### ABSTRACT

**Introduction:** Heart Failure (HF) is a complicated condition as well as a significant public health issue. Data processing is now required for machine and statistical learning techniques while it helps to identify key features and eliminates unimportant, redundant, or noisy characteristics, hence minimizing the feature space’s dimensions. A common cause of mortality in cases of heart disease is Dilated Cardiomyopathy (DCM).

**Methods:** The feature selection in this work depends on the Entropy Pelican Optimization Algorithm (EPOA). It is a recreation of pelicans’ typical hunting behaviour. This is comparable to certain characteristics that lead to better approaches for solving high-dimensional datasets. Then Deep Autoencoder (DAE) classifier has been introduced for the prediction of patients. DAE classifier is employed to compute the system’s nonlinear function through data from the normal and failure state.

**Results:** DAE was discovered to not only considerably increase accuracy but also to be beneficial when there is a limited amount of labelled data. Performance metrics like recall, precision, accuracy, f-measure, and error rate has been used for results analysis.

**Conclusion:** Publicly available benchmark dataset has been collected from Gene Expression Omnibus (GEO) repository to evaluate and contrast the suitability of the suggested classifier with other existing methods.

**Keywords:** Heart Failure (HF); Dilated Cardiomyopathy (DCM); Entropy Pelican Optimization Algorithm (EPOA); Gene Expression Omnibus (GEO); Deep Autoencoder (DAE); FS (Feature Selection).
Resultados: se descubrió que DAE no solo aumenta considerablemente la precisión, sino que también es beneficioso cuando hay una cantidad limitada de datos etiquetados. Para el análisis de resultados se han utilizado métricas de rendimiento como recuperación, precisión, exactitud, medida f y tasa de error. 
Conclusión: se recopiló un conjunto de datos de referencia disponibles públicamente del repositorio Gene Expression Omnibus (GEO) para evaluar y contrastar la idoneidad del clasificador sugerido con otros métodos existentes.

Palabras clave: Insuficiencia Cardíaca (HF); Miocardiopatía Dilatada (DCM); Algoritmo de Optimización de Entropía Pelícano (EPOA); Ómnibus de Expresión Genética (GEO); Codificador Automático Profundo (DAE); FS (Selección de Funciones).

INTRODUCTION

The prevalent cause for mortality is cardiovascular illness. Cardiovascular disorders remain to impact over 500 million individuals worldwide, accounting for around one-third of all fatalities. One of the cardiovascular diseases with an elevated risk for mortality as well as other uncomfortable symptoms and a poor standard of life is heart failure. It restricts regular physical and social activity and causes a range of emotional and bodily distress. The most prevalent reason for heart transplantation and a major contributor to heart failure is dilated cardiomyopathy (DCM). A homogeneous phenotypic classification that utilizes anomalies in Heart structure and function is used to make the diagnosis of DCM. But it is a disorder with a variety of aetiologies (genetic, clinical, concomitant, arrhythmia), clinical symptoms, and response to therapy. DCM is linked to an estimated 20% 5-year mortality rate. The terminology and medical category for DCM still are inconsistent. This shortcoming prevents clinicians from diagnosing and treating these clinically diverse heart failure symptoms in an efficient way.

Heart disease treatment options include genetic testing. It entails collecting an element of a person's DNA and examining it for alterations (also known as pathogenic mutations) that potentially result in genetic heart disease. There are three different categories into which DNA microarrays' main applications can be divided. For starters, scientists can identify genes that are differentially expressed in heart illness from normal and irregular cardiac rhythms by comparing expression levels among predetermined groups of samples. A second use is class prediction, which necessitates classifying the sample's members (for example, to determine the patient's heart rate). It is important to identify subgroups that have similar characteristics. Techniques for removing the informative genes linked to heart disease thus become essential and crucial. Although machine learning-based prediction models are increasingly being used in healthcare, physicians still find it difficult to depend on these algorithms in clinical practice. Multiple forms of distinct or multi-functional large-scale data sets, including transcriptomic, genomic, and demographic data of patients, have been used for the categorization of diseases by machine learning (ML).

It gets challenging to train different ML models with a huge gene characteristic. In order to conquer the high input dimensions and relatively limited sample sizes, feature selection has become a standard approach in the analysis of HD (High-Dimensional) data, like the EG (Expressed Gene) information used in this investigation. Filter, wrapper, and embedding techniques can be used to do classic feature selection tasks. The filter approach resolves the optimal feature ranking using statistical calculation mode rather than a machine learning model. It moves quickly but inaccurately. A particular feature evaluator or machine learning model is used by the wrapper technique. Using the heuristic search technique, it is always seeking the best feature combination. The best feature subset for the feature classifier can be discovered using the fitness function as the evaluator’s return value. The wrapper technique's drawbacks still are high temporal complexity and local optimization. The feature selection and machine learning process are skilfully combined in the embedded approach, which then produces the feature subset using the model's weight parameters. This approach's effectiveness is dependent on the ML technique, and some frameworks cannot produce weight parameters.

The majority of present hybrid FS approaches either infer the link among the samples using the data model or provided that each sample is evenly distributed. Data from DNA microarrays differs from data from natural sources. Their main characteristics (genes) have complex connections on one another rather than being completely independent. Due to the issues of these methods, optimization is the procedure for choosing the best answer from a series of potential ones. The two main categories of methods to tackling optimization problems are deterministic and stochastic techniques. Algorithms for population-based optimization are among the most effective ones among stochastic approaches. In this paper, Entropy Pelican Optimization Algorithm (EPOA) is introduced to imitate the pelican's hunting style and tactics. In order to get a subset of characteristics that have minimal redundancy, great robustness, and practical significance, EPOA has been established. In order to forecast individual heart tissues and the way they are related to heart tissues, Deep Autoencoder (DAE) classifier has been developed. The Gene Expression Omnibus (GEO) repository's publicly accessible benchmark
data have been considered to evaluate and compare the suitability of the suggested hybridization approaches to the other available techniques.

**Literature review**

Schiano et al.\(^{(11)}\) suggested a ML method using Protein-Protein Interaction (PPI) network. Investigating transcriptome properties of individual’s cardiac tissue required combining a novel computational strategy with a bioinformatics framework for differential expression. Real-time PCR quantitative validation was used. Using tissue samples from transplants, initial studies identified particular DCM-related genes recognized to play a role in cardiomyopathies. A surprising link among Left Ventricular End-Diastolic Diameter (LVEDD), and NEAT1 as determined by severity classification (NYHA-class III), was discovered when these expression patterns were combined with clinical features. Conclusions: Employing the ML method, initial genes that can swiftly result in several molecular targets associated with DCM clinical characteristics were discovered.

Zhu et al.\(^{(12)}\) explored genes having immune-related characteristics in DCM cases. The ‘limma’ tool in R software was utilized to find Differentially EG (DEG) following data preprocessing. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) route enriched studies followed to establish the biological roles of the DEG.xCell were used to calculate the patterns of stromal and immunological cells. According to Protein-Protein Interaction (PPI) network analysed by STRING web tool and Cytoscape program, functional modules and hub genes were discovered. Hub genes were the subject of a correlation investigation among immune cell subsets.

Chen et al.\(^{(13)}\) performed data analysis in GSEA (Gene Set Enrichment Analysis) and PPI detections. The feature gene selection was supported by the Support Vector Machine (SVM), and Least Absolute Shrinkage and Selection Operator (LASSO) methods, and the effectiveness of the prediction model was assessed using ROC curve evaluation. RNA transcripts (CIBERSORT) and Non-negative Matrix Factorization (NMF) approach have both been developed for molecular typing and forecasting immune cell infiltration, respectively.

Xie et al.\(^{(14)}\) suggested GNN (Graph Neural Network) based FS method. For creating graph-structured data, the suggested technique makes advantage of the in-depth connections among characteristics and the Pearson Correlation Coefficient (PCC). To combine node data on graph structured data, data aggregation and dissemination processes using graph neural networks are used. The spectral clustering technique groups the redundant features. Later, each clustering sub-cluster is subjected to the action of a feature ranking aggregation framework utilizing 8 feature evaluation techniques. The suggested technique can successfully get rid of extra features. The algorithm’s result is more accurate and stable in its classifications, which could assist in the selection of prospective biomarkers.

Kang et al.\(^{(15)}\) suggested a 2-Stage Teaching-Learning-Based Optimization (TS-TLBO) technique to enhance the classification of bioinformatics info performance. In order to efficiently decrease the search space, both noisy and possibly informative features are selected during the selection reduction stage. The educator and the pupil with adaptive self-education along FS difficulties in the subsequent comparative self-learning stage to improve the exploitation capabilities. To increase the excellence, an opposition-based education technique is also used to develop early answers. Through vigorously modifying the mutation rate in accordance with the educator convergence ability, a self-adaptive mutation method is provided to enhance search performance.

Yang et al.\(^{(16)}\) suggested three DCM datasets (GSE9800, GSE3585, and GSE84796) from the Gene Expression Omnibus ( GEO) database were eliminated from the dataset and batches were combined into a single dataset. Weighted Gene Co-Expression Network Analysis (WGCNA) in R software was utilized to screen DEG and analyse the relationships among gene co-expression elements and medical characteristics. Common Gene (CG) refers to any DEG from the combined dataset that intersected with the significant module genes. The CGs’ enrichment study was carried out using the dataset. Utilizing Cytoscape 3.8.2 software, the CGs PPI network was displayed, and hub gene was discovered.

Jiang et al.\(^{(17)}\) identified DEG on two DCM gene expression datasets, and performed Disease Ontology (DO), GO, and GSEA functional optimization to find possible routes. To identify diagnostic markers, the SVM- Recursive Feature Elimination (SVM-RFE) technique and the LASSO method were utilized. A tool called Cell Type Analysis Tool (CIBERSORT) has been developed to analyse immune cell infiltration.

Dauda et al.\(^{(18)}\) suggested hybridized dimension reduction using Deep Neural Networks (DNN), Genetic Algorithms (GA), and Boruta Algorithms (BA). Numerous genes are related to the forecasting of cardiac rhythm individuals, which is one of the remaining inquiries by computing and biological scientists. To identify the key genes, five different micro-array data on cardiac disorders have been considered. In order to obtain highly differentially expressed genes and improve both prediction of parameters and model interpretation, three distinct sets of two-way hybrids among BA and NN (BANN), and BA and DNN (BADNN) and GA- DNN (GADNN) were introduced.

Chen et al.\(^{(19)}\) suggested microarray probe-sequence re-annotation of long non-coding RNA (IncRNA) expression profiles connected to DCM and non-failing hearts. An extremely strong association between a module and DCM

https://doi.org/10.56294/sctconf2024761
status was found by weighted gene co-expression network analysis. Apart from ZNF571-AS1, all hub lncRNA localization to the cytoplasm was expected. Hub lncRNA may behave as competitive endogenous RNA (ceRNA) as a potential pathogenic mechanism for DCM. Effectively replicated in a different external dataset based on outcomes. Hub lncRNA was located in human aortic smooth muscle cells using RNA-Fluorescence In Situ Hybridization (FISH), and quantitative RT-PCR results verified that it was up-regulated in DCM individuals. The data suggested the possible involvement of eight candidate lncRNAs in DCM illness, possibly through competing endogenous RNA (ceRNA) crosstalk.

Song et al. (20) suggested the R package's Non-Negative Matrix Factorization (NMF) clustering to choose candidate genes for determining the DCM subgroup. The GEO database was utilized to obtain raw data on DCM and healthy heart tissue. The methods employed differential research, route enhancement, subtype identity, immunocytes infiltration, and single-cell RNA sequencing analysis. When it comes to immunological activation, heart failure in DCM, heparin sulphate (HS) is intimately correlated. Additional research was done on a molecular categorization of DCM using HS heterogeneity in DCM individuals.

Zhang et al. (21) suggested the WGCNA to find possible biomarkers and therapeutic targets. According to the Meta scape database, PPI networks of the important DEG. It was designed to investigate potential genes related with DCM and their dormant regulatory effects in both male and female patients. Hub genes in PPI networks, putative Transcription Factor (TF), microRNA, and highly enriched pathways were examined to determine the variations among female and male patients after the detection of critical DEG in an essential WGCNA module. Numerous important DEG have expression trends that are consistent with outcomes from earlier studies, demonstrating the validity of the database and data analysis findings.

**PROPOSED METHODOLOGY**

Here, Entropy Pelican Optimization Algorithm (EPOA) has been introduced for feature selection. This compares to several criteria that result in better approaches to solving high-dimensional datasets. Next, a Deep Autoencoder (DAE) classifier is used to forecast patients' heart tissues and how that tissue is related to them. The Gene Expression Omnibus (GEO) repository's publicly accessible benchmark data have been taken into account to evaluate and compare the suitability of the suggested hybridization approaches to current techniques.

https://doi.org/10.56294/sctconf2024761
Microarray dataset

Table 1 provides the outcomes of search in the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/) for RNA-seq research on DCM. The meta-analysis contained only research that used tissue models taken from the patients left ventricle.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Platform</th>
<th>Sample Size</th>
<th>Human Heart Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE116250</td>
<td>Illumina HiSeq 2500</td>
<td>37 DCM 14 NF</td>
<td>Left Ventricle</td>
</tr>
<tr>
<td>GSE57344</td>
<td>Illumina HiSeq 2500</td>
<td>2 DCM 14 NF</td>
<td>Left Ventricle</td>
</tr>
<tr>
<td>GSE71613</td>
<td>Illumina HiSeq 2500</td>
<td>2 DCM 4 NF</td>
<td>Left Ventricle</td>
</tr>
</tbody>
</table>

From the European Nucleotide Archive website (https://www.ebi.ac.uk/ena), raw data (FASTQ files) were obtained. Low read count genes were excluded utilizing the filtering process of DESeq2 with the default values, utilizing the filter statistic is the average of standardized counts. Fisher's united probability test for meta-analysis was carried out with the use of the Meta RNASeq package. Using the equation (1), this method integrates the p-values for each gene from many studies.

\[ F_g = -2 \sum_{s=1}^{S} \ln(p_{gs}) \]  

Where \( p_{gs} \) is the p-value for gene \( g \) in study \( S \). Fisher's test has a 2S degree of freedom Chi-squared distribution when p-values are independent. Larger \( F_g \) and null hypothesis rejection is produced by smaller p-values. Where p-values were modified to account for Benjamini-Hochberg False Discovery Rate (FDR), the raw p-value derived for gene \( g \) in a differential analysis for study \( S \) can be denoted as \( p_{gs} \). A statistically significant result was one where the modified p-value was less than 0.05. From the DEG list, genes whose expression patterns varied across different studies were eliminated.

Entropy pelican optimization algorithm (epoa)

Entropy Pelican Optimization technique (EPOA), a brand-new stochastic nature-inspired optimization technique, is presented for feature selection. It is a recreation of pelicans’ typical hunting behaviour. In EPOA, pelicans serve as search agents and look for food sources by selecting the best set of attributes from the dataset. The pelican can seize and consume target with the support of its long beak and huge pouch in its throat. This bird is sociable and frequently gathers in groups of lots of pelicans. Each population member in population-based procedures is a possible answer for the best possible feature selection. Depending on where they are in the search space, every population member offers values for the best qualities. Equation (2) is employed to arbitrarily generate population members according to upper and lower bounds of the feature range.

\[ x_{i,j} = l_j + \text{rand.}(u_j - l_j), i = 1, \ldots, N, j = 1, \ldots, m \]  

Where \( l_j \) is the jth lower bound and \( u_j \) is the jth upper bound of the problem feature, \( x_{i,j} \) is the value of the jth feature delivered by the ith candidate solution, \( m \) express problem features numbers, \( N \) express population members numbers, \( X_{i,j} \) is a random number within the inteval [0, 1].

Equation (3) uses a population matrix to recognize the pelican population members within the planned POA. The matrix columns reflect the feature range for the ideal solution, and each row indicates a sample from the gene dataset.

\[
X = \begin{bmatrix}
X_1 \\
\vdots \\
X_l \\
\vdots \\
X_N \\
\end{bmatrix}_{N \times m} = \begin{bmatrix}
x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{l,1} & \cdots & x_{l,j} & \cdots & x_{l,m} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,m} \\
\end{bmatrix}_{N \times m}
\]

If \( X_i \) is the ith pelican and \( X \) is the pelican population matrix. Each population member in the suggested POA serves as a pelican in the feature selection scheme for the gene dataset. As a result, depending on the potential solutions, the given problem’s OF (Objective Function) can be assessed. Equation (4) uses a vector to derive the values acquired for the OF.

https://doi.org/10.56294/sctconf2024761
Here $F$ is the fitness function vector and $F_i$ is the OF value of the $i$th dataset. $F$ is computed depending on the entropy range of the feature in the gene dataset.

Entropy is a system measurement. In the area of information, Shannon put out the concept of information entropy. Weight distribution could be accomplished by using information entropy. The better for a gene, the higher the entropy. Initially the definition of the uncertainty function shown in following:

$$g = \log\left(\frac{1}{p}\right) = -\log p$$  \hspace{1cm} (5)$$

Here $P$ is related probability. Next, by computing the uncertainty function’s expectation, the information entropy can be measured.

$$H(G) = E(G) = \sum_{g \in G} g \cdot p = -\sum p \log p$$ \hspace{1cm} (6)$$

Where $G$ is the set of all possible genes in the dataset. The suggested EPOA updates the feature selection approach by imitating the strategies and actions used by pelicans to assault and pursue prey. The two parts of hunting technique: (i) Reaching the target in the exploring phase. (ii) Water surface winging (exploitation phase).

**Phase 1: Moving towards Prey (Exploration Phase)**

The pelicans determine the prey location during the initial phase and subsequently fly in that direction. By simulating this pelican’s approach, researchers may scan the gene search space and demonstrate the suggested EPOA’s capacity to investigate a variety of search spaces. The crucial aspect of EPOA is prey’s gene position in the gene search space is generated randomly. Equation (7) details the pelican’s method to its target area of prey.

$$X_{i,j}^{P1} = \begin{cases} x_{i,j} + \text{rand.}(p_j - 1.x_{i,j}) \cdot E_p & \text{if } F_p < F_i \\ x_{i,j} + \text{rand.}(x_{i,j} - p_j), & \text{else} \end{cases}$$ \hspace{1cm} (7)$$

Here $X_{i,j}^{P1}$ is the $i$th pelican’s new position in the $j$th gene, according to phase 1 analysis, $P$ is the position of prey in the $j$th dimension, and $E_p$ is its OF value computed depending on the entropy of the gene. $I$ is a random number equal either $1$ or $2$. When this parameter’s value is $2$, it causes a member to experience more dislocation, which may result in that member’s selection of a fresher gene in the dataset. It significantly affects how precisely the EPOA may examine the gene search space. If the value of the entropy is increased at the novel gene position for a pelican, the suggested EPOA will approve it. This sort of updating, sometimes referred to as effective updating, prevents the model from moving to less-than-ideal areas. Equation (8) is used to simulate this procedure.

$$X_i = \begin{cases} X_{i}^{P1}, & \text{if } F_i^{P1} < F_i \\ X_{i}, & \text{else} \end{cases}$$ \hspace{1cm} (8)$$

Where $X_{i}^{P1}$ is the novel position of the $i$th pelican and $F_i^{P1}$ is its fitness function value depending on phase 1.

**Phase 2: Winging on the Water Surface (Exploitation Phase)**

The fish are propelled forward by the pelicans’ expanded wings as they approach the water’s surface for the second phase before being captured in their throat pouches. This procedure improves POA’s capacity for...
exploitation and local gene search. For gene solution, it checks the points nearby the pelican position. Equation (9) employs the mathematical computation to imitate the pelican’s chasing behavior.

\[ x^{P2}_{i,j} = x_{i,j} + R \left(1 - \frac{t}{T}\right) \cdot (2 \cdot \text{rand} - 1) \cdot x_{i,j} \]  

(9)

Where \( X^{P2}_{i,j} \) is the novel gene solution of the ith pelican in the jth gene depending on phase 2, \( R \) is a constant, which is equal to 0.2, \( R \cdot (1 - t/T) \) is the neighbourhood radius of \( X_{i,j} \) while, \( T \) is the maximum amount of iterations and \( t \) is the iteration counter. The coefficient “\( R \cdot (1 - t/T) \)” is used to search locally for the closest gene in order to arrive at an improved DCM solution. It displays the population’s neighbours’ neighbourhood’s radius. It works well with EPOA exploitation to come even closer to the greatest possible gene solution. Because of the high value of this coefficient in the initial iterations, a larger area surrounds every member is taken into account. The “\( R \cdot (1 - t/T) \)” coefficient lowers as the method replicates more, resulted in decreasing radii of each member’s neighbourhood. Given the dataset, EPOA can converge to conclusions that are nearer to the overall optimal gene. Equation (10) has been applied to reject or accept the revised pelican position.

\[ X_i = \begin{cases} X^{P2}_i, & F^{P2}_i < F_i \\ X_i, & \text{else} \end{cases} \]  

(10)

Where \( X^{P2}_i \) is the novel status of the ith pelican and \( F^{P2}_i \) is its objective function value. Every population member and the OF’s values will be updated after which the ideal candidate solution will be modified. The system moved to the following iteration, repeating the stages of the suggested EPOA using equations (2) through (10) until the full execution is complete. The pseudo-code of EPOA is represented in Algorithm 1.

Algorithm 1. Pseudo-code of EPOA

Start EPOA
Enter the dataset for the gene selection task.
Calculate the iterations number (\( T \)) and EPOA population (\( N \)) size.
Determine the fitness function using equations (4-6) and the original position of the pelicans.
For \( t = 1: T \)
Create the prey’s gene position at arbitrary.
For \( I = 1: N \)
Phase 1: approaching prey (exploration phase)
For \( j = 1: m \)
Use equation (7) to determine the jth gene’s new position.
End
Equation (8) is used to update the ith population member.
Phase 2: Winging on the water surface (exploitation phase)
For \( j = 1: m \)
Utilizing equation (9) evaluate the jth gene’s newest position.
End
Equation (10) should be updated for the ith population member.
End
Update the recommended result
End
Optimal gene output discovered by EPOA
End EPOA

Deep autoencoder (dae) classification

The reliable unsupervised learning model for the classification of HF with DCM is the autoencoder that utilizes a deep neural network. This method is extensively suited for classification since it may be used to extrapolate informative and differentiating characteristics from a large unlabelled dataset.\(^{(24,25)}\) An autoencoder essentially consists of hidden, three-layer NN with IL (Input Layer), and OL (Output Layers) for the encoder and Hidden Layer HL and OL for the decoder. The encoder converts a feature vector with a lower dimension from an input gene dataset with a higher dimension. The decoder then transforms the dataset into the input dimension. Thus, assist with the autoencoder correctly recreate the decoder’s output, it is used to establish a complex nonlinear connection among the input datasets. Because of this, significant characteristics will be
preserved in the concealed layer and the reconstruction error will continuously reduce over the training time. The result of the hidden layer will also demonstrate the success of the autoencoder’s categorization. The basic DAE arrangement is presented in figure 2. 

**Figure 2. Deep autoencoder (dae) framework**

Gene data instances for the n-dimension input $X=[x_1, x_2, ..., x_n]$, Equation (11) can be used to represent the output/activation of the hidden layer $h$ with m-dimension ($m < n$).

$$h = f^h(W^{(1)}x + b^{(1)})$$  

(11)

Where, $b^{(1)}$, $f^h$ and $W^{(1)}x$ indicate the bias, activation, and weight matrixes connecting the input and hidden layers, accordingly. Following the decoding procedure, the output layer’s reconstruction of the data, $x$ can be stated as follows.

$$\hat{x} = f^a(W^{(2)}h + b^{(2)})$$  

(12)

Here, $b^{(2)}$, $W^{(2)}$ indicate the output layer’s bias vector and weight matrix. Typically, a sigmoid function is employed as the AF (Activation Function) for the decoder and encoder elements. $f(t) = 1/(1+e^{-t})$ or given data type, any different activation function may be used. Initial weight and bias values are used in the training process. The settings must be changed during training in order to minimize the reconstruction error among the input data and reconstructed output. The mean squared error (MSE), which is indicated in equation (13) and used in evaluation, is used to quantify the reconstruction error.

$$L(W, b) = \frac{1}{N} \sum_{n=1}^{N} (x - \hat{x})^2$$  

(13)

Further instances, BCEL (Binary Cross-Entropy Loss) may be estimated as the reconstruction error utilizing equation (14) if the input values are between 0 and 1.

$$L(W, b) = -\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [x_j^{(i)} \log \hat{x}_j^{(i)} + (1 - x_j^{(i)}) \log (1 - \hat{x}_j^{(i)})]$$  

(14)

In this investigation, the activation function for the DAE’s output and hidden layers is Scaled Exponential Linear Unit (SELU). In addition to being an ideal option for the DCM classification and addressing the vanishing gradient issue in the DAE design, SELU is a non-saturating form of AF. By combining the DAE, SELU has been utilized to speed up the training process. Equation (15) can be used to describe SELU.

https://doi.org/10.56294/sctconf2024761
Thus, the values for the coefficients $\lambda$ and $\alpha$ are fixed to around 1,05 and 1,6731, accordingly. Adaptive moment estimation (Adam), which is utilized in this study, is the optimizer for updating the weight. Due to its capacity to learn previous gradients and also previous squared gradients significantly declining loss function average, this optimizer technique is quickly gaining popularity. In this analysis, a DAE is applied to classify DCM samples and the architecture is provided in table 2.

<table>
<thead>
<tr>
<th>Table 2. Structure of the designed deep autoencoder (dae)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer #</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>

RESULTS AND DISCUSSION

This part displays the outcomes of the classifier study of the NCBI GEO dataset. From the NCBI GEO database, five distinct datasets on heart rhythm were obtained. From the European Nucleotide Archive website (https://www.ebi.ac.uk/ena), raw data (FASTQ files) were obtained. Then, various techniques have been suggested for results comparison. Using classification and feature selection provides the numerous performance indices that are currently in usage. In the dataset, 70 % of samples are used training, and 30 % of samples has been employed for testing.

Performance measures: recall, precision, F-measure, accuracy, and error have been used to evaluate the classifiers in this study. The outline of the confusion matrix is revealed in table 3.

<table>
<thead>
<tr>
<th>Table 3. Confusion matrix in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Class</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>P</td>
</tr>
<tr>
<td>N</td>
</tr>
</tbody>
</table>

The performance of classifiers was assessed using the following metrics: accuracy, f-measure, recall, specificity, sensitivity, and precision. With the direction of the following equations (16–20).

\[
\text{Precision}(P) = \frac{TP}{TP + FP} \tag{16}
\]

\[
\text{Recall}(R) = \frac{TP}{TP + FN} \tag{17}
\]

\[
\text{F - measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}
\]

\[
\text{Accuracy}(\text{Acc}) = \frac{TP + TN + FP + FN}{FP + FN} \tag{19}
\]

\[
\text{Error} = \frac{TP + TN + FP + FN}{FP + FN} \tag{20}
\]
True Positive (TP) signifies that the positive samples’ diagnoses were accurate. False Negative (FN) denotes an inaccurate diagnosis of the positive samples. False Positive (FP) denotes erroneous diagnosis of the non-positive samples. True Negative (TN) signifies that the non-positive samples diagnoses were accurate.

Table 4. Gene selection results

<table>
<thead>
<tr>
<th>DATASET</th>
<th>GSE116250</th>
<th>GSE57344</th>
<th>GSE71613</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE116250</td>
<td>ENSG00000000005 TNMD</td>
<td>GSR1247</td>
<td>ENSG00000000003</td>
</tr>
<tr>
<td></td>
<td>ENSG00000000003 TSPAN6</td>
<td>GSR125B1</td>
<td>ENSG000000000938</td>
</tr>
<tr>
<td></td>
<td>ENSG000000004575CYL3</td>
<td>MFSD11</td>
<td>ENSG000000001084</td>
</tr>
<tr>
<td></td>
<td>ENSG0000000938FGR</td>
<td>METTL2A</td>
<td>ENSG000000001461</td>
</tr>
<tr>
<td></td>
<td>ENSG00000001167NFYA</td>
<td>ABCA10</td>
<td>ENSG000000001561</td>
</tr>
<tr>
<td></td>
<td>ENSG00000001460STPG1</td>
<td>ACAD10</td>
<td>ENSG000000002933</td>
</tr>
</tbody>
</table>

Some of the selected genes results of proposed system are discussed in table 4. Accuracy results comparison of classifiers before and after feature selection is discussed in table 5.

Table 5. Accuracy comparison of classifiers (before and after fs algorithm)

<table>
<thead>
<tr>
<th>DATASET</th>
<th>Classifiers</th>
<th>Feature Selection Accuracy (%)</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE116250</td>
<td>GNN+SVM</td>
<td>75,15</td>
<td>84,47</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>78,58</td>
<td>86,93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>80,47</td>
<td>88,49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>85,18</td>
<td>91,06</td>
<td></td>
</tr>
<tr>
<td>GSE57344</td>
<td>GNN+SVM</td>
<td>77,17</td>
<td>86,42</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>80,38</td>
<td>88,32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>82,66</td>
<td>91,55</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>84,49</td>
<td>93,44</td>
<td></td>
</tr>
<tr>
<td>GSE71613</td>
<td>GNN+SVM</td>
<td>71,88</td>
<td>80,18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>73,49</td>
<td>83,65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>76,77</td>
<td>86,48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>81,19</td>
<td>89,72</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Comparative performance of the suggested classifiers with existing methods

<table>
<thead>
<tr>
<th>DATASET</th>
<th>Classifiers</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F - Measure (%)</th>
<th>Accuracy (%)</th>
<th>Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSE116250</td>
<td>GNN+SVM</td>
<td>82,66</td>
<td>80,04</td>
<td>81,35</td>
<td>84,47</td>
<td>15,53</td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>84,95</td>
<td>82,19</td>
<td>83,57</td>
<td>86,93</td>
<td>13,07</td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>86,44</td>
<td>84,82</td>
<td>85,63</td>
<td>88,49</td>
<td>11,51</td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>90,21</td>
<td>86,29</td>
<td>88,25</td>
<td>91,06</td>
<td>8,94</td>
</tr>
<tr>
<td>GSE57344</td>
<td>GNN+SVM</td>
<td>80,19</td>
<td>82,27</td>
<td>81,23</td>
<td>86,42</td>
<td>13,58</td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>82,26</td>
<td>84,46</td>
<td>83,36</td>
<td>88,32</td>
<td>11,68</td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>84,39</td>
<td>86,59</td>
<td>85,49</td>
<td>91,55</td>
<td>8,45</td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>86,61</td>
<td>88,29</td>
<td>87,45</td>
<td>93,44</td>
<td>6,56</td>
</tr>
<tr>
<td>GSE71613</td>
<td>GNN+SVM</td>
<td>78,18</td>
<td>80,74</td>
<td>79,46</td>
<td>80,18</td>
<td>19,82</td>
</tr>
<tr>
<td></td>
<td>GADNN</td>
<td>81,11</td>
<td>83,33</td>
<td>82,22</td>
<td>83,65</td>
<td>16,35</td>
</tr>
<tr>
<td></td>
<td>BADNN</td>
<td>83,45</td>
<td>85,87</td>
<td>84,66</td>
<td>86,48</td>
<td>13,52</td>
</tr>
<tr>
<td></td>
<td>EPOA+DAE</td>
<td>85,72</td>
<td>86,64</td>
<td>86,18</td>
<td>89,72</td>
<td>10,28</td>
</tr>
</tbody>
</table>

https://doi.org/10.56294/sctconf2024761
Figure 3 presents the precision comparison of classifiers like GNN+SVM, GADNN, BADNN, and proposed classifier (EPOA+DAE) with respect to gene datasets (GSE116250, GSE57344, and GSE71613). EPOA+DAE classifier has produced the highest precision results of 90.21%, 86.61%, and 85.72% (GSE116250, GSE57344, and GSE71613). GNN+SVM, GADNN, BADNN has lowest precision of 82.66%, 84.95%, and 86.44% for GSE116250 dataset (Refer Table 6).

Figure 4. Recall comparison vs. Classifiers

Figure 5. F-measure comparison vs. Classifiers

https://doi.org/10.56294/sctconf2024761
Recall comparison of classifiers like GNN+SVM, GADNN, BADNN, and proposed classifier (EPOA+DAE) with respect to gene datasets (GSE116250, GSE57344, and GSE71613) has been illustrated in figure 4. EPOA+DAE classifier has produces highest recall results of 86.29 %, 88.29 %, and 86.64 % (GSE116250, GSE57344, and GSE71613). GNN+SVM, GADNN, BADNN has lowest recall of 80.04 %, 82.19 %, and 84.82 % for GSE116250 dataset (Refer Table 6).

F-measure comparison of classifiers like GNN+SVM, GADNN, BADNN, and proposed classifier (EPOA+DAE) by gene datasets (GSE116250, GSE57344, and GSE71613) are illustrated in figure 5. EPOA+DAE classifier has produced highest F-measure results of 88.25 %, 87.45 %, and 86.18 % (GSE116250, GSE57344, and GSE71613). GNN+SVM, GADNN, BADNN has lowest F-measure of 81.35 %, 83.57 %, and 85.63 % for GSE116250 dataset (Refer Table 6).

![Figure 6](https://example.com/figure6.png)

**Figure 6.** Accuracy comparison vs. Classifiers

Accuracy comparison of classifiers like GNN+SVM, GADNN, BADNN, and proposed classifier (EPOA+DAE) by gene datasets (GSE116250, GSE57344, and GSE71613) has been illustrated in figure 6. EPOA+DAE classifier has produces highest accuracy results of 91.06 %, 93.44 %, and 89.72 % (GSE116250, GSE57344, and GSE71613). GNN+SVM, GADNN, BADNN has lowest accuracy of 84.47 %, 86.93 %, and 88.49 % for GSE116250 dataset (Refer Table 6).

![Figure 7](https://example.com/figure7.png)

**Figure 7.** Error rate comparison vs. Classifiers

Error rate comparison of classifiers like GNN+SVM, GADNN, BADNN, and proposed classifier (EPOA+DAE) of gene datasets (GSE116250, GSE57344, and GSE71613) are illustrated in figure 7. EPOA+DAE classifier has produces lowest error rate results of 8.94 %, 6.56 %, and 10.28 % (GSE116250, GSE57344, and GSE71613). GNN+SVM, GADNN, BADNN has highest error rate of 15.53 %, 13.07 %, and 11.51 % for GSE116250 dataset (Refer Table 6).

**CONCLUSION AND FUTURE WORK**

This study, EPOA, a novel stochastic nature-inspired optimization approach is introduced for feature selection. EPOA is a simulation of pelicans’ normal hunting activity. In POA, pelicans serve as search agents...
by seeking food sources. In the EPOA, fitness is computed depending on the entropy range of the feature in the gene dataset. To identify the primary optimum region of the search space, POA exploration is employed. Then Deep Autoencoder (DAE) classifier is used to classify the samples into normal / DCM for approximating the behaviour of the data. While the EPOA approach eliminated noisy genes, the outcomes were found to be excellent. With three datasets, the suggested technique was seen to pick the fewest number of genes during the wrapper process. Additionally, the empirical analysis findings showed that EPOA+DAE functioned better than other suggested algorithms, with a prediction accuracy of 91.06 %, 93.44 %, and 89.72 % (GSE116250, GSE57344, and GSE71613). The utility, adaptability, and applicability of feature selection in the DAE domain have been successfully proved in this research. Future clinical use of DCM-related genes is used as targets for medicine and diagnostic indicators.

REFERENCES


FINANCING
"The authors did not receive financing for the development of this research".

CONFLICT OF INTEREST
"The authors declare that there is no conflict of interest".

https://doi.org/10.56294/sctconf2024761
AUTHORSHIP CONTRIBUTION

Conceptualization: T. Sangeetha.
Data curation: K. Manikandan.
Formal analysis: D. Victor Arokia Doss.
Research: T. Sangeetha.
Methodology: D. Victor Arokia Doss.
Drafting - original draft: K. Manikandan.
Writing - proofreading and editing: T. Sangeetha.