https://doi.org/ 10.33472/AFJBS.6.4.2024.429-445

AFJBS

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Automatic Biological Data Analysis for Breast Cancer Detection and Classification

S. Nathiya[1]*

[1]PhD Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu-641049, India

Email: nathiyasri24@gmail.com

Dr. J. Sumitha[2]

[2]Assistant Professor, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamil Nadu-641049, India

Article History

Volume 6,Issue 4, Feb 2024

Received:17 Feb 2024

Accepted : 01 Mar 2024

doi: 10.33472/AFJBS.6.4.2024.429-445

Abstract: Biological data such as genomic, transcriptomics, proteomic, metabolomics, or clinical data are used for analyzing the patterns in breast cancer. Early diagnosis and identification are vital in precise therapy for the adverse breast cancer effects. Recent studies have focused on utilizing Deep Learning (DL) algorithms like deep convolutional neural Networks (CNN) architectures for breast cancer classification tasks with high efficiency from gene expression profiles. However, the biological gene expression datasets have smaller sample sizes but high dimensionality problems, which reduce the suitability of the DL methods. Similarly, the problem of model complexity in CNN-based methods is also challenging. Therefore, this paper presents a lightweight hybrid classifier-based breast cancer detection and classification model called Convolutional Support Vector Machine (CSVM), developed by integrating the benefits of the CNN architecture with the Hybrid Kernel-based SVM classifier. This hybridized CSVM classifier model is formulated by replacing the softmax classifier in the heterogeneous CNN architecture with the SVMs to handle the high dimensional features of gene expression datasets efficiently. After pre-processing the data, they are initially clustered using K-means Clustering to improve the learning of patterns and relationships between the disease features. Then, the features are learned using the convolutional layer and the final classification by SVMs. This proposed CSVM model parameters are trained together to improve the sequence-level feature learning. Experimented on biological datasets related to breast cancer gene (BRCA) sourced from Mendeley data, the efficiency of the proposed CSVM-based model is validated by overcoming the model complexity issues and achieved 97.89%, 96.85%, and 98.11% accuracies for breast cancer detection with minimized processing time.

Keywords: Biological Data, Gene expression analysis, Breast cancer, Convolutional Neural Networks, Convolutional Support Vector Machine, K-means Clustering, Mendeley Data.

1. INTRODUCTION

Biological data analysis involves the application of various computational and statistical techniques to interpret large-scale biological datasets with the goal of identifying patterns, biomarkers, or signatures associated with breast cancer [1]. Initially, the biological data, such as genomic and clinical data are collected. These biological data are obtained from patient samples, and cell lines. These raw datasets are cleaned and pre-processed by applying quality control, data imputation, batch correction and denoising methods to remove noise, correct errors, and normalize the data for ensuring consistency. Then, the relevant features from these pre-processed biological data are extracted and the dimensionality is reduced using suitable statistical or feature selection techniques. Finally, the statistical tests or ML algorithms are used to analyze the processed data for finding the potential biomarkers, genetic variants, gene expression signatures, or other molecular features associated with disease presence, progression, or response to treatment. These patterns are used to exactly identify the stage of the disease and then initiate the treatment process. Breast cancer is a tumour form of cancer caused due to abnormal lumps in tissues that are characterized by complicated metabolic and immune system abnormality, which involves complex interactions between various biological processes in the breast tissues. Mostly, women in the age of 40-plus category are often affected by breast cancer. In 2020, a total of 19.3 million cancer cases were diagnosed worldwide, with breast cancer accounting for 11.7% of these cases. The most common symptoms are the presence of breast lump, breast pain or discomfort, differences in the size of the breast or shape, skin changes colour of the breast, breast and under-arm swelling, and abnormal nipple discharge. Sometimes, itching and burning sensations may also be experienced in the breast or at the site of the lump.

One of the main reasons for the increased number of breast cancer patients in recent years is the significant lifestyle changes with the consumption of highly saturated fats taking over healthy food habits. Menstruation problems and obesity, especially after menopause, lactation problems in feeding mothers, family history and stress can also be major factors contributing to breast cancer. It is often identified by self-examinations or physical examinations by physicians followed by biopsy, histopathology examinations, X-ray Mammography, Ultrasonography, Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Computed Tomography (CT) scan, Thermography and Hormone Receptor testing. Once the diagnosis is accurately completed, the treatment process is initiated through medications and surgery with optional additional treatments in the forms of chemotherapy, hormone therapy or radiation [2]. Although the above diagnosis methods are universally acclaimed, most are intensive, costly, and complex. In the field of medical, the decision support system for clinical analysis by using big data and AI methods has been largely developed for disease detection in recent years. These decision support systems utilize clinical data, including scan images. Gene expression analysis is one of the most important and efficient clinical data analysis methods that utilize the gene expression profiles of patients in diagnosing diseases. Due to the invading cancer cells, the patient's genomic data is characterized by the analysis of biomarkers and genetic information related to the specific mutation, variations or patterns used for the detection of breast cancer. Based on the genetic information, the risk analysis and the possible treatment can be suggested [3]. Understanding the extracted genes and effectively distinguishing them from the normal cells for disease diagnosis plays an important role in the detection. Data mining and big data analysis techniques enhance the gene expression profiles and help in identifying breast cancer in patients distinctively.

The ML and DL methods have many advanced techniques for data mining applications [4]. The Clinical decision support systems have higher effectiveness in learning disease patterns from clinical data using large-scale ML and DL methods. To analyze complex datasets such as gene expression datasets, traditional data mining methods were limited and time-consuming. For breast cancer detection, traditional statistical measure-based data mining methods reduce the effective detection method. By utilizing the advanced feature learning strategies, the limitations of ML and DL can be overcome [5]. To extract the specific features and for predictive learning, ML algorithms are easier to train on the gene expression dataset to aid in breast cancer detection. Due to the shallow learning property, the ML algorithms fail to learn the hidden patterns and extract the long-term correlations within the genomic data. The DL algorithms are used to upgrade the drawback of the shallow learning property and upgrade the ML algorithm by using deep feature learning. To detect breast cancer from the images and genomic datasets, DL algorithms such as CNN and RNN have been used previously [6]. These algorithms topped at learning the features and hidden complex patterns from gene expression profiles. Hence, this research aims to detect breast cancer by using advanced DL gene expression data analysis. For the DL algorithms, a larger amount of dataset is required for training and learning sufficient features which is difficult in smaller datasets. Compared to ML algorithms, DL algorithms provide highly effective results despite minor disadvantages. Developing a lightweight DL model can help to overcome these limitations and acquire accurate results.

This paper has presented a hybrid classifier lightweight Convolutional Support Vector Machine (CSVM) for breast cancer prediction and classification method by integrating the SVM within the CNN architecture. The model used CNN's Convolution layer for feature learning, and SVM is used in the classification layer by replacing the softmax classifier function. CNN has the better feature extraction capability in breast cancer diagnosis and classification; it can enhance the visibility of malignancy in breast cancer profiles and add in the early treatment before further progression. For handling the dimensionality in gene expression, the softmax classifier in the final layer of the standard CNN architecture has limitations. Also, the softmax classifier is highly sensitive to input values and can cause over-fitting issues. Therefore, in the proposed model, the softmax classifier is replaced with SVMs with a hybrid kernel to improve high-dimensional data classification. The proposed model initially utilizes the K-means clustering algorithm in the pre-processing stage to cluster the gene expression data. K-means clustering is utilized to enhance learning patterns and relationships among disease features. Hence, it clusters the genes within the breast cancer dataset by identifying potential biomarkers to be incorporated into the proposed model for accurate prediction. In the CSVM method, both the parameters of SVM and CNN are jointly trained instead of training separately. This improves the sequence-level feature learning. In the presented method, the evaluation is performed grounded on the BRCA datasets from the Mendeley data. The paper is structured as: literature survey works in part 2. The presented CSVM-based breast cancer detection model is explained in part 3; the detailed analyzed results are in part 4, while part 5 presents a summary of the proposed research with suggestions for future studies.

2. RELATED WORKS

Algorithms such as Decision tree [7], SVM [8] and ANN were the baseline methods for breast cancer detection from various biological datasets utilized by many researchers recently before the introduction of DL methods. Elbashir [9] presented a method of lightweight CNN architecture for breast cancer prediction. This method pre-processes gene expression data and transforms it into a 2D image. Then, the outlier removal was done using the Array-Array Intensity Correlation (AAIC) technique, and CNN was used for the classification process. By using the RNA-seq gene expression data, F-Score, Accuracy, Precision, sensitivity and specificity of 0.955, 98.76%, 100%, 91.43% and 100%, respectively, were obtained. However, applying CNNs to gene expression data increased the computational demands. Jazayeri and Sajedi [10] proposed a Non-negative Matrix Factorization (NMF) and an Extreme Learning Machine (ELM) algorithm for classifying breast cancer. This method combined NMF with column splitting for dimension reduction, and ELM was used for the classification process. Experimented on the NCBI dataset, this model reduced the classification error rate by 2.7%, but it has problems handling feature redundancy, noises and irrelevant data. Arya and Saha [11] suggested a two-stage stacked ensemble framework for predicting breast cancer, with CNN used for extracting the features in the first stage and a stacked ensemble model using these features for final classification in the second stage. Tested on a multi-model dataset and obtained a 90.2% accuracy and 0.93 AUC value. However, the CNN used in this model increased the complexity when stacked as an ensemble. Jia [12] proposed a DL-based model for the detection of breast cancer with gene selection using Weighted Gene expression network Analysis (WGCNA) and Differential Expression Analysis (DEA). The 23 genes were screened using Protein-Protein Interaction (PPI) and utilized different classifiers. ANN performed better with average accuracy, F1 value, sensitivity, specificity, and AUC values of 97.36%, 0.8535, 98.32%, 89.59%, and 0.99 for GSE15852 and TCGA-BC datasets.

Lamba [13] presented a DNN-based classification for cancer in the breast. In this method, minority class balancing was performed using the SMOTE algorithm and BFS Best-First Search was used for the selection of features and CFS before classifying using DNN. This model achieved 93% accuracy for GSE15852 datasets but has also suffered from over-fitting issues due to a smaller sample size. Cheng [14] developed a DNN-based breast cancer detection model and combined ensemble learning with Systems biology feature selection methods. This model obtained AUC values of 0.7677 and 0.7836 between genes and clinical features and a concordance index (CI) of 0.6683 for the METABRIC dataset. Liu [15] proposed a hybrid DNN for predicting breast cancer based on multi-modal data that combines the gene model data with the image model data. The feature extraction network works based on weighted linear aggregation to improve the DNN performance in this method. This hybrid model obtained 88.07% accuracy for the TCGA-BRCA dataset but suffers from a high processing time of 40 minutes. Alromema [16] introduced a sequential model by combining minimal Redundancy-Maximum Relevance (mRMR), a two-tailed unpaired t-test, and meta-heuristics for hybrid Feature Selection (FS). This framework predictor selects the biomarkers gene for the ML classifiers. Evaluated on GSE22820, the XGBoost-based model with this hybrid FS framework achieves higher 0.976 accuracy, 0.974 of F1-Score, and 0.961 of AUC values, respectively. However, this FS framework requires large memory capacity and is time-consuming.

Kayikci and Khoshgoftaar [17] proposed an attention-based deep learning model in multi-modal for breast cancer prediction. Initially, the features which are stacked are created using attention on sigmoid-gated CNN, and then, the flattened, dropout and dense processes are used for bi-modal attention. Experimented with multi-modal data combining the METABRIC and TCGA-BRCA datasets, the model gained a 0.95 AUC, accuracy of 0.912, precision of 0.841, and sensitivity of 0.798. However, this model increased the complexity of handling the multi-modal data. Mustafa [18] presented an ensemble model using multi-modal data and multiple neural networks for breast cancer survivability prediction. Here, CNN is used for clinical modalities. To handle data in multi-dimensional data and modalities in gene expression, LSTM is utilized and DNN is used for CNV effectively. This model obtained 98% accuracy, 99% F1-score, 98% precision, and 100% sensitivity for the METABRIC dataset, but the memory complexity is higher than other DL-based methods. Wang and Lee [19] proposed deep auto-encoders and K-means clustering for detecting the sub-groups of breast cancer. In this model, the deep auto-encoders extracted the latent features and are used in K-Means clustering to detect the two forecasting subgroups, namely BPS-LumA and WPS-LumA. The deep auto-encoders obtained MSE of 0.02 and 0.075 for METABRIC and TCGA datasets. However, training these deep auto-encoders was computationally expensive. Mohamed [20] proposed an Ebola optimization search (EOSA)-based CNN model used to diagnose cancer in the breast. Evaluations were performed using the TCGA dataset, and this model obtained accuracy, precision, recall, f1-score, kappa, specificity, and sensitivity values of 98.3%, 99%, 99%, 99%, 90.3%, 92.8%, and 98.9%, respectively. However, the imbalanced data issue has greatly reduced this EOSA-CNN model's performance. Table 1 summarizes the performance of the literature methods over different BRCA datasets.

Table.1. Comparison of Literature Methods

Methods	**Dataset**	**Accuracy (%)**	**Precision (%)**	**F-measure (%)**
SVM [8]	METABRIC	89.59	-	-
CNN [9]	TCGA	98.76	100	95.5
NMF-ELM [10]	GSE32393	99.28	-	-
Two-stage Stacked Ensemble [11]	METABRIC	90.02	84.01	-
	TCGA	88.01	94.9	-
ANN-WGCNA-DEA [12]	GSE15852	97.36	-	85.35
	TCGA	97.36	-	85.35
DNN [13]	GSE15852	93	-	-

DNN ensemble [14]	METABRIC	71.79	-	-
Hybrid DNN [15]	TCGA	88.07	-	-
XGBoost-Hybrid FS [16]	GSE22820	97.06	-	96.01
Sigmoid gated attention CNN[17]	GSE15852	91.02	84.01	-
	BC-TCGA	91.02	84.01	-
Ensemble DL [18]	METABRIC	98	98	99
Deep Autoencoder-K-means [19]	METABRIC	92.05	-	-
	TCGA	98	-	-
EOSA-CNN [20]	TCGA	98.3	99	90.3

The literature study shows that the recent advanced methods developed for breast cancer detection have significantly improved the prediction performance. However, some limitations are still needed to be considered. The results obtained by these models are obtained for different BRCA datasets. Comparing their performance using these results will be unfair since a method can work better for a dataset while underperforming for another dataset. The smaller sample size and high dimensionality of the gene expression datasets have significantly reduced the performance of ML and DL methods. Similarly, the complexity issues in DL-based methods are also a challenging concern. Therefore, a lightweight hybrid classifier-based breast cancer detection and classification model called CSVM has been developed in this paper.

3. METHODOLOGY

The proposed CSVM-based breast cancer detection model is illustrated in Fig. 1. Initially, the standard pre-processing techniques are applied on the input biological datasets, particularly for outlier removal. Then, these biological data are clustered using the K-means algorithm to improve the pattern-learning process. Finally, these clustered data are fed to the CSVM classifier, which learns the gene features and classifies the data into respective classes to identify the breast cancer samples.

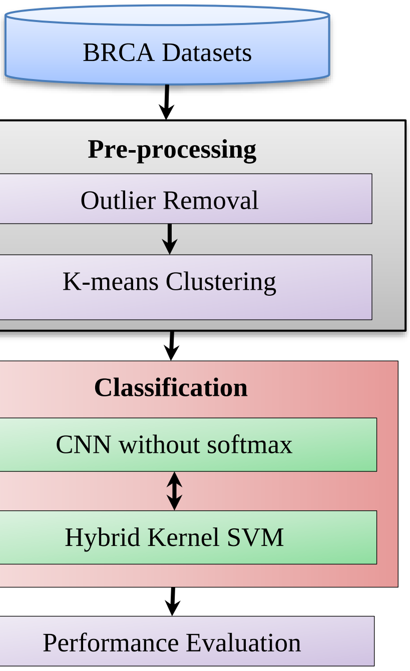

Fig.1. Overall Workflow of the Proposed Model

3.1. Datasets

The publically available benchmark biological datasets for BRCA are collected from Mendeley Data [21]. The three biological datasets, BC-TCGA, GSE2034, and GSE25066, are used for evaluation. BC-TCGA consists of 61 normal samples, 529 breast cancer samples, and a total of 590 samples. GSE25066 contains 100 pathologic complete response (PCR) samples and 392 residual diseases (RD) samples among the 492 total samples. GSE2034 contains 286 samples (107 recurrences and 179 no-recurrence samples). The number of genes in BC-TCGA, GSE2034, and GSE25066 are 17814, 12634 and 12634, respectively. Table 2 illustrates the distribution of these biological datasets.

Table.2. BRCA Gene Expression Data Distribution

Datasets	**Genes count**	**Samples count**		
		Overall	**Healthy Class**	**Tumor Class**
BC-TCGA	17814	590	61	529
GSE2034	12634	286	179	107
GSE25066	12634	492	100	392

3.2. Pre-processing Stage

The pre-processing stage performs two vital tasks: outlier removal and initial data clustering. The BRCA biological data consists of outlier data that deviate largely from the other ranges. Hence, an outlier removal technique using a standard Z-score is applied to refine the input data. The refined data are clustered using the k-means algorithm to obtain similar genes for feature representation from the raw data.

Outlier removal using Z-score: To estimate the standard deviations of data from the mean of the data group Z-score, a statistical measure is used. In BRCA gene expression datasets, the Z-score can be used to estimate and remove the outliers [22]. The Z-score is computed for each data point.

$$Z_i = \frac{x_i - mean}{standard\ deviation} \tag{1}$$

Here, Z_i is the Z-score representation between (-3, 3) and x_i denotes the data point. A threshold value of Z-score is determined beyond which the data point will be considered an outlier. The threshold can be positive or negative based on the used dataset. To indicate the data point following mean Z-score is used; the mean above the data point is considered a negative Z-Score, and the mean below the data point is considered a positive Z-score. This study sets the threshold as 2 as the BRCA datasets are not normally distributed. The data points above this threshold are considered outliers, and the decision to remove them is based on the number of outliers and their impact on the problem objectives.

K-means algorithm for Data clustering: K-means clustering is employed to group the data samples according to the similar expression values of different genes. The clustering of expression profiles is used for grouping both genes and samples and can also be used to promote new markers in specific types of cells along with the recognition of tumor sub-types [23]. Using k-means reduces the dimensionality of the samples and produces tighter clusters, especially if the clusters are globular with faster processing. The data is characterized by its expression levels across the samples, resulting in an n-dimensional numerical vector for each gene. K number of clusters is chosen as a user-defined parameter. The cluster centroids are randomly assigned in the gene expression space. For each sample, the Euclidean distance is calculated between the sample and the centroid. The centroid position for each cluster is updated by calculating the mean expression values for genes, and the algorithm then iteratively assigns each sample to the nearest centroid. The iteration is continuous until the centroid reaches the convergence point or until there is no significant change can be made for the centroid. The k-means algorithm is mathematically expressed by considering, $X_1, \dots., X_n$ as the dimensional point set–d into the K clusters that are to be clustered. Let $C^{(k)} = \{C_k, k = 1, \dots, K\}$ is the K clusters partition and μ_k is denoted to be the mean clusterC_k:

$$\mu_k \coloneqq \frac{1}{|C_k|} \sum_{i \in C_k} X_i \tag{2}$$

Here, the cluster k cardinality is denoted as $|C_k|$. The k-means approach aims to diminish the addition of the squared errors (SSE) for clusters $C^{(k)}$ of each set as

$$SSE\left(C^{(k)}\right) \coloneqq \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - \mu_k\|_2^2 \tag{3}$$

Here, $i \in C_k$ if $\|X_i - \mu_k\|_2 = min_{k'} \|X_i - \mu_{k'}\|_2$. By using the selection method, the cluster numbers for the k-means are calculated. This is a condemnation of criteria penalized on asymptotic for all positive integer K.

$$crit(K) \coloneqq \sum_{k=1}^{K} \sum_{i \in C_k} \|h(X_i) - \mu_{h,k}\|_2^2 + pen\ (K) \tag{4}$$

$$crit(K) = \sum_{i=1}^{n} \min_{k=1,\dots,K} \|h(X_i) - \mu_{h,k}\|_2^2 + \quad pen\ (K) \tag{5}$$

Here, for compositional data, transformation is denoted as h. $pen: N \to R_+$is the function of penalty is defined by

$$pen\ (K) \coloneqq a_h \sqrt{Knd} \tag{6}$$

Until the multiplication factors are constant the term of penalty is determined. Therefore, the selected cluster number is given by

$$\widehat{K} \coloneqq arg \min_{K \le n} crit(K) \tag{7}$$

Thus, k-means clustering for gene expression data analysis is used to discover patterns and relationships among genes based on their expression profiles, effectively enhancing breast cancer classification.

3.3. Classification using the CSVM Method

The proposed breast cancer detection model utilizes the Convolutional Support Vector Machine (CSVM) method to identify and classify cancer in breast samples. Different CSVM models have been developed in recent years [24], but this study develops a CSVM model using a hybrid kernel. In this model, the CNN and hybrid kernel SVM classifiers are integrated by removing the softmax classifier in the last layer of the CNN model and employing the SVM in its place. Fig.2 shows the architecture of the proposed CSVM model.

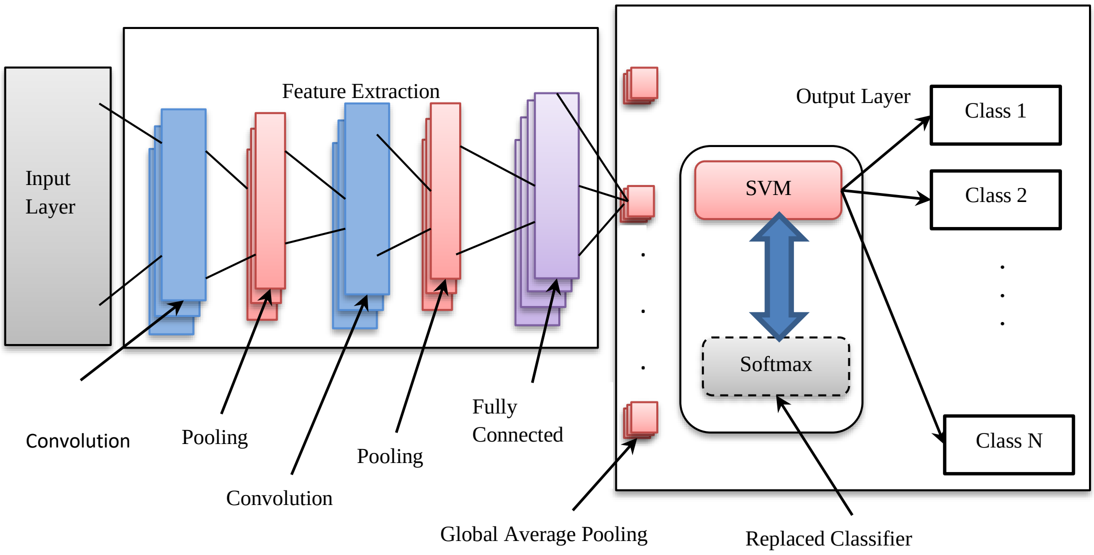

Fig.2. Proposed CSVM Architecture

In this proposed model, the CLs are used to identify the hidden features within the gene expression data, and these features are subsequently employed in SVM for the effective classification of cancer in breast and non-cancer samples. The significant advantage of the CL lies in its capability to learn the features that remain consistent despite translation, rotation, and shifting. A typical CNN comprises an input layer and numerous CLs linked together through pooling and output layers. The CL extracts the features by learning the characteristics of the samples within the expression data of the gene. The pooling layer focuses on the most informative data from the gene expression while discarding redundant data after the extracted features. For each convolution layer, the convolution operation is defined as

$$h_{ij}^{k} = f((W^{k} * x)_{ij} + b_{k}) \tag{8}$$

Here, the activation function is defined as f, W^{k} is the weight of the feature k^{th} map, and bias b_k for k^{th} map.

SVM possesses good non-linear mapping and linear regression can be performed in the feature space, which can also take high-dimensionality feature space to map the data. The regression is denoted as

$$f(x) = w^{T}\varphi(x) + b \tag{9}$$

Here, φ is defined as the non-linear function of mapping, w is for weight, and b refers to the bias, respectively. Therefore, optimizing SVM is performed as follows:

$$F_{SVM} = \min_{w,b,\xi,\xi^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{10}$$

$$s.t. \begin{cases} y_i - (\langle w, x_i \rangle + b) \leq \varepsilon + \xi_i \\ (\langle w, x_i \rangle + b) - y_i \leq \varepsilon + \xi_i^* \ (i = 1,2, \dots l) \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Here, F_{SVM} denotes the optimization function of the SVM, l is sample numbers, the input x_i and output y_i for the training data, n is the sample number, the upper and lower for training error are denoted as ξ, and ξ_i^*; ε is the constant for regularization, and C is said to be the loss-insensitive factor. The prediction function for this CSVM model is estimated as

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*)\, K(x, x_i) + b \tag{11}$$

Here, a_i, a_i^* represents the Lagrange multipliers, and $K(x, x_i)$ denotes the function of a kernel for SVM. The major Kernel functions are linear, polynomial, radial basis function (RBF) and sigmoid functions. These kernels are formulated for input (x, x_i) as follows:

Sigmoid kernel:

$$K_S(x, x_i) = tanh\ (\eta \times (x, x_i) + \delta) \tag{12}$$

Polynomial kernel:

$$K_P(x, x_i) = (\eta \times (x.x_i) + \delta)^d \tag{13}$$

Gaussian kernel:

$$K_G(x, x_i) = exp\left(-\frac{\|x - x_i\|^2}{2 \times \sigma^2}\right) \tag{14}$$

Here, σ is the standard deviation, and η, δ and d are the kernel parameters. These parameters determine the filter size to be used in the classifier model. This study aims to improve SVM's strong learning ability and generalization performance by developing a hybrid kernel. Therefore, the hybrid kernel called the Sigmoid-Polynomial-Gaussian (SPG) kernel is formulated as

$$K_{SPG} = \beta_1 .exp\ \left(-\beta_2\ tanh\ \left(\frac{(\eta \times (x.x_i) + \delta)^d}{2 \times \sigma^2}\right)\right) + \beta_3 .(x.x_i) \tag{15}$$

Here, $\beta = [\ \beta_1, \beta_2, \beta_3]$ is a vector with $\beta_1 + \beta_2 + \beta_3 = 1$ and $\eta, \delta, d > 0$.

The difference between the softmax and the SVM lies in the parameterized function for the weighted matrices w. SVM aim to increase the separation circumference data points belonging to different classes, whereas the softmax classifier focuses on maximizing log-likelihood and minimizing the cross-entropy.

The proposed CSVM model comprises three CLs and two fully connected layers (FCL). The CL bundles the pooling layers and batch normalization function with an optional dropout function. The foremost function of the input layer is to load the data and produce an output vector that serves as the input for the convolutional layer. The convolutional layer follows the input layer. To extract various features, different convolutional filters or kernels can be employed. The filters in the CL are 4, 8 and 16. A crucial feature of the convolutional operation is its ability to enhance the original features while reducing noise. The activation function Rectified Linear Unit (ReLU) and the Adam Optimization algorithm are employed for faster learning. The max-pooling is applied with a pool size of 2 and batch normalization [0,1]. The optional dropout is set at 5% to 20% whenever the feature learning process slows.

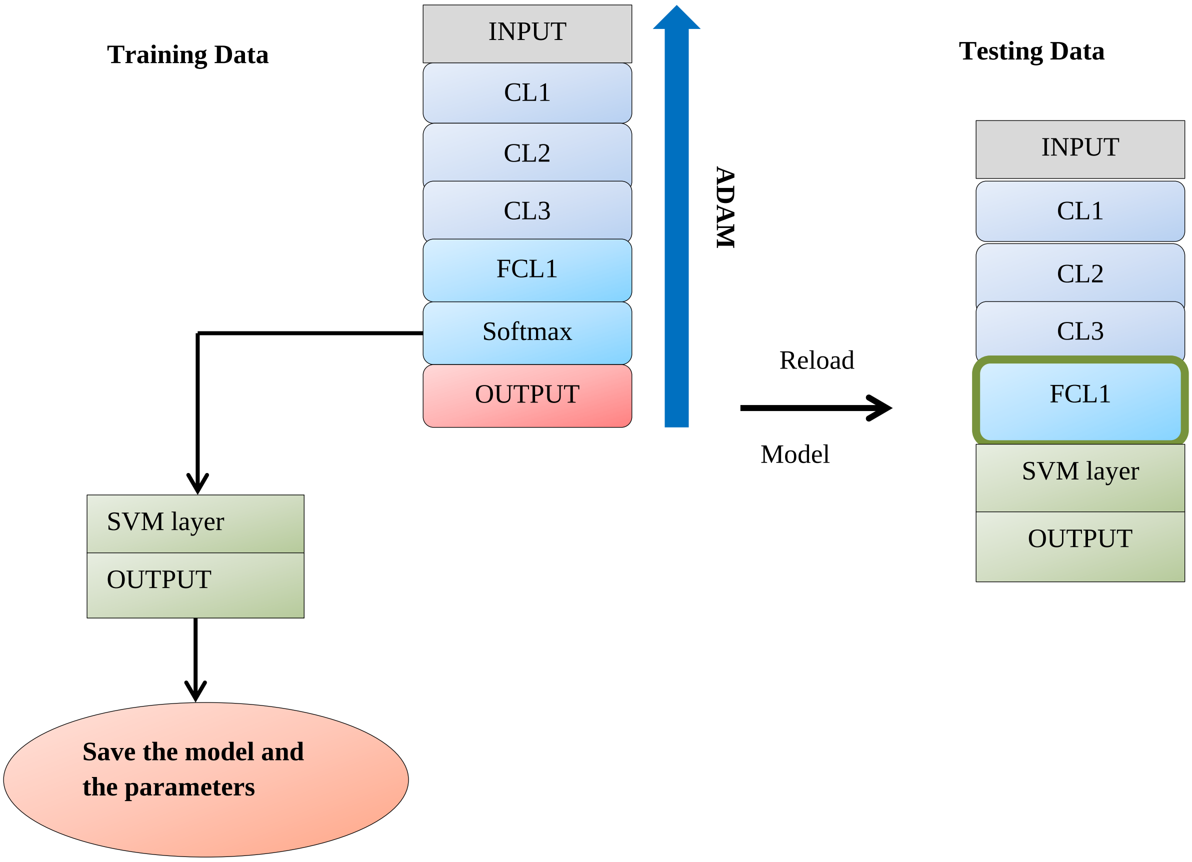

Fig.3. Computational Process of CSVM

The model employs six activation functions, namely, f_1, f_2, f_3, f_4, f_5 and f_6. The function f_1 transfers the input to the CL1, f_2 transfers the CL1 to the CL2, f_3 transfers the CL2 to the CL3, f_4 transfers the CL3 to the FCL1, f_5 transfers the FCL1 to the FCL2 and finally, the function f_6 transfers the FCL2 (softmax) to the output layer. The proposed model takes features as the input for the SVM, which is built and trained by extracting and loading the FCL1. This model is saved and reloaded, and the model for testing is built. The FCL1 is directly associated with the SVM layer, and the parameters are shared. Here, φ , the non-linear mapping function, transfers the FCL1 to the SVM layer instead of the FCL2 containing the softmax classifier. Fig.3 illustrates the computational process of CSVM for the BRCA datasets.

4. RESULT AND DISCUSSION

The suggested CSVM-based breast cancer detection model is evaluated using benchmark biological gene expression datasets for BRCA from Mendeley Data. The implementations use an i5 processor in Intel of Windows 10 OS in a controlled environment along with the MATLAB tool (R2021a), with RAM of 8GB and SSD of 512GB. The evaluation parameters are accuracy, Precision, Recall, F-Measure, and Processing Time. The results obtained in Table 1 were obtained for different BRCA datasets with different numbers of samples and features. Comparing the performance using these results will be unfair since a method in literature can work better for one biological dataset while underperforming for another biological dataset. Therefore, the methods in the literature are also implemented similarly to the proposed CSVM model using the benchmark biological datasets for BRCA and enforced in the same environment as the presented model to ensure fair comparisons. Table 3 demonstrates the performance comparisons of the proposed CSVM-based model against the existing methods for the BC-TCGA dataset.

Table.3. Performance comparison for BC-TCGA

Methods Used	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-measure (%)**	**Processing time (s)**
Decision Tree [7]	82.34	83.56	78.77	81.09	9.57
SVM [8]	89.59	86.54	89.91	88.19	12.62
CNN [9]	95.77	93.21	90.61	91.89	9.44
NMF-ELM [10]	91.83	88.94	85.17	87.01	12.93
Two-stage Stacked Ensemble [11]	92.25	89.10	84.28	86.62	14.35
ANN-WGCNA-DEA [12]	91.50	86.71	82.15	84.37	11.09
DNN [13]	92.18	90.45	87.36	88.88	10.67
DNN ensemble [14]	95.67	93.33	90.81	92.05	13.05
Hybrid DNN [15]	94.91	91.72	89.65	90.67	10.45
XGBoost-Hybrid FS [16]	94.45	92.13	93.56	92.84	11.98
Sigmoid gated attention CNN[17]	96.33	92.22	90.02	91.11	14.59
Ensemble DL [18]	94.76	90.35	88.76	89.55	9.92
Deep Autoencoder-K-means [19]	93.23	93.59	90.43	91.98	12.91
EOSA-CNN [20]	96.47	91.0	92.89	91.94	8.87
Proposed CSVM	**97.89**	**98.11**	**93.80**	**95.91**	**8.32**

The evaluation of the proposed CSVM-based classification model for cancer in the breast executes better when compared to other models for the BC-TCGA dataset. There have been accuracy improvements in the CSVM, approximately by 1.42%, 4.66%, 3.13%, 1.56%, 3.44%, 2.98%, 2.22%, 5.71%, 6.39%, 5.64%, 6.06%, 2.12%, 8.3%, and 15.55% higher than

EOSA-CNN, Deep Autoencoder-K-means, Ensemble DL, Sigmoid gated attention CNN, XGBoost-Hybrid FS, Hybrid DNN, DNN ensemble, DNN, ANN-WGCNA-DEA, Two-stage Stacked Ensemble, NMF-ELM, CNN, SVM, and Decision Tree methods, respectively. Similarly, better evaluation is obtained in terms of recall, F-measure and precision parameters. The time for processing the CSVM model is also less than the other methods. The performance comparisons made for the GSE2034 dataset are shown in Table 4.

Table.4. Performance comparison for GSE2034

Methods Used	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-measure (%)**	**Processing time (s)**
Decision Tree [7]	84.97	86.8	83.99	85.37	9.91
SVM [8]	89.21	88.64	84.38	86.46	14.52
CNN [9]	90.98	87.93	83.37	85.59	15.35
NMF-ELM [10]	93.76	92.74	93.09	92.92	9.35
Two-stage Stacked Ensemble [11]	90.89	90.62	91.54	91.08	8.17
ANN-WGCNA-DEA [12]	94.01	91.98	84.32	87.98	12.55
DNN [13]	90.87	89.58	83.17	86.26	8.78
DNN ensemble [14]	93.19	91.84	88.90	90.35	8.36
Hybrid DNN [15]	90.84	90.11	95.22	92.55	15.56
XGBoost-Hybrid FS [16]	91.54	94.97	88.81	91.79	9.46
Sigmoid gated attention CNN[17]	94.94	94.26	86.62	90.28	11.78
Ensemble DL [18]	94.69	88.19	94.06	91.03	13.88
Deep Autoencoder-K-means [19]	92.95	91.74	93.90	92.81	10.13
EOSA-CNN [20]	95.65	93.89	94.53	94.21	8.30
Proposed CSVM	**96.85**	**95.50**	**96.78**	**96.14**	**7.92**

The evaluation for the GSE2034 dataset also shows that the CSVM-based model performs best than the extant methods. CSVM model achieved accuracy of 96.85% which is 1.2%, 3.9%, 2.16%, 1.91%, 5.31%, 6.01%, 3.66%, 5.98%, 2.84%, 5.96%, 3.09%, 5.87%, 7.64%, and 11.88% higher than the EOSA-CNN, Deep Autoencoder-K-means, Ensemble DL, Sigmoid gated attention CNN, XGBoost-Hybrid FS, Hybrid DNN, DNN ensemble, DNN, ANN-WGCNA-DEA, Two-stage Stacked Ensemble, NMF-ELM, CNN, SVM, and Decision Tree methods, respectively. Similarly, better evaluation is obtained in terms of F-measure, recall precision and processing time. Similarly, the performance comparisons made for the GSE25066 dataset are exhibited in Table 5.

Table.5. Performance comparison for GSE25066

Methods Used	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-measure (%)**	**Processing time (s)**

Decision Tree [7]	83.83	84.41	80.9	82.62	12.52
SVM [8]	88.74	86.49	81.85	84.11	11.90
CNN [9]	91.35	88.5	82.65	85.48	11.17
NMF-ELM [10]	93.66	91.13	89.19	90.15	10.59
Two-stage Stacked Ensemble [11]	92.42	92.97	90.95	91.95	9.76
ANN-WGCNA-DEA [12]	94.91	88.16	83.83	85.94	12.22
DNN [13]	93.51	92.84	87.27	89.97	10.15
DNN ensemble [14]	92.95	92.23	85.63	88.81	12.50
Hybrid DNN [15]	92.8	91.03	89.3	90.16	12.97
XGBoost-Hybrid FS [16]	96.67	90.19	88.85	89.52	10.56
Sigmoid gated attention CNN[17]	89.91	95.23	84.88	89.76	13.98
Ensemble DL [18]	94.68	97.24	90.16	93.57	12.15
Deep Autoencoder-K-means [19]	95.9	92.58	93.17	92.87	11.37
EOSA-CNN [20]	97.32	95.79	92.85	94.30	12.54
Proposed CSVM	**98.11**	**97.67**	**94.48**	**96.05**	**9.31**

For GSE25066 dataset, the proposed CSVM model achieved 96.85% accuracy, which is 0.79%, 2.21%, 3.43%, 8.2%, 1.44%, 5.31%, 5.16%, 4.6%, 3.2%, 5.69%, 4.45%, 6.76%, 9.37%, and 14.28% higher than the EOSA-CNN, Deep Autoencoder-K-means, Ensemble DL, Sigmoid gated attention CNN, XGBoost-Hybrid FS, Hybrid DNN, DNN ensemble, DNN, ANN-WGCNA-DEA, Two-stage Stacked Ensemble, NMF-ELM, CNN, SVM, and Decision Tree methods, respectively. Similarly, better performance is obtained in terms of processing time, F-measure, recall and precision. These improved outcomes of the CSVM-based model improved convergence and advanced learning-based classification. It concludes that the proposed CSVM has better examined and obtained the accurate classification of cancer in the breast with less complexity on the biological gene expression data.

5. CONCLUSION

This research has developed an advanced hybrid classifier of CSVM with a hybrid kernel for improving the classification of the expression data in genes for better cancer in breast detection through biological data analysis. The proposed CSVM classifier overcomes the high computation and model complexity issues for dissecting the biological gene expression data. It is used along with the K-means data clustering approach and evaluated over biological datasets for BRCA from the Mendeley repository. The proposed CSVM-based breast cancer detection model obtained classification accuracies of 97.89%, 96.85%, and 98.11%, respectively, for BC-TCGA, GSE2034, and GSE25066 datasets, with less time for processing. Thus, the proposed model has achieved significantly improved performance for breast cancer detection, but still, there is room for improvement. The class imbalance problem and variability of expression levels of the biological datasets will be investigated in future research.

REFERENCES

1. Azim Jr, H. A., Michiels, S., Bedard, P. L., Singhal, S. K., Criscitiello, C., Ignatiadis, M., ... & Loi, S. (2012). Elucidating prognosis and biology of breast cancer arising in young women using gene expression profiling. *Clinical cancer research*, *18*(5), 1341-1351.
2. I. Mittra, G.A. Mishra, R.P. Dikshit, S. Gupta, V. Y. Kulkarni, H. K. A. Shaikh and R. A. Badwe, "Effect of screening by clinical breast examination on breast cancer incidence and mortality after 20 years: prospective, cluster randomized controlled trial in Mumbai". *Bmj*, vol. *372*, February 2021.
3. C. Horr, and S. A. Buechler, "Breast Cancer Consensus Subtypes: A system for subtyping breast cancer tumors based on gene expression". NPJ breast cancer, vol. 7(1), pp. 136, October 2021.
4. A. U. Mazlan, N. A Sahabudin, M. A. Remli, N. S. N. Ismail, M. S. Mohamad, H. W. Nies, and N. B. Abd Warif, "A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data". Processes, vol. 9(8), pp. 1466, August 2021.
5. M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu, and M. O. Agyeman, "A survey of machine learning approaches applied to gene expression analysis for cancer prediction". *IEEE Access, vol. 10, pp.* 27522-27534, January 2022
6. R. A. Dar, M. Rasool, and A. Assad, "Breast cancer detection using deep learning: Datasets, methods, and challenges ahead". *Computers in biology and medicine, vol. 149, pp.* 106073, October 2022.
7. J. Quist, H. Mirza, M. C. Cheang, M. L. Telli, J. A. O'Shaughnessy, C. J. Lord and A. Grigoriadis, "A four-gene decision tree signature classification of triple-negative breast cancer: implications for targeted therapeutics". *Molecular cancer therapeutics, vol.* 18(1), pp. 204-212, *January 2019.*
8. M. E. Ozer, P. O. Sarica, and K. Y. Arga, "New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines". *Omics: a journal of integrative biology, vol.* 24(5), pp. 241-246, May 2020.
9. M. K. Elbashir, M. Ezz, M. Mohammed, and S. S. Saloum," Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data". *IEEE Access*, vol. *7*, pp. 185338-185348, December 2019.
10. N. Jazayeri, and H. Sajedi, "Breast cancer diagnosis based on genomic data and extreme learning machine". *SN Applied Sciences*, vol. *2*, pp. 1-7, December 2020.
11. N. Arya, and S. Saha, "Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. *19*(2), pp. 1032-1041, August 2020.

12. D. Jia, C. Chen, C. Chen, F. Chen, N. Zhang, Z. Yan, and X. Lv, "Breast cancer case identification based on deep learning and bioinformatics analysis". *Frontiers in genetics*, vol. *12*, pp. 628136, May 2021.
13. M. Lamba, G. Munjal, and Y. Gigras, "A hybrid gene selection model for molecular breast cancer classification using a deep neural network". *International Journal of Applied Pattern Recognition*, vol. *6*(3), pp. 195-216, August 2021.
14. L. H. Cheng, T. C. Hsu, and C. C. Lin, "Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction". *Scientific Reports*, vol. *11*(1), pp. 14914, July 2021.
15. T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng, "A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multi-modal data". *Irbm*, vol. *43*(1), pp. 62-74, Feburary 2022.
16. N. Alromema, A. H. Syed, and T. Khan, "A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data". *Diagnostics*, vol. *13*(4), pp. 708, February 2023
17. S. Kayikci, and T. M. Khoshgoftaar, "Breast cancer prediction using gated attentive multi-modal deep learning". *Journal of Big Data*, vol. *10*(1), pp. 1-11, May 2023.
18. E. Mustafa, E. K. Jadoon, S. Khaliq-uz-Zaman, M. A. Humayun, and M. Maray, "An Ensembled Framework for Human Breast Cancer Survivability Prediction Using Deep Learning". *Diagnostics*, vol. *13*(10), pp. 1688, May 2023.
19. S. Wang, and D. Lee, "Identifying prognostic subgroups of luminal-A breast cancer using deep autoencoders and gene expressions". *PLOS Computational Biology*, vol. *19*(5), pp. e1011197, May 2023.
20. T. I. Mohamed, A. E. Ezugwu, J. V. Fonou-Dombeu, A. M. Ikotun, and M. Mohammed, "A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data". *Scientific Reports*, vol. *13*(1), pp. 14644, September 2023.
21. Xie, Haozhe; Li, Jie; Jatkoe, Tim; Hatzis, Christos, "Gene Expression Profiles of Breast Cancer", Mendeley Data, V1, doi: 10.17632/v3cc2p38hb.1, 2017.
22. N. B. Chikodili, M. D. Abdulmalik, O. A. Abisoye, and S. A. Bashir, "Outlier detection in multivariate time series data using a fusion of K-medoid, standardized Euclidean distance and Z-score". In *International Conference on Information and Communication Technology and Applications* (pp. 259-271). Cham: Springer International Publishing, November 2020.
23. R. Jothi, S. K. Mohanty, and A. Ojha, "DK-means: a deterministic k-means clustering algorithm for gene expression analysis". *Pattern Analysis and Applications*, vol. *22*, pp. 649-667, December 2019.
24. V. Passricha, and R. K. Aggarwal, "Convolutional support vector machines for speech recognition". *International Journal of Speech Technology*, vol. *22*, pp. 601-609, December 2019.