Received: February 28, 2022. Revised: May 25, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Psoriasis Identification from Gene Expression Analysis Using Relative Entropy Fuzzy Score Normalization, Genetic Weighted K-Nearest Neighbors Imputation and Hybrid Machine Learning Classifier

Ganapathy Kavitha[1]* **Kandasamy Vanitha[1]**

[1]*Department of Computer Science, Dr G R Damodaran College of Science, Tamil Nadu 641014, India*
* Corresponding author's Email: kavithagstudy@gmail.com

Abstract: Psoriasis diagnosis from gene expression data requires efficient machine learning classifiers with the finest pre-processing methods to overcome the outlier and missing value problems. This paper developed relative entropy fuzzy score (REFS) normalization and genetic weighted k-nearest neighbors (GWKNN) imputation for enhancing pre-processing and hybrid multi-kernel universum support vector machines (MKUSVM) - radial basis function extreme learning machine (RBFELM) classifier for accurate psoriasis recognition. REFS normalization utilizes improved fuzzification and relative entropy in the gene fuzzy score estimation to reduce the skewness and outlier gene data. GWKNN imputation performs symbolic regression of weighted k-nearest neighbors (KNN) with genetic programming (GP) for missing value estimation. Finally, the hybrid MKUSVM-RBFELM classifier classifies the psoriasis gene data with features selected using mutual-information parameter. Evaluated over public dataset GSE55201, the proposed model achieved 91.67 % accuracy, 92.58 % precision, 90.94 % recall, 91.75 % F-measure, and 0.8452 MCC values with a reduced processing time of 0.373 seconds.

Keywords: Psoriasis classification, Skin diseases, Machine learning, Relative entropy fuzzy score, Genetic weighted k-nearest neighbors, Multi-kernel universum support vector machines, Radial basis function extreme learning machine.

1. Introduction

Skin diseases are common chronic diseases that vary differently in terms of severity and symptoms [1]. While most of the skin diseases are less threatening and minor, some are serious and side-effects of other major diseases. These diseases might cause rashes, inflammation, itchiness or other skin changes. Although these disorders are less threatening, some of these disorders might be the primary symptoms of diseases like skin cancer, toxic epidermal necrolysis [2], staphylococcal scalded skin syndrome [3], etc. Therefore efficient analysis must be performed to diagnose the skin diseases for early treatments and prevention of serious issues. Psoriasis is a common chronic autoimmune or inflammatory skin disease that keeps flaring for a longer period with hyper-proliferative expressions on the skin and joints. Although psoriasis is not considered life-threatening, it is associated with heart and diabetes issues such as amplified hyperlipidemia, hypertension, coronary artery disease (CAD), diabetes (type II), stroke and fatness [4]. Latest studies have shown that psoriasis is also allied with angst and emotional disorders and also isolates the patients from having close relationships in their social life. Psoriasis has certain phenotypic indicators such as epidermal hyperplasia and angiogenesis with infiltrations by dendritic cells, lymphocytes, chemokine and cytokine [5].

Diagnosis of psoriasis is often performed using the visual investigation of cutaneous lesion biopsy, but it is an unproductive process due to the possibility of errors. The genetic and molecular analysis of psoriasis must be performed well enough to obtain useful revelations about the disease [6]. Gene expression profiles and protein-protein interaction data are valid sources and the quantitative polymerase chain reaction (qPCR) and microarray (high-throughput) are the procedures to perform these investigations. Examining the molecular basis and the genetic expression patterns will help in diversifying the normal and diseased patients. Therefore, gene expression and molecular analysis are demonstrated as the main approaches for extracting and recognizing psoriasis disease using efficient classification models.

Recent years have seen the increase in usage of machine learning (ML) based classifiers for many applications including medical data mining and classification. Many ML algorithms like naïve bayes (NB), SVM, artificial neural network (ANN), etc. were established to analyse and diagnose diseases [7]. The gene expression data might contain noisy outlier data and missing values due to poor data compilation [8]. It also contains redundant and inconsistent data instances which increase the difficulty of the search process in ML-based classifiers. When handling missing values, the traditional ML classifiers face high computation complexity and less accurate classification [9]. Among the ML classifiers, SVM has shown higher throughput gene expression classification because of their kernel functions and boundary parameter. SVM fits the hyper-plane to categorize the group of genes by projecting the input space into a higher space using effective kernel functions. However, the architecture of standard SVM becomes complex when the size of the input gene expression data increases. Similarly, SVM has limitations in selecting relevant genes for classification and the selection is upper bound limited by the training data size leading to selecting only a few genes in highly correlated genes [10]. Therefore, this paper is aimed at developing an efficient gene expression analysis model using efficient pre-processing with better normalization and missing value imputation (MVI) and advanced ML classifier.

This paper presents an efficient pre-processing technique by proposing relative entropy fuzzy score (REFS) normalization and genetic weighted k-nearest neighbors (GWKNN) imputation. REFS based normalization uses an improved fuzzification process in which the relative entropy is included in the gene fuzzy score. This method is based on GP and weighted KNN to calculate the incomplete data attributes through the probability of nearest informative genes. Additionally, this paper also presents an advanced SVM based hybrid classifier by combining MKUSVM-RBFELM algorithms. The evaluation of the proposed models is performed using public gene expression datasets of psoriasis disease from gene expression omnibus (GEO) repository. The rest of this paper is arranged as: relevant literature studies in section 2, proposed pre-processing and classification algorithms for gene expression classification in section 3, their evaluations and results in section 4 and finally the inferences to conclude the research in section 5.

2. Related works

Various techniques have been presented in recent years for pre-processing using normalization and imputation and classifying gene expression data. Belorkar and Wong [11] developed gene fuzzy score (GFS) as a pre-processing transformation technique for the normalization task. This GFS focussed on reducing the obscuring variation of the gene expressions using the fuzzy scores derivative from rank values of genes in the distinct data samples to perform the normalization. This method has reduced the batch effects and also increases the interpretability of the transformed outcomes without any negative impact on the sample size variation. Yet, this method is adaptable only for microarray gene expression while do not support other high throughput gene expressions. Franks et al. [12] presented feature specific quantile normalization (FSQN) for improving the molecular subtype classification of the gene data. This method achieved robust performance with 98 % and 97 % on the BRCA and CRC gene data, respectively, but has increased the computation time. Tang et al. [13] developed Bayesian normalization (bayNorm) for the normalization of single-cell RNA-sequencing data. This approach utilized the scaling based probability function of the binomial model for estimating the expression values using the priors. This model preserved low false-positive rates but reduced AUC significantly. Breda et al. [14] proposed a Bayesian inference method called sanity (Sampling-Noise-corrected Inference of Transcription activity) for the RNA-seq data. This sampling removed the Poisson sampling fluctuations, thus obtaining the true variation of the genes without the need for parameter tuning. However, this approach does not improve the accuracy when the sample size is varied. Lause et al. [15] developed analytic pearson residuals (APR) from the negative binomial regression for the normalization of RNA-seq data. This method reduces the biased per-gene over-dispersion estimation through the use of negative control data without biological variability. This model provided a high F1-score, precision and recall with minimized runtimes, but has exaggerated the training time in the classifiers.

Chen and Zhou [16] developed variability-preserving imputation for expression recovery (VIPER) for estimating the missing values. This method employed the cell data of comparable gene expressions through the selection of a sparse set of local neighbourhood cells by a sparse non-generative regression model. This progressive selection reduces the complexity in estimating the imputation weights with better computation stability and reduced errors. Yet, this method fails to consider the impact of over-dispersion of the dropout events. Howey et al. [17] utilized a Bayesian network (BN) approach for estimating the missing data in complex biological analysis problems. This approach utilized the pseudo-Bayesian nearest neighbour based estimation of the best fit BN through the sampling process. This approach increased the precision and recall values without increasing the model complexity. But this approach supports only when the missing data is below 10 %. Hu et al. [18] developed weighted decomposition of gene expression (WEDGE) to assign the missing gene expression values using the biased low-rank matrix decomposition method (bLRMD). This WEDGE algorithm reproduced the missing values through effective cell-wise and gene-wise correlations even in several cell type datasets with high dropout rates. However, this method reduced the overall accuracy through higher error rates for high dimensional gene expression data. Li et al. [19] presented a hybrid MVI algorithm of jointly fuzzy C-means and VQNN (Vaguely Quantified Nearest Neighbor) imputation (JFCM-VQNNI) along with an extended fitted model. These models utilized fuzzy matrixes, tolerance relations, and fuzzy membership relations to obtain the potential closest values to fill the missing or zero values. These models reduced the RMSE and MAE for gene expression data, but these models also consume high running time. Keerin and Boongoen [20] introduced an improved k-nearest neighbour (IKNN) imputation method for solving the missing values problem. This improved method utilized the perception of the ordered weighted averaging (OWA) operator to enhance the summarization in KNN for estimating the missing or zero values. This method achieved 0.8 and 0.84 normalized RMSE values for 55 and 20 % missing values, respectively. However, this method has limited performance with high errors when the sample size is larger.

Ahmed et al. [21] demonstrated the robustness of the NB classifier for microarray gene expression analysis. Evaluated over cancer gene expression data HNC, this NB classifier achieved 76 % accuracy, 0.9 MCC values and zero error rates. However, this classifier performed poorly for the large datasets due to the location and scale parameters of gene data. Cahyaningrum and Astuti [22] introduced ANN and genetic algorithm (GA) based microarray gene expression classification for cancer discovery. This ANN model obtained accuracies of 83.33 %, 76.47 % and 89.93 % for colon tumor, prostate tumor and lung cancer in the gene expression data. However, the training time of

Table 1. List of notations

Notations	**Description**
θ_1 and θ_2	quantile thresholds for assigning fuzzy scores
g_i	Genes of the patients
p_j	Patient to whom the genes are collected
$r(g_i, p_j)$	rank of gene expression
$q(p_j, \theta)$	rank equivalent to the upper quantile threshold
$s(g_i, p_j)$	fuzzy score allocated to g_i in patient p_j
$A = s(g_A, p_A)$	fuzzy value of g_A in patient p_A
μ_i	supporting conditional variable
v_i	opposing conditional variable
π_i	non-available group conditional variable
λ	risk preference index
δ_A	REFS function
p, q	real-valued gene expression instances
$distance(p, q)$	Euclidean distance between p, q
r_i	range of the i-th feature
$X_{i,j}^K$	Set of input instances
$R_{i,j}$	Set of reference instances
$D_{i,j}^K$	Distances between instances
$W_{i,j}^K$	distance-based weights
$I_{i,j}$	instances
$J_{i,j}$	feature indexes of the data subset
$V_{i,j}$	reference instances of non-missing values
$\{G_g\}_{g=1}^{N}$	GP regression function
$T_{i,j}$	target missing values
$WRSE_{i,j}^K$	weighted relative squared error
$(R, \bar{R})$	possible classes-relevant and irrelevant
$p(R)$	probability of the relevant class
$p(\bar{R})$	probability of the irrelevant class
$x_j^* \in R^n$	Universum samples in R^n search space
C_t	margin parameter of SVM
C_u	margin parameter of MKUSVM
ψ_t, ψ_t^*	lack variables of MKUSVM
ε	insensitive loss function
$K(x_i, x_j)$	kernel function
γ	output weight of the hidden layer
$H^{\dagger}$	Moore–Penrose generalized inverse of H

ANN is higher. Tapak et al. [23] proposed a GA-based SVM for effective feature selection and gene expression data classification for psoriasis recognition. This model utilized the GEO repository dataset GSE55201 from which select best gene features were extracted using GA and applied to the SVM classifier. This model obtained an accuracy of 62.5 % for whole features while 79.17 % for selected 27265 features. However, the local optimum problem of GA and limited correlated genes in upper bound training of SVM degrades the classification performance. Chatra et al. [24] developed cancer classification from gene data by binary bat optimization (BBO) based feature selection and ELM classifier. The ELM model utilized the genes nominated by the BBO to obtain an accuracy of 89 % to 100 % over different cancer types. However, the ELM provides huge variance for the degenerate gene expression data.

From the above literature studies, the major research problems are identified for this research paper. The primary problem of pre-processing due to the less effective normalization and MVI methods are considered. The proposed pre-processing methods aim at reducing the error rates associated with the normalized and imputed values of gene expression data. Then the classification errors and complexity in traditional and recent methods are considered. Based on the limitations, this paper focuses on developing a hybrid classification approach using advanced SVM and ELM classifiers.

3. Methods

The proposed psoriasis gene expression classification model has been aimed at providing high accurate psoriasis prediction with analysis of quality gene expression profiles by efficient pre-processing methods.

Table 1 lists the notations used in this study.

The proposed pre-processing method has two major steps: REFS normalization and GWKNN imputation process. Then the features/gene signatures are extracted and the top features/gene signatures are selected using mutual information measures. Lastly, the carefully chosen gene signatures are fed as input to the hybrid classifier of the MKUSVM-RBFELM for psoriasis identification. The functional illustration of the proposed approach is given in Fig. 1.

3.1 Data description

The gene expression profiles related to the psoriasis treatment were utilized from the publically available psoriasis whole blood transcriptome dataset named GSE55201 created using the affymetrix human genome U133 plus (microarray) with platform ID GPL570 [25]. This open-source dataset is available in the GEO repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55201 [23]). It contains samples collected from 81 patients who have volunteered for interleukin IL-7 treatments. This dataset includes gene expression data of 30 normal people, 44 psoriasis patients at baseline and 7 psoriasis patients after two weeks of treatment. The differently expressed genes (DEGs) amongst the 30 healthy and 44 diseased baseline patients were determined using the data analysis models. This dataset has been selected for evaluation because of the number of gene signatures (54675 probe features) exceeding the expected range.

3.2 REFS normalization method

The proposed REFS normalization process

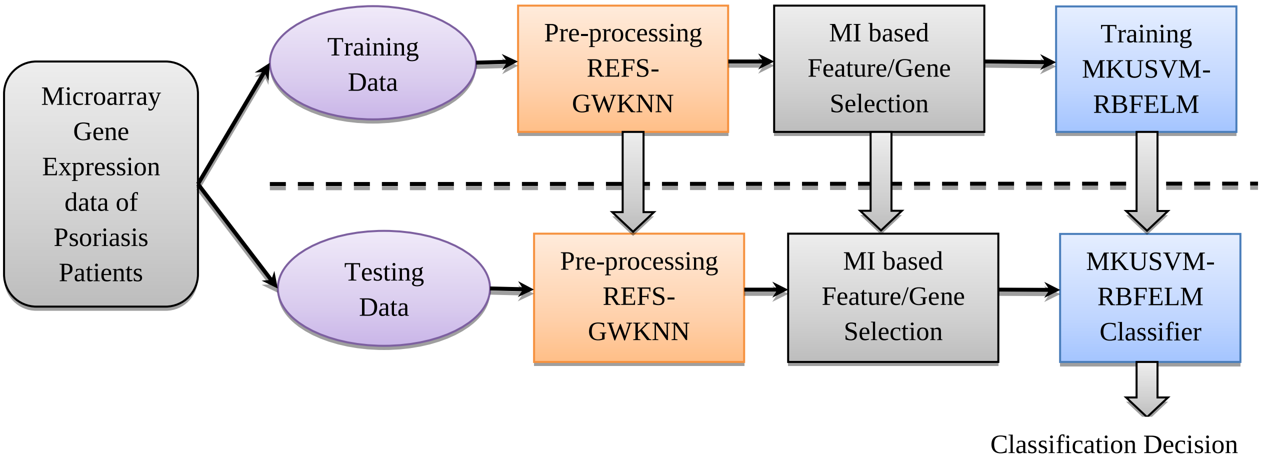
Microarray Gene Expression data of Psoriasis Patients
Training Data
Pre-processing REFS-GWKNN
MI based Feature/Gene Selection
Training MKUSVM-RBFELM
Testing Data
Pre-processing REFS-GWKNN
MI based Feature/Gene Selection
MKUSVM-RBFELM Classifier
Classification Decision

Figure. 1 Representative illustration of the proposed approach

integrates the relative entropy score functions into the gene fuzzy score computation. The fuzzy score is generally a positive value but in some cases, most of the genes are nearer to the negative fuzzy values. Relative entropy is a natural generalization score function that considers two difference constraints, one between the degree of membership and the degree of hesitance while the other between the degree of membership and the degree of no-membership. This relative entropy uses degree of objectivity for reducing the error rates.

In REFS, the raw gene expression matrix of each gene expression profile is transformed based on the rank values of the genes within each microarray. It uses two quantile thresholds namely θ_1 and θ_2 for assigning a fuzzy score to each gene. The genes with ranks below the θ_2 threshold value are reduced to zero scores while the genes with values above the θ_1 threshold values are assigned a score of 1. The intermediate valued genes are assigned a score between 0 and 1 based on their rank. Let $r(g_i, p_j)$ denote the rank of gene expression of a gene g_i in patient p_j and $q(p_j, \theta)$ denote the rank equivalent to the upper quantile threshold θ_1 of gene expression in patient p_j.

The fuzzy score $s(g_i, p_j)$ allocated to a gene g_i in patient p_j can be computed as

$$s(g_i, p_j) = \begin{cases} 1, & if\ q(p_j, \theta_1) < r(g_i, p_j) \\ \frac{r(g_i, p_j) - q(p_j, \theta_2)}{q(p_j, \theta_1) - q(p_j, \theta_2)}, & if\ q(p_j, \theta_1) \geq r(g_i, p_j) \geq q(p_j, \theta_2) \\ 0, & if\ r(g_i, p_j) > q(p_j, \theta_2) \end{cases} \tag{1}$$

This equation denotes the fuzzy values assigned to the gene based on the rank of the gene expression. However, as described above, this score function can also result in zero fuzzy values when the noise in the genes is very high. In such cases, the normalized values will be nearer to zero and do not offer much information on the classification of psoriasis skin genes. So to overcome this limitation, the relative entropy score function is added. It is computed as the expected score function of these fuzzy values. It is achieved by applying natural generalization.

Let $A = s(g_A, p_A)$ be a fuzzy value of a gene g_A in patient p_A and rank $r(g_A, p_A)$. The relative entropy score function δ_A for this gene is given as

$$\delta_A = \mu_i - v_i + \left(\lambda \mu_i \log \frac{\mu_i}{v_i} + (1 - \lambda) \mu_i \log \frac{\mu_i}{v_i}\right) \pi_i \tag{2}$$

μ_i denotes the supporting conditional variable, v_i denotes the opposing conditional variable and π_i denotes the non-available group conditional variable. λ denotes the risk preference index which determines the negative, neutral or positive risks associated with each sample.

To apply the relative entropy score function, the non-decreasing property of the general relative entropy must be proved with respect to the membership grade values. By applying the fuzzy scores to the relative entropy score function, generality is gained with the effective preservation of the important genes. The REFS function is given by

$$\delta_A = \frac{r(g_i, p_j)}{q(p_j, \theta_1)} + \left[\mu_i - v_i + \left(\lambda \mu_i \log \frac{\mu_i}{v_i} + (1 - \lambda) \mu_i \log \frac{\mu_i}{v_i}\right) \pi_i\right] \tag{3}$$

This function of the proposed REFS is effective in normalizing the gene expression profiles of psoriasis skin gene datasets.

3.3 GWKNN imputation method

This proposed imputation method is developed by integrating the genetic operators to the weighted KNN. The first step in this model is to employ the weighted KNN to abstract the K instances of gene data nearest to the missing value. In the next step, the genetic operators are used to predict the imputation value based on the extracted K instances. The K instances are selected using a distance-based similarity measure namely normalized Euclidean distance which is expressed as

$$distance(p, q) = \sqrt{\sum_{i=1}^{n} \frac{(q_1 - p_1)^2}{r_i^2}} \tag{4}$$

Here p, q is the real-valued gene expression instances and r_i denotes the range of the i-th feature. The usage of weighted KNN will increase the prominence of the instances that are closer to the missing values. Firstly, weighted KNN extracts the K nearest instances based on the distance value computed using Eq. (4). The weighted KNN will be used to obtain the $X_{i,j}^{K}$ which includes the K nearest instances to the reference instances $R_{i,j}$. The corresponding distances $D_{i,j}^{K}$ and the distance-based weights $W_{i,j}^{K}$ are also computed. The value of K is computed as

$$K \leftarrow \min(\max\left|J_{i,j}\right|, \left[\frac{|I_{i,j}|}{4}\right], \left|I_{i,j}\right|) \quad (5)$$

Here $I_{i,j}$ denotes the instances and $J_{i,j}$ denotes the feature indexes of the whole data subset $X_{i,j}$. $\left[\frac{|I_{i,j}|}{4}\right]$ refers to the one-fourth of the number of retrieved instances which is used as the upper bound of K. Using the distance computation in Eq. (4), the values of $D_{i,j}^{K}$ and $W_{i,j}^{K}$ are calculated as

$$D_{i,j}^{K}[k] = distance\left(X_{i,j}^{K}[k], V_{i,j}\right), \quad k = 1,2,3, \ldots., K \quad (6)$$

$$W_{i,j}^{K} = \frac{D_{i,j}^{K}[k]}{D_{i,j}}, \qquad k = 1,2,3, \ldots, K \quad (7)$$

Here $V_{i,j}$ is the new reference instances of non-missing values and $D_{i,j} = \max_{k=1,2,\ldots,K} D_{i,j}^{K}[k]$.

Then the genetic programming is used to build the imputation process model with these K instances as input instances for estimating the missing values. For performing this operation, the GP regression function $\left\{G_g\right\}_{g=1}^{N}$ with the target variable (missing value) as $T_{i,j}$. The weighted relative squared error (WRSE) is computed to be the fitness function

$$WRSE_{i,j}^{K} = \frac{\sum_{k=1}^{K} \frac{1}{W_{i,j}^{K}} (Y_{i,j}[k] - T_{i,j}[k])^2}{\sum_{k=1}^{K} (T_{i,j}[k] - \hat{T}_{i,j})^2} \quad (8)$$

Using this fitness function, the temporary and final best solutions for G_g^{tmp} is obtained through iterative ranking. These obtained results in this process will help in replacing the missing values.

3.4 Feature selection using mutual information

Feature selection is the process of choosing the most informative features or gene signatures while avoiding the less informative and irrelevant features. In this study, the wrapper method is utilized by estimating the mutual information (MI) for the available pairs of gene signatures. MI is computed for each feature subset and the process will be terminated if the maximum MI value is obtained. MI is estimated as the amount of information via the reduction in entropy which measures the diversity in the attributes and helps in obtaining the impurity of information to quantify the uncertainty of the given variables. 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.

$$H(y) = -p(R) \log p(R) - p(\bar{R}) \log p(\bar{R}) \quad (9)$$

Here $(R, \bar{R})$ denotes the possible classes-relevant and irrelevant, $p(R)$ indicate the probability of the sample being $y \in (R)$ and $p(\bar{R})$ represent 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,

$$H(Y|X) = \sum_{x \in X} p_x(x)\, H(Y|X = x) = \sum_{x \in X} \sum_{y \in Y} p_{xy}(x, y) \log p_y(y|x) \quad (10)$$

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 value of entropy is obtained when all the samples belong to the same class. MI of y can be computed using the entropy and conditional entropy from a feature x as

$$IG(y|x) = H(y) - H(y|x) \quad (11)$$

Larger IG defines the higher discriminative power for the decision process and determines the relevance of the gene signatures with respect to the classification problem.

3.5 MKUSVM

Universum samples are included in the SVM so that the classifier learns the patterns of unknown classes and also the known classes effectively with prior knowledge. USVM constructs the data-dependent architecture of SVM based on the set of tolerable functions using the universum samples. It is more appropriate to obtain the set of universum samples for the SVM learning process instead of defining the data distributions explicitly. Since the universum samples do not belong to any of the pre-defined classes, the MKUSVM hyper-plane will fall inside the margin borders determined by C due to the usage of the maximal margin procedure. Therefore, the MKUSVM must utilize a maximal soft-margin procedure and maximum number of universum samples that are distributed around the hyper-plane. The procedure of MKUSVM follows the process of SVM with the universum definitions.

A training set with given Universum samples is defined as

$$S = \{(x_1, y_1), \dots .(x_s, y_s)\} \cup \{x_1^*, x_2^*, \dots, x_u^*\} \quad (12)$$

Here $x_j^* \in R^n, j = 1,2,\dots,u$ denote the universum samples in R^n search space, $x_i \in R^n, i = 1,2,\dots,s$ and $y_i \in \{1,-1\}$ for binary classification and $y_i \in R^n, i = 1,2,\dots,s$ 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 MKUSVM with the maximal soft-margin procedure is given as

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C_t \sum_{i=1}^{s} \xi_i + C_t \sum_{t=1}^{u} (\psi_t + \psi_t^*) \quad (13)$$

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

Here C_t denotes the margin parameter or penalty parameter of SVM, C_u denotes the margin parameter of MKUSVM, ψ_t, ψ_t^* denotes the slack variables of MKUSVM and ε represents the in-sensitive loss function for Universum samples. Eq. (13) of MKUSVM maximizes the margin between the classifying hyper-planes and also maximizes the amount of Universum samples to be distributed around the hyper-plane. If $C_u = 0$, Eq. (13) will be equivalent to the standard SVM equations

$$\min_{\alpha,\mu,v} \frac{1}{2} \sum_{i=1}^{s} \sum_{j=1}^{s} y_i y_j \alpha_i \alpha_j \; K(x_i, x_j) + \frac{1}{2} \sum_{t=1}^{u} \sum_{z=1}^{u} (\mu_t - v_t)(\mu_z - v_z) K(x_t^*, x_z^*) + \sum_{i=1}^{s} \sum_{t=1}^{u} y_i \alpha_i (\mu_t - v_t) K(x_i, x_t^*) - \sum_{i=1}^{s} \alpha_i + \varepsilon \sum_{t=1}^{u} (\mu_t + v_t) \quad (14)$$

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

Here μ_i and v_i are Lagrangian multipliers similar to α_i. Instead of selecting the linear kernel, this MKUSVM selects the appropriate kernel function $K(x_i, x_j)$ from the multiple kernels namely linear, polynomial, RBF and Sigmoid Tanh kernels. By using these equations, $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_s^*)^S$, $\mu^* = (\mu_1^*, \mu_2^*, \dots, \mu_u^*)^S$ and $v^* = (v_1^*, v_2^*, \dots, v_u^*)^S$ are obtained.

Then the optimal classifying hyper-plane with the Universum prior knowledge is estimated as

$$g(x) = \sum_{i=1}^{s} y_i \alpha_i^* \; K(x_i, x) - \sum_{t=1}^{u} (v_t^* - \mu_t^*) K(x_t^*, x) + b^* \quad (15)$$

$$b^* = y_i - \sum_{i=1}^{s} y_i \alpha_i^* \; K(x_i, x_j) + \sum_{t=1}^{u} (v_t^* - \mu_t^*) K(x_t^*, x_j) \quad (16)$$

3.6 RBFELM

The proposed RBFELM uses the RBF kernel for optimal estimation of the weights and bias. The input features are mapped to the hidden layer H using the mapping function $H = a(Wx + b)$ where W denotes the input weight matrix, $a(.)$ denote the activation function and b represent the bias vector. The hidden layer is plotted into the remodelled input vector $\hat{x} = a(WH + b)$. The training process of the HH-ELM parameters is performed by minimizing the remodelling errors between the actual input and encoded outcomes. For the N training features with input X and output O with varying dimensions, the estimated function can be learned through the computation of the output weights. The training stage performs random mapping and least-squares constraints solving. Random mapping forms the hidden layer with neurons mapped by RBF function

$$\sum_{i=1}^{N} \gamma_i \, \theta_i(x_j) = o_j \quad (17)$$

Here, γ_i denotes the output weight vector of the i-th kernel, θ_i denotes the output of kernel and o_j denotes the output of the hidden layer. In this hidden layer, the output vector is modelled as $H(x) = R^{N \times \xi}$, where r is the dimension of variables and ξ denotes the number of hidden nodes. The output can be modelled as

$$\hat{o}_n = H(x_n)\gamma, \quad n = 1,2, \dots N \quad (18)$$

Here γ is the output weight of the hidden layer obtained by reducing the cost function Γ_{ELM}

$$\min_{\gamma \in R^{H \times r}} \Gamma_{ELM} = \|O - \hat{O}\|^2 = \|O - H\gamma\|^2 \quad (19)$$

The output weight can be modelled using least-squares constraints solving to obtain the modified output weights γ'.

$$\gamma' = H^{\dagger} O \quad (20)$$

Where $H^{\dagger}$ denote the moore–penrose generalized inverse of H.

Table 2. Comparison of pre-processing methods

Normalization methods			MVI methods		
Methods	**SS**	**p-value**	**Methods**	**PC**	**p-value**
GFS	0.811	0.75	**VIPER**	0.8241	0.7692
FSQN	0.835	0.72	**BN**	0.8341	0.7810
bayNorm	0.803	0.81	**WEDGE**	0.8034	0.7214
Sanity	0.830	0.79	**JFCM**	0.8238	0.7571
APR	0.819	0.78	**IKNN**	0.8412	0.7526
REFS	**0.867**	**0.84**	**GWKNN**	**0.8661**	**0.7961**

Table 3. Comparison of psoriasis classification methods

Method	A (%)	P (%)	R (%)	F (%)	MCC	T (s)
NB	79.17	80.18	80	80.09	0.654	0.549
ANN	83.33	84.52	83.33	83.92	0.714	0.501
SVM	87.5	88.64	86.96	87.79	0.777	0.440
ELM	90.67	91.58	89.91	90.74	0.812	0.473
MKUSVM-RBFELM	91.67	92.58	90.94	91.75	0.845	0.373

3.7 MKUSVM-RBFELM classifier model

The hybrid classifier model of MKUSVM-RBFELM is built by a simple process of selecting the optimal number of RBFELM nodes using MKUSVM. Firstly, the selected gene signatures by the MI method are fed as input to the MKUSVM and RBFELM separately to obtain individual classification results. Then the results of MKUSVM are used as targets for deriving the mean error rate of the trained RBFELM model. These training errors are projected to the MKUSVM dimensions and the hidden nodes that are inactive are pruned out by the MKUSVM to obtain the final classification results.

4. Results and discussion

The proposed MKUSVM-RBFELM classifier with the REFS-GWKNN pre-processing method is evaluated over the psoriasis gene expression dataset obtained from the GEO repository. The evaluations are conducted using the MATLAB tool (R2016b version 9.1) installed on the computer with specifications of i5 processor, 8GB RAM and 512GB SSD with Windows 10 operating system. The performance of all three proposed methods is evaluated and compared with state-of-the-art models.

4.1 Evaluation of pre-processing methods

The proposed REFS method is evaluated and compared with existing GFS [11], FSQN [12], bayNorm [13], Sanity [14] and APR [15]. The comparisons are made in terms of the Silhouette score (SS) and p-value. The proposed GWKNN imputation method is evaluated and compared with existing VIPER [16], BN [17], WEDGE [18], JFCM [19] and IKNN [20]. The comparisons are made in terms of pearson correlation (PC) and p-value. Table 2 shows the obtained results for REFS-GWKNN against the existing methods over the GSE55201 dataset.

From the results obtained in Table 2, it is concluded that the suggested REFS has better performance than the other implemented methods. REFS achieved 4.8 %, 3.7 %, 6.4 %, 3.2 % and 5.6 % higher silhouette scores than APR, sanity, bayNorm, FSQN and GFS methods. Similarly, it has achieved 6 %, 5 %, 3 %, 12 % and 9 % higher p-values than the existing APR, sanity, bayNorm, FSQN and GFS methods. This improvement is attributed to the use of the relative entropy score function to the gene fuzzy scores.

It is also concluded that the GWKNN has better performance than the implemented existing methods. GWKNN imputation method has achieved 2.49 %, 4.23 %, 6.27 %, 3.2 % and 4.2 % higher pearson correlation than the IKNN, JFCM-VQNNI, WEDGE, BN and VIPER methods. Similarly, it has achieved 4.35 %, 3.9 %, 7.47 %, 1.51 % and 2.69 % higher p-values than the existing IKNN, JFCM-VQNNI, WEDGE, BN and VIPER methods. This better performance of the GWKNN imputation method is because of the use of benefits from genetic programming in selecting the features of the KNN imputation method which has increased the effectiveness of missing values prediction.

4.2 Evaluation of classifier methods

The proposed MKUSVM-RBFELM classifier method is evaluated and compared with existing NB [21], ANN [22], SVM [23] and ELM [24]. The comparisons are made in terms of accuracy (A), precision (P), recall (R), f-measure (F) and mathew correlation coefficient (MCC). The processing time (T) for each classifier is also estimated in seconds (s). Table 3 shows the obtained results for MKUSVM-RBFELM and other classifier methods for psoriasis identification from the GSE55201 dataset.

For a fair comparison, the existing classifiers are implemented similar to the proposed MKUSVM-RBFELM under similar environmental settings over the GSE55201 dataset. From the results obtained in Table 3, it is concluded that the MKUSVM-RBFELM has improved performance than the other implemented classifier methods. The MI based feature selection method has removed 5088 features from the total 54675 features and the remaining features are utilized for classification. MKUSVM-RBFELM achieved 1 %, 4.17 %, 8.34 % and 12.5 % higher accuracy than the existing ELM, SVM, ANN and NB classifiers. It has also achieved 1 %, 3.94 %, 8.06 % and 12.4 % higher precision, 1.03 %, 3.98 %, 7.61 % and 10.94 % higher recall, and 1.01 %, 3.96 %, 7.83 % and 11.66 % higher f-measure, than the existing ELM, SVM, ANN and NB classifiers, respectively. In terms of processing time, the proposed MKUSVM-RBFELM classifier consumed 0.373 seconds for the GSE55201 dataset which is 21.14 %, 15.25 %, 25.6 % and 32.1 % lesser than the time consumed by the existing ELM, SVM, ANN and NB classifiers. MKUSVM-RBFELM has achieved a higher MCC value of 0.8452 indicating the perfect classification. The MCC values of MKUSVM-RBFELM is 3.3 %, 6.78 %, 13.09 % and 19.05 % higher than the existing ELM, SVM, ANN and NB classifiers. This better performance of the proposed classifier is because of the integration of the two classifier models with minimized error rates.

5. Conclusion

This paper presented an efficient pre-processing method using REFS normalization and the GWKNN imputation method for increasing the proficiency of psoriasis classification. In addition, an advanced classification model of MKUSVM-RBFELM has been developed. This proposed approach of pre-processing has improved the data quality and reduced the negative impacts of noise and outliers. The proposed REFS-GWKNN pre-processing and MKUSVM-RBFELM classifier models are evaluated using a publically available gene expression dataset for psoriasis classification. The experimental results were compared with the state-of-the-art methods and it implied that the suggested approaches have significantly improved the identification of psoriasis disease from the gene expression data with 91.67 % accuracy, 92.58 % precision, 90.94 % recall, 91.75 % F-measure, and 0.8452 MCC values with a reduced processing time of 0.373 seconds. In future, the possibility of evaluating multi-source gene expression data will be investigated. Also, the gene signature (feature) selection and classification models will be examined in parallel to the fast-rising deep learning methods.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

This work is a contribution of the authors: “Conceptualization, Ganapathy Kavitha and Kandasamy Vanitha; methodology, Ganapathy Kavitha; software, Ganapathy Kavitha; validation, Ganapathy Kavitha and Kandasamy Vanitha; formal analysis, Ganapathy Kavitha; writing—original draft preparation, Ganapathy Kavitha; writing—review and editing, Ganapathy Kavitha and Kandasamy Vanitha.

References

[1] G. Kavitha and K. Vanitha, “A Study of Dermatological Diseases using Classification Approaches”, *International Journal for Science and Advance Research in Technology*, Vol. 7, No. 11, pp. 104-109, 2021.

[2] V. Harris, C. Jackson, and A. Cooper, “Review of toxic epidermal necrolysis”, *International Journal of Molecular Sciences*, Vol. 17, No. 12, pp. 2135-2146, 2016.

[3] M. Z. Handler and R. A. Schwartz, “Staphylococcal scalded skin syndrome: diagnosis and management in children and adults”, *Journal of the European Academy of Dermatology and Venereology*, Vol. 28, No. 11, pp. 1418-1423, 2014.

[4] M. A. Lowes, M. S. Farinas, and J. G. Krueger, “Immunology of psoriasis”, *Annual Review of Immunology*, Vol. 32, No. 1, pp. 227-255, 2014.

[5] A. Rendon and K. Schäkel, “Psoriasis pathogenesis and treatment”, *International Journal of Molecular Sciences*, Vol. 20, No. 6,

pp. 1475-1503, 2019.

[6] J. T. Elder, A. T. Bruce, J. E. Gudjonsson, A. Johnston, P. E. Stuart, T. Tejasvi, and R. P. Nair, “Molecular dissection of psoriasis: integrating genetics and biology”, *Journal of Investigative Dermatology*, Vol. 130, No. 5, pp. 1213-1226, 2010.

[7] K. Yu, M. N. Syed, E. Bernardis, and J. M. Gelfand, “Machine learning applications in the evaluation and management of Psoriasis: A systematic review”, *Journal of Psoriasis and Psoriatic Arthritis*, Vol. 5, No. 4, pp. 147-159, 2020.

[8] O. Toka and M. Çetin, “Imputation and deletion methods under the presence of missing values and outliers: a comparative study”, *Gazi University Journal of Science*, Vol. 29, No. 4, pp. 799-809, 2016.

[9] S. Chan, V. Reddy, B. Myers, Q. Thibodeaux, N. Brownstone, and W. Liao, “Machine learning in dermatology: current applications, opportunities, and limitations”, *Dermatology and Therapy*, Vol. 10, No. 3, pp. 365-386, 2020.

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

[11] A. Belorkar and L. Wong, “GFS: fuzzy pre-processing for effective gene expression analysis”, *BMC Bioinformatics*, Vol. 17, No. 17, pp. 169-184, 2016.

[12] J. M. Franks, G. Cai, and M. L. Whitfield, “Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data”, *Bioinformatics*, Vol. 34, No. 11, pp. 1868-1874, 2018.

[13] W. Tang, F. Bertaux, P. Thomas, C. Stefanelli, M. Saint, S. Marguerat, and V. Shahrezaei, “bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data”, *Bioinformatics*, Vol. 36, No. 4, pp. 1174-1181, 2020.

[14] J. Breda, M. Zavolan, and E. V. Nimwegen, “Bayesian inference of gene expression states from single-cell RNA-seq data”, *Nature Biotechnology*, Vol. 39, No. 8, pp. 1008-1016, 2021.

[15] J. Lause, P. Berens, and D. Kobak, “Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data”, *Genome Biology*, Vol. 22, No. 1, pp. 1-20, 2021.

[16] M. Chen and X. Zhou, “VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies”, *Genome Biology*, Vol. 19, No. 1, pp. 1-15, 2018.

[17] R. Howey, A. D. Clark, N. Naamane, L. N. Reynard, A. G. Pratt, and H. J. Cordell, “A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships”, *PLoS Genetics*, Vol. 17, No. 9, pp. 1-28, 2021.

[18] Y. Hu, B. Li, W. Zhang, N. Liu, P. Cai, F. Chen, and K. Qu, “WEDGE: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition”, *Briefings in Bioinformatics*, Vol. 22, No. 5, pp. 1-32, 2021.

[19] D. Li, H. Zhang, T. Li, A. Bouras, X. Yu, and T. Wang, “Hybrid missing value imputation algorithms using fuzzy c-means and vaguely quantified rough sets”, *IEEE Transactions on Fuzzy Systems*, Vol. 30, No. 3, pp. 1-15, 2021.

[20] P. Keerin, and T. Boongoen, “Improved KNN Imputation for Missing Values in Gene Expression Data”, *CMC-Computers Materials and Continua*, Vol. 70, No. 2, pp. 4009-4025, 2022.

[21] M. Ahmed, M. Shahjaman, M. Rana, M. Mollah, and N. Haque, “Robustification of Naïve Bayes classifier and its application for microarray gene expression data analysis”, *BioMed Research International*, Vol. 2017, No. 1, pp. 1-17, 2017.

[22] K. Cahyaningrum and W. Astuti, “Microarray gene expression classification for cancer detection using artificial neural networks and genetic algorithm hybrid intelligence”, In: *Proc. of 2020 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, pp. 1-7, 2020.

[23] L. Tapak, S. Afshar, M. Afrasiabi, M. K. Ghasemi, and P. Alirezaei, “Application of Genetic Algorithm-Based Support Vector Machine in Identification of Gene Expression Signatures for Psoriasis Classification: A Hybrid Model”, *BioMed Research International*, Vol. 2021, No. 1, pp. 1-10, 2021.

[24] K. Chatra, V. Kuppili, D. R. Edla, and A. K. Verma, “Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function”, *Medical and Biological Engineering and Computing*, Vol. 57, No. 12, pp. 2673-2682, 2019.

[25] C. Q. Wang, C. Q. M. S. Farinas, K. E. Nograles, C. A. Mimoso, D. Shrom, E. R. Dow, and J. G. Krueger, “IL-17 induces

inflammation-associated gene products in blood monocytes, and treatment with ixekizumab reduces their expression in psoriasis patient blood", *The Journal of Investigative Dermatology*, Vol. 134, No. 12, pp. 2990, 2014.