces dimensions of data and to evaluate accuracy of traditional algorithms after applying WBC for earlier detection of for early detection of pediatric cardiomyopathy from Gene dataset

(ii) Data reduction is applied and strong relationship analysis between the genes for particularly for pediatric cardiomyopathy disease is identified through RMSE values between the genes and validated for accuracy in prediction

(iii) The pediatric cardiomyopathy is detected at early stage using Window based correlation method based on data reduction and RMSE. The proposed method evaluated for runtime and space complexity analysis

2. Literature Survey

The Table 1 Shows the Literature review of Gene Prediction using different algorithms and their advantages and disadvantages are analysed.

3. Methodology

Earlier detection of pediatric cardiomyopathy is predicted using Window Based Correlation Method (WBCM). In this method OMIM (Online Mendelian Inheritance in Man) Gene Micro array data used for pediatric cardiomyopathy disease gene prediction. Biological features analysis done by dividing the number of features using Block Processing method (2^n). Each divided block pre-processed with two methods such as Normalization and Standardization and features constructed by analysing the variances between these methods.

3.1. Block Processing Method. Block processing methods, input training samples process with block. Block of input training samples are treated as vector transformed, output vector samples by transformation 'H' as in equation (1).

$$X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \text{-----H-----} \rightarrow \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix} = Y \quad (1)$$

where x – number of input samples y – transformed output vector samples. H- denotes transformation.

In Block processing method from the 1163 of training samples splited into 300 and given to block process. Each 300 considered as one block and 10 attributes. Correction

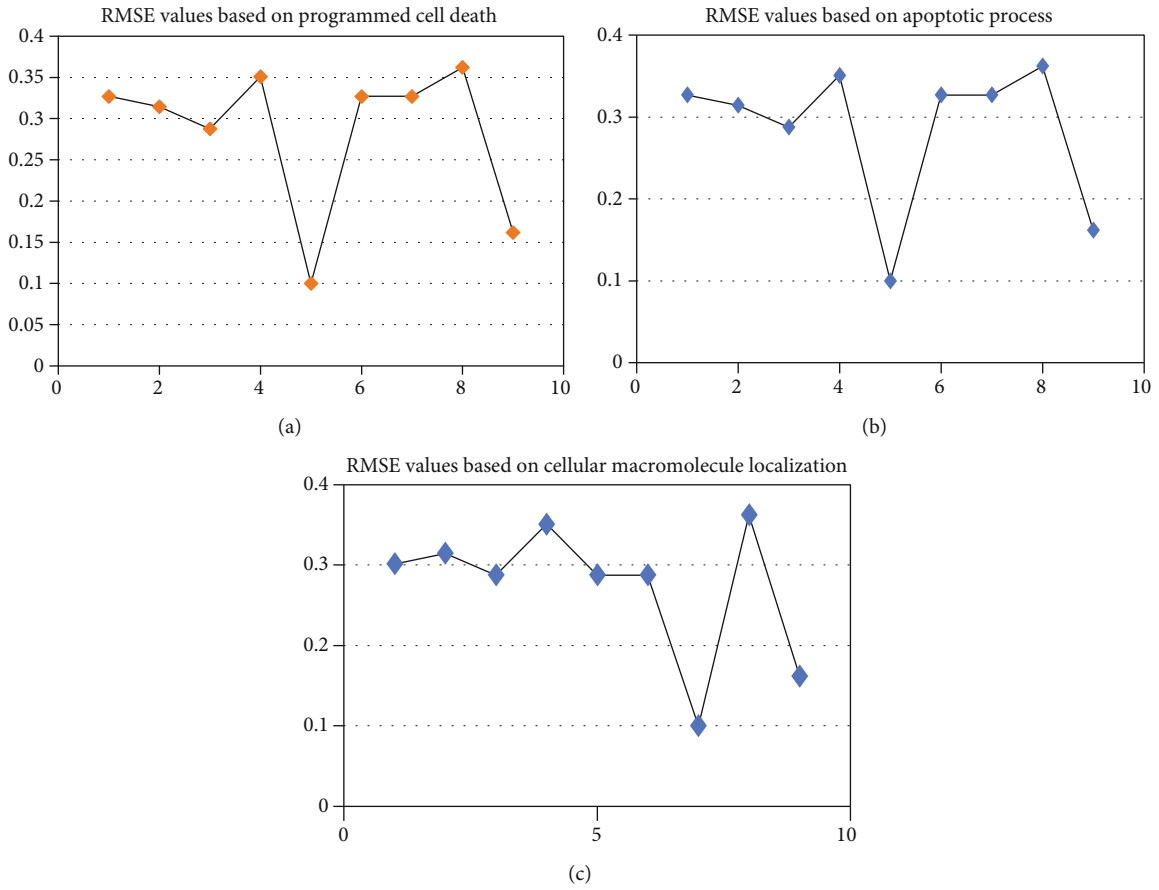

Figure 2: Support Vector Machine classifier detects disease genes with the biological features.

between each attribute is compared and attribute more than 0.5 taken as features and from new block of training samples generated. The training samples further used for training the model The Figure 1 shows overview of proposed methodology in predicting of cardiomyopathy disease genes using Window Based Correlation Method.

3.2. Normalization. Normalization transforms features in similar scale and improves stability of model. In this paper Min-Max Normalization transforms biological features into similar range of data. [0 – 1]. Min – Max Normalization for Biological Features Compared with Protein transport which provides RMSE value as 0.31 for Cellular macromolecule localization and positive regulation of molecular function.

3.3. Standardization. Training data set having different set of ranges, finds the correlation between biological features standardization, relevant features and supports for scaling the range of training data. Standardization provides different scales of features. From this organic substance catabolic process, protein – containing complex assembly, programmed cell death, apoptotic process with maximum similarity such as 0.8,0.9,1.0 ranges and find the correlation between the biological features.

3.4. FEATURIZATION. Comparing the above two methods of pre-processing the standardization identify the range of features and finds the relevant features to be used in the proposed method Window Based Correlation Method (WBCM).

3.5. Window – Based Correlation Method. WBCM method belongs to filter model which grades the feature based on correction using heuristic evaluation. Irrelevant feature ignored due to weak correlation. Redundant features are excluded due to strong correlation with other features. Window – Based correlation method measures correlation matrix of class with feature and corrects features on Training, and searching the subset of feature space with best first search.

Best subset feature is evaluated, and score defined as

$$Merit_S = \frac{P\overline{r}_{cf}}{\sqrt{p + p(p-1)\overline{rff}}} \quad (2)$$

Merit $_S$ - heuristic merit subgroup 'S' con of 'p' features. Correction between feature and class output $\overline{r}_{cf}$ is average of correlation between feature and class (f ∈ S), and $\overline{r_{ff}}$ is average of correction between feature the correlation between feature with class uses Symmetrical Uncertainty (SU) as

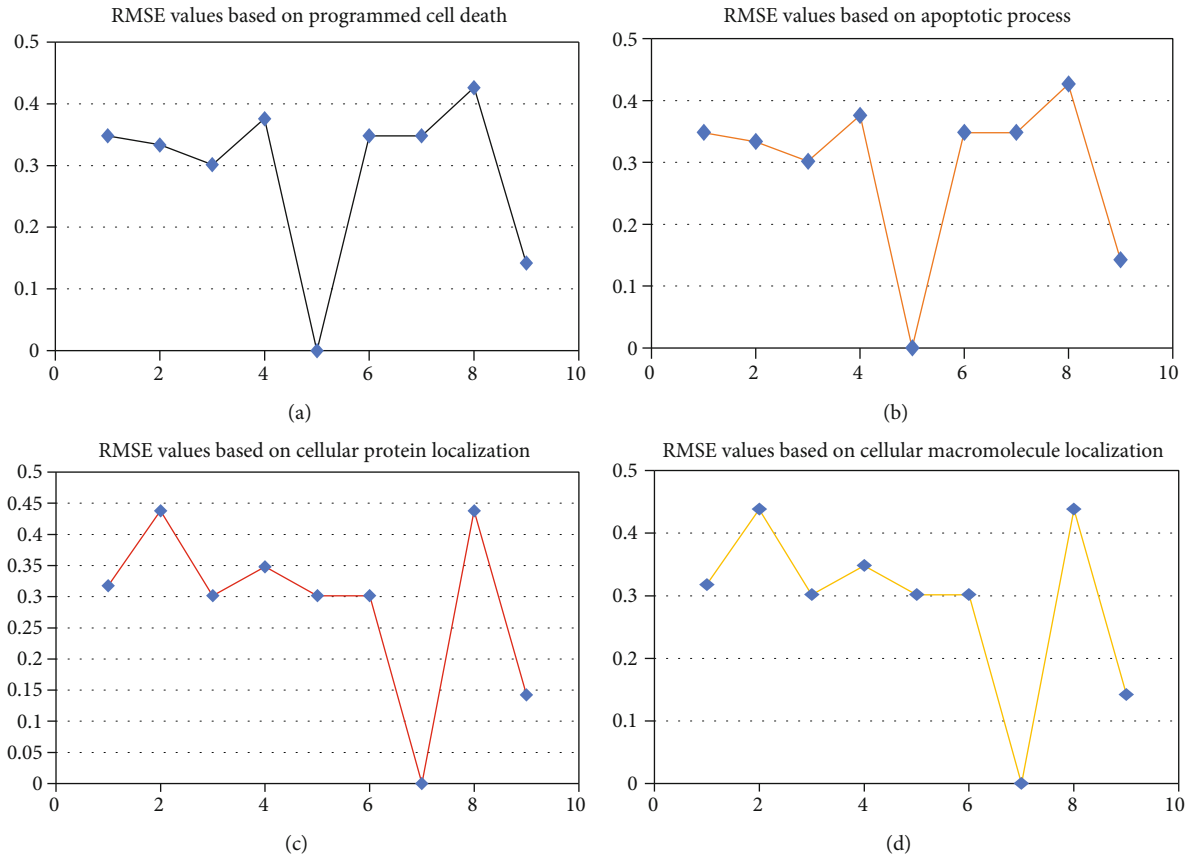

Figure 3: RMSE values based on biological features predicted with Naïve Bayes classifier to identify disease genes.

follows

$$\mathrm{SU}(\mathrm{X}, \mathrm{Y}) = \left[\frac{IE(x \mid Y)}{H(x) + H(y)}\right] \tag{3}$$

Where X – number of input samples Y – Transformed output vector samples H(X) – transformed input samples H(Y) – transformed output vector samples.

3.6. Classification Models

3.6.1. Naive Bayes. Naive Bayes is classified as a set of independent values that are not linked to one another. The status of one feature in class has no significance on the status of other features. The Naive Bayes algorithm, in its most basic form, is based on conditional probability. It utilizes classification of pediatric cardiomyopathy disease gene. It supports biological data to determine the gene which is associated with pediatric cardiomyopathy.

The Bayes theorem is used to build the Nave Bayes machine learning method. It depends on training data and posterior likelihood. It finds the pattern class using Bayes' theorem. Posterior probability assumption treats every feature as a disease gene class and follows conditionally independent rules. [5].

If the naive bayes classifier's assumption is right, the disease genes for pediatric cardiomyopathy will be identified; otherwise, the class will mean that the genes are not associated with the disease. The key benefit of the naive bayes classifier is that it is less complex for large databases and that the results are determined quickly and accurately [9]. The highest posterior probability has the lowest error rate.

$$\frac{\mathrm{P}(\mathrm{class} : \mathrm{yes} \mid \mathrm{no}) \mid (\mathrm{B1}, \mathrm{B2} \cdots \cdots \mathrm{Bd})}{\mathrm{P}(\mathrm{Class} \mid \mathrm{Pattern})} \tag{4}$$

B1, B2, and Bd are properties of gene's process. In this paper, Naive Bayes is used to predict the probabilities of associated to disease-related gene. Finally, best hypothesis of a biological data is extracted using the naive Bayes classifier algorithm.

3.6.2. Random Forest. Since dataset used in this research paper represented in tabular data. Using Random Forest classifier for the prediction of cardiomyopathy disease gene classification will leads to improve the process of identifying the disease genes classification. Random forest is one of the Meta classifiers, which means it is made up of several trees (Individual learner).

The random forest is composed of several randomly chosen trees, each random tree will help to identify pediatric cardiomyopathy disease genes by Biological features on a specific formation are focused on. The random forest would ensure that the right features are used in prediction of

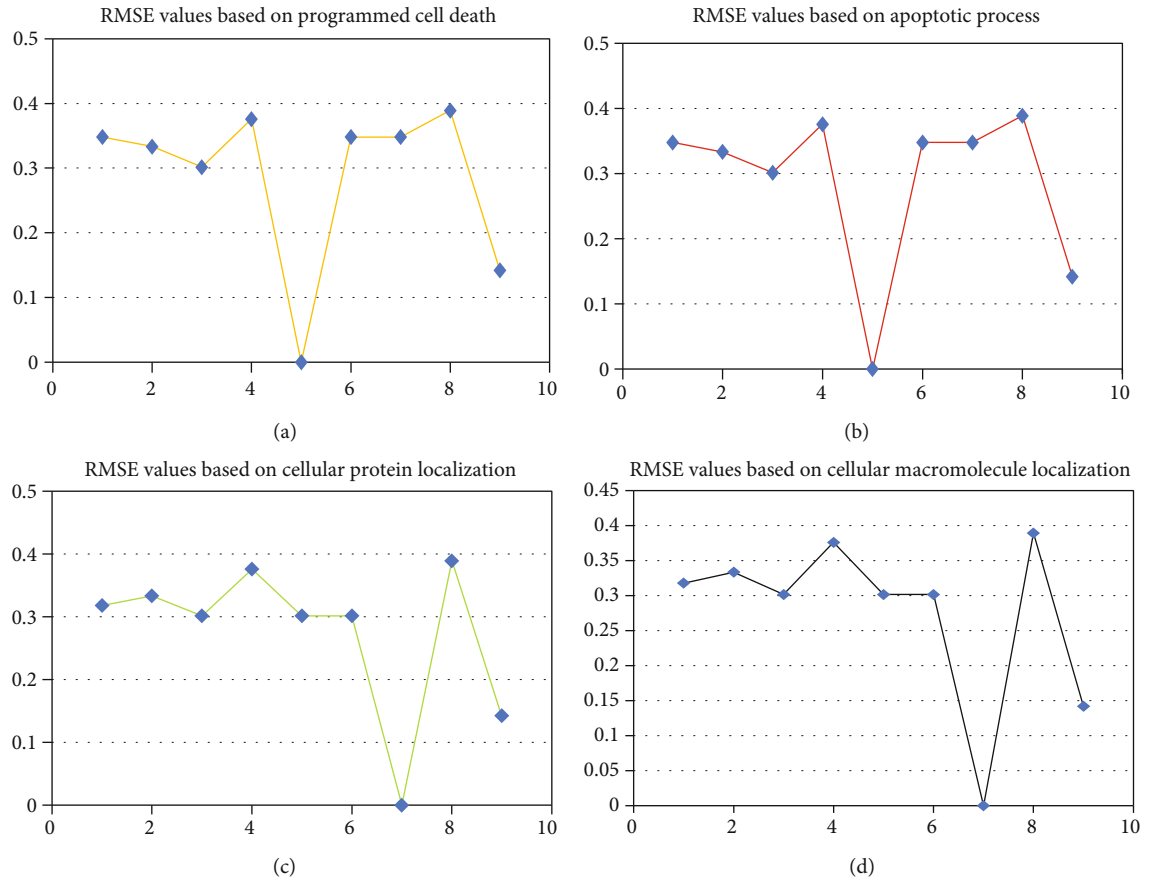

Figure 4: RMSE values based on biological features predicted with Logistic Regression to identify disease genes.

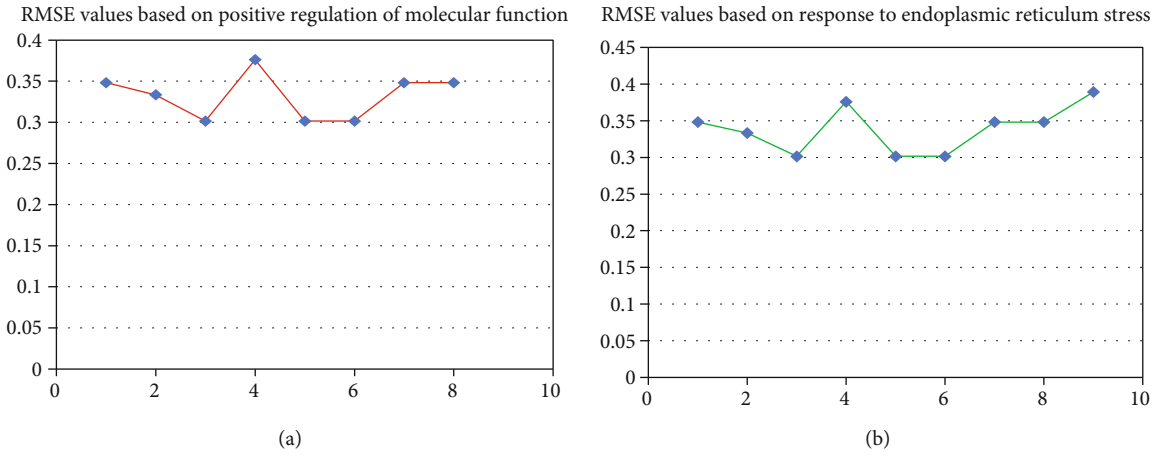

Figure 5: RMSE values based on biological features predicted with Random Forest to identify disease genes.

cardiomyopathy disease associated genes since each vote has the similar weight.

Three steps are used to predict genes related to cardiomyopathy disease. Using Random Forest classifier. First, using random selection, a collection of decision trees was created from biological features of pediatric cardiomyopathy disease training data. Second, the various votes received from all the biological feature decision trees are illustrated. Third, the pediatric cardiomyopathy disease class is defined by most votes cast in each of the decision trees that have been developed.

The random forest algorithm builds multiple decision trees by selecting labels and features at random. This allows to identify disease-based genes with greater accuracy. In this paper based on Root Mean Square value obtained from equation (2) of each block identified the features and that

RF	NB	LR	SVM
1.44469	1.42572	1.32904	1.31803
0.444685	0.425717	0.329038	0.318028
0.44402	0.411715	0.328116	0.318028
0.335973	0.238059	0.275294	0.318028
0.285401	0.227781	0.274461	0.318028
0.27155	1.63E-07	0.222151	0.318028
0.239184	1.54E-07	0.22143	0.318028
0.171612	6.48E-08	0.180533	0.318028

Figure 6: Threshold values of disease genes predicted with classification algorithms.

is used for training samples of Random Forest algorithm.

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\boldsymbol{observed\ gene\ labels - predicted\ gene\ labels})^2}{\boldsymbol{N}}} \tag{5}$$

Where N- Number of Samples.

3.6.3. Support Vector Machines (SVM). Pediatric cardiomyopathy disease gene classification is Bi- classification problem, to solve this Support vector Machine classifier used. Since the dataset is linearly separable, direct classification involves isolating disease genes from non-disease genes using an isolating hyper plane s(x) that runs through the majority of the 2 categories.

Genes which are related to pediatric cardiomyopathy treated as positive class (Yes). Non disease genes are randomly selected which is referred as Negative class. (No). Both Positive and negative class are same in size. The outputs of WBCM used for evaluation test. The number of linear hyper planes is higher. By finding the maximum between the two groups (Yes/No), SVM selects the best function. This margin is referring to as the hyperplane's separation of the two groups. Using equation (3), the SVM classifier gives a geometric margin predict disease associated with gene from a biological based gene feature.

$$\mathbf{Svmh}(\mathbf{disease\ gene\ labels}) = \begin{cases} +\mathbf{1}, \text{ if } \boldsymbol{w}.\boldsymbol{x} + \boldsymbol{b} \geq \mathbf{0} \\ -\mathbf{1}, \text{ if } \boldsymbol{w}.\boldsymbol{x} + \boldsymbol{b} < 0 \end{cases} \tag{6}$$

Where x- input features w- weights b- bias.

3.6.4. Logistic Regression. Logistic regression is the most used predictive analysis algorithm. In this paper, logistic Regression used to identify the relationship between one biological feature with more and biological feature of pediatric cardiomyopathy disease gene data such as protein transport as taken as dependent variable that compared with cellular catabolic process, organic substance catabolic process, protein – containing complex assembly, Programmed cell death, apoptotic process, cellular protein localization, cellular macromolecule localization, positive regulation of molecular function, response to endoplasmic reticulum stress as independent variables.

Logistic regression is a nonlinear transformation that uses a sigmoid function called the defined logistic distribution function to evaluate it. We predict that the features associated with pediatric cardiomyopathy disease genes are not (Yes/No) in this paper. Basically, Logistic Regression is associated with the data's probabilistic statistics values [4]. The sigmoid function can be used to map probabilities from expected values. Any real value between 0 and 1 is mapped by this function.

3.6.5. Voting Based Classification Model Selection. Comparing the threshold values of output array from the classification algorithms voting based training model has been selected. In this paper Naïve Bayes provides the best accuracy when compare with other classification algorithms based on threshold values.

Figure 2 shows that the RMSE values of programmed cell death (a), apoptotic process (b) and cellular macromolecule localization (c) values will be in constant range in between 1 to 5 and increased to 5 and remains same for other feature values. Features ranges varies constantly in the range that identifies pediatric cardiomyopathy disease genes such as Interacting protein in protein transport (SEC23, XPO4), Programmed cell death protein 5 (PDCD5), positive regulation of molecular function (AKIRIN1), Positive Regulation of RNA (PSMC2).

From the Figure 3, RMSE values of programmed cell death (a) and apoptotic process (b) values are increasing constantly. Cellular protein localization (c) and cellular macromolecule localization (d) the ranges will get varied between the ranges 8 to 10. In the positive regulation of molecular function and positive regulation of programmed cell death and apoptotic process (HTATIP2) cellular protein localization (LMNA) RMSE values remains same till reaches to 5 and finally increased to 9. These biological features locate the genes related to pediatric cardiomyopathy disease. Figure 3 shows the RMSE values of Naive Bayes which used to identify the disease associated genes.

In Figure 4 shows that Programmed cell death (a), apoptotic process (b) cellular protein localization (c) and cellular macromolecule localization (d) are in 2 to 5 and 7 to 9 leads to identify Positive regulation of protein localization (PLK1), apoptotic process (CD14) genes are related with biological features of pediatric cardiomyopathy disease genes which represented from logistic regression.

In Figure 5, Correlation between the positive regulations of molecular function (a) and response to endoplasmic reticulum stress (b) RMSE values are gradually increases in range of 1 to 9. Biological features of apoptotic process (LTBR), positive regulation of molecular function (AKIRIN1), reticulum stress (PLCG1) represents the pediatric cardiomyopathy disease related gene which is identify by Random Forest classifier. Figure 5 shows RMSE values of Random Forest classifier based on the biological features of the genes.

Ablation test in classification algorithms is according to the threshold values comparing the values naive bayes

Table 2: Comparison WBC algorithm with Classifiers.

Algorithms	Accuracy without WBC	Accuracy With WBC	Advantages and disadvantages of WBC	Computational complexity	Run time complexity
Random Forest	58%	65%	Identifying the relationship between the features consumes time and increases computational complexity random Forest with window-based correlation method.	**O (k)** (where k- number of trees)	**O (k log n)** (where n- number of datapoints in dataset.)
Logistic regression	64%	70%	Complex relationships of features are never identified with WBC based logistic regression.	**O(d)** (where d- number of dimensions)	**O(d)**
Support vector machines	76%	80%	Correlation of features found never provides accurate result for bigger window size for larger data set.	**O** $(\mathbf{n} * \mathbf{d})$ (where n- number of datapoints in dataset d- number of dimensions)	**O** $(\mathbf{k} * \mathbf{d})$ (where k- kernel for distance measure d- number of dimensions)
Naïve Bayes	81%	85%	Relationship between the features of genes accurately predicted with window based correlation method. The accuracy in perdition is proportional to window size	**O**$(\mathbf{c} * \mathbf{d})$ (where c- number of classes d- number of dimensions)	**O**$(\mathbf{c} * \mathbf{d})$

provides the best accuracy to classify the disease associated genes of pediatric cardiomyopathy.

In Figure 6 the pediatric cardiomyopathy disease genes identified by the classification algorithms (Random Forest, Naïve Bayes, Logistic Regression and Support Vector Machines). which based on threshold values in the range of series from 1.5 and above correctly classify the pediatric cardiomyopathy disease related genes. Naïve Bayes provides the best classification of disease genes with above 85% of Accuracy. Table 2 shows the Comparison of WBC algorithm with Classifiers.

4. Conclusion

Global data-based prediction of gene never provides the relationship between individual gene parameters. For example, gene data such as SNHG9, ACTG1, and EXTL3 needs the relational analysis for their protein transport, cellular catabolic process, organic substance catabolic process, protein – containing complex assembly, Programmed cell death, apoptotic process, cellular protein localization, cellular macromolecule localization, positive regulation of molecular function, and response to endoplasmic reticulum stress.

Global Analysis of pediatric cardiomyopathy disease can detect the features of prognosis and characteristics of disease. The Global data analysis has more time complexity and accuracy of prediction is less due to size of the data and structure of the data.

Global data is reduced to spatial data using block reduction technique (2^n) reduces dimensions of data. Data reduction is applied and strong relationship analysis between the genes is identified through RMSE values between the genes. The pediatric cardiomyopathy is detected at early stage using Window based correlation method based on data reduction and RMSE. Window size should be changed according to the data size and window size fixed based on trial and error method.

Data Availability

All data analyzed during this study are included in this research article.

Conflicts of Interest

The author(s) declare(s) that they have no conflicts of interest.

References

[1] A. Vasighizaker, A. Sharma, and A. Dehzangi, "A novel one-class classification approach to accurately predict disease-gene association in acute myeloid leukemia cancer," *PLoS One*, vol. 14, no. 12, article e0226115, 2019.

[2] P. Luo, Y. Li, L. P. Tian, and F. X. Wu, "Enhancing the prediction of disease-gene associations with multimodal deep learning," *Bioinformatics*, vol. 35, no. 19, pp. 3735–3742, 2019.

[3] D. H. Le, "Machine learning-based approaches for disease gene prediction," *Briefings in Functional Genomics*, vol. 19, no. 5–6, pp. 350–363, 2020.

[4] D.-H. Le and V.-T. Dang, "Ontology-based disease similarity network for disease gene prediction," *Vietnam Journal of Computer Science*, vol. 3, no. 3, pp. 197–205, 2016.

[5] A. Tran, C. J. Walsh, J. Batt, C. C. dos Santos, and P. Hu, "A machine learning-based clinical tool for diagnosing myopathy using multi-cohort microarray expression profiles," *Journal of Translational Medicine*, vol. 18, no. 1, pp. 1–9, 2020.

[6] J. Zahoor and K. Zafar, "Classification of microarray gene expression data using an infiltration tactics optimization (Ito) algorithm," *Genes (Basel)*, vol. 11, no. 7, pp. 1–28, 2020.

[7] V. Malik, Y. Kalakoti, and D. Sundar, "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer," *BMC Genomics*, vol. 22, no. 1, pp. 1–11, 2021.

[8] O. Aromolaran, T. Beder, M. Oswald, J. Oyelade, E. Adebiyi, and R. Koenig, "Essential gene prediction in *Drosophila*

melanogaster using machine learning approaches based on sequence and functional features," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 612–621, 2020.

[9] R. K. Barman, A. Mukhopadhyay, U. Maulik, and S. Das, "Identification of infectious disease-associated host genes using machine learning techniques," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[10] P. Popov, I. Bizin, M. Gromiha, A. Kulandaisamy, and D. Frishman, "Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure," *PLoS One*, vol. 14, no. 7, pp. 1–13, 2019.

[11] J. Peng, J. Guan, and X. Shang, "Predicting Parkinson's disease genes based on node2vec and autoencoder," *Frontiers in genetics*, vol. 10, pp. 1–6, 2019.

[12] X. Chen, Q. Huang, Y. Wang et al., "A deep learning approach to identify association of disease-gene using information of disease symptoms and protein sequences," *Analytical Methods*, vol. 12, no. 15, pp. 2016–2026, 2020.

[13] Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman, and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins," *Bioinformatics*, vol. 27, no. 13, pp. i645–i652, 2010.

[14] X. Wang, N. Gulbahce, and H. Yu, "Network-based methods for human disease gene prediction," *Briefings in Functional Genomics*, vol. 10, no. 5, pp. 280–293, 2011.

[15] X. Zeng, N. Ding, A. Rodríguez-Patón, and Q. Zou, "Probability-based collaborative filtering model for predicting gene-disease associations," *BMC Medical Genomics*, vol. 10, Supplement 5, p. 76, 2017.

[16] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "PGCN: disease gene prioritization by disease and gene embedding through graph convolutional neural networks," pp. 1–9, 2019, https://www.biorxiv.org/content/10.1101/532226v1/.

[17] T. Williams, "Genomics offers new possibilities for global health through international collaboration," *Disease Models & Mechanisms*, vol. 3, no. 3–4, pp. 131–133, 2010.

[18] P. Luo, L. P. Tian, B. Chen, Q. Xiao, and F. X. Wu, "Ensemble disease gene prediction by clinical sample-based networks," *BMC Bioinformatics*, vol. 21, Supplement 2, pp. 79–112, 2020.