CAAI Transactions on Intelligence Technology

Research Article

Deep learning approach for microarray cancer data classification

ISSN 2468-2322
Received on 31st March 2019
Revised on 12th November 2019
Accepted on 2nd December 2019
doi: 10.1049/trit.2019.0028
www.ietdl.org

Hema Shekar Basavegowda, Guesh Dagnew ✉

Department of Studies and Research in Computer Science, Mangalore University, Mangalore, Karnataka, India
✉ *E-mail: guesh.nanit@gmail.com*

Abstract: Analysis of microarray data is a highly challenging problem due to the inherent complexity in the nature of the data associated with higher dimensionality, smaller sample size, imbalanced number of classes, noisy data-structure, and higher variance of feature values. This has led to lesser classification accuracy and over-fitting problem. In this work, the authors aimed to develop a deep feedforward method to classify the given microarray cancer data into a set of classes for subsequent diagnosis purposes. They have used a 7-layer deep neural network architecture having various parameters for each dataset. The small sample size and dimensionality problems are addressed by considering a well-known dimensionality reduction technique namely principal component analysis. The feature values are scaled using the Min–Max approach and the proposed approach is validated on eight standard microarray cancer datasets. To measure the loss, a binary cross-entropy is used and adaptive moment estimation is considered for optimisation. The performance of the proposed approach is evaluated using classification accuracy, precision, recall, *f*-measure, log-loss, receiver operating characteristic curve, and confusion matrix. A comparative analysis with state-of-the-art methods is carried out and the performance of the proposed approach exhibit better performance than many of the existing methods.

1 Introduction

Cancer is a group of diseases characterised by the abnormal growth of cells. In a healthy body, the growth of the cells is in control such that they grow and die systematically. The internal and environmental factors may damage the cells' genetic make-up which results in the continuous growth of cells to form tumours [1]. Factors such as improper cell division and damage to the deoxyribonucleic acid are the major internal factors whereas, exposure to substances such as chemicals in tobacco smoke, radiation, and an ultraviolet ray of the sun are significant environmental factors leading to cancer [2, 3].

The cancer disease is diagnosed and differentiated using gene expression profiles. In the field of Computational Biology, Genomics, Statistics, and Pattern Classification, microarray gene expression data analysis is one of the challenging research domains. The core challenge with microarray cancer analysis is associated with the high curse of dimensionality and small sample size which exists due to irrelevant and redundant genes [4–6]. In addition, medical datasets are typically noisy, have variations in feature values and an imbalanced number of classes which results in over-fitting and lower classification accuracy [7, 8]. The need to conduct research in microarray data analysis, specifically cancer classification helps to identify and understand the features which contribute to the development of cancer. The significant role played by microarray data classification approach is the identification of genes which contributes to certain biological outcome and usage of such genes to predict new observations. This helps in the detection of cancer in its early stage so that the domain experts can make a treatment plan to enhance the survival rate of cancer patients [4, 9, 10]. Hence, the problem requires careful construction of a model that takes an input pattern which represents objects and predicts the category of the object under consideration and hence, there is a need to develop an accurate prediction model on the given test data [7, 11, 12].

Microarray cancer data classification consisting of major tasks such as collecting the data from its source, pre-processing, feature selection, classification, and post-classification analysis. The feature selection is the process of selecting important genes from the tens of thousands of highly correlated and informative genes and providing these filtered data elements to a classifier to achieve better classification accuracy [13, 14]. Feature selection plays a vital role in the classification of cancer data in order to identify optimal and relevant subset of features, thereby enhancing the classification accuracy and computational stability [15–18]. Analysis of microarray cancer data analysis plays a key role in getting better insight about the disease that ultimately helps in planning decisive measures and improve the cancer diagnosis procedure [19].

In our work, we propose to employ a deep learning-based classifier to classify microarray cancer data. Deep learning takes a large volume of data to learn the behaviour of features during training and predicts the class of unseen data. To validate the proposed method, we have considered eight standard microarray cancer datasets namely Prostate, Colon, Central Nervous System (CNS), Ovarian, Leukaemia, and Lung cancer datasets. Feature values are scaled using the Min–Max approach to overcome the bias of decision in favour of high valued features.

The remaining part of the paper is organised as follows. We discuss related research works in Section 2. Section 3 presents the proposed methodology. The experimental set-up and results analysis are presented in Section 4. Section 5 covers discussion and comparative analysis, and the concluding remarks along with future works are presented in Section 6.

2 Related works

Microarray data analysis is attracting the attention of researchers across several disciplines. There has been considerable concern in developing classification methods for microarray cancer data. Some of the latest works proposed for microarray data analysis in the field of artificial intelligence, machine learning, pattern recognition, and other related areas are discussed below.

Recently, due to the advancement in biomedical and information technologies, several research works are going on, leading to different algorithms that are helpful for cancer diagnosis using different data-driven diagnosis methods [20, 21]. Mabu *et al.* [22] proposed cluster-based feature selection and artificial neural network method for classification of gene expression datasets. Zeebaree *et al.* [12] proposed a gene selection and classification approach for microarray cancer data using convolution neural network. The authors do not reveal the procedure to obtain dimensionality reduced features from the original dataset. Hou *et al.* [23] introduced a diagnostic prediction model by integrating an optimised genetic algorithm with artificial neural networks and the model was validated on prostate cancer dataset. Mohapatra *et al.* [4] suggested ridge regression (RR) with a single hidden layer feed-forward network and feature weights were randomly chosen. To validate the method, binary microarray datasets namely Breast, Prostate, Colon tumour, and Leukaemia were used. It is observed that the standard train/test protocol suit is not followed with respect to breast cancer dataset.

Salem *et al.* [3] suggested genetic programming-based cancer classifier with the help of information gain (IG) for feature selection. Lin *et al.* [11] introduced genetic algorithm with silhouette statistics for feature selection and classification on SRBCT dataset. We have observed that the feature selection method is non-optimal as it generates thousands of features which result in the over-fitted model. Sharbaf *et al.* [16] proposed a hybrid approach for gene selection and classification of microarray datasets using cellular learning automata and ant colony optimisation. They have examined the impact of various rank-based feature selection methods and they use three classifiers namely support vector machine (SVM), k-nearest neighbour (KNN), and Naive Bayes for validation. Kumar *et al.* [17] built a feature selection and classification algorithms based on the MapReduce concept along with KNN classifier. Nguyen *et al.* [24] proposed an aggregate gene selection for microarray data classification and experimented their model on four standard datasets namely DLBCL, Leukaemia, Prostate, and Colon datasets. To validate the method, five existing classifiers namely linear discriminant analysis, KNN, probabilistic neural network, SVM, and multilayer perceptron (MLP) were used and they claimed that the proposed method has stability across different classifiers but they could not reveal the claimed stability beyond five classifiers. Lofti and Keshavarz [25] introduced a hybrid of Principal Component Analysis (PCA) and brain emotional learning for microarray cancer data classification. They validated the work on three datasets which are not enough to confirm about the generalisation of the method. Ravi *et al.* [26] have carried out an extensive review to reveal the potential of deep learning models in health informatics. They have illustrated different deep learning architectures such as deep feed-forward, convolutional networks, and recurrent networks applicable to problems across several problem areas. Kar *et al.* [27] proposed particle swarm optimisation-based feature selection for classification of microarray cancer data. The acute lymphoblastic leukaemia-acute myeloid leukaemia (ALL-AML) and SRBCT datasets were used to validate the method. They conduct an experiment ten times on each dataset and average of these ten runs are reported as final results. Garcia and Sanchez [28] proposed a two-stage method for microarray classification. In the first stage, they have performed a feature selection using the ReliefF ranking algorithm followed by training a classifier on the lower-dimensional feature space to classify each sample into their respective classes. They have used three linear classification models namely Fisher linear discriminant, SVM, and MLP neural network classifiers while the ReliefF algorithm is used for feature selection and experimental results are reported on eight different cancer datasets. Chen *et al.* [29] proposed particle swarm optimisation-based feature selection and the C4.5 decision tree has been applied for classification. Experimental results with 5-fold cross-validation approach on different tumour cancer datasets are reported. An adaptive rule-based classifier is proposed by Farid *et al.* [30] for big biological datasets considering decision tree and KNN. The limitation of this work is that it is not adaptive with respect to the number of neighbours. Lyu *et al.* [31] have proposed a filter-based feature selection method based on maximum information coefficient and Gram-Schmidt Orthogonalisation approach. Li *et al.* [32] constructed an overlapped grouping strategy and data-driven weights based on information theory for lung cancer classification. Piao *et al.* [33] introduced a feature subset-based ensemble method that learns from the different projection of the original feature space to classify multi-class microarray cancer data. Wang *et al.* [34] proposed an integrated Markov blanket technique and Wrapper-based feature selection to tackle the higher computational complexity due to redundant features during feature selection. Hoque *et al.* [35] introduced a greedy feature selection technique that uses mutual information that combines feature-feature and feature-class mutual information. The method was validated on three base classifiers namely KNN, RF, and SVM.

Thus, we have seen a noticeable amount of research works that are being carried out in the domain of microarray data classification. Some of the existing methods found to work on a few datasets and possess little lesser accuracy. These issues motivate us to explore deep learning method for the classification of microarray cancer data preceded by pre-processing operation using PCA to achieve better classification accuracy. The details are brought out in the following sections.

3 Proposed methodology

In this section, we present the proposed work which includes various phases such as feature scaling, dimensionality reduction, and deep feed-forward Neural Network-based classification method which also includes parameter settings. The framework of the proposed approach consists of the major tasks such as loading of the raw microarray cancer data, followed by normalisation using the Min–Max method, dimensionality reduction, and deep learning-based classification as presented in Fig. 1.

3.1 Feature scaling

In the field of pattern recognition and machine learning, it is a known fact that the feature scaling technique is explored to normalise the data. This process brings all the data elements down to the same scale in order to avoid outliers and hence enhances the quality of prediction. Since the features in microarray cancer datasets are having high variance, we propose to explore feature scaling as one of the pre-processing techniques for data normalisation. Feature scaling is carried out using the Min–Max approach in order to fit the sigmoid activation function as it considers values between 0 and 1 with a threshold value of 0.5 for binary classification during model training (see (1)).

$$X = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where X_i is the original feature value, X stands for the normalised data, $X_{\min}$ is the minimum value, and $X_{\max}$ stands for the maximum value in the original dataset before scaling.

3.2 Dimensionality reduction method

In this work, we propose to explore a dimensionality reduction technique namely PCA as a pre-processing technique to obtain a relatively more compact form of data. The PCA-based dimensionality reduction linearly transforms the features into a lower dimensional space. The PCA works by linearly mapping the high dimensional microarray cancer data to a lower dimensional space with the intent of maximising the variance of the data in the lower dimensional space [28, 36]. It is a well-known fact that dimensionality reduction helps to overcome over-fitting, enhances accuracy, and maintains the simplicity of the model thereby enhances the classification accuracy. Each sample

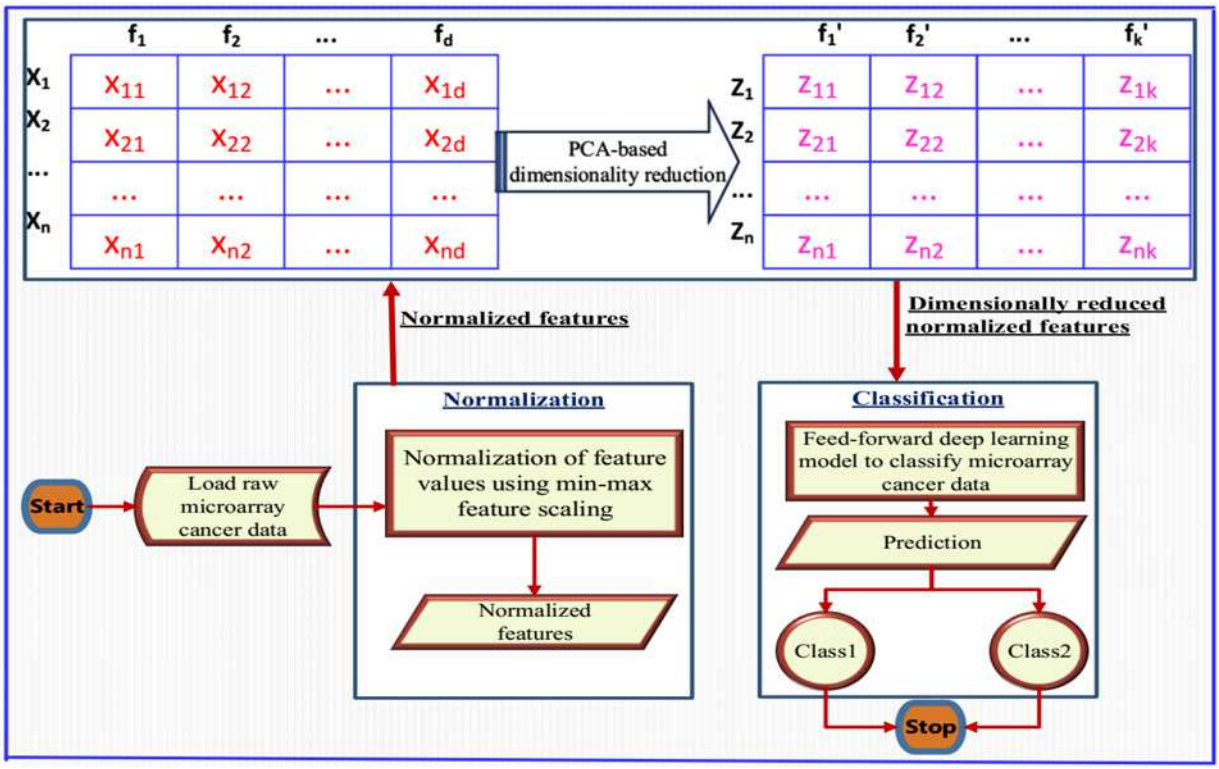

Fig. 1 *Framework of the proposed deep learning approach*

in the original dataset is represented based on (2) where X is any sample in the dataset with a dimension d.

$$X = [X_1, X_2, \ldots, X_n], \quad X \epsilon \mathbb{R}^{n \cdot d} \tag{2}$$

Given an input dataset that can be expressed as $X_i = x_1, x_2, \ldots, x_d$ where X_i is any sample in the input dataset, the dimension of the input dataset is expressed based on (2). The PCA constructs a compact feature set Z with dimension k that retains the maximum information by keeping maximum variance among features in the new dataset. The eigenvectors $\boldsymbol{W}$ generated by the PCA model are multiplied with each sample vector to create the new dataset.

$$\boldsymbol{Z} = X\boldsymbol{W} = [z_1, z_2, \ldots, \boldsymbol{z}_k], \quad \boldsymbol{Z} \epsilon \mathbb{R}^{n \cdot k}, \quad \boldsymbol{W} \epsilon \mathbb{R}^{d \cdot k}, \quad k \ll d \tag{3}$$

where the dimension of the original dataset d is much higher than the dimension of the new matrix.

3.3 Deep learning-based microarray cancer data classification

We propose to explore deep feed-forward neural network-based model to classify the microarray data. The deep feed-forward network is a vital deep learning model. The model is called feed-forward neural network because information flows through the function being evaluated from $\boldsymbol{z}_k$, through the intermediate computations used to define the function f, and finally to the output y_p, where $\boldsymbol{z}_k$ is the input feature vector and y_p is the predicted class label. In a deep-forward learning approach, the output of a given node is not connected to the node itself. Rather, the feed-forward neural network is represented by different directed (linked) functions hence, the model is a directed graph describing how the functions are connected together. For instance, let us consider three functions f_1, f_2, and f_3 connected in a chain to form the network and hence to define $f(x) = (f_1(x), f_2(x), f_3(x))$. These chain structures are the most commonly used structures in neural networks. In this case, f_1 is the first layer of the neural network, f_2 is the second layer, and f_3 is the third layer and so on [37]. The proposed model takes an input vector $\boldsymbol{Z} = z_1, z_2, z_3, \ldots, z_n$ where $\boldsymbol{Z} \varepsilon \mathbb{R}^{n \cdot k}$ and each input vector is multiplied by its corresponding weight. Hence, the weight vector for the input data is represented as $\boldsymbol{w} = w_1, w_2, w_3, \ldots, w_n$ and the bias b is added to the weighted input vectors. In each layer of the model, the weighted input vectors are multiplied by the sigmoid activation function to yield the intermediate probabilistic results in the hidden layers based on (4).

$$y_p = f\left(\sum_{i=1}^{n} \boldsymbol{w}_{n_i} \cdot \boldsymbol{z}_{k_i} + b\right) \tag{4}$$

where y_p is the dependent variable to be predicted, $\boldsymbol{w}_n$ are the weight matrices, $\boldsymbol{z}_k$ are feature vectors, and f is the sigmoid activation function based on (5).

We propose to use the sigmoid activation function during the training of the deep feed-forward neural network model to predict the class belongingness for a new data element. The sigmoid function is employed as an activation function has bounded output between 0 and 1 which handles the prediction of the class labels as the probability (see (5)).

$$f(z_i) = \frac{1}{1 + \mathrm{e}^{-(z_n \cdot \boldsymbol{w}_n)}} \tag{5}$$

Equation (6) handles multi-features across the hidden layers by learning the interesting behaviour of the data along the way to the output layer.

$$y_p = \begin{cases} 1, & \text{if} \quad w_0 + w_1 \cdot z_1 + w_2 \cdot z_2 + , \cdots, + \boldsymbol{w}_n \cdot \boldsymbol{z}_n >= 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\boldsymbol{w}_n$ is a weight vector, w_0 is a bias which is initialised to 1, $\boldsymbol{z}_n$ is a feature vector and y_p is the predicted class label.

Equation (6) is used to find the relation between the target-dependent variable and one or more of the independent

CAAI Trans. Intell. Technol., 2020, Vol. 5, Iss. 1, pp. 22–33

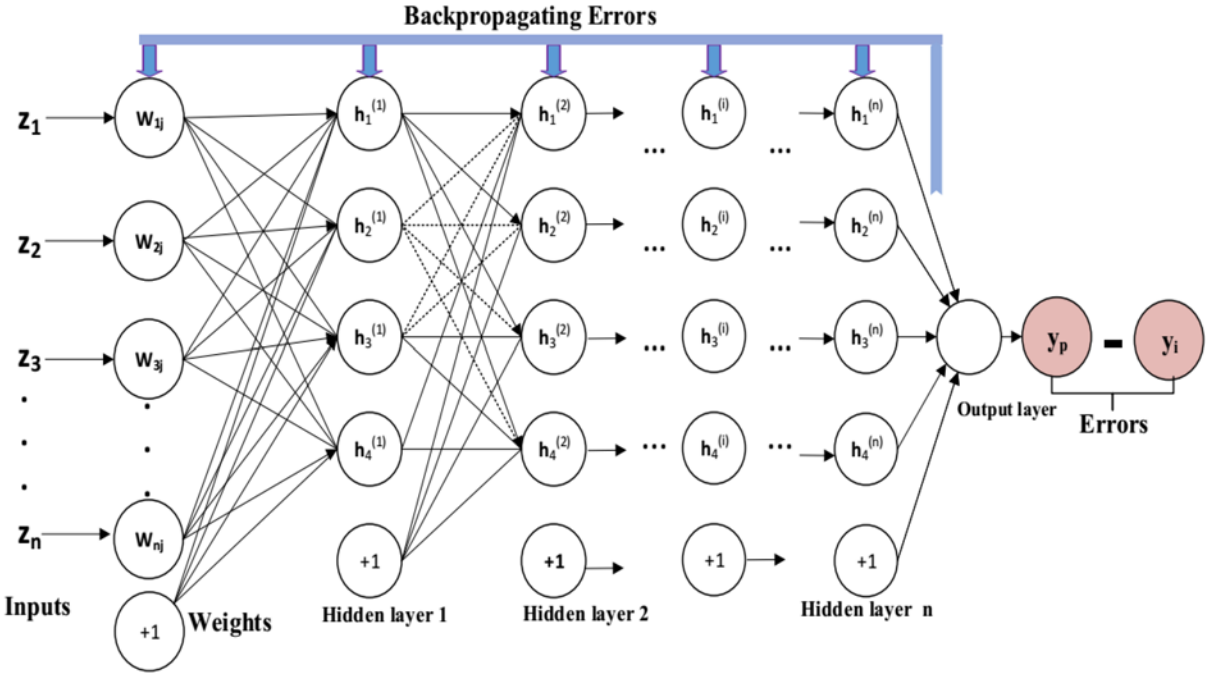

Fig. 2 *Deep learning model and its computational units*

features. Weight initialisation is carried out randomly and uniformly, and then the weight update is performed during the course of the training period. The proposed model compares the probability of the actual and predicted class and assigns a class value based on the threshold value of a binary class and assigns the prediction to either of the classes (see (6)).

In a deep learning-based classification, the error of the final prediction is computed as the difference between the predicted class label y_p and the given class label y_i using the predefined objective function namely cross-entropy. The errors are then backpropagated back through the entire network to optimise the weights for minimal error value [36].

The deep feed-forward approach and its computational units applicable for classification purpose is presented in Fig. 2. It consists of input features and weights, activation function, output, error computation, and error back propagation. Each input feature is multiplied with its weight to give a single output which will be also an input to the next layer and finally gives the prediction. The evaluation takes place by subtracting the actual class label from the predicted class and the difference is triggered back as back-propagation error. The back-propagation of errors is meant to update the weights and finally gives a maximum possible prediction, and hence the minimum error is registered.

As shown in Fig. 3, the activation function a defines the output of the dot product of the input features $[z_1, z_2, \ldots, z_n]$ and corresponding weights $[w_1, w_2, w_3, \ldots, w_n]$ in the first hidden layer h_1. The output of an activation function passes to the next node of the hidden layer and finally, the output is triggered at the node of the output layer. Weights are updated by propagating the weights back to the previous nodes until optimal prediction is achieved.

3.3.1 Parameter settings: We have used a seven-layer deep feed-forward model for all of the datasets and is presented in Table 1. The number of parameters in each layer is computed based on (7), where C_i is the number of neurons at the current layer i, P_i is the number of neurons in the previous layer, and 1 is the bias. The None parameter in the deep feed-forward learning approach indicates the batch size at each layer that may vary. It is noted that the number of neurons in the input layer is the input shape for the next layer. We have used an epoch size of 1000 with respect to seven datasets namely Breast, CNS, Colon, Leukaemia,

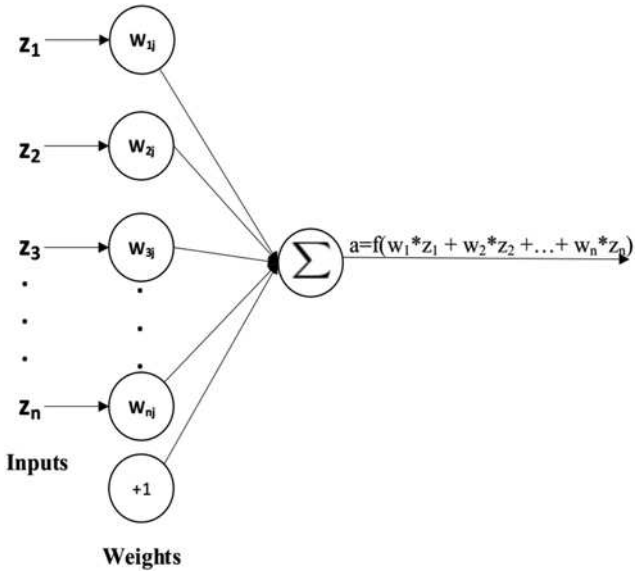

Fig. 3 *Activation function*

Table 1 Proposed seven-layer deep feed-forward model

Type of layers	Output shape	# Parameters
Dense Layer_1	(None, 200)	4600
Dropout_1	(None, 200)	0
Dense Layer_2	(None, 100)	20,100
Dropout_2	(None, 100)	0
Dense Layer_3	(None, 50)	5050
Dense Layer_4	(None, 40)	2040
Dense Layer_5	(None, 30)	1230
Dense Layer_6	(None, 20)	620
Dense Layer_7	(None, 1)	21
total number of trainable parameters		33,661

Ovarian, Prostate, Lung-Michigan cancers, and 1200 epochs in the case of Lung-Harvard2 cancer dataset. The total number of parameters $T_{\text{parameters}}$ in one model is the summation of all

parameters in each layer. Hence, the total number of trainable parameters in our work sums up to 33,661 which is the result of the sum of all the parameters in all the layers

$$T_{\text{parameters}} = \sum_{i=1}^{7} C_i(P_i + 1) \tag{7}$$

The parameters employed are sigmoid activation function, a binary cross-entropy to compute the loss on training and test data, and adaptive moment estimation (ADAM) optimiser [38]. ADAM optimiser works by computing the adaptive learning rate for each parameter. It computes and keeps decaying average of past gradients momentum m_t as indicated in (8). It stores an exponentially decaying average of the past squared gradients variance (v_t) as shown in (9), where t is time and g is the gradient.

$$m_t = \beta_1.m_{t-1} + (1 - \beta_1).g_t \tag{8}$$

$$v_t = (1 - \beta_2)\sum_{i=1}^{t} \beta_2^{t-i}.g_i^2 \tag{9}$$

In our work, we pass the ADAM optimiser by name, and parameters will take default values. The default values of ADAM optimiser parameters are $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 10^{-8}$ where β_1 and β_2 are decay rates [38].

Since (8) and (9) are biased [38], another mathematical expression which corrects the biases is required. Hence, (10) and (11) are used to alleviate the problem of bias.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{10}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{11}$$

Finally, the model updates its parameters by back-propagation as shown in (12), where θ stands for the updated parameters that enable the model to converge at timestamp t.

$$\theta_t = \theta_{t-1} - \frac{\eta.\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{12}$$

Since the model predicts based on the probabilistic approach, binarisation of the class level is carried out to get the Boolean results at a threshold value of 0.5 for either of the classes. The cost function in our model is defined in terms of binary cross-entropy. The proposed model computes probabilistic estimation of the target value y_i for a given class c and an input vector $\boldsymbol{z}_k$ in the output layer to determine class belongingness as expressed in (13) where $\boldsymbol{z}_k$ is a feature, y_i is the target, and c is the class which is either one or zero.

$$f(z_i) = p(y_i = c|z_n) \tag{13}$$

In binary class problem, the distance between the two probability vectors which are the predicted and actual class are compared using the distance function as shown in (14).

$$D(f(z_i, y_p, y_i)) = \frac{1}{N}\sum_i ((\boldsymbol{w}_n.z_n + b), y_p, y_i) \tag{14}$$

The distance of these vectors is averaged over the entire training set N for all the input values z_n, D is the distance function between the predicted, say y_p and actual, say y_i class labels. Here, $\boldsymbol{w}_n$ is the weight matrix and b is the bias. To maximise the probability of the correct target y_i, given input features z_i, we apply negative log-likelihood minimisation function as shown in (15) which maximises the estimation of a given sample $\boldsymbol{z}_n$ belonging to a class label y_i

$$L(f(z_i), y_i) = -\sum_c 1_{(y_i=c)} \log f(\boldsymbol{z}_n)_c = -\log f(\boldsymbol{z}_n) \cdot y_i \tag{15}$$

where L is the loss, $\boldsymbol{z}_n$ is a feature vector, y_i is the target class label corresponding to all features, and c is the class.

4 Experimental setup and results analysis

In this section, we present the description of datasets used in our experimentation, performance metrics used in evaluating the performance of the proposed method followed by experimental results and analysis. As a development tool, we have used Anaconda Python 3.5, Keras Deep Learning Library as a front-end [39] and Tensor-flow open-source deep-learning library as the back-end to construct our model.

4.1 Dataset description

The proposed approach is evaluated using eight different standard microarray datasets obtained from ELVIRA Biomedical Dataset Repository for high dimensional biomedical datasets http://leo.ugr.es/elvira/DBCRepository/index.html. The CNS cancer dataset contains 7129 features and 60 samples of which 21 are survivors and 39 are failures. The Colon cancer dataset contains 2000 genes and 62 samples, 40 of the samples are tumours which are labelled as positive and 22 are normal cases. The Ovarian cancer has 15,153 genes and 253 samples, out of which 162 samples are cancers and 91 samples are normal cases. The Leukaemia is a bone marrow cancer containing 7129 features and 72 samples. It has two classes where 47 samples of Acute Lymphoblastic Leukaemia (ALL) and 25 of them are Acute Myeloid Leukaemia (AML). The Prostate cancer contains 12,600 features and 102 samples of which 52 observations are tumours and the remaining 50 observations are normal. The Breast cancer contains 24,481 genes and 97 samples in which 46 cases are relapse and 51 samples are non-relapse. It is noted that both Lung-Harvard2 and Lung-Michigan cancers have highly imbalanced class distributions. The Lung-Harvard2 dataset has 181 samples with 150 Adenocarcinoma (ADCA) and 32 Malignant Pleural Mesothelioma (MPM) classes. The Lung-Michigan dataset has 96 samples out of which 86 belongs to Adenocarcinomas class and the remaining 10 samples are to non-neoplastic class.

Table 2 presents the description of eight standard datasets in terms of the original number of features, the selected features using PCA, the percentage of discarded features by PCA, the sample size,

Table 2 Dataset description

Dataset name	#Features	# Selected features	% of discarded features	Sample size	Training size	Test size	#Classes
CNS cancer	7129	108	98.49	60	36	24	2
Colon tumours	2000	104	94.80	62	37	25	2
Leukaemia cancer	7129	53	99.26	72	39	33	2
Prostate tumours	12,600	76	99.40	102	61	41	2
Ovarian cancer	15,154	24	99.84	253	202	51	2
Breast cancer	24,481	60	99.75	97	78	19	2
Lung-Michigan	7129	45	99.37	96	57	39	2
Lung-Harvard2	12,533	77	99.39	181	32	149	2

CAAI Trans. Intell. Technol., 2020, Vol. 5, Iss. 1, pp. 22–33

training and test size, and the number of classes. The proposed PCA-based dimensionality reduction is giving minimum but informative features that maximise the performance of the proposed classifier by neglecting the irrelevant features. We note that most of the features in the original datasets are not relevant for the prediction of a class label.

4.2 Performance measures

To validate the performance of the proposed approach, we have used several standard performance measures such as classification accuracy, precision, recall, *f*-score, and receiver operating characteristic (ROC) curve where it is summarised by area under the curve (AUC), log-loss, and confusion matrix. The accuracy is used to evaluate the overall predictive capability of the model that considers four parameters namely True positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), as shown in (16). It works by considering the number of correctly classified samples to the ratio of the total number of test samples

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{16}$$

Recall also known as sensitivity is a true positive rate that is the ratio of true positive (TP) to the sum of true positives (TP) and false negatives (FN) as shown in (17)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

Another performance measure used in this work is precision, which is also referred to as positive predictive value as shown in (18)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{18}$$

As shown in (19), F_{Measure} is used to neutralise the bias in precision and recall since it considers the harmonic mean of both precision and recall

$$F_{\text{score}} = 2 \cdot \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{19}$$

The error score due to the proposed method is computed using the log-loss function (see (20)), where N is the number of samples, y_i is the actual class label, and p_i is the probability that the ith sample belongs to either of the classes. Log-loss measures the performance of a model by computing the prediction as probability values between zero and one. A better classifier has to score a minimum log-loss error value with the intention of minimising to zero in the case of a perfect classifier.

$$\text{Log Loss} = \frac{-1}{N} \sum_{i=1}^{N} y_i \cdot (\log(p_i) + (1 - y) \cdot \log(1 - p_i)) \tag{20}$$

4.3 Analysis of experimental results

The detailed analysis of the experimental results obtained due to the proposed method on different datasets with various performance measures is presented in this sub-section.

In the proposed deep feed-forward neural network-based microarray cancer classification method, each of its input layers takes features and passes it for further manipulation in the next layer with each neuron densely connected throughout the model. The model is fully connected neural network which takes the selected features as an input passes the features through its hierarchical hidden layers to give a class prediction on the output layer, which has a single neuron that produces an output of probabilistic value between 0 and 1 inclusive, whereby if that value is binarised it will give the predicted result based on the specified threshold value of 0.5 which is a default value for binary classification problem.

In our work, to verify the generalisation capability of the proposed method, we have used separate training and test samples during pre-processing, i.e. in carrying out the scaling of features, dimensionality reduction, and classification. Some of the datasets such as Breast cancer, Lung-Harvard2, Leukaemia, have a standardised set of training and test samples, hence we use these standards in our work. For some of the datasets such as CNS, Colon, and Prostate, since there are no separate set of training and test samples, we have divided the datasets into training and test samples based on the ratio of 60:40 for training and test cases before pre-processing, and 80:20 ratio for Ovarian cancer data. The rationale behind dividing the datasets in the specified ratio for training and testing is that it is set as the standard used in many of the machine learning algorithms [40–42].

The experimental result of the proposed method is presented in Table 3, indicating the sample size of each dataset, the training and test sample sizes. The classification accuracy, precision, recall, *f*-score, AUC, and log-loss error are also presented. The proposed method shows perfect classification performance scoring 1.00 on four of the datasets namely, Leukaemia, Ovarian, Prostate, and Lung-Michigan cancer datasets. On the remaining four datasets, the proposed method has scored a classification accuracy of 0.99 on Lung-Harvard, 0.95 on Breast cancer, and 0.96 on both CNS, and Colon cancer datasets.

To validate the performance of the proposed method, we have used a confusion matrix that shows correctly and/or wrongly classified test samples. The confusion matrix depicts correctly classified samples along the diagonals and misclassified test samples along the off-diagonal elements. In Breast cancer data, out of 19 test samples, 1 case from the relapse class goes wrong to the non-relapse class (see Fig. 4*a*). Similarly, out of 24 test samples, 1 sample from failures class is misclassified as survivors class in CNS dataset (see Fig. 4*b*).

Moreover, in the case of the Colon dataset, out of 25 test cases, 1 negative case is misclassified as a positive case as shown in Fig. 4*c*. Out of 149 test cases in the Lung-Harvard dataset, 1 ADCA class is misclassified as MPM class as shown in Fig. 4*e*. In the other datasets, the approach exhibits perfect classification accuracy, as shown in Figs. 4*d*, *f*, *g*, and *h* for Leukaemia, Lung-Michigan, Ovarian, and Prostate cancer datasets, respectively.

Table 3 Experimental results on all datasets

Dataset	Training set size	Test set size	Training accuracy	Test accuracy	Precision	Recall	*F*-Measure	AUC	Log-loss error
breast cancer	78	19	1.00	0.95	0.95	0.95	0.95	0.96	0.410
CNS	36	24	0.96	0.96	0.96	0.96	0.96	0.97	0.219
Colon	37	25	1.00	0.96	0.97	0.96	0.96	0.97	0.189
Leukaemia	39	33	1.00	1.00	1.00	1.00	1.00	1.00	0.000
Ovarian	202	51	1.00	1.00	1.00	1.00	1.00	1.00	0.000
Prostate	61	41	1.00	1.00	1.00	1.00	1.00	1.00	0.003
Lung-Harvard2	32	149	1.00	0.99	0.99	0.99	0.99	1.00	0.032
Lung-Michigan	57	39	1.00	1.00	1.00	1.00	1.00	1.00	0.000

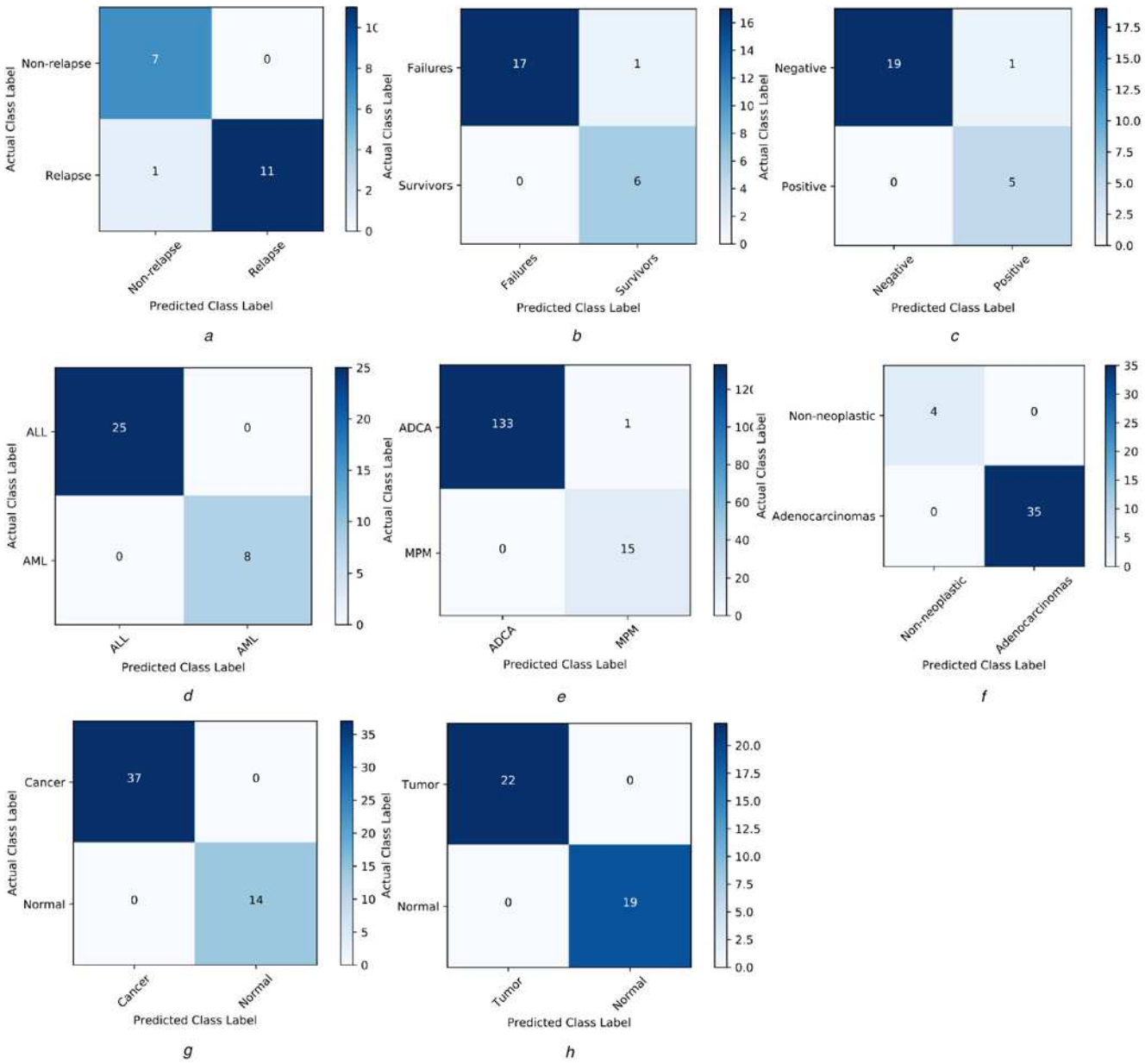

Fig. 4 *Confusion matrix on eight microarray cancer datasets*

a Breast dataset
b CNS dataset
c Colon dataset
d Leukaemia dataset
e Lung-Harvard2 dataset
f Lung-Michigan dataset
g Ovarian dataset
h Prostate dataset

To evaluate the performance of the proposed method, we have computed the classification accuracy. Fig. 5*a* shows the classification accuracy due to the proposed method on Breast cancer data. Similarly, Figs. 5*b*, *c*, and *e* show the classification accuracy of the proposed method on CNS, Colon, and Lung-Harvard2 datasets that shows the minimum gap between the lines of training and test cases. The classification accuracy of the proposed method on Leukaemia, Lung-Michigan, Ovarian, and Prostate datasets is shown in Figs. 5*d*, *f*, *g* and *h*, respectively. Due to the perfect classification accuracy on these datasets, there is no significant gap between lines associated with training and test samples.

We have also validated our work using the log-loss error function. As it can be observed from Fig. 6*a*, the loss on the training set is close to zero, however, the reported loss on test case is 0.410 (see Table 3) and this infers that there is still a need of further study on this particular dataset to minimise the error.

The proposed method shows better performance by scoring minimum error in the test cases of CNS, Colon, and Lung Harvard datasets as shown in Figs. 6*b*, *c* and *e* scoring 0.219, 0.189, and 0.032, respectively, (see the last column of Table 3). Moreover, there is no loss incurred on Leukaemia, Lung-Michigan, and Prostate cancer datasets as shown in Figs. 6*d*, *f* and *h* and a very negligible error that is 0.003 is scored on Ovarian cancer dataset as shown in Fig. 6*g*. It is noted that the gap between the lines of training and test cases is an indicator of whether the model is over-fitting or not.

In a classification problem, the ROC curve is widely used to check the performance of the model. It works by computing the AUC at various threshold settings. It is one of the metrics used to measure

CAAI Trans. Intell. Technol., 2020, Vol. 5, Iss. 1, pp. 22–33

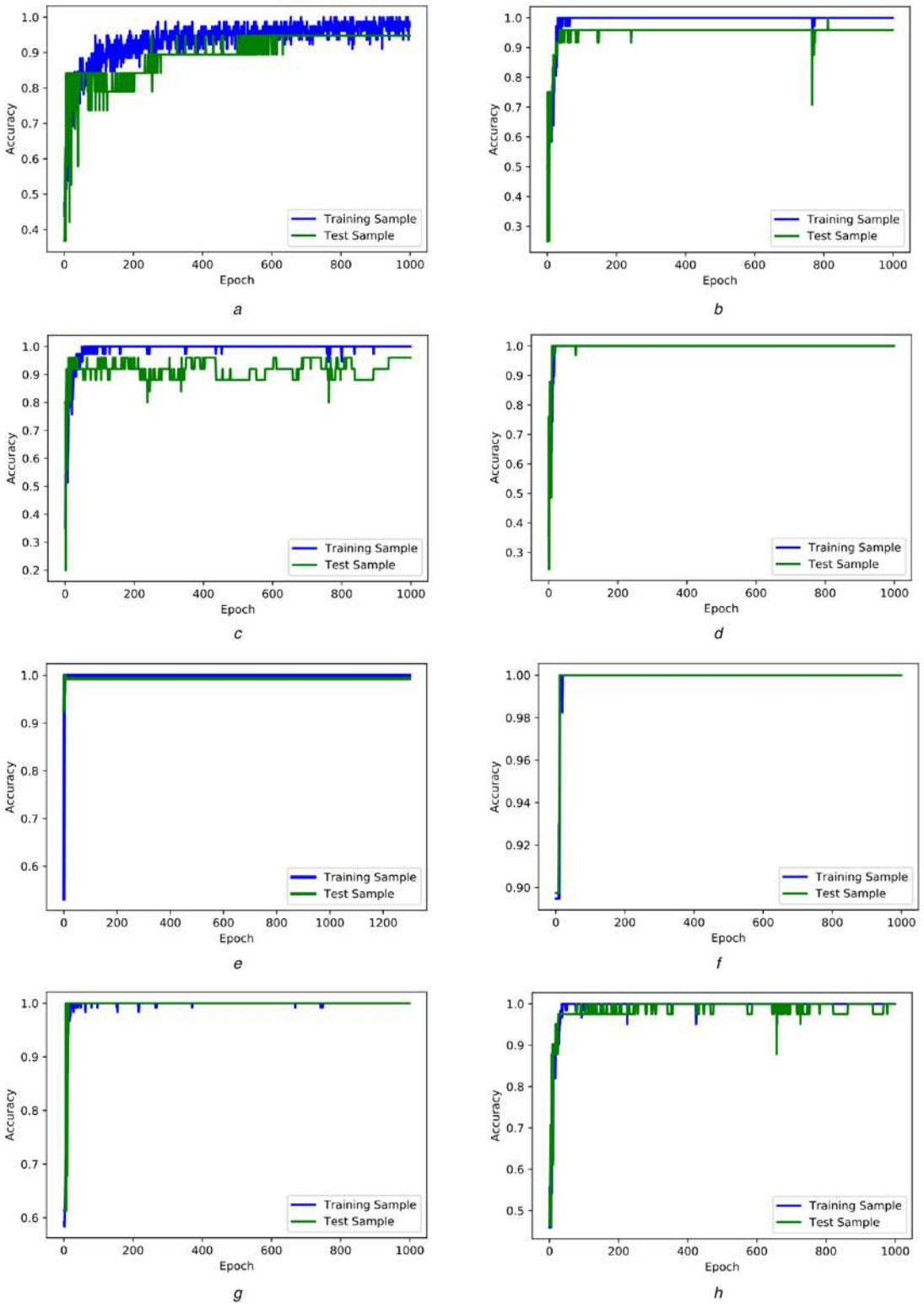

Fig. 5 *Classification accuracy on eight microarray cancer datasets*

a Beast dataset
b CNS dataset
c Colon dataset
d Leukaemia dataset
e Lung-Harvard2 dataset
f Lung-Michigan dataset
g Ovarian dataset
h Prostate dataset

the performance of any classifier and possess a high degree of tolerance in classifying data with lower-class imbalance. Unlike accuracy and other performance measures, the ROC curve shows the area coverage in terms of AUC. The intuition behind ROC and AUC is that, the more the curve moved to the left top corner, the better the classification accuracy it will be and this is achieved by

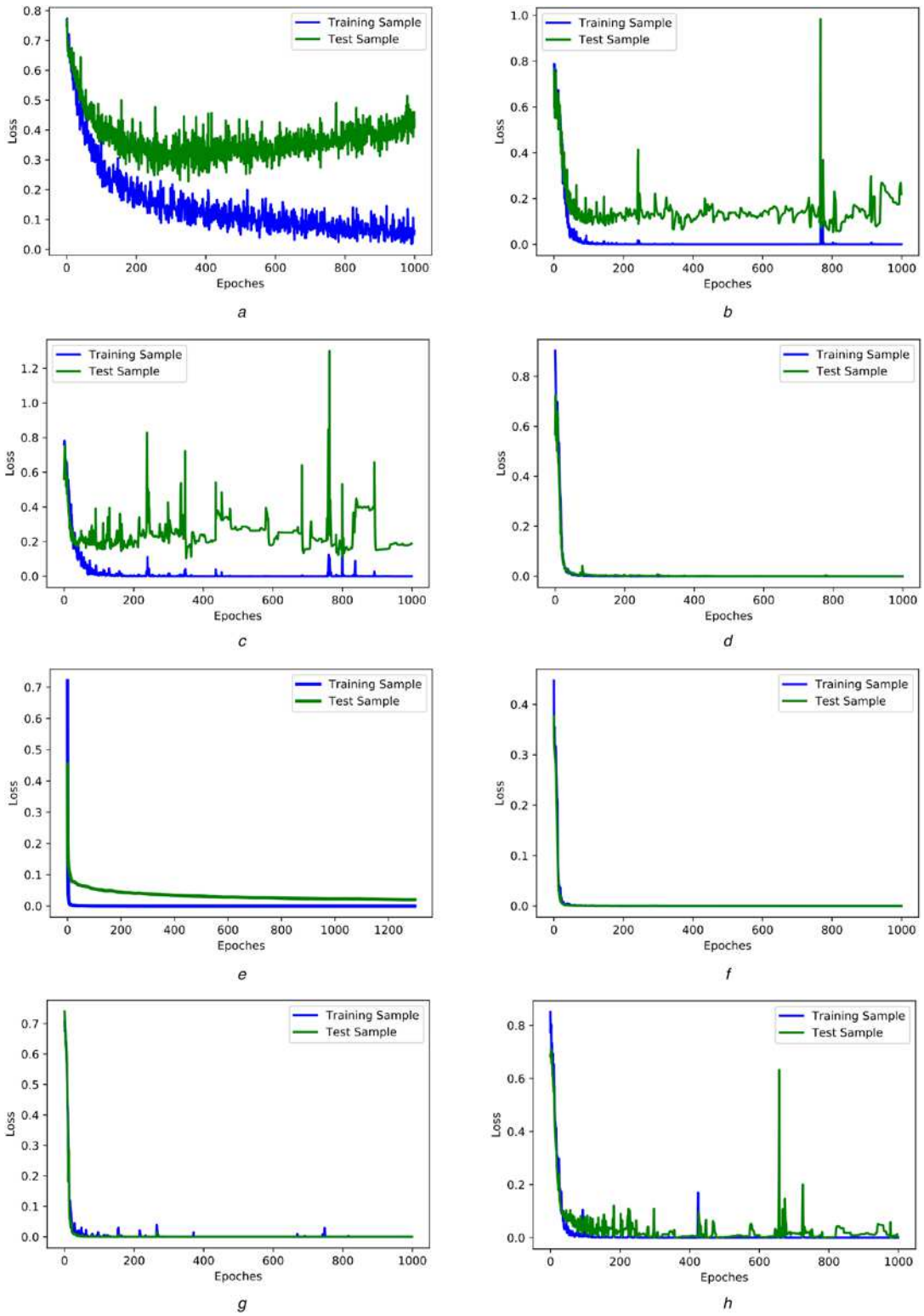

Fig. 6 *Loss due to the proposed method on eight microarray cancer datasets*

a Beast dataset
b CNS dataset
c Colon dataset
d Leukaemia dataset
e Lung-Harvard2 dataset
f Lung-Michigan dataset
g Ovarian dataset
h Prostate dataset

the proposed method. The ROC curve obtained due to the proposed methodology on the Breast cancer dataset is shown in Fig. 7*a*, scoring an AUC of 0.96 and is acceptable as per the rating of acceptance of results from the literature. Figs. 7*b* and c show the ROC curve for CNS and Colon cancer data scoring an AUC of 0.97 each which is rated as highly acceptable area coverage

CAAI Trans. Intell. Technol., 2020, Vol. 5, Iss. 1, pp. 22–33

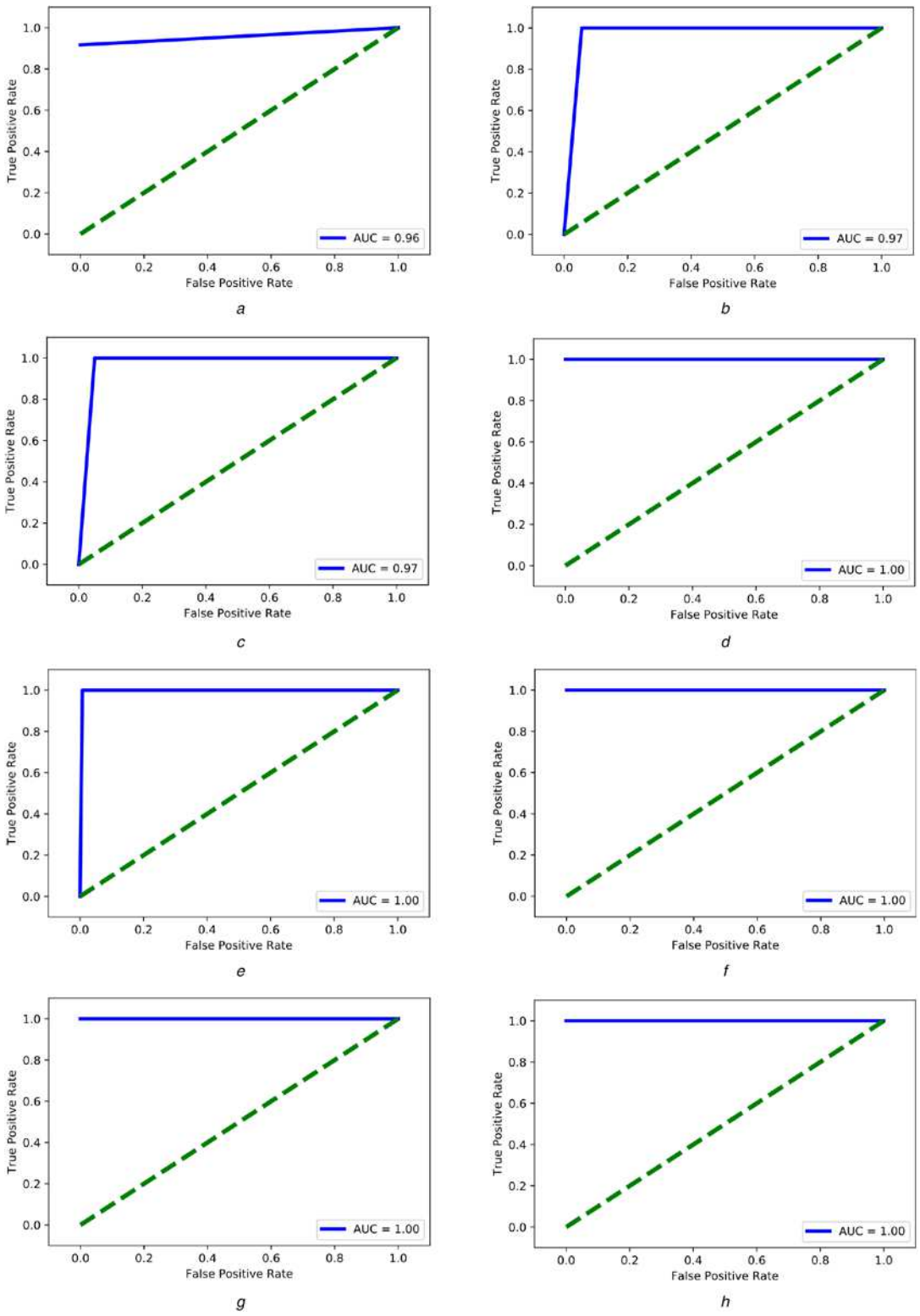

Fig. 7 *ROC curve with area under the curve (AUC) on eight microarray cancer datasets*

a Beast dataset
b CNS dataset
c Colon dataset
d Leukaemia dataset
e Lung-Harvard2 dataset
f Lung-Michigan dataset
g Ovarian dataset
h Prostate dataset

according to the literature [43, 44]. Similarly, the AUC obtained due to the proposed method on the Lung-Harvard2 dataset is 0.99 as shown in Fig. 7*e*. Since the proposed method is achieving a perfect classification accuracy on Leukaemia, Lung-Michigan, Ovarian, and Prostate datasets, the AUC score for each of these datasets is 1.00 as shown in Figs. 7*d*, *f*, *g*, and *h*, respectively.

Table 4 Comparison of classification accuracy of the proposed model with some of the latest related research works on CNS, Colon, Ovarian, Prostate, Leukaemia, Lung-Harvard2, Lung-Michigan, and Breast cancer datasets

References	Datasets							
	CNS	Colon	Ovarian	Prostate	Leukaemia	Lung-Harvard2	Lung-Michigan	Breast
Salem *et al.* [3]	0.87	0.85	–	**1.00**	0.97	–	**1.00**	–
Mohapatra *et al.* [4]	–	0.93	–	0.99	0.99	–	–	0.76
Singh and Sivabalakrishnan [5]	0.79	0.72	–	–	**1.00**	**1.00**	–	–
Medjahed *et al.* [9]	–	**0.97**	0.98	–	0.96	**0.99**	–	0.86
Tarek *et al.* [10]	–	0.65	–	0.92	**1.00**	–	0.72	–
Nguyen *et al.* [24]	–	0.88	–	0.91	0.94	–	–	–
Kar *et al.* [27]	–	–	–	–	0.97	–	–	–
Garcia and Sanchez [28]	0.73	0.84	0.99	0.	–	0.98	0.98	0.65
Chen *et al.* [29]	–	–	–	0.94	–	–	–	–
proposed method	**0.96**	**0.96**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**0.95**

Table 5 Comparison of Recall of the proposed model with IG/SGA [3]

References	CNS	Colon	Ovarian	Prostate	Leukaemia	Lung-Harvard2	Lung-Michigan	Breast
Salem *et al.* [3].	—	0.83	—	**1.00**	0.97	—	**1.00**	—
proposed method	**0.96**	**0.96**	**1.00**	**1.00**	**1.00**	**0.99**	**1.00**	**0.95**

Table 6 Comparison of the proposed model with [4] on data dimensionality, Training size, Test Size, Classification Accuracy (CA) and AUC parameters

Authors and Method	Dataset	Dimension	Training size	Test size	Accuracy	AUC
Mohapatra *et al.* [4] (WKRR)	Breast	97 * 24,481	70	27	0.81	0.89
	Colon	62*2000	40	22	0.95	0.79
	Leukaemia	76*7129	50	26	0.95	0.87
proposed model (deep learning)	Breast	97*24,481	78	19	**0.95**	**0.96**
	CNS	60*7129	36	24	**0.96**	**0.97**
	Colon Tumour	62*2000	37	25	**0.96**	**0.97**
	Leukaemia	72*7129	39	33	**1.00**	**1.00**
	Ovarian	253*15,154	202	51	**1.00**	**1.00**
	Prostate	102*12,600	61	41	**1.00**	**1.00**
	Lun-Michigan	96*7129	57	39	**1.00**	**1.00**
	Lung Harvard	181*12,533	32	149	**0.99**	**1.00**

5 Discussion and comparative analysis

This section provides a detailed discussion about the proposed method based on the results achieved and comparisons with state-of-the-art methods. An optimal performance by the proposed deep learning-based classifier is achieved when we use a PCA-based dimensionality reduction strategy to obtain informative features. The proposed method classifies perfectly with 1.00 accuracy on four of the datasets namely Leukaemia, Lung-Michigan, Ovarian, and Prostate datasets. Moreover, an accuracy of 0.99 is obtained on the Lung-Harvard dataset. We have got an accuracy of 0.96 on two datasets namely CNS and Colon, and 0.95 on Breast cancer. This shows that the proposed method is performing better than many of the state-of-the-art methods.

This part of the paper presents a comparative analysis of the proposed method with some selected latest works with respect to classification accuracy. Table 4 demonstrates a comparison of the classification accuracy of the proposed method with nine latest methods. The hyphen (-) in the particular cells of the table depicts that the authors did not consider the dataset in their work.

As presented in Table 4, the proposed method achieves better classification accuracy, which is 1.00 in four datasets namely Leukaemia, Lung-Michigan, Ovarian, and Prostate datasets. In the case of CNS and Colon datasets, we get better results comparing to the other works which are 0.96. In the case of Breast cancer and Lung-Harvard2, an accuracy of 0.95 and 0.99 is achieved, respectively. Generally, the proposed approach exhibits better performance when compared to the other methods. Furthermore, we suggest that the proposed method can be extended to multi-class datasets and other binary class datasets such as Brain cancers to show its validity which is our future work. Table 5 shows a comparison of the proposed method with the IG/SGA method [3]. It is shown that the proposed approach perform better than the IG/SGA method in terms of recall on two datasets namely Colon and Leukaemia and the similar results are achieved on Prostate and Lung-Michigan datasets. The hyphen (-) symbol in this table is to indicate that the authors do not consider datasets along that column.

We have also made a comparative study of our proposed deep learning method with other state-of-the-art work introduced by Mohapatra *et al.* [4]. We compare the methods in terms of dimensionality, training and test size, accuracy, and AUC. Based on the empirical evidence in Table 6, it shall be noticed that work exhibits better performance when measured in terms of classification accuracy and AUC.

6 Conclusion

In our work, we propose a deep feed-forward neural network approach for the classification of binary class microarray datasets. To validate the proposed method, eight standard microarray cancer datasets namely CNS, Colon, Prostate, Leukaemia, Ovarian, Lung-Harvard2, Lung-Michigan, and Breast cancers are used. To overcome the curse of dimensionality and other problems associated with the nature of the data, the PCA is used as a dimensionality reduction technique. Feature scaling is carried out using the Min–Max approach. To compute the magnitude of error during training and testing, the binary cross-entropy is applied since it is a standard loss function and is recommended for binary classification problems. For optimisation purposes, we have adapted the ADAM. A comparative study of the proposed method with state-of-the-art methods is carried out. Experimental results

CAAI Trans. Intell. Technol., 2020, Vol. 5, Iss. 1, pp. 22–33

on these standard microarray datasets and comparative analysis with state-of-the-art methods reveal that the performance of the proposed method is highly acceptable. To measure the performance of the proposed method, we have contributed the performance measures namely classification accuracy, precision, recall, *f*-measure, ROC curve, confusion matrix, and log-loss. The classification accuracy of the proposed method on four datasets namely Leukaemia, Lung-Michigan, Ovarian, and Prostate is 1.00, which depicts a perfect classification performance. Moreover, the proposed method scores an accuracy of 0.99 on Lung-Harvard2, 0.96 on CNS and Colon and 0.95 on Breast cancers. Furthermore, the ROC curve is illustrated for each of the datasets. The Area Under Curve (AUC) of the proposed method is 1.00 for Leukaemia, Lung-Michigan, Lung-Harvard2, Ovarian, and Prostate datasets. The AUC for CNS and Colon cancers is 0.97 and 0.96 for Breast cancer. As a future work, we are planning to extend the proposed method and apply it to multi-class microarray cancer datasets. We are also aiming to improve the classification accuracy on those binary datasets that scores less classification accuracy.

7 References

[1] Alberts, B., Bray, D., Hopkin, K., *et al*.: 'Essential cell biology', (Garland Science, 2013, Oct 15, Available at. http://dx.doi.org/10.1201/9781315815015)

[2] Cotter, T.G.: 'Apoptosis and cancer: the genesis of a research field', *Nat. Rev. Cancer*, 2009, **9**, (7), pp. 501–507

[3] Salem, H., Attiya, G., El-Fishawy, N. 'Classification of human cancer diseases by gene expression profiles', *Appl. Soft Comput.*, 2017, **50**, pp. 124–134

[4] Mohapatra, P., Chakravarty, S., Dash, P.K.: 'Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system', *Swarm. Evol. Comput.*, 2016, **28**, pp. 144–160

[5] Singh, R.K., Sivabalakrishnan, M.: 'Feature selection of gene expression data for cancer classification: a review', *Procedia Comput. Sci.*, 2015, **50**, pp. 52–57

[6] Moayedikia, A., Ong, K.L., Boo, Y.L., *et al*.: 'Feature selection for high dimensional imbalanced class data using harmony search', *Eng. Appl. Artif. Intell.*, 2017, **57**, pp. 38–49

[7] Wang, Z., Zineddin, B., Liang, J., *et al*.: 'cDNA microarray adaptive segmentation', *Neurocomputing*, 2014, **142**, pp. 408–418

[8] Chandra, B., Gupta, M.: 'An efficient statistical feature selection approach for classification of gene expression data', *J. Biomed. Inf.*, 2011, **44**, (4), pp. 529–535

[9] Medjahed, S.A., Saadi, T.A., Benyettou, A., *et al*.: 'Kernel-based learning and feature selection analysis for cancer diagnosis', *Appl. Soft Comput.*, 2017, **51**, pp. 39–48

[10] Tarek, S., Elwahab, R.A., Shoman, M.: 'Gene expression based cancer classification', *Egypt. Inform. J.*, 2017, **18**, (3), pp. 151–159

[11] Lin, T.C., Liu, R.S., Chen, C.Y.:*et al*.: 'Pattern classification in DNA microarray data of multiple tumor types', *Pattern Recognit.*, 2006, **39**, (12), pp. 2426–2438

[12] Zeebaree, D.Q., Haron, H., Abdulazeez, A.M.: 'Gene selection and classification of microarray data using convolutional neural network'. 2018 Int. Conf. on Advanced Science and Engineering (ICOASE), Duhok Technical University, Iraq, October 2018, pp. 145–150

[13] Jiang, D., Tang, C., Zhang, A.: 'Cluster analysis for gene expression data: a survey', *IEEE Trans. Knowl. Data Eng.*, 2004, **16**, (11), pp. 1370–1386

[14] Sasikala, S., Alias Balamurugan, S.A., Geetha, S.: 'A novel adaptive feature selector for supervised classification', *Inf. Process. Lett.*, 2017, **117**, pp. 25–34

[15] Khashei, M., Hamadani, A.Z., Bijari, M.: 'A fuzzy intelligent approach to the classification problem in gene expression data analysis', *Knowl.-Based Syst.*, 2012, **27**, pp. 465–474

[16] Sharbaf, F.V., Mosafer, S., Moattar, M.H.: 'A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization', *Genomics*, 2016, **107**, (6), pp. 231–238

[17] Kumar, M., Rath, N.K., Swain, A., *et al*.: 'Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor', *Procedia Comput. Sci.*, 2015, **54**, pp. 301–310

[18] Prieto, A., Prieto, B., Ortigosa, E.M., *et al*.: 'Neural networks: an overview of early research, current frameworks and new challenges', *Neurocomputing*, 2016, **214**, pp. 242–268

[19] Bashiri, A., Ghazisaeedi, M., Safdari, R., *et al*.: 'Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review', *Iran J. Public Health*, 2017, **46**, (2), p. 165

[20] Wang, H., Zheng, B., Yoon, S.W., *et al*.: 'A support vector machine-based ensemble algorithm for breast cancer diagnosis', *Eur. J. Oper. Res.*, 2018, **267**, (2), pp. 687–699

[21] Kourou, K., Exarchos, T.P., Exarchos, K.P., *et al*.: 'Machine learning applications in cancer prognosis and prediction', *Comput. Struct. Biotechnol. J.*, 2015, **13**, pp. 8–17

[22] Mabu, A.M., Prasad, R., Yadav, R.: 'Gene expression dataset classification using artificial neural network and clustering-based feature selection', *Int. J. Swarm Intell. Res. (IJSIR)*, 2020, **11**, (1), pp. 65–86

[23] Hou, Q., Bing, Z.T., Hu, C., *et al*.: 'Rankprod combined with genetic algorithm optimized artificial neural network establishes a diagnostic and prognostic prediction model that revealed C1QTNF3 as a biomarker for prostate cancer', *EBioMedicine*, 2018, **32**, pp. 234–244

[24] Nguyen, T., Khosravi, A., Creighton, D., *et al*.: 'A novel aggregate gene selection method for microarray data classification', *Pattern Recognit. Lett.*, 2015, **60**, pp. 16–23

[25] Lotfi, E., Keshavarz, A.: 'Gene expression microarray classification using PCA–BEL', *Comput. Biol. Med.*, 2014, **54**, pp. 180–187

[26] Ravi, D., Wong, C., Deligianni, F., *et al*.: 'Deep learning for health informatics', *IEEE. J. Biomed. Health. Inform.*, 2016, **21**, (1), pp. 4–21

[27] Kar, S., Sharma, K.D., Maitra, M.: 'Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique', *Expert Syst. Appl.*, 2015, **42**, (1), pp. 612–627

[28] Garcia, V., Sanchez, J.S.: 'Mapping microarray gene expression data into dissimilarity spaces for tumor classification', *Inf. Sci.*, 2015, **294**, pp. 362–375

[29] Chen, K.H., Wang, K.J., Wang, K.M., *et al*.: 'Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data', *Appl. Soft Comput.*, 2014, **24**, pp. 773–780

[30] Farid, D.M., Al-Mamun, M.A., Manderick, B., *et al*.: 'An adaptive rule-based classifier for mining big biological data', *Expert Syst. Appl.*, 2016, **64**, pp. 305–316

[31] Lyu, H., Wan, M., Han, J., *et al*.: 'A filter feature selection method based on the maximal information coefficient and gram-Schmidt orthogonalization for biomedical data mining', *Comput. Biol. Med.*, 2017, **89**, pp. 264–274

[32] Li, J., Wang, Y., Song, X., *et al*.: 'Adaptive multinomial regression with overlapping groups for multi-class classification of lung cancer', *Comput. Biol. Med.*, 2018, **100**, pp. 1–9

[33] Piao, Y., Piao, M., Ryu, K.H.: 'Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles', *Comput. Biol. Med.*, 2017, **80**, pp. 39–44

[34] Wang, A., An, N., Yang, J., *et al*.: 'Wrapper-based gene selection with Markov blanket', *Comput. Biol. Med.*, 2017, **81**, pp. 11–23

[35] Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: 'MIFS-ND: A mutual information-based feature selection method', *Expert Syst. Appl.*, 2014, **41**, (14), pp. 6371–6385

[36] Daoud, M., Mayo, M.: 'A survey of neural network-based cancer prediction models from microarray data', *Artif. Intell. Med.*, 2019, **97**, pp. 204–214

[37] Goodfellow, I., Bengio, Y., Courville, A.: 'Deep learning' (MIT Press, Cambridge, MA, USA, 2016), 1, Nov 10

[38] Kingma, D.P., Ba, J.: 'Adam: A method for stochastic optimization', arXiv preprint arXiv: 2014, 1412.6980

[39] Chollet, F.: 'Keras', https://keras.io, accessed January 2019

[40] Dobbin, K.K., Simon, R.M.: 'Optimally splitting cases for training and testing high dimensional classifiers', *BMC Med. Genet.*, 2011, **4**, (1), pp. 31–38

[41] Guyon, I.: 'A scaling law for the validation-set training-set size ratio', *AT&T Bell Lab.*, 1997, pp. 1–11

[42] Larsen, J., Goutte, C.: 'On optimal data split for generalization estimation and model selection'. Neural Networks for Signal Processing IX: Proc. of the IEEE Signal Processing Society Workshop, Cat. No. 98TH8468, Madison, WI, USA, 1999, pp. 225–234

[43] Park, S.H., Goo, J.M., Jo, C.H.: 'Receiver operating characteristic (ROC) curve: practical review for radiologists', *Korean J. Radiol.*, 2004, **5**, (1), pp. 11–18

[44] Gonen, M.: 'Receiver operating characteristic (ROC) curves', *SAS Users Group Int. (SUGI)*, 2006, **31**, pp. 210–231