NATURALISTA CAMPANO
ISSN: 1827-7160
Volume 28 Issue 1, 2024

Integrated Variational Autoencoder Model for Genome Transcription Analysis of Gene Expression Profiles Towards Prediction and Classification of Cardiomyopathy Disease

Ms. T. Sangeetha[1*], Dr. K. Manikandan[2], Dr. D. Victor Arokia Doss[3]

[1*]Research Scholar, Department of Computer Science, PSG College of Arts & Science, Coimbatore - 641014, Tamil Nadu, India.
[2]Associate Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore - 641014, Tamil Nadu, India.
[3]Associate Professor, Department of Biochemistry, PSG College of Arts & Science, Coimbatore 641014, Tamil Nadu, India

Email: [3]victordoss64@gmail.com, [2]prof.k.manikandan@gmail.com
Corresponding Email: [1]thangarajusangeetha@gmail.com

Abstract: Cardiomyopathy is one of important cause of chronic heart failure which makes heart muscle harder to pump blood to other part of the body which leads to high mortality rate. Hence it is becomes mandatory to diagnosis and predict the disease in order to prevent the person against heart failure. However manual analysis of the disease is highly complex and leads to poor prognosis. In order to alleviate those challenges and predict the disease in early stage, many risk assessment methods has been modeled using machine learning and deep learning paradigms using genome wide association studies. Especially Cardiomyopathy risk assessment through gene expression from microarray data provides excellent results. In this article, microarray data containing gene expression data are preprocessed using missing value imputation through factor analysis and normalization through Z score normalization. Preprocessed gene expression data is employed to dimensionality reduction process through feature extraction and feature selection technique. In this model, linear discriminant analysis is employed as feature extraction method to extract differentially expressed gene (transcription of the RNA molecules that coded and non-coded for protein) which is represented as mutation chromosomes. Those genes are employed to feature selection technique to extract the targeted genes (type of variant and its score at specified location in genome of DNA) with respect to protein synthesized value (gene protein value) or molecular value of the gene using ant colony optimization. Optimal target genes contains the mutated chromosomes is selected. Finally target genes is employed to unsupervised deep learning model entitled as Integrated variational Autoencoder model for Genome transcription Analysis. It classifies the target gene representing miRNA on comparison with core set of target genes extracted from the diseased patient of the mutated chromosomes related to Cardiomyopathy which is considered as ground truth data into various classes of cardiomyopathy disease as ischemic Cardiomyopathy, dilated Cardiomyopathy and neurofibromatosis. Experimental analysis of various classifier employed to classify the core set of genes into type of classes of Cardiomyopathy is carried out on interfering the results of the classifier on the cross fold validation. Performance evaluation of the architectures on the mentioned dataset is performed using performance measure.

Keywords: Cardiomyopathy, Classification, Microarray data, Target Genes, Gene Profiling, Normalization, mRNA, Genome transcription Analysis**.**

1. Introduction

Cardiomyopathy is leading cause of chronic heart failure which increases the risk of prognosis of the patient. Hence it is becoming mandatory to diagnosis and predict the disease earlier to prevent the disease's negative effects. Cardiomyopathy is a heart muscle disease which is experienced in the left ventricles and leading to systolic dysfunction and contractile functions of the ventricle. Further it leads to poor blood circulation around the body. In order to alleviate those challenges and predict the disease in early stage, many risk assessment

methods have been modeled using machine learning and deep learning paradigms using genome wide association studies [1]. Especially Cardiomyopathy risk assessment through gene expression from microarray data provides excellent results. However, that architecture will produces discordant results due to genetic susceptibility.
In this article, microarray data containing gene expression data are preprocessed using missing value imputation through factor analysis and normalization through Z score normalization. Preprocessed gene expression data is employed to dimensionality reduction process through feature extraction and feature selection technique [2]. In this model, linear discriminant analysis is employed as feature extraction method to extract differentially expressed gene (transcription of the RNA molecules that coded and non-coded for protein) which is represented as mutation chromosomes [3]. Those genes are employed to feature selection technique to extract the targeted genes (type of variant and its score at specified location in genome of DNA) with respect to protein synthesized value (gene protein value) or molecular value of the gene using ant colony optimization [5]. Optimal target genes contain the mutated chromosomes is selected [4]. Finally target genes is employed to unsupervised deep learning model entitled as Integrated variational Autoencoder model[6] for Genome transcription Analysis. It classifies the target gene representing miRNA on comparison with core set of target genes extracted from the diseased patient of the mutated chromosomes related to cardiomyopathy which is considered as ground truth data into various classes of cardiomyopathy disease as ischemic cardiomyopathy, dilated cardiomyopathy and neurofibromatosis [7].
The article is sectioned into following section, literature similar to the high dimensional data classification section 2 provides a detailed analysis and description of the deep learning model utilizing machine learning. The proposed architecture of deep regularized variational Autoencoder is designed and implemented to classify Cardiomyopathy in section 3 and performance evaluation of experimental outcomes using benchmark dataset produces high effectiveness of classification model against conventional approaches is demonstrated in section 4. Finally, article has been summarized in section 5.

2. Related work

In this part, gene expression data is processed to classify the Cardiomyopathy diseases using deep learning and machine learning methods have been analysed in detail on its generated discriminate classes along the processing of the data along the preprocessing and feature extraction process to generate the efficient classes. Technique which performs similar to proposed model for microarray data classification is described as follows

2.1. Support Vector Machine

In this architecture, Support Vector Machine classifier is employed to classify the Cardiomyopathy into ischemic Cardiomyopathy, dilated Cardiomyopathy and neurofibromatosis on processing the target genes from the differentially expressed genes of the gene expression data. Initially gene expression data is represented in the cellular matrix which is further processed using the protein synthesis values. Extracted target gene has been updated against reconstructing an error to identify the key genes for categorization. Classifier model uses the hyper plane and margin to classify the core set gene into the classes of the disease with support vector machine. It is machine learning classifier with decision boundaries [8].

2.2. Convolution Neural Network

In this architecture, Convolution Neural Network is employed to classify the Cardiomyopathy into ischemic Cardiomyopathy, dilated Cardiomyopathy and neurofibromatosis on processing the target genes from the differentially expressed genes of the gene expression data. Classifier composed of convolution layer, max pooling layer, flatten layer and fully connected layer. Each layer process the core set of feature selected by generation of feature map with high level features in pooling layer. Feature map is classified in the fully connected layer. To determine the lowest reconstruction error via feature refinement, the features are further adjusted to the hyperparameter of different deep learning model network layers [9]. In the feature space used for genome analysis, the Softmax layer reduces intra- and inter-gene variance.

3. Proposed Model

In this section, architecture of proposed Integrated variational Autoencoder model for microarray data containing gene expression data are processed on inclusion of fine tuning of RMSProp generating output from the encoder and decoder layers the disease classes on the basis base pair of gene in the protein synthesis has been illustrated as follows.

a. **Data Pre-processing**

Genome Transcription architecture handles microarray data which contains the irrelevant gene variants such as cystic fibrosis and mitochondria with missing value. Missing value prediction and data normalization utilizes factor analysis in its execution [10] as it generates the normalized gene expression data. Data normalization is employed with Z score normalization [11].

- **Factor Analysis for Missing Value Imputation**

To imputation of missing values, factor analysis is used. The highest common variation on the specific amino acid residues is found using factor analysis. Using the Eigenvalue, it follows the Kaiser criteria. To fill in the missing value of the gene expression data, it calculates the variance score of a specific amino acid residue. Based on the data correlation of the missing amino acid residue for protein synthesis, it calculates the greatest probability value [11].

- **Z Score Normalization**

Normalization is to produce inside a certain range, the numbers of the decreased nucleotide values. Z-Score normalization technique employs the standard deviation and mean measures to normalize (transform or rescale) each input values of resultant nucleotides such that subsequent nucleotides have unit variance and a zero mean [12]. Each sample, $X_{i,n}$ in the dataset is changed into $X'_{i,n}$. It is also known as zero-mean normalization. The normalized nucleotides is given by

$$X'i,n = \frac{Xi,n-\mu}{\sigma_i} \text{....Eq.1}$$

Where μ, represents the mean and standard deviation value of i^{th} nucleotides respectively

- **Variational Autoencoder - Dimensionality Reduction**

Variational Autoencoder [12] is employed to the preprocessed gene expression data as it capable of minimizing the high dimensional gene expression to low dimensional gene expression on learning the dependencies and non-dependency amino acid on the sequence of message RNA (mRNA). Variational autoencoder are feed forward acyclic neural network eliminate the non-dependent gene expression. Dependent gene expressions are represented as latent space in vector form containing the probability distribution.

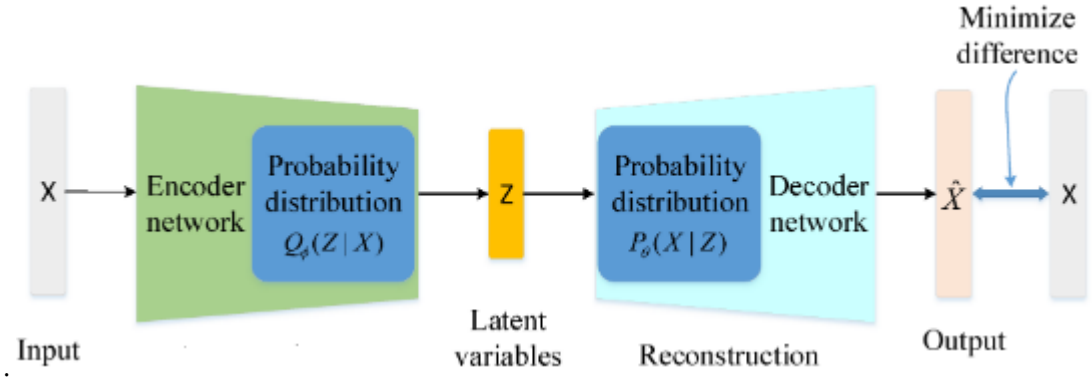

Figure 1: Autoencoder based Dimensionality Reduction

Variational Autoencoder model considered as best transformation vector containing the encoder block, decoder block and latent spaces contain the latent sequence of nucleotides. The optimal gene vector for classification is extracted on reduction of the high dimensional gene expression patterns on employment of encoding process as latent sequence of nucleotide**.**

Original gene expression X= {$Nucleotides_1$, $Nucleotides_2$…….. $Nucleotides_n$}...Eq.2

Latent Sequence of Nucleotide L= { $Nucleotides\ probability_1$, $Nucleotides\ probability_2$… Nucleotides probability}

Encode Operation $Z = h(W_{encoding} X + b_{encoding})$ ...Eq.3

Where $W_{encoding}$ and $b_{encoding}$ is the parameter and h is the activation function

Optimal Nucleotides = Decode(z)...Eq.4

$Decode(z) = h(W_{decoding}Z + b_{decoding})$..Eq.5

Where $W_{decoding}$ and $b_{decoding}$ is the parameter and h is the activation function

Probability distribution of the hidden attributes after decoding operation is given by KL divergence by

Probability distribution of attribute vector $P(z|x) = \frac{p(z)p(z)}{p(x)}$..Eq.6

The variational Nucleotides selected on inference to approximate the conditional probability the latent variables' distribution. KL divergence determines the disparity among the Nucleotides and it minimizes the difference as best as possible. Model parameter and activation function on the probability distributions on the latent vectors with respect to loss function and shuffling of the Nucleotides reconstructed on decoding operation [13].

3.2. Feature Extraction - Linear Discriminant Analysis

Linear discriminant analysis is employed as feature extraction method to extract differentially expressed gene (transcription of the RNA molecules that coded and non-coded for protein) which is represented as mutation chromosomes. Further it is used to compute the linear gene and non-linear gene on the gene expression data on mean and standard deviation measurement through scatter matrix. Scatter matrix contains the non-linear genes [14]. Given pre-processed gene expression data, initially the mean vector for the genome transcription is carried out.

Mean protein transcription for Single gene Mi $= \frac{1}{n} \sum_{x\varepsilon C}^{n} \blacksquare x, n_i$ Eq.7

Total protein transcription for entire genome Vector M $= \frac{1}{n} \sum_{x\varepsilon C}^{1} \blacksquare x, N$ ...Eq.8

Scatter Matrix for protein transcription of genome vector is $S_w = \sum_{i=1}^{c} \blacksquare s_i \frac{ni}{n}$.... Eq.9

Scatter Matrix between nucleotides of genome $S_b = \frac{1}{2} \sum_{i=1}^{c} \blacksquare P_i P_j (m_i - m_j)(m_i - m_j)^T$

where $(m_i - m_j)$ is vector nucleotides difference.

Covariance to the nucleotides of two gene of nucleus is calculated as $\mathbf{S_i} = \frac{1}{n} \sum_{x\varepsilon C}^{1} \blacksquare (x - m_i)(x - m_i)^T$ Eq.10

To extract the high variance nucleotides, the total covariance of the scatter matrixes is $\mathbf{S_T} = \frac{1}{N} \sum (m_i - m_j)(m_i - m_j)^T$ ... Eq.11

S_w reduction and S_b maximization are two important LDA method goals. Increasing the ratio $|S_b| / |S_w|$ will help achieve this. A further enhancement in ratio weight occurs when the projection matrix's column vector is formed by the eigenvectors of $S_b S_w$.

$W = [W_1, W_2 \ldots\ldots\ldots W_{c-1}]$ ...Eq.12

where the set of decreasing eigenvalues λ_i is represented by the eigenvectors of the S_b and S_w, W_i.

The best set of nucleotides has been found using linear discriminant analysis, which reduces dimensionality when estimating the linear combination of nucleotides on two or more genes and normalizes pair-wise similarities of the correlated nucleotides extracted on the variance of the object.

3.3. Feature Selection - Ant Colony Optimization

Ant Colony Optimization algorithm is to select the gene using biomarkers. Core set of feature is selected on basis of fitness function on the operation of cross over and mutation operation of the chromosomes on the available feature extracted which represents the gene [15]. Figure 3 represent the gene selection of the extracted genes. It selects the mutated chromosomes

Initialize n = 0 Initialize the variational gene as population as P(n)
Evaluate fitness function for the gene P(n)
While { The condition for termination is not met }
Do n = n+1 {iterate}
fitness function_Mutation ()
{Extract gene Nucleotides with large variance as better fitness}
End while
Return {fittest gene as Target Gene}

b. **Integrated variational Autoencoder**

Integrated variational Autoencoder classification is deep learning architecture is employed to classify the selected differentially expressed gene on the mRNA Nucleotides code into various classes of the Cardiomyopathy disease. In this part, variational Autoencoder employs the fine tuning of the activation function with exponential linear unit towards class generation with objective functions of generative adversarial network [16].

- **Embedding loss function**

It includes the non-linear loss components to eliminate the physical and functional structures of the chromosome 7 genes. In embedding loss function are latent representations of Nucleotides which contains the maximum Euclidean distance and reduce the base pair of gene for the protein synthesis.

Embedding loss of the cell structures = Min L(X, X|) +$\sum_{k=0}^{n}$ $G(z)$.....Eq.13

- **Activation layer**

Activation layer uses the ReLu function and exponential linear unit to produce the physical and structural cell structures. It preserves the Inter-point distance. Activation layer of the model is capable of approximation of the inherent Nucleotides of the gene variants. It is capable of computing the properties of the gene variants. Regularization of the cellular matrix for cell structure of different disease biomarker

Cell structure of the gene biomarker is Min g(z) =$\frac{1}{2}\sum_{i<j}^{n}$ ■ $||z_i||$.....Eq.14

Table 1: Hyper parameter of the Variational Autoencoder

Hyper Parameter	**Values**
Sample Size	629
Autoencoder learning Rate	0.04
Hidden layer	2 layers
Encoding dimensions	50
Dataset dimension	45
Epoch	50
Dropout	0.2
Optimizer	RMSprop
Loss function	Poisson Function

- **Drop Out Layer**

The dropout layers eliminate the temporal Nucleotides which contains the non-coded DNA sequences. Drop layer set the value of 0.2 which eliminate the overfitting issues. Regularized Autoencoder model regularizes the physical and functional cell structures. It is highly effective in discriminating the Nucleotides of the disease biomarkers. It contains the gene preserving constraints to high level Nucleotides.

High level Nucleotides is given as W_{ij} = { exp($\frac{x_i - x_j}{\alpha^2}$}...Eq.15

Disease preserving constraints to high level Nucleotides is given as

P_{ij} = {1 0 1 represent high level and 0 represent low level

The disease preserving approach heuristically computes the weight of the Nucleotides and represents it learns the discriminative features using few parameters and progresses from low-level to big abstract features. Pre-training the parameters has been used varied settings, while entire model is trained in an in fixed settings procedures for enabling faster training process.

- **Optimizer function**

In this work, RMSProp is used as optimizer to increase the gene discriminant and overfitting issue of the biomarker containing Nucleotides. It is considered loss reduction function. It contains high learning rate and it has good convergence rate. It reaches the global minima on the cost function with least possible value. Genome Transcription representation depends on the weight and bias value of the gradient for the biomarker. Gradient is given as

Class Gradient G_w = β w +(1-β)w...Eq.16

Class gradient minimize the distance of the biomarker in the Nucleotides containing high level Nucleotides of the cell on effective decomposition of the Nucleotides or transfer of the Nucleotides to the outlier class as approximation.

Table 2: Genome transcription Analysis for Normal Gene and Differentially Expressed Gene to different Cardiomyopathy disease

Gene Transcription	**Gene Type**	**Disease type**

AGAAGATGTTCAACTTCATAAAGATGTTGAAGAAGAAATGGAGA	Differentially Expressed Gene	Ischemic Cardiomyopathy
TTGACACTGGAGAATGATACCTACCCTGAAATAACTCACTTCCT	Differentially Expressed Gene	Dilated Cardiomyopathy
CAGCTGAGGCTGTGGTGGGAGAAGAATGACAAGAACTGA	Differentially Expressed Gene	Neurofibromatosis

- **Epoch Layer**

It is considered as no of the iteration for the high level Nucleotides structure. It determines the effective convergence of the model. Epoch layer generate the final weight of the similarity computation. It correlates the Nucleotides into classes. It learns the deep structure correlation on the adjustable parameters of the batch sizes.

Algorithm 1: Deep Adversarial regularized Variational Autoencoder

Input: Microarray Dataset
Output: High Representative Data Clusters
Process
Data Pre-process ()
Compute gene expression samples containing Missing value ()
Factor analysis ()
Kaiser Criteria on the gene expression to insert the Nucleotides of the missing field
Dimensionality Reduction ()
Feature Extraction_Autoencoder ()
Encoder ()
Assign Latent Nucleotides
G= {probability of the Nucleotides)
Encoded feature Set F_s containing high level Nucleotides
Decoder ()
Z= {encoded (g)}
Decoded Nucleotides Set F_s containing high level Nucleotides
Apply Deep Learning of GAN ()
Transfer learning ()
Activation function ()
Local properties of the Nucleotides
Embedding Loss Layer ()
Eliminate Physical and functional structures of the organelle
Dropout Layer ()
Biomarker preserving of the high level Nucleotides on eliminating the overfitting features
Epoch()
Output Layer ()
Softmax () --- Representation of the Class
High Representative disease Classes of cardiomyopathy

Algorithm produces the class with the differentially expressed gene in form mRNA as target genes. It establishes the class with more weighted with minimized reconstruction error and class losses on the various Nucleotides of the gene.

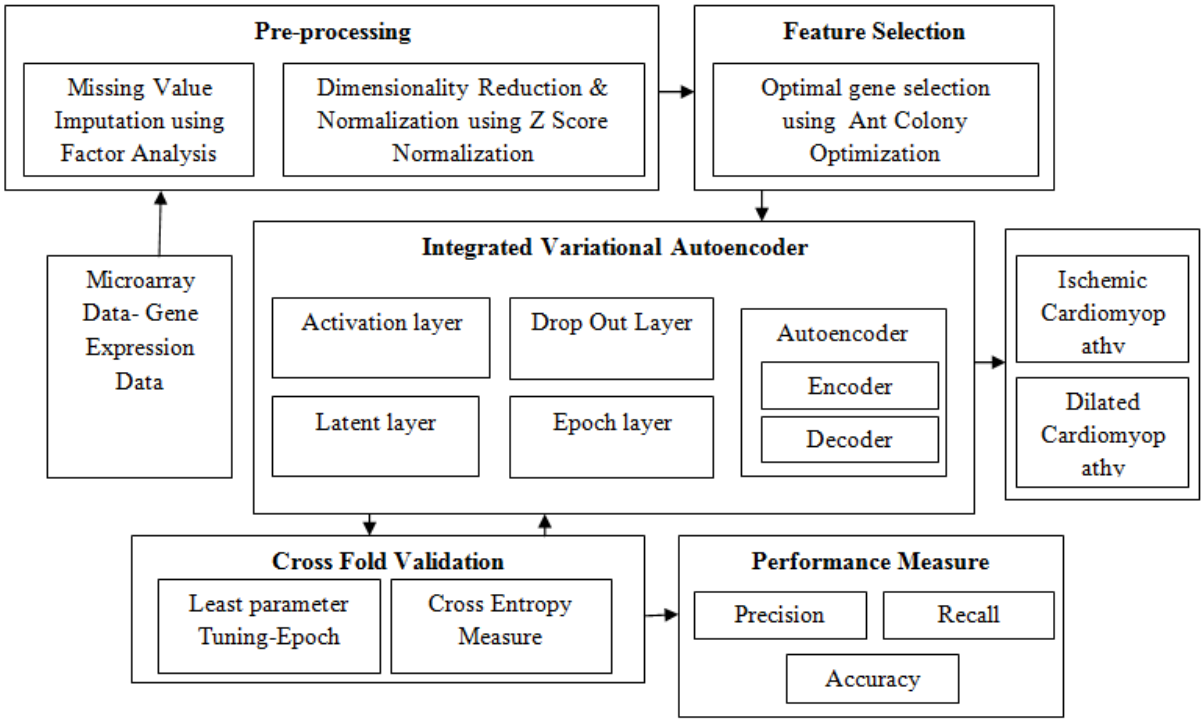

Figure 2: Architecture of the Integrated variational Autoencoder approach

Figure 2 represents the Integrated variational Autoencoder model for Classification of the microarray data. Model incorporates nucleotides preserving on encoding operation to preserve the high level amino acid chains. Activation layer uses the exponential linear unit to compute the high representative classes. Proposed architecture guarantees produces high convergence and learning rate of the class generation. However, nucleotides of the target genes are fine-tuned using stochastic gradient decent to yield the better results.

The activation function of the model makes use of transfer learning and can manage the non-linear dependence of the nucleotides. Finally, it avoids curse of dimensionality and effectively addressing the data sparsity difficulties in the feature weight update. Variational Autoencoder learning algorithm has been employed to produce the cluster with high nucleotides distance and maximizes the accuracy of classification in addition to minimizing the reconstruction error on optimizing loss function.

On increase of the training epoch, the reconstruction error and classification loss will slowly reduce and the accuracy of approach is enhanced [17]. Hence, weight updates details of the training phase of the variational Autoencoder in the GAN have been detailed and Cross validation has been carried out on the test data to validate the model.

4. Experimental Results

Experimental analysis of variational Autoencoder architecture for high representative class formation to Cardiomyopathy diseases has been formulated out on the GEO database which is mentioned as GSE138678 which was obtained from GEO database repository (http://www.ncbi.nlm.nih.gov/geo/) especially in CSV pattern [18]. Proposed performance has been evaluated on measure such as precision, recall and F-measure. In this study, the proposed model on cross-fold validation is validated using 20% of the dataset after 60% of the geo dataset has been trained.

Finally, a 10-fold validation is used to cross-examine the suggested model's performance. The performance assessment of the suggested architecture in terms of accuracy using the Twitter dataset is shown in Figure 2. Table 2 provides information about the training parameter for complete encoder learning.

Table 2: Training parameters

Parameter	**Value**
Learning rate of the Model	10^{-6}

Loss Function	Poisson Function
Batch size	15
Epoch	45

Dataset Description: Geo Database
Geo data set contains 1 lakh gene expression data of cardio respiratory illness patients. Data is represented in CSV format.

4.2. Performance Evaluation

The proposed architecture has been computed using confusion matrix on 10-fold validation. In this work, proposed model compute the distance target genes on the dataset mentioned. The process of the activation function, dropout layer, embedding loss layer, optimizer function, and model hyperparameters all affect how well the deep learning model performs.

- **Precision**

Positive predictive value is measured by it. The further illustrated as the fraction of relevant nucleotides among each cell structures projected employing the model. In other words, Precision is termed as ratio of number of optimal nucleotides to the number of all subspace of the nucleotides set extracted [19]. The suggested architecture's accuracy performance assessment on the geo dataset is shown in Figure 3.

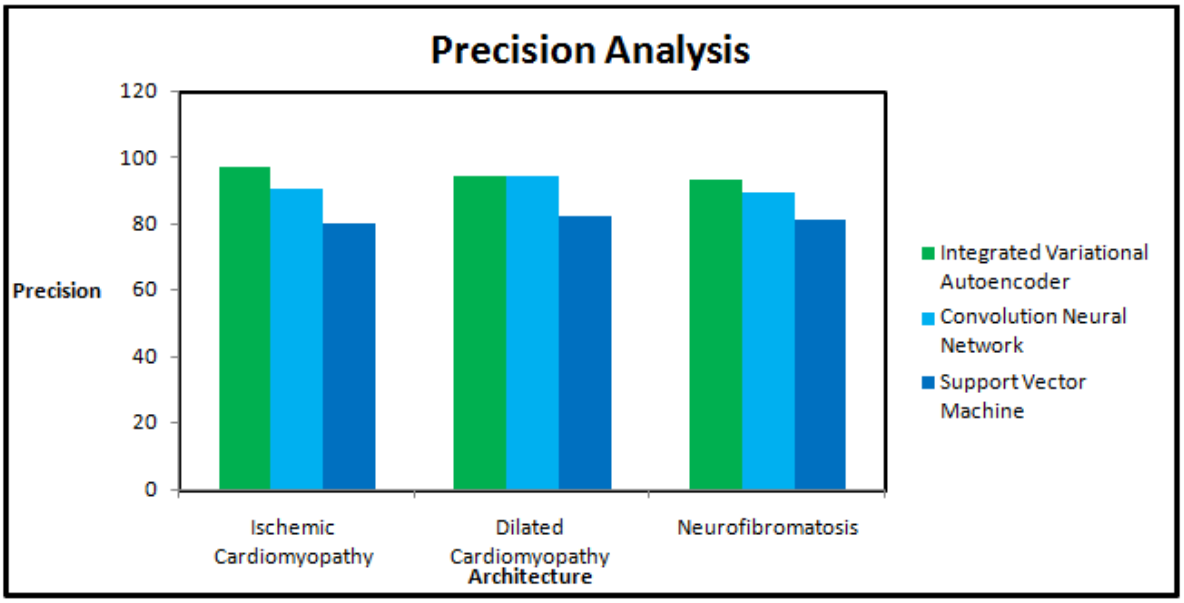

Figure 3: Performance Evaluation of the methodology with respect to Precision

Hyperparameters optimization is the reason for effectiveness.

$$\text{Precision} = \frac{True\ positive}{True\ positive + False\ Positive}$$

The number of comparable sites in the data is known as a true positive, while the number of really distinct nucleotides in the gene is known as a false negative [20]. High nucleotide gene similarity is a common characteristic of strong classification performance. Recall measure may be used to compute it.

- **Recall**

From the decoder portion of the model to the total number of similar data points gathered for the whole cluster, recall is the portion of the cluster's similar data points that are extracted. Recall is the evaluation of the grouped relevant features.

$$\text{Recall} = \frac{True\ positive}{True\ positive + False\ negative}$$

A true positive is the number of nucleotides in the data that are similar, and a false negative is the number of nucleotides in the dataset that are similar. The performance assessment of the suggested design on recall measures using traditional methods is shown in Figure 4.

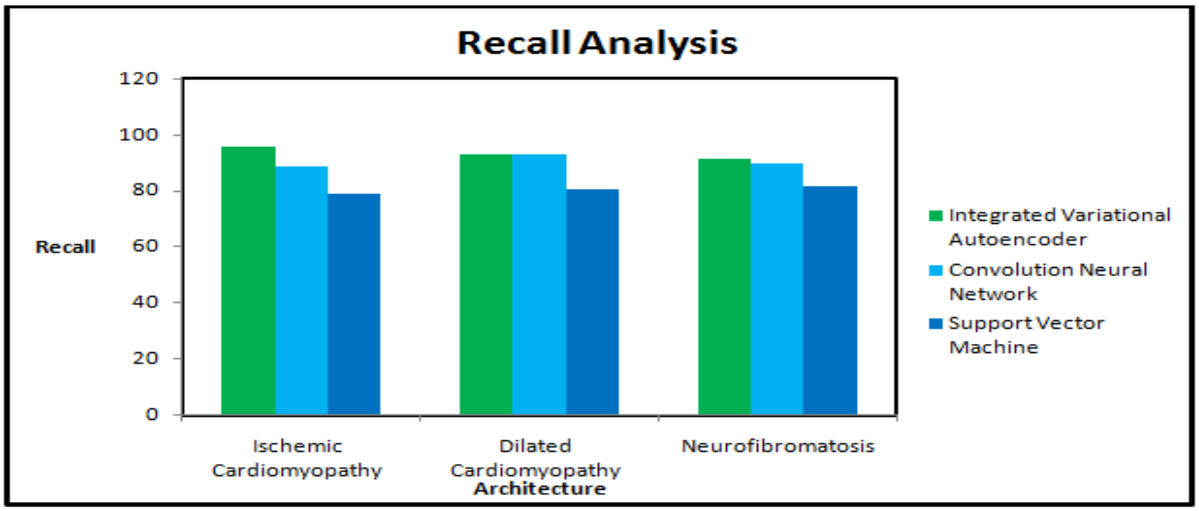

Figure 4: Performance Evaluation of the methodology with respect to Recall

Quality of the disease class depends on activation function and embedded loss layer of the model. Encoder dimension calculates the high level nucleotides to create subspace and dropout layer minimize the low level nucleotides of the gene. F measure is a measure of the class quality using epoch and batch size of the class on eliminating the over fitting issue and efficient in generating the outlier to the nucleotides.

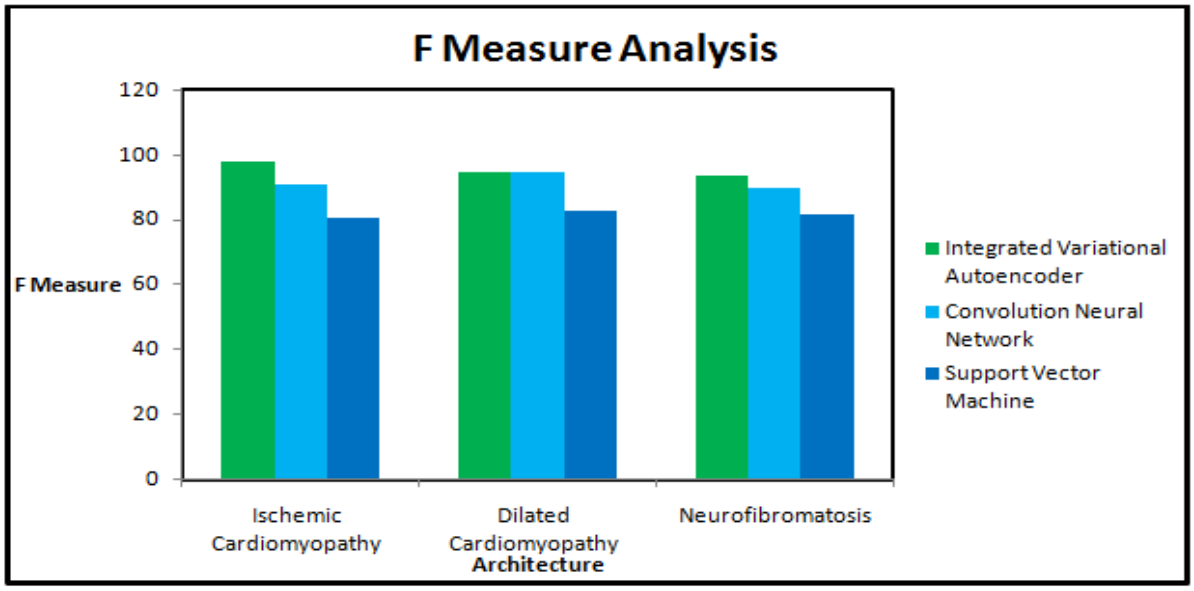

Figure 5: Performance Evaluation of the methodology with respect to F- Measure

- **F measure**

It is the number of relevant nucleotides in the class of gene expression data in the learning model. It is considered as accuracy. Accuracy is given by

$$Accuracy = \frac{True\ positive + True\ Negative}{True\ positive + True\ Negative + false\ positive + False\ negative}$$

Even though different nucleotides may affect the creation of distinct classes for Cardiomyopathy condition. The suggested model's f-measure performance in comparison to the techniques techniques for microarray data categorization is shown in Figure 5. Nevertheless, the curse of dimensionality reduces this gene vector eventually. Conversely, less separable classes are created for nucleotides with significant variability.
The suggested encoder function maps the input into a distribution by acting as an approximation function. Then using conditional probability, the generative probabilistic decoder attempts to produce the original sample. The technique's performance value for class analysis of various cardiomyopathy diseases is shown in Table 3.

Table 3: Performance Analysis of Autoencoder architecture against conventional approaches

Disease Class	**Method**	**Precision %**	**Recall %**	**F – Measure %**
Ischemic Cardiomyopathy	Support Vector Machine	80.78	78.78	82.78
	Integrated variational Autoencoder-Proposed	82.48	92.78	94.89
	Convolution Neural Network-Existing 2	81.48	88.78	80.48
Dilated Cardiomyopathy	Support Vector Machine -Existing 1	98.48	80.48	98.88
	Integrated variational Autoencoder - Existing 1	94.82	90.82	94.78
	Convolution Neural Network-Existing 2	90.75	88.75	90.58
Neurofibromatosis	Support Vector Machine -Existing 1	81.46	79.46	83.26
	Integrated variational Autoencoder - Proposed	93.58	88.58	91.78
	Convolution Neural Network-Existing 2	89.87	87.87	91.78

Finally, proposed classification approach not only for identifies high representative class to the Cardiomyopathy disease and it is capable of identifying stage of the diseases.

5. Conclusion

Integrated Variational Autoencoder technique is designed and implemented in this work to classify the microarray data into classes representing Cardiomyopathy disease. On processing the proposed model, it is capable of exploring deeply about the latent structure of the nucleotides and computes the associations of the nucleotides to construct the differentially expressed gene structures to microarray data. Further nucleotides are approximated using the variational Autoencoder to preserve the biomarker structures. Euclidean distance is employed in embedding function of the Autoencoder to generate the efficient disease classes. Finally, the suggested approach demonstrates its effectiveness and great scalability on high-dimensional data to achieve a high degree of correctness in the class structures that are produced.

6. References

1. 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.
2. D. H. Le, “Machine learning-based approaches for disease gene prediction,” Briefings in Functional Genomics, vol. 19, no. 5–6, pp. 350–363, 2020.
3. D.-H. Le and V.-T. Dang, “Ontology-based disease similarity network for disease gene prediction,” Vietnam Journal of ComputerScience, vol. 3, no. 3, pp. 197–205, 2016.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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
9. 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.
10. Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, “PGCN: disease gene prioritization by disease and gene embedding through graph convolution neural networks,” pp. 1–9, 2019, https://www.biorxiv.org/content/10.1101/532226v1/
11. Dizaji KG, Herandi A, Cheng,” Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017, 5747–56.
12. Xie J, Girshick R, Farhadi A ”Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning. New York City, NY, USA: ICMLR, 2016, 478–87.
13. K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, “INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification,” Proc. 15th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), Part II, pp. 149-160, 2011.
14. L. Parsons, E. Haque, and H. Liu, “Subspace clustering for high dimensional data: a review,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 90–105, 2004
15. N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, “The role of hubness in clustering high-dimensional data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 739–751, 2014.
16. H. Kremer, P. Kranen, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, “An effective evaluation measure for clustering on evolving data streams,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 868–876.
17. P. Huang,Y. Huang,W.Wang, and L.Wang, ``Deep embedding network for clustering," in Proc. 22nd International. Conference. Pattern Recognition. (ICPR), Aug. 2014, pp. 1532-1537.
18. P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, ``Deep subspace clustering networks," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 23-32.
19. W. Harchaoui, P. A. Mattei, and C. Bouveyron, ``Deep adversarial Gaussian mixture auto-encoder for clustering," in Proc. ICLR, 2017, pp. 1-5.
20. N. Dilokthanakul et al. (2016). ``Deep unsupervised clustering with Gaussian mixture variational autoencoders." [Online]. Available: https:// arxiv.org/abs/1611.02648