Breast Cancer Classification by Gene Expression Analysis using Hybrid Feature Selection and Hyper-heuristic Adaptive Universum Support Vector Machine

Original Scientific Paper

V. Murugesan
Department of Computer Science, VLB Janakiammal College of Arts and Science, Coimbatore,
Tamil Nadu- 641042, India
murugeshvlb2020@gmail.com

P. Balamurugan
Department of Computer Science, Government Arts College, Coimbatore,
Tamil Nadu-641018, India
spbalamurugan@rediffmail.com

Abstract *– Comprehensive assessments of the molecular characteristics of breast cancer from gene expression patterns can aid in the early identification and treatment of tumor patients. The enormous scale of gene expression data obtained through microarray sequencing increases the difficulty of training the classifier due to large-scale features. Selecting pivotal gene features can minimize high dimensionality and the classifier complexity with improved breast cancer detection accuracy. However, traditional filter and wrapper-based selection methods have scalability and adaptability issues in handling complex gene features. This paper presents a hybrid feature selection method of Mutual Information Maximization - Improved Moth Flame Optimization (MIM-IMFO) for gene selection along with an advanced Hyper-heuristic Adaptive Universum Support classification model Vector Machine (HH-AUSVM) to improve cancer detection rates. The hybrid gene selection method is developed by performing filter-based selection using MIM in the first stage followed by the wrapper method in the second stage, to obtain the pivotal features and remove the inappropriate ones. This method improves standard MFO by a hybrid exploration/exploitation phase to accomplish a better trade-off between exploration and exploitation phases. The classifier HH-AUSVM is formulated by integrating the Adaptive Universum learning approach to the hyper-heuristics-based parameter optimized SVM to tackle the class samples imbalance problem. Evaluated on breast cancer gene expression datasets from Mendeley Data Repository, this proposed MIM-IMFO gene selection-based HH-AUSVM classification approach provided better breast cancer detection with high accuracies of 95.67%, 96.52%, 97.97% and 95.5% and less processing time of 4.28, 3.17, 9.45 and 6.31 seconds, respectively.*

Keywords*: Gene expression analysis, Breast cancer, hybrid gene selection, Mutual Information Maximization, Improved Moth Flame Optimization, Support Vector Machine, Adaptive Universum learning, Hyper-heuristic algorithm*

1. INTRODUCTION

Breast cancer is widespread among women with a high global death rate and has multiple causes, including genetic and hereditary factors [1]. The genetic factors can be estimated only through learning the gene expression data for molecular analysis of breast cancer pathogenesis. This data is a part of the cancer transcriptome, including the different RNA sequencing data types. The transcriptome of an organism is measurable using RNA-seq or DNA microarrays. The molecular analysis of this genetic information from DNA can provide all the information about the features and functions of all the body cells [2]. The genes also provide the vital specification of the phenotypes, which can be identified by analyzing the gene expression profiles of all diseased tissues and healthy tissues for obtaining the genetic variables for the pathological process [3]. The gene expression data can provide information about the cancer cells, the impact of drugs on the tissues and genetic variations in the diseased cells. Therefore, the gene expression data helps obtain the different features associated with breast cancer, which can be analyzed using advanced computational methods to

identify the gene targets, detect the disease's presence and develop suitable drugs [4]. Many studies have used gene expression data for obtaining deep tumor characteristics, which provide options for treating, caring and monitoring cancer patients. Detection of the genes more highly expressive of the tumor characteristics than the normal cell characteristics is often challenging when selecting the best computation method [5]. The gene expression data analysis also has challenges, such as high dimensionality of gene features, moderately smaller sample size and a higher noise ratio.

Numerous studies have utilized supervised and unsupervised learning systems for cancer identification from gene expression data. The unsupervised category includes cluster-based methods and decision tree classifiers, while the supervised category includes statistical and machine learning (ML) algorithms [6]. ML algorithms are predominantly utilized for the classification of disease data. They have often produced efficient results using the algorithms such as support vector machines (SVM), Random forests (RF), artificial neural networks (ANN), etc. The latest studies have used deep learning (DL) methods, the so-called complicated algorithms of the ML family, for cancer classification tasks. Algorithms such as Deep Neural Networks (DNN), Convolutional Neural networks (CNN), etc., have better learning rates and improved deterministic powers than ML algorithms. Still, these algorithms require more training data to learn the deep features. They are also mostly limited by the high dimensionality and the sparse sample size of the gene expression data [7]. Considering such limitations of DL methods and the extensive research still needed to integrate them for the genomics data analysis, ML algorithms are suggested for breast cancer classification from gene expression data.

SVM has provided better performance for breast cancer classification with reduced training and testing time [8]. Yet, SVM also has its share of limitations, namely the limited ability to handle high dimensions, underperformance when the target classes are overlapping or unbalanced, and, most importantly, the parameters of SVM do not adapt automatically to the given problem [9]. This paper has focused on developing an efficient breast cancer classification approach by improving the SVM classifier's performance and establishing a hybrid algorithm for improved gene selection so that the high-dimension problem is solved. The proposed approach has developed a feature selection method in which the filter-based method of MIM is combined with the wrapper method of IMFO to form the MIM-IMFO method. This proposed approach is based on two stages; the first is selecting the most important gene features using Mutual information. The second stage reduces the irrelevant features by the optimal feature subset selection using IMFO. Here, IMFO is developed by introducing a hybrid exploration/exploitation phase to the standard MFO [10] to achieve a good trade-off between the exploration and exploitation phases. The proposed classifier HH-AUSVM is an improved model of SVM in which the Adaptive Universum (AU) learning approach is applied to provide prior knowledge adaptively about the optimal classification problem and minimize the class imbalance problem. The Universum samples are data added to the imbalance classes as false data to balance the data distribution for easy computation without impacting the final output. Additionally, the SVM parameters are tuned using hyper-heuristics to form optimal SVM configuration with high accuracy and reduced model complexity. Evaluation of the MIM-IMFO and HH-AUSVM is performed using gene expression datasets from Mendeley Data Repository for breast cancer.

2. RELATED WORKS

Recent studies have presented various feature selection and classification methods for breast cancer analysis from gene expression data. Feature or Gene selection methods can be filter, wrapper or embedded methods. Statistical measures such as correlation coefficient and mutual information are used in filter methods to select genes based on relevancy. Vanitha et al. [11] computed MI between genes and class labels to select the best genes and applied SVM for classification. Recently, Rahmanian and Mansoori [12] developed an unsupervised gene selection method using multivariate normalized MI (MNMI) with higher classification accuracy. Wrapper methods have mostly utilized one or more metaheuristic algorithms. Several algorithms, such as GA [13], PSO [14], GSA [15], etc., provided a training-based selection of genes. Embedded methods are a hybrid of the wrapper and filter methods and select genes based on relevancy and training accuracy. In [16], the authors combined MI and GA to form a hybrid selection method and achieved 90% accuracy. Sun et al. [17] presented hybrid gene selection using ReliefF and ant colony optimization (RFACO) and achieved an average accuracy of 94.3%. Although the embedded methods provide a better performance, the complexity of these methods must be reduced.

Zhang et al. [18] proposed an efficient feature selection strategy using an SVM based on recursive feature elimination and parameter optimization (SVM-RFE-PO). Evaluated on GEO and TCGA datasets, this model achieved an AUC of 96% but also increased the complexity of training. Kong and Yu [19] presented Forest Deep Neural Network (fDNN) model using RF and DNN to extract features to increase the classification accuracy of gene expression data. This model used GEO repository datasets, namely GSE99095 and GSE106291, for evaluation and achieved testing AUC of 0.986 and 0.778 for the two datasets. But this model has limited performance when there are overlapping classes in the dataset. Zhang et al. [20] developed an ensemble classifier based on the principal component analysis (PCA), deep Auto-Encoders and AdaBoost algorithm (PCA-AE-Ada). This model obtained 0.714 AUC and increased the

accuracy from 75% to 85%. Yet, this model is prone to over-fitting owing to the sparse gene datasets.

Elbashir et al. [21] introduced Lightweight CNN with selected hyper-parameters for breast cancer classification with Array-Array Intensity Correlation (AAIC) outlier removal, filtering and normalization. This model obtained 98.76% accuracy on the TCGA dataset for breast cancer gene expression data. However, this method has higher computational complexity. Mondal et al. [22] developed an entropy-based supervised learning method of SVM, RF, k-nearest neighbor (KNN) and naive Bayes for cancerous breast genes. Among them, SVM achieved 91.5% classification accuracy on the GSE349 & GSE350 datasets from the GEO repository. Yet, this model suffers from a class imbalance problem. AbdElNabi et al. [23] developed a cancer classification approach using information gain (IG)-grey wolf optimization (GWO) feature selection and SVM classifier to overcome the over-fitting problems. Evaluated on skewed cancer datasets from Kent Ridge Bio-Medical Data websites, this approach obtained 94.87% accuracy for breast cancer and 95.935% for colon cancer. However, this approach has generated high false positives due to reduced feature selection.

Pham et al. [24] established a model for subtyping breast cancer from gene expression data using an SVM-RFE classifier with GS (SVM-RFE-GS). This model achieved an accuracy of 89.40% with an improvement of 5.44% on TCGA datasets but did not consider the class imbalance problem. Gupta and Gupta [25] presented an improved SVM-RFE gene selection scheme with the Least Absolute Shrinkage Selector Operator (LASSO) and Ridge regression for classifying breast cancer genes. This method reduced the RMSE values from 0.15 to 0.24. Yet, this method has limited learning capability. Hosseinpour et al. [26] developed a Hybrid High-order Type-2 Fuzzy Cognitive Map Improved RF classification (HHTFCMIRF) approach. This approach utilized the improved RF for classifying the breast cancer gene data and achieved 93.5% on the TCGA dataset. But this method also increased the false positives in the presence of the class imbalance problem. Wei et al. [27] proposed generative adversarial networks (GAN) model with data augmentation to detect breast cancers with an accuracy of 92.6%. Still, the class imbalance problem or overlapping class labels is not considered.

Few inferences are obtained from the studies in the literature. The main inference is that the ML algorithms can be more suited for the sparsely sampled gene expression data irrespective of the emergence of complicated DL methods. Though DL methods have better feature learning, ML methods can avoid over-fitting and class imbalance problems more effectively. The other main inference is that feature learning methods can improve classification accuracy with reduced complexity in high-dimensional data. Considering these inferences, the proposed approach has developed the MIM-IMFO feature selection and HH-AUSVM classifier.

3. METHODS

The overview of the proposed approach for breast cancer classification is shown in Fig. 1.

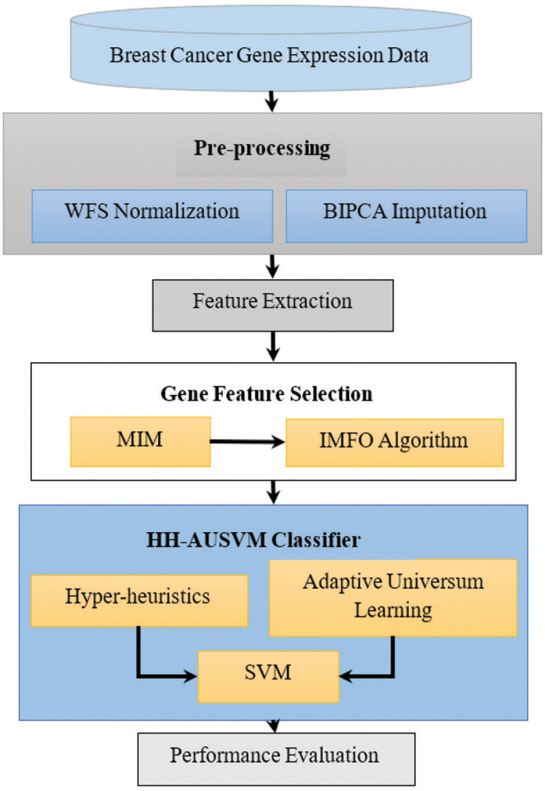

Fig. 1. Overview of the Proposed Breast Cancer Classification model

The proposed approach for breast cancer classification using gene expression analysis includes three main steps: pre-processing, feature extraction and selection, and classification. The pre-processing step intends to offer high-quality gene expression data for investigating breast cancer physiognomies. The pre-processing method has two major processes, namely Weighted Fuzzy Score (WFS) based data normalization and Bayesian Independent Principal Component Analysis (BIPCA) based missing value imputation [28]. After pre-processing, the gene feature vectors are extracted from the datasets, and then the hybrid feature selection method is utilized. First, the mutual information values are computed for the gene vector pairs, and the best gene pairs are selected based on the maximum mutual information values. Then the gene pairs are rearranged using the hybrid phase, and the best genes are ranked in descending order with respect to the fitness values using the IMFO algorithm. Finally, the selected pivotal gene features are fed to the HH-AUSVM to train the classifier for obtaining accurate breast cancer classification in the testing stage.

3.1. DATASETS

The breast cancer gene expression profiles are collected from the Mendeley Data repository (https://data.mendeley.com/datasets/v3cc2p38hb/1).

The main dataset contains four separate datasets: BC-TCGA, GSE2034, GSE25066 and Simulation Data. Table 1 shows the data distribution in these datasets.

Table 1. Distribution of Breast Cancer gene expression datasets.

Datasets	Number of genes	Number of samples		
		Total	Normal Class	Cancer Class
BC-TCGA	17,814	590	61	529
GSE2034	12,634	286	179	107
GSE25066	12,634	492	100	392
Simulation Data	10,000	200	100	100

Table 2 shows examples of normal and tumor classes based on the selected five gene values from the input datasets.

Table 2. Example of Normal and Tumor data.

Tissue	Gene Expressions (Gene Names)					Class
	ELMO2	PNMA1	MMP2	ZHX3	CHD8	
1	0.2043	0.5385	0.7076	-0.117	-0.160	Normal
2	0.0645	0.2335	0.99	-0.468	-0.146	Normal
3	0.2887	0.2327	1.3211	-0.376	-0.162	Normal
4	1.2194	-0.2187	-0.148	0.2108	0.8762	Tumor
5	1.2426	-0.026	-1.073	0.3796	0.3047	Tumor
6	1.1717	-0921	-0.435	0.6231	0.2454	Tumor

3.2. PRE-PROCESSING

The gene expression datasets for breast cancer are raw data with limitations in terms of missing values and outliers. The high-quality data will ensure efficient disease classification. Therefore, data normalization and missing value imputation techniques are applied to the input datasets. WFS-based normalization was developed by integrating the Minkowski Weighted Score Functions into the gene fuzzy score computation. This WFS method used these weighted fuzzy scores to transform the gene expression data values without large variations. Similarly, the missing value problem is solved using the BIPCA imputation method, which utilizes Bayesian theory applied to a fusion model of principal component analysis (PCA) and independent component analysis (ICA) to replace the missing values through likelihood values of informative genes.

3.3. GENE FEATURE SELECTION USING MIM-IMFO

Gene feature selection is choosing the most informative features while eliminating the less informative and irrelevant features. Feature selection can be performed by wrapper, filter and hybrid methods. This study uses a hybrid method by combining the MIM method and the IMFO algorithm. The gene features are selected quickly and efficiently by collaboratively using these two methods. In this hybrid model, the Mutual Information (MI) is used inside the IMFO algorithm and as the metric to estimate the importance of the features. Then the IMFO-based optimization strategy ranks the features based on the fitness values and returns the top-ranked features for the classification. In this model, the IMFO is developed to overcome the limitation of MFO, i.e., the degeneration of the global search capability and slow convergence. To improve the MFO, a hybrid phase is added between exploration and exploitation, which can improve the search process and convergence speed.

Initially, the moths' population is assembled, and the initial parameters are defined along with the maximum number of iterations. In standard MFO, the moths are updated based on the flames, and flames are generated by sorting the best moths. Yet, this process will lead to poor population diversity leading to slow convergence and reduced global search capability. The moth population is initialized, and the elite individual information of moths is protected to eliminate the loss rate. The best solution for the entire population will be stored in a matrix form

$$H = \begin{bmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,d} \\ h_{2,1} & \dots & \dots & h_{2,d} \\ \dots & h_{i,j} & \dots & \dots \\ h_{n,1} & h_{n,2} & \dots & h_{n,d} \end{bmatrix} \quad (1)$$

Here $h_{i,j}$ refers to the best position of the *i-th* moth at the *j-th* dimension. This equation is an enhanced form of the standard MFO initialization step. When the iteration k=1, this equation will equal the initialization function of MFO. This ensures all the moth individuals have a set initial position. These initial positions are updated by the logarithmic spiral as

$$M_i^k = D_i^{k-1} e^{bt} \cos(2\pi t) + F_i^{k-1} \quad (2)$$

Here, M_i^k denotes the position of *i-th* moth at the iteration k, $D_i^{k-1} = |F_i^{k-1} - M_i^{k-1}|$ denotes the distance between *i-th* moth and *i-th* flame at *k-1* iteration and F_i^{k-1} denotes the position of the *i-th* flame at iteration k-1. The parameter b defines the spiral shape, and t denotes a random number between [r,1], with r being a linearly decreasing function from −1 to −2.

Then the MI is computed for these feature subsets, which are mapped with the gene features using Eq. (5), and the process will be terminated if the maximum MI value is obtained. MI is used to compute the effectiveness of the features from high dimensional data to obtain higher classification accuracy. It is estimated as the amount of information via the reduction in entropy. Entropy can measure the diversity in the attributes and helps in obtaining the impurity of information to quantify the uncertainty of the prediction results using the given variable. Hence the entropy is first formulated to compute the MI. Let y denote the discrete random variable attribute with two possible outcomes, i.e., relevant (R) and irrelevant ($\bar{R}$) to the ideal features. The binary function H can be expressed as a logarithmic value.

$$H(y) = -p(R)\log p(R) - p(\bar{R})\log p(\bar{R})$$

Here (R, $\bar{R}$) denotes the possible classes- relevant and irrelevant, $p(R)$ denote the probability of the sample being $y\in(R)$ and $p(\bar{R})$ denote the probability of the sample being $y\in(\bar{R})$. Conditional entropy defines the quantity of the uncertainties of each feature in the decision process, and it is computed between two events, X and Y, where X has the value of feature x,

$$\begin{gathered} H(Y \mid X)=\sum_{x\in X} p_x(x)\, H(Y \mid X=x)= \\ \sum_{x\in X} p_x(x) \sum_{y\in Y} p(y \mid x)\ log\, p_y(y \mid x)= \\ \sum_{x\in X}\sum_{y\in Y} p_{xy}(x,y)\, logp_y(y|x) \end{gathered} \quad (4)$$

The smaller values of the impurity will result in more skewed class distributions. The values of entropy and the misclassification errors will be the highest when the class distribution is uniform and the minimum when all the samples belong to the same class.

The MI of y can be computed using the entropy and conditional entropy from a feature x as

$$MI(y|x) = H(y) - H(y|x) \quad (5)$$

A larger MI defines the higher discriminative power for the decision process and determines the relevance of the features with respect to the classification problem. The gene pairs are rearranged in the hybrid phase to obtain improved gene features.

IMFO includes a new step compared to the MFO. As stated before, the IMFO performs three phases. The primary phase pledges an optimal exploration, the intermediary phase is a hybrid exploration/exploitation, and the final phase improves the exploitation. The iterations determine it, and hence the iterations are divided into $I1$, $I2$ and $I3$ for each phase. It is defined as

$$I_1 = [1, \delta_1] \cap N, I_2 = [\delta_1, \delta_2] \cap N, I_3 = [\delta_2, K] \cap N \quad (6)$$

Here $\delta1=\alpha K$, $\delta2=\beta K$ with $\alpha,\beta\in[0,1]$ and N denotes the set of numbers.

The hybrid phase has been introduced to avoid an abrupt transition between the exploration and exploitation phases. A weighted factor is added to the fitness function (accuracy or error rate) to improve the exploitation phase without downgrading the exploration capability. Weight factor w is given as

$$w = \left|\frac{f(M_{best})}{f(M_i^k)}\right| \quad (7)$$

Here, $f(M_{best})$ denotes the fitness value of the best solution, and $f(M_i^k)$ denotes the fitness values of the i-th moth at iteration k. The exploitation is improved by using this factor during the hybrid phase without influencing the exploration phase. The moth positions are updated as

$$M_i^k = D_i^{k-1} e^{bt}\cos(2\pi t) + w.F_i^{k-1} + (1-w).M_{best} \quad (8)$$

Thus, a good balance is obtained in the trade-off between exploration and exploitation. The fitness value is computed for the new features obtained after the hybrid phase using these steps. Then these features are ranked based on their relevance with respect to the fitness value (accuracy or error rate) computed in IMFO. Thus, the MIM-IMFO gene feature selection helps deep explore and decide the best gene subsets.

3.4. Classification using HH-AUSVM

The benefits of using SVM-based classifiers are that they have universal optimization and a great simplification facility to make the grouping precise. Moreover, it resolves over-fitting problems and reduces computational complications. However, the standard SVM cannot handle the noises and unknown class samples effectively. Hence Adaptive Universum learning is utilized with the SVM so that the classifier learns the patterns of unknown classes and the known classes effectively with prior knowledge. AUSVM constructs the data-dependent architecture of SVM based on the set of tolerable functions for ensuring adaptability. It is more appropriate to obtain the set of Universum samples for the SVM learning process instead of defining the data distributions explicitly. This Universum Learning solves the class imbalance problem better than the other sampling-based methods with good regularization and generalization for the data. The Universum samples are the additional samples generated based on current data but do not belong to the current classes. These data are added to the imbalance classes as false data to balance the data distribution for easy computation without impacting the final output. Since the Universum samples do not belong to any predefined classes, the AUSVM hyper-plane will fall inside the margin borders determined by C due to the usage of the maximal margin procedure. Therefore, the AUSVM must utilize a maximal soft-margin procedure and maximize the number of Universum samples distributed around the hyper-plane.

A training set with given Universum samples is defined as

$$S = \{(x_1, y_1), (x_2, y_2), \ldots . (x_s, y_s)\} \cup \{x_1^*, x_2^*, \ldots, x_u^*\} \quad (9)$$

Here $x_j^* \in R^n$,j=1,2,…,u denote the Universum samples in R^n search space, x_i R^n,i=1,2,…,s and $y_i\in\{1,-1\}$ for binary classification and $y_i R^n$, i=1,2,…,s for multi-class classification. As the Universum samples provide the prior knowledge of the network traffic classification by approximating the hyper-plane $g(x)$=0, the primal optimization algorithm of the AUSVM with the maximal soft-margin procedure is given as

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C_t\sum_{i=1}^{s}\xi_i + C_u\sum_{t=1}^{u}(\psi_t + \psi_t^*) \quad (10)$$

Subject to $y_i\,(w.\Phi\,(x_i) + b) \geq 1\text{-}\xi_i$, - ε - $\psi_t^* \leq w.\Phi\,(x_t^*) + b \leq \varepsilon + \psi_t$, $\xi_i \geq 0$, i=1,2,...,s and ψ_t, $\psi_t^* \geq 0$, t=1,2,...,u.

Here C_t denotes the margin parameter or penalty parameter of SVM, C_u denotes the margin parameter of AUSVM, ψ_t, ψ_t^* denotes the slack variables of AUSVM and ε represents the in-sensitive loss function for Universum samples. Eq. (13) of AUSVM maximizes the margin between the classifying hyper-planes and the amount of Universum samples to be distributed

around the hyper-plane. If C_u=0, then Eq. (13) will become equivalent to the standard SVM equation. This dual problem of USVM is formulated as

$$\begin{gathered}\min_{\alpha,\mu,v}\frac{1}{2}\sum_{i=1}^{s}\sum_{j=1}^{s}y_iy_j\alpha_i\alpha_j\ K(x_i,x_j)+\\ \frac{1}{2}\sum_{t=1}^{u}\sum_{z=1}^{u}(\mu_t-v_t)(\mu_z-v_z)K(x_t^*,x_z^*)+\\ \sum_{i=1}^{s}\sum_{t=1}^{u}y_i\alpha_i(\mu_t-v_t)K(x_i,x_t^*)-\sum_{i=1}^{s}\alpha_i+\\ \varepsilon\sum_{t=1}^{u}(\mu_t+v_t)\end{gathered} \tag{11}$$

Subject to $\sum_{i=1}^{s} y_i \alpha_i \sum_{t=1}^{u} (\mu_z - v_z) = 0;\ 0 \le \alpha_i \le C_t$, $i=1,2,...,s$; $0 \le \mu_t, v_t \le C_u t=1,2,...,u$.

Here μ_i and v_i are Lagrangian multipliers similar to α_i.

When the classifier is non-linear, the problem in the input space from Eq. (13) is defined as

$$f(x) = sign\left(\sum_{i=1}^{m}\alpha_i^* \times y_i \times K(x_i, y_i) + b^*\right) \tag{12}$$

Here $f(x)$ symbolizes the objective function, α_i^* denotes control parameter, and b^* indicates the bias. $K(x_i, y_i)$ represents the kernel function that creates the central product of the feature space. The parameters namely C_u, $K(x_i, y_i)$ and its parameters are selected optimally using HH. The aim is to choose an SVM structure that diminishes the miscalculation error and increases the accuracy without manipulating the complication. It is exhibited as a non-convex optimization issue conveyed as a tuple form

$$< AUSVM, \Theta, \mathcal{D}, \mathcal{C}, S > \tag{13}$$

Where AUSVM is the constructed system, Θ is the exploration area of the conceivable SVM structures, D is the dissemination of the set of cases, C is the fitness utility, and S is the arithmetic data. The objective is to reduce C to achieve the resolution set over a set of issue circumstances to discover

$$\theta^* \in \arg\min_{\theta\in\Theta}\frac{1}{|\mathcal{D}|}.\sum_{\pi\in\mathcal{D}}\mathcal{C}(\theta,\pi) \tag{14}$$

Each $\theta \in \Theta$ characterizes one conceivable structure of the SVM, and the result C is attained while analyzing the SVM through numerous illustrations. The multi-constraint optimization is expressed as

$$\min F(X) = |f_1(x), f_2(x)| \tag{15}$$

Where $f_1(x)$=Accuracy; $f_2(x)$ = Model Complexity

Here $f_1(x)$, $f_2(x)$ are the two objective functions of SVM, and Model complexity is expressed as the Number of Support Vectors (NSV). The HH algorithm reduces this function to acquire a lightweight SVM with high accuracy and less training time.

Hyper-heuristic optimization consists of high-level and low-level heuristics in the SVM design structure optimization. HH is the multi-level algorithm that accomplishes heuristic interpretations and generates low-level policies based on the problem obligation from prevalent solutions. The low-level heuristics follow the solution space and regulate the current solutions to create new solutions for assessment.

The high-level policy chooses the appropriate low-level heuristics as an alternative to probing the solutions so that the low-level policy can execute its functions without any disruptions from the advanced search policy.

4. RESULTS AND DISCUSSION

The proposed MIM-IMFO feature selection and HH-AUSVM classification approach for the breast cancer classification problem is evaluated over the Mendeley gene expression datasets. The evaluations are conducted using the MATLAB tool (R2016b version 9.1) on an Intel Core i7 processor, Windows 10 OS with 8GB RAM and 512GB SSD. The performance metrics, namely Accuracy, Precision, Recall, F-Measure, Pearson Correlation Coefficient (PCC) and Processing Time, are used for the evaluation. These parameters are chosen for evaluation since they can help determine the correctness of the models and also detect the linear relationship of the variables.

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(Total\ Classes)} \tag{16}$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{17}$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \tag{18}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{19}$$

$$PCC = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{20}$$

Here, x_i and y_i denote the x and y variables of the sample, $\bar{x}$ and $\bar{y}$ denote the mean of the values of the x and y variables.

The performance of the proposed gene selection method is evaluated with other existing gene selection methods based on the number of selected genes. All simulations are performed with the same amount of data for fair comparisons. Table 3 shows the performance of gene selection methods for the four datasets.

Table 3. Comparison of Gene Selection Methods

Method	BC-TCGA	GSE2034	GSE25066	Simulation Data
Total Genes	17814	12634	12634	10000
MI [11]	12540	9875	9833	8101
MNMI [12]	11345	8603	8628	6957
GA [13]	9790	7884	7752	5880
PSO [14]	9176	7542	7267	5541
GSA [15]	9042	7125	6823	5179
MI-GA [16]	7920	6043	6043	4222
RFACO [17]	7775	5779	5450	3980
Proposed MIM-IMFO	7108	5369	5122	3338

Table 4 shows the obtained results for the proposed methods over the testing sets of the datasets. The existing method SVM is used as the base classifier along with Adaptive Universum SVM. The proposed HH-AUSVM classifier is used without the feature selection method and with the MIM-IMFO method.

Table 4. Performance comparison of Breast Cancer Gene Expression Classification.

Accuracy				
Method	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	92.36	91.86	90.02	93.89
AUSVM	93.17	92.99	92.58	94.31
HH-AUSVM	94.49	94.26	95.67	94.94
MIM-IMFO + HH-AUSVM	95.67	96.52	97.97	95.5
Precision				
Methods	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	91.66	96.15	89.32	95.36
AUSVM	93.27	97.49	92.33	96.67
HH-AUSVM	95.88	98.67	95.69	98.5
MIM-IMFO + HH-AUSVM	98.75	99.24	96.38	100
Recall				
Methods	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	93.94	92.22	86.25	92.48
AUSVM	94.22	94.67	90.49	94.31
HH-AUSVM	94.89	95.55	92.18	95.96
MIM-IMFO + HH-AUSVM	95.72	96.2	95.04	97.78
F-measure				
Methods	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	92.79	94.14	87.76	93.9
AUSVM	93.74	96.06	91.4	95.48
HH-AUSVM	95.38	97.08	93.9	97.21
MIM-IMFO + HH-AUSVM	97.21	97.7	95.71	98.88
Pearson Correlation Coefficient				
Methods	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	0.9012	0.9147	0.812	0.8813
AUSVM	0.9225	0.9381	0.8892	0.9156
HH-AUSVM	0.9467	0.9448	0.9218	0.9271
MIM-IMFO + HH-AUSVM	0.9588	0.9633	0.9692	0.9524
Processing time (seconds)				
Methods	**BC-TCGA**	**GSE2034**	**GSE25066**	**Simulation Data**
SVM	11.98	10.96	13.56	10.55
AUSVM	10.26	10.18	12.40	9.89
HH-AUSVM	8.78	8.11	10.87	8.25
MIM-IMFO + HH-AUSVM	4.28	3.17	9.45	6.31

The results in Table 1 show that the proposed HH-AUSVM classifier with the MIM-IMFO feature selection has provided better breast cancer classification than the SVM, AUSVM and HH-AUSVM models. Among the compared methods, the proposed HH-AUSVM with MIM-IMFO feature selection has achieved 1.18%, 2.5% and 3.31% higher accuracy, 2.87%, 5.48% and 7.09% higher precision, 0.83%, 1.5% and 1.78% higher recall, 1.83%, 3.47% and 4.42% higher f-measure, 1.21%, 3.63% and 5.76% higher Pearson Correlation Coefficient, and 51.25%, 58.29% and 64.27% reduced processing time than the HH-AUSVM, AUSVM and SVM methods for the BC-TCGA dataset.

For GSE2034 data, it has achieved 2.26%, 3.53% and 4.66% higher accuracy, 0.57%, 1.75% and 3.09% higher precision, and 0.65%, 1.53% and 3.98% higher recall, 0.62%, 1.64% and 3.56% higher f-measure, 1.85%, 2.52% and 4.86% higher Pearson Correlation Coefficient, 60.91%, 68.86% and 71.08% reduced processing time than the HH-AUSVM, AUSVM and SVM methods. For GSE25066 data, HH-AUSVM with the MIM-IMFO has achieved 2.3%, 5.39% and 7.95% higher accuracy, 0.69%, 4.05% and 7.06% higher precision, 2.86%, 4.55% and 8.79% higher recall, 1.81%, 4.31% and 7.95% higher f-measure, 4.74%, 8% and 15.72% higher Pearson Correlation Coefficient and 13.06%, 23.79% and 30.31% reduced processing time than the HH-AUSVM, AUSVM and SVM methods.

For the simulation data, HH-AUSVM with the MIM-IMFO has achieved 0.56%, 1.19% and 1.61% higher accuracy, 1.5%, 3.33% and 4.64% higher precision, 1.82%, 3.47% and 5.3% higher recall, 1.67%, 3.4% and 4.98% higher f-measure, 2.53%, 3.68% and 7.11% higher Pearson Correlation Coefficient and 23.52%, 36.2% and 40.19% reduced processing time than the HH-AUSVM, AUSVM and SVM methods.

The confusion matrix evaluation obtained for the four datasets is shown in Table 5.

Table 5. Confusion Matrix of Proposed Method

Datasets	Total Data	True Positive	True Negative	False Positive	False Negative
BC-TCGA	590	59	506	11	14
GSE2034	286	173	103	3	7
GSE25066	492	98	384	4	6
Simulation Data	200	93	97	4	6

The proposed method's confusion matrix evaluation has shown a good trade-off ratio between the true and false values. In Simulation data, the 200 samples are classified correctly into 93 True Positives (Normal class) and 97 True Negatives (Tumor Class) and wrongly into 4 False Positives and 6 False Negatives. The confusion matrix complements the justification provided by the other evaluation metrics.

This better performance of the MIM-IMFO-based gene selection and HH-AUSVM classifier is because of the use of effective pre-processing by WFS-BIPCA, improved convergence and global search capability of gene feature selection and advanced learning-based optimized classification.

They are compared with the methods used in the literature studies to evaluate the proposed approach further. Since the literature methods have used different breast cancer gene expression datasets in different experimental conditions, comparing their results directly will not be ideal. Hence, for a fair comparison, the methods described in those studies are implemented in the same environment as the proposed approach over the GSE2034 and GSE25066 datasets with an equal amount of data. The comparisons are made in terms of accuracy and processing time. Table 6 shows the comparison of the proposed approach against the literature studies.

Table 6. Performance comparison against methods in the literature.

Ref. No.	Method	GSE2034		GSE25066	
		Accuracy (%)	Time (s)	Accuracy (%)	Time (s)
[18]	SVM-RFE-PO	91.5	7.54	91.87	15.76
[19]	fDNN	93.47	11.66	94.98	24.61
[20]	PCA-AE-Ada	90.88	12.89	91.45	22.55
[21]	CNN	95.25	15.63	95.79	21.8
[22]	Entropy-based SVM	91.31	6.55	92.67	17.17
[23]	IG-GWO + SVM	90.15	5.41	91.11	13.2
[24]	SVM-RFE-GS	88.91	7.88	89.28	15.78
[25]	SVM-RFE-LASSO	86.72	8.91	87.4	16.83
[26]	HHTFCMIRF	91.2	6.37	90.76	14.56
[27]	GAN	93.8	17.88	95.51	23.5
Proposed	**MIM-IMFO + HH-AUSVM**	**96.52**	**3.17**	**97.97**	**9.45**

The comparison of the proposed approach against the existing methods in the literature studies also shows that the proposed MIM-IMFO and HH-AUSVM-based breast cancer classification model performs better than the other methods for the GSE2034 and GSE25066 datasets. There have been accuracy improvements in the proposed approach by approximately 1 to 10%. The processing time of the proposed method is also less than the other methods. This concludes that the proposed approach of MIM-IMFO feature selection and HH-AUSVM classifier has provided a better analysis of the gene expression data for accurate breast cancer classification with less complexity.

5. CONCLUSION

This paper aimed to introduce a hybrid feature selection technique and advanced classifier for reducing the dimensionality of the breast cancer gene expression data to improve the classification performance. Considering the existing limitations, this paper presented an efficient classifier of HH-AUSVM to overcome the class imbalance problem, noisy data, sparse data and parameter tuning problems for analyzing the gene expression data. Utilized with the MIM-IMFO feature selection method, the HH-AUSVM classifier obtained breast cancer classification accuracies of 95.67%, 96.52%, 97.97% and 95.5% for BC-TCGA, GSE2034, GSE25066 and Simulation Data, respectively. It has also consumed about 1-10% less processing time for all four datasets than the existing methods. In the future, the possibility of improving the feature learning property of HH-AUSVM will be investigated. Although the evaluations have been made only on breast cancer gene expression datasets, the proposed method is also suitable for other cancer gene expression datasets. The efficiency of this method for classifying other types of cancer will be examined in the future.

6. REFERENCES

[1] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, F. Cardoso, "Breast cancer", Nature Reviews - Disease Primers, Vol. 5, No. 1, 2019, pp. 1-31.

[2] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer", Nature, Vol. 415, No. 6871, 2002, pp. 530-536.

[3] Z. Sun, Y. W. Asmann, K. R. Kalari, B. Bot, J. E. Eckel-Passow, T. R. Baker, E. A. Thompson, "Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing", Plos One, Vol. 6, No. 2, 2011, pp. 1-16.

[4] N. R. Latha, A. Rajan, R. Nadhan, S. Achyutuni, S. K. Sengodan, S. K. Hemalatha, P. Srinivas, "Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy", Critical Reviews in Oncology/Hematology, Vol. 151, No. 1, 2020, pp. 102964-102978.

[5] M. Abd-Elnaby, M. Alfonse, M. Roushdy, "Classification of breast cancer using microarray gene expression data: A survey", Journal of Biomedical Informatics, Vol. 117, No. 1, 2021, pp. 103764-103782.

[6] A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, H. Ehtesham, "Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review",

Iranian Journal of Public Health, Vol. 46, No. 2, 2017, pp. 165-179.

[7] R. Shahane, M. Ismail, C. S. R. Prabhu, "A survey on deep learning techniques for prognosis and diagnosis of cancer from microarray gene expression data", Journal of computational and theoretical Nanoscience, Vol. 16, No. 12, 2019, pp. 5078-5088.

[8] A. Statnikov, L. Wang, C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification", BMC Bioinformatics, Vol. 9, No. 1, 2008, pp. 1-10.

[9] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics", Cancer Genomics & Proteomics, Vol. 15, No. 1, 2018, pp. 41-51.

[10] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm", Knowledge-based Systems, Vol. 89, No. 1, 2015, pp. 228-249.

[11] C. D. A. Vanitha, D. Devaraj, M. Venkatesulu, "Gene expression data classification using support vector machine and mutual information-based gene selection", Procedia Computer Science, Vol. 47, No. 1, 2015, pp. 13-21.

[12] M. Rahmanian, E. G. Mansoori, "An unsupervised gene selection method based on multivariate normalized mutual information of genes", Chemometrics and Intelligent Laboratory Systems, Vol. 222, No. 1, 2022, pp. 104512-104522.

[13] S. Sayed, M. Nassef, A. Badr, I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets", Expert Systems with Applications, Vol. 121, No. 1, 2019, pp. 233-243.

[14] Y. Prasad, K. K. Biswas, M. Hanmandlu, "A recursive PSO scheme for gene selection in microarray data", Applied Soft Computing, Vol. 71, No. 1, 2018, pp. 213-225.

[15] P. K. Ram, P. Kuila, "GSA-based approach for gene selection from microarray gene expression data", Machine Learning Algorithms and Applications, Wiley Press, 2021, pp. 159-174.

[16] M. Jansi Rani, D. Devaraj, "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification", Journal of Medical Systems, Vol. 43, No. 8, 2019, pp. 1-11.

[17] L. Sun, X. Kong, J. Xu, Z. A. Xue, R. Zhai, S. Zhang, "A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification", Scientific Reports, Vol. 9, No. 1, 2019, pp. 1-14.

[18] Y. Zhang, Q. Deng, W. Liang, X. Zou, "An efficient feature selection strategy based on multiple support vector machine technology with gene expression data", BioMed Research International, Vol. 2018, No. 1, 2018, pp. 1-11.

[19] Y. Kong, T Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification", Scientific Reports, Vol. 8, No. 1, 2018, pp. 1-9.

[20] D. Zhang, L. Zou, X. Zhou, F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer", IEEE Access, Vol. 6, No. 1, 2018, pp. 28936-28944.

[21] M. K. Elbashir, M. Ezz, M. Mohammed, S. S. Saloum, "Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data", IEEE Access, Vol. 7, No. 1, 2019, pp. 185338-185348.

[22] M. Mondal, R. Semwal, U. Raj, I. Aier, P. K. Varadwaj, "An entropy-based classification of breast cancerous genes using microarray data", Neural Computing and Applications, Vol. 32, No. 7, 2020, pp. 2397-2404.

[23] M. L. R. Abd El Nabi, M. Wajeeh Jasim, H. M. EL-Bakry, H. N. Taha, N. E. M. Khalifa, "Breast and colon cancer classification from gene expression profiles using data mining techniques", Symmetry, Vol. 12, No. 3, 2020, pp. 408-422.

[24] T. C. Pham et al. "A New Feature Selection and Classification Approach for Optimizing Breast Cancer Subtyping Based on Gene Expression", Advances in Intelligent Information Hiding and Multimedia Signal Processing, Springer, Singapore, 2021, pp. 298-307.

[25] M. Gupta, B. Gupta, "A novel gene expression test method of minimizing breast cancer risk in reduced cost and time by improving SVM-RFE gene selection method combined with LASSO", Journal of Integrative Bioinformatics, Vol. 18, No. 2, 2021, pp. 139-153.

[26] M. Hosseinpour, S. Ghaemi, S. Khanmohammadi, S. Daneshvar, "A hybrid high-order type-2 FCM improved random forest classification method for breast cancer risk assessment", Applied Mathematics and Computation, Vol. 424, No. 1, 2022, pp. 127038-127052.

[27] K. Wei, T. Li, F. Huang, J. Chen, Z. He, "Cancer classification with data augmentation based on generative adversarial networks", Frontiers of Computer Science, Vol. 16, No. 2, 2022, pp. 1-11.

[28] V. Murugesan, P. Balamurugan, "Weighted Fuzzy Score Normalization and Bayesian Independent Principal Component Analysis Imputation for Breast Cancer Gene Expression Analysis", International Journal of Intelligent Engineering and Systems, Vol. 15, No. 3, 2022, pp. 80-89.