Received: December 18, 2021. Revised: February 14, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Weighted Fuzzy Score Normalization and Bayesian Independent Principal Component Analysis Imputation for Breast Cancer Gene Expression Analysis

Velusamy Murugesan[1]* **Ponnunaicker Balamurugan[2]**

[1]*Department of Computer Science, VLB Janakiammal College of Arts and Science,*
Coimbatore, Tamil Nadu- 641042, India
[2]*Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu-641018, India*
* Corresponding author's Email: murugeshvlb2020@gmail.com

Abstract: Breast cancer classification through Gene expression analysis is very challenging due to the large variations and missing values problem. However, the traditional normalization and missing value imputation (MVI) methods perform poorly in the occurrence of significant batch effects, heterogeneity, artefacts, and low resolution. Therefore, an efficient pre-processing technique is presented in this paper using Weighted Fuzzy Score (WFS) and Bayesian Independent Principal Component Analysis (BIPCA) for upholding the quality of the expression analysis. In this proposed WFS-BIPCA technique, the large variations are reduced through sustainable transformations by WFS. BIPCA impute replaces the missing expression values without degrading the quality, consistency, and coherence of the output results without repeating all microarray experiments. The proposed WFS-BIPCA technique is evaluated with different classifiers over breast cancer gene expression datasets from Mendeley Data. Results showed that WFS-BIPCA with Support Vector Machine (SVM) classifier achieved high accuracy of 92.36%, 91.86%, 90.02% and 93.89% for BC-TCGA, GSE2034, GSE25066 and Simulation Datasets, respectively. Similarly, it achieved 91.66%, 96.15%, 89.32%, and 95.36% precision, 93.94%, 92.22%, 86.25%, and 97.48% recall, 92.79%, 94.14%, 87.76%, and 96.41% f-measure values and low processing time of 1.98, 0.96, 3.56 and 0.55 seconds for BC-TCGA, GSE2034, GSE25066 and Simulation Datasets, respectively.

Keywords: Gene expression analysis, Breast cancer, Normalization, Missing value imputation, Weighted fuzzy score, Bayesian independent principal component analysis.

1. Introduction

Microarray technology has developed into one of the most valuable tools in genetic and genome research studies [1]. Gene expression profiles produced using these tools are utilized in genetic analyses which are often collected from the gene clusters exhibiting variations or noises in specified expression values that include the gene cell status, growth stages, disease types, disease stages and response to certain interventions or diagnosis methods [2, 3]. After the collection process, the gathered data are mined for identifying the appropriate patterns of such variations relevant to the considered hypothesis under the specified time. Another major problem is the missing values in the gene expression data which has a similar negative impact on the downstream analysis [4]. Less than 1% of missing values is often inconsequential to the overall analysis and 1-5% of missing values can be easily controlled. Since, 90% of the gene expression data contains at least one missing value, up to 10% of missing values in expression data can be the maximum tolerable limit [5]. However, some data suffers from more than the tolerable rate and hence the quality of analysis becomes questionable. It is practically not feasible to achieve effective analysis without suitable normalization and missing value imputation techniques for the gene expression data.

The normalization techniques must be adequate to tackle all forms of noise and variations in the gene expression data [6]. The variations can occur due to various factors and processes. Firstly, the process of segregation and quantification of the Ribonucleic Acid (RNA) from genes might cause inaccuracies in the final measurements due to noise-related errors which are implied during the longer processing stages. Secondly, the changes in the experimental environment and settings including the system biases might lead to such variations in the gene expression data during the measurements in terms of batch effects [7]. Finally, the external factors including the manual errors also lead to variations in the gene expression measurements. All these factors must be considered by the normalization technique and it must provide effective recovery of meaningful biological information from the data without impacting non-noisy elements as well as eradicating the variations.

Similar to the variations, the missing value problem occurs due to artefacts in microarray, noise or variations, image corruption, insufficient resolution, logical errors, hybridizing botches, negative feedback intensity of the backgrounds, etc. The gene spot having negative background intensity is also termed as missing value since the value completely mismatches with the expected values. This missing value problem will lead to inconsequential or wrong measurements, negatively impact the feature selection and classification processes. The algorithms such as Support Vector Machines (SVM), Singular Value Decomposition (SVD) [8], principal component analysis (PCA) [9] and independent component analysis (ICA) have been greatly affected by the missing value problem when employed in gene expression analysis process. The imputation of these missing values is performed by gathering the probability of informative genes and it is vital in ensuring the quality of the gene analysis at a low cost. It will also help detect the genes of specified targets for a particular class. Hence the MVI is very important to reduce the repeated experiments in trial-and-error methods performed to determine probable values for these missing elements.

To mitigate the performance degradation resulting from the expression data variations and missing values problems, previous studies suggested many pre-processing techniques. However, the sub-par performances of those methods have increased the need for efficient normalization and MVI methods. This paper presents an efficient pre-processing technique by proposing Weighted Fuzzy Score (WFS) and Bayesian Independent Principal Component Analysis (BIPCA). WFS based normalization uses weighted fuzzy scores to transform the gene expression data values without large variations. BIPCA based MVI is presented by combining the Bayesian theory with a hybrid analysis model of PCA and ICA to replace the missing values through the probability of informative genes. Evaluation of the proposed technique is performed using public gene expression datasets from Mendeley Data for breast cancer. The remainder of this article is structured as follows: related literature studies in Section 2, proposed pre-processing technique in Section 3, their evaluations and results in Section 4 and conclusions in Section 5.

2. Related works

Recent studies have presented various techniques for pre-processing with effective normalization and imputation processes. Yasrebi [10] used Z-score normalization techniques in breast cancer gene data for assessing the survival rate and risks. This standard method achieved higher performance improvement with better similarity values. However, the standard z-score normalization has limitations of assuming a normal distribution leading to unequal and skewed origin gene lines. Zhao et al. [11] utilized Quantile normalization for perfectly normalizing gene-expression datasets. However, the quantile normalization has drawbacks of a large number of undetected genes due to inconsistent median across the gene cells. Belorkar and Wong [12] developed Gene Fuzzy Score (GFS) as a pre-processing transformation method for the normalization task. GFS used the fuzzy score derived from rank values of gene expression and has reduced the batch effects and also increases the interpretability of the transformed outcomes without any negative impact on the sample size variation. Tang et al. [13] developed Bayesian Normalization (bayNorm) for normalization of single-cell RNA-sequencing data. This model preserved low false positive rates, but reduced AUC significantly. Borella et al. [14] proposed power-law Pareto distribution parameter estimate based normalization (PsiNorm). This model provided good trade-off between accuracy and scalability and also does not need a reference to normalize the new out-of-sample data. Though the performance is improved, there is a small increase in computational time.

De Silva and Perera [15] developed Evolutionary k-nearest neighbor (E-KNN) imputation for missing value problems in gene expression data. This E-KNN impute is an improved model of KNN impute in which the genetic algorithm is used to select the similarity matrix and k-parameter of the KNN impute algorithm. This improved model has high efficiency in solving the datasets with higher missing rates. Wang et al. [16]

resolved the missing value problem using Elastic-net Regularized Local Least Squares imputation (E-RLLSI) models of Least Squares imputation techniques. This technique provided highly accurate missing values imputation which is justified by the low RMSE values and the lesser time complexity. However, this technique has not considered global and local structural information in estimating the missing values. Zhu et al. [17] proposed an ensemble method of single imputation models for MVI. In this ensemble, bootstrapping is applied to predict the missing values with the predictions are weighted optimally using minimization of the cost function for reduced imputation error. This ensemble included multiple KNN related imputation models which provide higher imputation performance. Shahjaman et al. [18] developed robust iterative MVI approach called rMisbeta based on the minimum beta divergence method. This method reduced the misclassification error rate and computation time while also improved the accuracy, sensitivity and specificity. Dubey and Rasool [19] proposed an MVI approach considering the local similarity structure that predicts the missing data using similarity-based spectral clustering and weighted nearest neighbour (SSC-WNN). This method predicted the missing values accurately even when the dataset has varying dimensionality and characteristics. But the pattern of MVI based on neighbour similarity does not utilize the missing read counts.

These literature methods are helpful in improving the breast cancer detection performance of the classifier models such as SVM, Naïve Bayes, Decision tree, Artificial Neural Networks (ANN) [20], Random Forests (RF) [21], Extreme Learning Machine (ELM) [22] and Deep learning Convolutional Neural Networks (CNN) [23].

From the literature studies, it is understood that the existing normalization methods suffer from the limitations of random assumptions and inconsistent threshold determinations. Likewise, the imputation methods have limitations of handling higher missing rates due to the reduced computation abilities. Considering these limitations, the proposed approach developed two models for normalization and MVI that perform better than the existing methods.

3. Methodology

The proposed pre-processing method has two major steps: WFS and BIPCA. The breast cancer gene expression datasets from Mendeley Data are used for evaluations.

Table 1. Distribution of breast cancer gene expression datasets

Datasets	**Number of genes**	**Number of samples**		
		Total	**Class 1**	**Class 2**
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

3.1 Datasets

Publically available breast cancer gene expressions are obtained from Mendeley Data [24] (www.data.mendeley.com/datasets/v3cc2p38hb/1).

The main dataset contains four sub-datasets namely, BC-TCGA, GSE2034, GSE25066 and Simulation Data. BC-TCGA consists of 17,814 genes and 590 samples (including 61 normal tissue samples and 529 breast cancer tissue samples). GSE2034 includes 12,634 genes and 286 breast cancer samples (including 107 recurrence tumor samples and 179 no recurrence samples). GSE25066 has 492 breast cancer samples available (including 100 pathologic complete response (PCR) samples and 392 residual diseases (RD) samples) and 12,634 genes. Simulation Data includes 100 positive samples and 100 negative samples with 10,000 features, and each feature in SData follows normal distributions: $N(0, 0.1)$ and $N(0 \pm r, 0.1)$ for positive and negative samples, respectively, where $r \in [-0.125, 0.125$. The collected datasets contain independent pairs of micro-array expressions. Table 1 shows the description of the evaluation datasets. All the datasets were split into training and testing samples in the ratio of 7:3, i.e. 70% training samples and 30% testing samples.

3.2 Weighted fuzzy score based normalization

The proposed WFS normalization process is developed by integrating the Minkowski Weighted Score Functions to the gene fuzzy score computation [25]. Minkowski Score is a natural generalization of the expected score functions. The weighted function of this score is used to assign a better score for the fuzzy values to migrate them farther from the negative fuzzy values.

In WFS, the raw gene expression matrix of each gene expression profile is transformed based on the rank values of the genes within each microarray. As in GFS, this method 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 values 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 inpatient p_j and $q(p_j, \theta)$ denote the rank corresponding to the upper quantile threshold θ_1 of gene expression inpatient p_j. The fuzzy score $s(g_i, p_j)$ assigned to a gene g_i in patient p_j can be computed as

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

This equation denotes the fuzzy values assigned to the gene-based on the rank of the gene expression of a gene. 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 breast cancer genes. So to overcome this limitation, the Minkowski Weighted Score Function is added. Minkowski Score 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 Minkowski Score function δ_A for this gene is given as

$$\delta_A = \frac{t_A - f_A + 1}{2} \tag{2}$$

t_A denote the membership grade of gene A while f_A denote the non-membership grade of gene A.

To apply the Minkowski weighted score function, the non-decreasing property of the general Minkowski score function must be proved concerning the membership grade values. Let $A = (t_A, f_A)$ and $B = (t_B, f_B)$ denote two fuzzy values of genes g_A and g_B respectively. The normalized Minkowski distance can be defined as

$$D_n(A, B) = \left(\frac{|t_A - t_B|^n |f_A - f_B|^n}{2}\right)^{1/n} \tag{3}$$

Applying the normalized Minkowski distance between A and transformed A denoted as A^*,

$$D_n(A, A^*) = \left(\frac{t_A^n + |1 - f_A|^n}{2}\right)^{1/n} = s_n(A) \tag{4}$$

when applying $n = 1$,

$$s_1(A) = \frac{t_A - f_A + 1}{2} = \delta_A \tag{5}$$

when $n \geq 1$,

$$s_n(A) = s_n(t_A, f_A) = \left(\frac{t_A^n + |1 - f_A|^n}{2}\right)^{1/n} \tag{6}$$

Applying partial derivative function of $s_n(A)$, we get

$$\frac{\partial s_n}{\partial t_A} = \frac{t_A^{n-1}}{2n} \left(\frac{t_A^n + |1 - f_A|^n}{2}\right)^{\frac{1-n}{n}} \geq 0 \tag{7}$$

This concludes that the Minkowski score function $s_n(t_A, f_A)$ is non-decreasing concerning t_A.

Now applying the weights to the Minkowski score function, the generality is gained with effective preservation of the important gene properties. Minkowski weight function is given by

$$s_n^w(A) = \frac{1}{n}\left((1 - w)t_A^n + w|1 - f_A|^n\right)^{\frac{1-n}{n}} \tag{8}$$

If $n = 1$, the Eq. (8) becomes Hamming weighted score function of the fuzzy score, i.e.

$$s_1^w(A) = \left((1 - w)t_A + w|1 - f_A|\right) \tag{9}$$

If $n = 2$, the Eq. (8) becomes Euclidean weighted score function of the fuzzy score, i.e.

$$s_2^w(A) = \frac{1}{2}\left((1 - w)t_A^2 + w|1 - f_A|^2\right)^{-1/2} \tag{10}$$

To prove that this Minkowski weighted score function is suitable for the normalization process of WFS, the partial derivative is obtained to prove that it is non-decreasing with respect to t_A.

$$\begin{aligned} s_n^w(A) &= s_n^w(t_A, f_A) \\ &= \frac{1}{n}\left((1 - w)t_A^n + w|1 - f_A|^n\right)^{\frac{1-n}{n}} \end{aligned} \tag{11}$$

Applying partial derivative,

$$\begin{gathered} \frac{\partial s_n^w}{\partial t_A} = \frac{1}{n}(1 - w)t_A^{n-1} \\ \left((1 - w)t_A^n + w|1 - f_A|^{n(n-1)}\right)^{\frac{1-n}{n}} \geq 0 \end{gathered} \tag{12}$$

This concludes that the proposed WFS using the Minkowski weighted fuzzy score function is effective in normalizing the gene expression profiles of breast cancer datasets.

3.3 BIPCA based imputation

BIPCA impute is formed by integrating the Bayesian PCA and ICA methods. The integration of these algorithms solves the limitation of BPCA for extracting the irrelevant features and linear transformation leading to inappropriate imputation. Initially, the process is performed similarly to the BPCA imputation. Then the ICA is integrated to form the BIPCA impute. The BIPCA algorithm represents the D-dimensional gene expression vectors Y as a linear combination of K with $(K < D)$ principal axis vectors $\alpha_l, (1 \leq l \leq K)$ as

$$y = \sum_{l=1}^{K} x_l \alpha_l + \varepsilon \tag{13}$$

Here x_l denotes the factor score, α_l denote the principal vector and ε represent the residual error. The principal vectors are attained by calculating the eigenvalues and eigenvectors of the covariance matrix of the dataset Y. As there are missing values in the original matrix Y, the principal vectors are divided into two fragments as $\alpha = (\alpha^{obs}, \alpha^{miss})$, corresponding to the observed value and missing value, respectively. Factor scores $x = (x_1, x_2, \ldots, x_k)$ are achieved by reducing the residual error of the practical value.

$$\varepsilon = \left\| y^{obs} - \alpha^{obs} x \right\|^2 \tag{14}$$

As the factor scores x and residual error ε follow normal distributions, BIPCA utilizes probabilistic PCA to estimate the parameters α. Along with α, the values of eigenvalues μ and eigenvectors τ are obtained simultaneously to form the parameter set $\gamma = \{\alpha, \mu, \tau\}$. The missing values can be estimated from this parameter set, but limitations of PCA linear transformation leads to inappropriate imputation and convergence to local optima. To solve these problems, the ICA is integrated which reflect the internal structure of the gene expression data to reduce noise and missing values.

The standard BPCA used second-order statistics which led to the inappropriate imputation and hence the BPICA uses higher-order statistics to recover the statistically independent signal from the observations of an unknown linear mixture.

Let $X(n \times m)$ denote the centred data matrix formed by principal components and $C(n \times m)$ be the matrix containing the independent components. BIPCA problem is reduced by using $G(n \times n)$

$$X = GC \tag{15}$$

The mixing matrix G designates how the independent components of C are linearly joined to build X. It can be rewritten as

$$G = UX \tag{16}$$

Here $U(n \times n)$ denotes the reverse mixing matrix that describes the inverse process of mixing the independent components. In training BIPCA, it is very useful to whiten the data matrix X, i.e., to obtain $Cov(X) = I$. Therefore, $Cov(GC) = I$ and $GG^T = CC^T = I$ where I denote the unit matrix. The orthogonality of the matrix also allows some parameters to be assessed. If we can rewrite the standard PCA matrix, then

$$L^T = D^{-1} O^T X^T \tag{17}$$

Here L is a $n \times m$ matrix whose columns are uncorrelated, O is an $n \times n$ orthogonal matrix, and D is a $n \times n$ diagonal matrix. Since the columns of O are orthonormal, the rows of L^T are uncorrelated and have zero mean. To complete the whitening step, we can multiply L^T by $\sqrt{n-1}$, so that the rows of L^T have unit variance. The independent principal components obtained in this stage are estimated as

$$G = UL \tag{18}$$

Eq. (14) can be solved easily in BIPCA to find the missing values. By utilizing the G and factor scores x and α^{miss}, the missing part of the dataset is predicted when there is an ideal error i.e. no residual error occurred.

$$y^{miss} = \alpha^{miss} x. G \tag{19}$$

These relevant variables should have important weights in the loading vectors while other irrelevant or noisy variables should have very small weights. In this manner, the missing values are imputed.

3.4 Feature selection and classification

For feature selection or gene selection, the standard approach of mutual information is used. It computes the mutual dependence rate between the two features and the features with higher mutual information scores are used in classification. The SVM classifier is one of the most common and effective classifiers for categorizing the breast cancer gene expression profiles. Multi-class SVM is used in this study since the performance accuracy of this algorithm is significantly higher than the other methods.

4. Results and discussion

The proposed WFS-BIPCA pre-processing method is evaluated over the breast cancer datasets obtained from the Mendeley data repository [24]. The data descriptions are provided in section 3.1. The evaluations are conducted using the MATLAB tool (R2016b version 9.1). The evaluations are conducted in three stages. First, the WFS is implemented and compared with existing methods. Secondly, the BIPCA impute method is implemented and evaluated. Finally, the impact of the proposed pre-processing method on the classifier performance is evaluated.

4.1 Evaluation of normalization methods

The proposed WFS method is evaluated and compared with existing z-score normalization [10], quantile normalization [11], GFS [12], bayNorm [13] and PsiNorm [14]. The comparisons are made in terms of the Silhouette score and p-value. Table 2 shows the obtained results for WFS and other normalization methods over the testing datasets.

From the results obtained in Table 2, it is concluded that the proposed WFS has better performance than the implemented existing methods. For all four parts of the breast cancer dataset, the proposed WFS achieved higher values of the Silhouette score and p-value.

In the proposed model, the priors calculated within each individual, but across batches. This strategy allows for maintaining differences between individuals while minimizing batch effects. To quantify the result, a ratio is defined between the number of genes detected between each pair of batches within the same individual and the total number of genes. WFS also maintained differences between individuals. Efficient normalization and batch effect correction is expected to minimize false positive rate while maximizing accuracy values. Using WFS with the classifier has outperformed other methods in terms of correcting batch effects while maintaining meaningful biological information identified by the significant increase in the Silhouette score and p-value.

Table 2. Comparison of normalization methods

Silhouette score				
Method	**BC-TCGA**	**GSE 2034**	**GSE 25066**	**Simulation Data**
z-score	0.721	0.63	0.75	0.82
quantile	0.735	0.62	0.7665	0.836
GFS	0.811	0.75	0.812	0.845
bayNorm	0.813	0.825	0.8334	0.8588
PsiNorm	0.8082	0.79	0.8051	0.8367
WFS	0.8765	0.89	0.886	0.8755
p-value				
Method	**BC-TCGA**	**GSE 2034**	**GSE 25066**	**Simulation Data**
z-score	0.685	0.6100	0.689	0.6671
quantile	0.6675	0.6545	0.670	0.6921
GFS	0.7241	0.7234	0.7354	0.7345
bayNorm	0.7667	0.7410	0.7575	0.7456
PsiNorm	0.7533	0.7352	0.7441	0.7367
WFS	0.7907	0.7592	0.7963	0.750

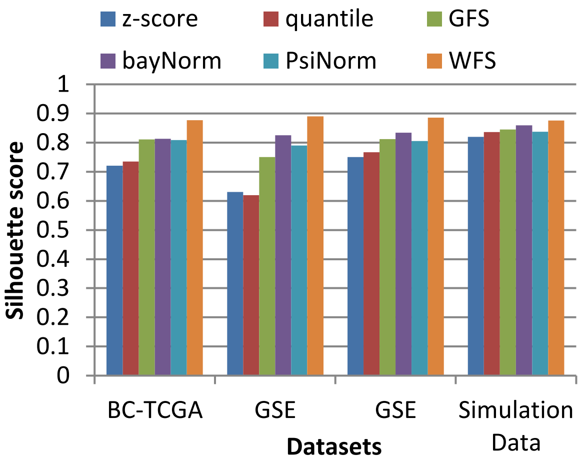

Figure. 1 Silhouette score of normalization methods

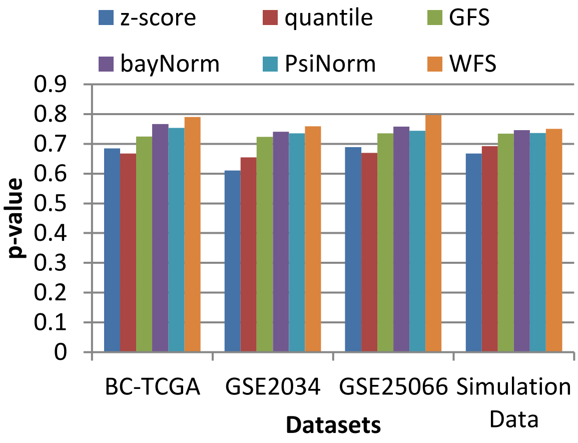

Figure. 2 P-values of normalization methods

Fig. 1 and 2 illustrate the silhouette score and p-value of the normalization methods. From Figure 1, it can be inferred that the proposed WFS normalization has achieved 6.83%, 6.35%, 6.5%, 14.15% and 15.55% higher silhouette scores than PsiNorm, bayNorm, GFS, Quantile and Z-score methods respectively, for the BC-TCGA dataset. Similarly, it has achieved 10%, 6.5%, 14%, 27% and 26% higher values for GSE2034 data, 8.09%, 5.26%, 7.4%, 12% and 13.6% higher values for GSE25066 data and 3.88%, 1.67%, 3.05%, 3.95% and 5.55% higher silhouette scores for the simulation data, respectively than the existing

PsiNorm, bayNorm, GFS, Quantile and Z-score methods. Likewise, in terms of p-value, the proposed WFS normalization has achieved 3.74%, 2.4%, 6.34%, 12.32% and 20.57% higher p-value than PsiNorm, bayNorm, GFS, Quantile and Z-score methods for the BC-TCGA dataset. It has also achieved 2.4%, 1.82%, 3.6%, 10.47% and 14.92% higher p-value for GSE2034 data, 5.22%, 3.88%, 6.09%, 12.63% and 10.7% higher p-value for GSE25066 data and 1.33%, 0.44%, 1.55%, 5.79% and 8.29% higher p-value for the simulation data, respectively than the existing PsiNorm, bayNorm, GFS, Quantile and Z-score methods. This improvement is attributed to the use of Minkowski weight scores for the gene fuzzy functions.

4.2 Evaluation of MVI methods

The proposed BIPCA impute method is evaluated and compared with existing E-KNN [15], E-RLLS [16], Ensemble imputation [17], rMisbeta [18] and SSC-WNN [19]. The missing data is maintained at high amount of 40% to 50%. The comparisons are made in terms of Pearson Correlation and p-value. Table 3 shows the obtained results for BIPCA and other imputation methods over the testing sets of the datasets.

From the results obtained in Table 3, it is concluded that the proposed BIPCA impute method has achieved better performance than the implemented existing methods. For all four parts of the breast cancer dataset, the proposed BIPCA achieved higher values of Pearson correlation and p-value.

The reason behind it is that as the number of components decreases, the participation of relevant and irrelevant genes also decreased, causing the prediction accuracy to be increased. The proposed technique's performance cannot be worst even if the number of components is kept high since the weighted function criterion does not allow the irrelevant gene to be considered in the imputation process.

Table 3. Comparison of MVI methods

Pearson Correlation				
Method	**BC-TCGA**	**GSE 2034**	**GSE 25066**	**Simulation Data**
E-KNN	0.8324	0.8018	0.7345	0.8100
E-RLLSI	0.8865	0.8032	0.7656	0.8356
Ensemble	0.8942	0.8675	0.7754	0.8543
rMisbeta	0.8885	0.8950	0.7816	0.8675
SSC-WNN	0.8976	0.8992	0.7934	0.8699
BIPCA	0.9012	0.9147	0.812	0.8813
p-value				
Method	**BC-TCGA**	**GSE 2034**	**GSE 25066**	**Simulation Data**
E-KNN	0.6545	0.6311	0.6012	0.6574
E-RLLSI	0.7011	0.6275	0.6363	0.7123
Ensemble	0.7329	0.6854	0.6901	0.7767
rMisbeta	0.7654	0.7232	0.7116	0.7876
SSC-WNN	0.7704	0.7210	0.7194	0.7914
BIPCA	0.7765	0.7354	0.7325	0.7961

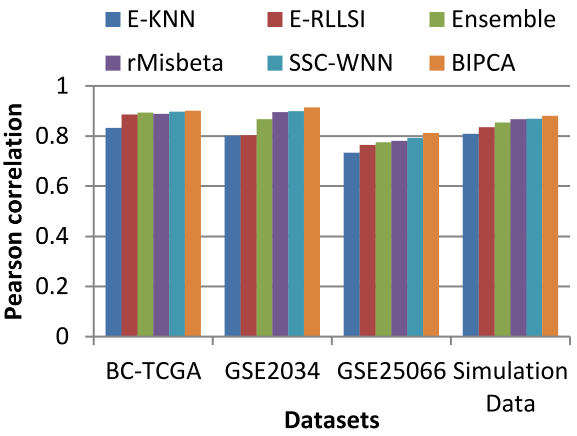

Figure. 3 Pearson correlation of MVI methods

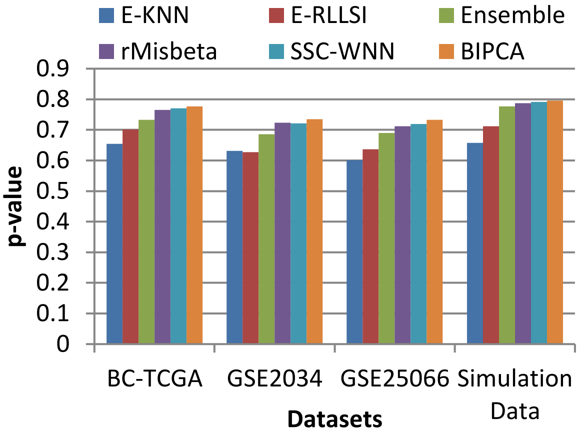

Figure. 4 P-values of MVI methods

Fig. 3 and 4 illustrate the Pearson correlation and p-value of the MVI methods. From Fig. 3, it can be inferred that the proposed BIPCA impute has achieved 0.36%, 1.27%, 0.7%, 1.47% and 6.88% higher Pearson correlation than SSC-WNN, rMisbeta, Ensemble, RLLS and E-KNN impute methods for the BC-TCGA dataset. Similarly, it has achieved 1.55%, 1.97%, 4.72%, 11.15% and 11.29% higher values for GSE2034 data, 1.78%, 2.96%, 3.6%, 4.58% and 7.67% higher values for GSE25066 data and 1.14%, 1.38%, 2.7%, 4.57% and 7.13% higher Pearson correlation for the simulation data, respectively than the existing SSC-WNN, rMisbeta, Ensemble, RLLS and E-KNN impute methods. Likewise, in terms of p-value, the proposed BIPCA impute has achieved 0.61%, 1.11%, 4.36%, 7.54% and 12.2% higher p-value than SSC-WNN, rMisbeta, Ensemble, RLLS and E-

KNN impute methods for the BC-TCGA dataset. It has also achieved 1.44%, 1.22%, 5%, 10.79% and 10.43% higher p-value for GSE2034 data, 1.31%, 2.09%, 4.24%, 9.62% and 13.13% higher p-value for GSE25066 data and 0.47%, 0.85%, 2%, 8.38% and 13.87% higher p-value for the simulation data, respectively than the existing SSC-WNN, rMisbeta, Ensemble, RLLS and E-KNN impute methods. This better performance of the proposed BIPCA method is because of the use of benefits from Bayesian models of PCA and ICA in the same model for the prediction of missing values.

4.3 Evaluation of the classification methods

The proposed WFS-BIPCA pre-processing method is implemented with mutual information feature selection and Multi-class SVM classifier. To evaluate the overall performance achieved by the proposed method, the classifier performance is evaluated and compared with existing classifier models of ANN [20], RF [21], Parallel feature selection based ELM (PFS-ELM) [22] and CNN [23]. The comparisons are made in terms of accuracy, precision, recall, f-measure and processing time. Table 4 shows the obtained results for the classifier methods over the testing sets of the datasets when utilizing the proposed WFS-BIPCA method.

Among the classifiers, SVM with proposed WFS-BIPCA has achieved 0.57%, 2.01%, 2.37%, and 1.29% higher accuracy than the CNN, PFS-ELM, RF and ANN classifier methods for the BC-TCGA dataset. Similarly, it has achieved 1.95%, 7.68%, 8.5% and 4.19% higher accuracy for GSE2034 data, 0.52%, 2.35%, 1.14% and 0.8% higher accuracy for GSE25066 data and 1.22%, 3.75%, 3.23% and 2.47% higher accuracy for the simulation data, respectively than the CNN, PFS-ELM, RF and ANN classifier methods. Likewise, in terms of precision, recall, and f-measure, the proposed WFS-BIPCA increased the performance of the SVM classifier by 1% to 20% than the CNN, PFS-ELM, RF and ANN classifier methods.

In terms of processing time, the SVM classifier models with the proposed WFS-BIPCA has achieved 0.34, 0.14, 0.67, and 0.9 seconds lesser time than the CNN, PFS-ELM, RF and ANN classifier methods for the BC-TCGA dataset. It has reduced the processing time by 0.46, 0.25, 0.49 and 0.78 seconds for GSE2034 data, 0.6, 0.42, 0.55 and 1.0 seconds for GSE25066 data and 0.175, 0.137, 0.104 and 0.066 seconds for the simulation data, respectively than the CNN, PFS-ELM, RF and ANN classifier methods.

Table 4. Performance improvement of Classifier results using proposed WFS-BIPCA method

Method	**BC-TCGA**	**GSE 2034**	**GSE 25066**	**Simulation Data**
Accuracy (%)				
ANN	92.17	87.67	89.22	91.42
RF	91.09	83.36	88.88	90.66
PFS-ELM	91.45	84.18	87.67	90.14
CNN	92.89	89.91	89.50	92.67
SVM	93.46	91.86	90.02	93.89
Precision (%)				
ANN	87.75	94.56	86.50	94.93
RF	83.36	89.89	84	94.87
PFS-ELM	85.12	92.35	85.67	94.81
CNN	89.39	95.27	87.70	95.08
SVM	91.66	96.15	89.32	95.36
Recall (%)				
ANN	92.05	91.80	85.92	96.67
RF	91.7	91.91	85.87	96.34
PFS-ELM	91.88	90.16	85.99	95.82
CNN	92.89	91.92	86.18	96.99
SVM	93.94	92.22	86.25	97.48
F-measure (%)				
ANN	89.85	93.16	86.21	95.79
RF	87.33	90.89	84.92	95.60
PFS-ELM	88.37	91.24	85.83	95.31
CNN	91.11	93.57	86.93	96.03
SVM	92.79	94.14	87.76	96.41
Processing time (seconds)				
ANN	2.88	1.74	4.56	0.616
RF	2.65	1.45	4.11	0.654
PFS-ELM	2.12	1.21	3.98	0.687
CNN	2.32	1.40	4.16	0.725
SVM	1.98	0.96	3.56	0.55

From the results obtained in Table 4, it is concluded that the proposed WFS-BIPCA method has achieved better performance for the classifiers. For all four parts of the breast cancer dataset, the SVM classifier with mutual information (MI) gene selection and the proposed WFS-BIPCA achieved higher values of accuracy, precision, recall, and f-measure and low processing time. This better performance of the MI gene selection and SVM classifier is because of the use of effective pre-processing methods in the form of WFS-BIPCA.

This experimental study shows that the importance of data normalization for improving data quality and subsequently the performance of machine learning classifiers has been improved with the utilization of WFS normalization. Also, BIPCA imputation with the tested methods improves classification accuracy when compared to classification without imputation. Although the results show that there is no universally best imputation method, BIPCA imputation is shown to give the best results for all the compared classifiers with SVM classifier outperforming for datasets with high amount (i.e., 40% and 50%) of missing data.

5. Conclusion

This paper was aimed at developing an efficient pre-processing approach. Two main stages of pre-processing are performed using hybrid techniques. First, the data normalization is performed using WFS normalization. This approach is an improved fuzzification process in which the weight parameters are included in the gene fuzzy score. This approach is intended to reduce the skewness and reduce the outlier gene data. Secondly, the missing value problem is handled by employing BIPCA. This MVI method employs the hybrid of independent component analysis and Bayesian principal component analysis to estimate the missing gene value. This proposed pre-processing approach is evaluated by implementing them with Mutual Information based gene selection and standard SVM based classification. Evaluations on Mendeley data for breast cancer detection showed that the proposed model achieved better performance with increased silhouette scores by 1-15%, Pearson correlation 1-12%, p-values 1-14%, accuracy by 0.5-9%, precision by 0.2-7%, recall by 0.1-3% and f-measure by 0.2-4% while also reducing the processing time by 0.06 to 1.0 seconds than the existing methods. In future, the performance of gene expression analysis for breast cancer classification can be improved by developing advanced gene selection and classification models. Likewise, the impacts of environmental factors and rare gene variations can also be investigated for improving the quality of gene expression profiles.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

This work is a contribution of the authors: "Conceptualization, Velusamy Murugesan and Ponnunaicker Balamurugan; methodology, Velusamy Murugesan; software, Velusamy Murugesan; validation, Velusamy Murugesan and Ponnunaicker Balamurugan; formal analysis, Velusamy Murugesan; writing—original draft preparation, Velusamy Murugesan; writing—review and editing, Velusamy Murugesan and Ponnunaicker Balamurugan.

References

[1] M. J. Heller, "DNA microarray technology: devices, systems, and applications", *Annual Review of Biomedical Engineering*, Vol. 4, No. 1, pp. 129-153, 2002.

[2] J. Seita, D. Sahoo, D. J. Rossi, D. Bhattacharya, T. Serwold, M. A. Inlay, and I. L. Weissman, "Gene Expression Commons: an open platform for absolute gene expression profiling", *PloS One*, Vol. 7, No. 7, pp. 40321-40334, 2012.

[3] J. B. D. Kok, R. W. Roelofs, B. A. Giesendorf, J. L. Pennings, E. T. Waas, T. Feuth, and P. N. Span, "Normalization of gene expression measurements in tumor tissues: comparison of 13 endogenous control genes", *Laboratory Investigation*, Vol. 85, No. 1, pp. 154-159, 2005.

[4] A. W. C. Liew, N. F. Law, and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information", *Briefings in Bioinformatics*, Vol. 12, No. 5, pp. 498-513, 2011.

[5] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments", *Biostatistics*, Vol. 19, No. 4, pp. 562-578, 2018.

[6] R. Huis, S. Hawkins, and G. Neutelings, "Selection of reference genes for quantitative gene expression normalization in flax (Linum usitatissimum L.)", *BMC Plant Biology*, Vol. 10, No. 1, pp. 1-14, 2010.

[7] R. Hornung, D. Causeur, C. Bernau, and A. L. Boulesteix, "Improving cross-study prediction through add-on batch effect adjustment add-on normalization", *Bio-informatics*, Vol. 33, No. 3, pp. 397-404, 2017.

[8] R. Wei, J. Wang, M. Su, E. Jia, S. Chen, T. Chen, and Y. Ni, "Missing value imputation approach for mass spectrometry-based metabolomics data", *Scientific Reports*, Vol. 8, No. 1, pp. 1-10, 2018.

[9] A. F. Fortuny, F. Arteaga, and A. Ferrer, "Assessment of maximum likelihood PCA missing data imputation", *Journal of Chemo-Metrics*, Vol. 30, No. 7, pp. 386-393, 2016.

[10] H. Yasrebi, "Comparative study of joint analysis of microarray gene expression data in survival prediction and risk assessment of breast cancer patients", *Briefings in Bioinformatics*, Vol. 17, No. 5, pp. 771-785, 2016.

[11] Y. Zhao, L. Wong, and W. W. B. Goh, "How to do quantile normalization correctly for gene

expression data analyses", *Scientific Reports*, Vol. 10, No. 1, pp. 1-11, 2020.

[12] A. Belorkar and L. Wong, "GFS: fuzzy preprocessing for effective gene expression analysis", *BMC Bioinformatics*, Vol. 17, No. 17, pp. 169-184, 2016.

[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] M. Borella, G. Martello, D. Risso, and C. Romualdi, "PsiNorm: a scalable normalization for single-cell RNA-seq data", *Bioinformatics*, Vol. 38, No. 1, pp. 164-172, 2022.

[15] H. M. D. Silva and A. S. Perera, "Evolutionary k-nearest neighbor imputation algorithm for gene expression data", *ICTer*, Vol. 10, No. 1, pp. 1-12, 2017.

[16] A. Wang, J. Yang, and N. An, "Regularized Sparse Modelling for Microarray Missing Value Estimation", *IEEE Access*, Vol. 9, No. 1, pp. 16899-16913, 2021.

[17] X. Zhu, J. Wang, B. Sun, C. Ren, T. Yang, and J. Ding, "An efficient ensemble method for missing value imputation in microarray gene expression data", *BMC Bioinformatics*, Vol. 22, No. 1, pp. 1-25, 2021.

[18] M. Shahjaman, M. R. Rahman, T. Islam, M. R. Auwul, M. A. Moni, and M. N. H. Mollah, "rMisbeta: A robust missing value imputation approach in transcriptomics and metabolomics data", *Computers in Biology and Medicine*, Vol. 138, No. 1, pp. 104911-104922, 2021.

[19] A. Dubey and A. Rasool, "Efficient technique of microarray missing data imputation using clustering and weighted nearest neighbour", *Scientific Reports*, Vol. 11, No. 1, pp. 1-12, 2021.

[20] A. A. Yousef and S. Samarasinghe, "A Novel Computational Approach for Biomarker Detection for Gene Expression-Based Computer-Aided Diagnostic Systems for Breast Cancer", *Artificial Neural Networks*, pp. 195-208, 2021.

[21] J. Quist, L. Taylor, J. Staaf, and A. Grigoriadis, "Random forest modelling of high-dimensional mixed-type data for breast cancer classification", *Cancers*, Vol. 13, No. 5, pp. 991-1004, 2021.

[22] S. Hira and A. Bai, "A Novel Map Reduced Based Parallel Feature Selection and Extreme Learning for Micro Array Cancer Data Classification", *Wireless Personal Communications*, Vol. 122, No. 3, pp. 1-23, 2021.

[23] M. Mostavi, Y. C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression", *BMC Medical Genomics*, Vol. 13, No. 5, pp. 1-13, 2020.

[24] H. Xie, J. Li, T. Jatkoe, and C. Hatzis, "Gene Expression Profiles of Breast Cancer", *Mendeley Data*, Vol. 1, 2017.

[25] F. Feng, Y. Zheng, J. C. R. Alcantud, and Q. Wang, "Minkowski weighted score functions of intuitionistic fuzzy values", *Mathematics*, Vol. 8, No. 7, pp. 1143-1156, 2020.