A hybrid gene selection model for molecular breast cancer classification using a deep neural network

Monika Lamba*

Department of Computer Science and Engineering (CSE),
The NorthCap University,
Gurugram, India
Email: missmonikalamba@gmail.com
*Corresponding author

Geetika Munjal

Amity School of Engineering and Technology,
Amity University,
Noida, Uttar Pradesh, India
Email: munjal.geetika@gmail.com

Yogita Gigras

Department of Computer Science and Engineering (CSE),
The NorthCap University,
Gurugram, India
Email: yogitagigras@ncuindia.edu

Abstract: Microarray-based gene expression outlining portrays a dominant part in a healthier understanding of breast cancer. From the large quantum of data, a powerful technique is required to understand and extract the required information. The molecular subtype extraction is one of such important information regarding breast cancer, which is very crucial in defining its treatment strategy. This manuscript has formulated a deep neural network-based model for molecular classification of breast cancer. The proposed model exploits pre-processing steps along with the hybrid approach of filter and wrapper-based feature selection to extract relevant genes. The extracted genes are evaluated using various machine learning approaches where it is observed that selected features are successful in solving this multiclass problem. Using the proposed hybrid model, we have achieved the highest accuracy with six microarray datasets. The model outperforms magnificently in standings of sensitivity, f-measure, specificity, MCC and recall. Hence, deep neural network is identified as the best efficient classifiers concluding brilliant performance with all the selected microarray gene expression datasets for a range of selected genes.

Keywords: breast cancer; deep neural network; DNN; molecular subtype; feature selection; CFS; best first search; BFS; SMOTE.

Reference to this paper should be made as follows: Lamba, M., Munjal, G. and Gigras, Y. (2021) 'A hybrid gene selection model for molecular breast cancer classification using a deep neural network', *Int. J. Applied Pattern Recognition*, Vol. 6, No. 3, pp.195–216.

Biographical notes: Monika Lamba graduated in BTech specialised in CSE from Ansal Institute of Technology (AIT), affiliated to the Guru Gobind Singh Indraprastha University (GGSIPU), Dwarka, New Delhi, India in 2014. She earned her MTech in Computer Science and Engineering from University School of Information, Communication and Technology (USICT), affiliated to the GGSIPU, Dwarka, New Delhi, India in 2016. Presently, she is pursuing her PhD in Computer Science and Engineering from The NorthCap University, Gurugram, India. Her academic experience includes functioning as a visiting faculty and her current examination works incorporate the use of machine learning and deep learning.

Geetika Munjal has a teaching experience of more than 13 years in various esteemed institutions. She holds a BTech from Kurukshetra University, received her MTech CSE degree from Punjab Technical University and PhD from The NorthCap University, Gurugram. She has worked on a project titled 'Phylogenetic model for cancer classification', funded by Department of Science and Technology. Her areas of research include data mining, pattern recognition, machine learning and software engineering. She has guided around 12 BTech projects, seven MTech theses. She has published 21 papers in peer reviewed international journals with good indexing and reputed national/international conference proceedings.

Yogita Gigras is currently working as an Assistant Professor (Sel. Grade) in the Department of CSE and IT, School of Engineering and Technology, NCU. She has more than nine years of extensive teaching experience at both post and undergraduate level. She is a committed researcher in the field of soft computing and has completed her PhD in the same area. She has done her MTech in Computer Science and Engineering from Banasthali University, Rajasthan in 2009 with honours. She is a reviewer and assistant editor of various international journals. She is a lifetime member of ISTE.

1 Introduction

The breast cancer is second extremely common cancer amongst females, as per the statistics from 2019 more than 1.7 million cases have originated (Centers for Disease Control and Prevention, https://www.cdc.gov/cancer/breast/young_women/index.htm). The risk of breast cancer increases with age, females who are 50 years and above are generally found suffering from this disease, but almost 11% of the breast cancer is now found among females below 45 years of age. Although there have been cases of developing breast cancer at an even younger age which is a cause of major concern and thus creating physical and psychological burden Centers for Disease Control and Prevention, https://www.cdc.gov/cancer/breast/young_women/index.htm; BreastCancer.org, https://www.breastcancer.org/symptoms/types/molecular-subtypes). Classification of breast cancer to its correct subtype is necessary to prescribe the best possible treatment to patients.

Table 1 Details regarding molecular subtypes of breast cancer

Subtypes	*Description*	*Hormonal status*	*Treatment*	*Prognosis*	*Percentage*
Luminal A	Is hormone receptor positive Most common subtype for every age and race Grow slowly	ER+ PR+ HER2– Low Ki-67 Tumour grade 1 or 2	Hormonal therapy	Best Low recurrence rate (Voduc et al., 2010; Harris et al., 2012; Foukakis and Bergh, 2016; Arvold et al., 2011; Metzger-Filho et al., 2013; McGuire et al., 2017)	About 30%–45% (Voduc et al., 2010; Harris et al., 2012; Foukakis and Bergh, 2016)
Luminal B	Grow more quickly than luminal A Diagnosed at younger age than luminal A (Metzger-Filho et al., 2013; Partridge et al., 2016)	ER+ PR+ HER2+/– High Ki-67 Poorer tumour grade Larger tumour size Lymph node positive	Chemotherapy Hormone therapy Medication targeting HER2	Slightly worse (Clark et al., 2011; Voduc et al., 2010; Metzger-Filho et al., 2013)	About 10% (Clark et al., 2011; Voduc et al., 2010; Harris et al., 2012)
HER2–	Is hormone receptor negative Detected at younger age than luminal A and luminal B	ER– PR– HER2+ Lymph node positive	Combination of surgery, radiation therapy and chemotherapy	Poor	About 5%–15% (Clark et al., 2011; Arvold et al., 2011)
Triple negative/basal-like	More in women suffering from *BRCA1* gene mutations Grow faster than luminal cancers	ER– PR– HER2– (Clark et al., 2011; McGuire et al., 2017)	Chemotherapy Radiation therapy Medication targeting non-HER2	Worse (Clark et al., 2011)	About 15%–20% (Clark et al., 2011; Voduc et al., 2010; Harris et al., 2012)
Normal-like	Analogous to luminal A Low-level Ki-67	PR+ and/or ER + HER2– Low-level Ki-67 (BreastCancer.org, https://www.breastcancer.org/symptoms/types/molecular-subtypes)		Good Slightly worse than luminal A	

Detection of breast cancer has been improved using a computer-aided diagnosis (CAD) system (Gilbert et al., 2018; Lehman et al., 2015; Doi, 2007; Gromet, 2008; Ko et al., 2006). CAD seems to be very useful for breast radiologists to promote the performance in diagnosing cancer in case of correctness and time (Jung et al., 2014). In CAD, an appropriate classifier is critical to help medical experts in the foremost discovery of breast cancer subtype. A supplementary and trustworthy way to classify breast cancer gene expression depending on molecular features, determined by a test known as PAM50 and given by American Cancer Society (https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html).

Initially in 1970, breast cancer started to be divided into two categories established on oestrogen receptor (ER) status. In addition, the other clinicopathological parameters like the presence of tumour size, histological grade, lymph node metastasis, and three known markers PR, HER2 and ER also played a pivotal role in medication choice. Over the last two decades, the discovery of breast cancer research depends on a detailed study of intrinsic breast cancer subtypes including luminal as LumA and LumB. Luminal tumours are ER_positive where LumA are PR_positive, low-grade, and HER2_negative tumours, LumB have high-grade value, PR_negative, PR_positive, HER2_negative or HER2_positive and have high Ki-67 score (Clark et al., 2011; Voduc et al., 2010; Harris et al., 2012; Foukakis and Bergh, 2016; Arvold et al., 2011; Metzger-Filho et al., 2013; McGuire et al., 2017). These subtypes differ in their genomic (complexity), key genetic alternations, and prognosis. The survival rate of LumA is better than the remaining groups, as the low grade is the persistent sign in most of the tumours. These subtypes also occur in ductal carcinoma in situ (DCIS) (Clark et al., 2011). Details regarding molecular subtypes are illustrated in Table 1.

Correct molecular classification is an essential step towards the breast cancer severity identification due to:

a deficiency of standardised molecular class forecasting

b how many molecular classes exist for breast cancer

c the number of types can be correctly classified with presently accessible data.

Thus, this manuscript proposes an effective and innovative way to identify molecular subtype. The current research is divided into following sections, Section 2 is state-of-the-art, Section 3 presents the proposed model describing strategy adopted in detail, Section 4 consists of classification methods, followed by performance measure (Section 5), experimental outcomes (Section 6), discussion (Section 7) and conclusions and future scope (Section 8).

2 State-of-the-art

A right and reliable approach is needed to study breast cancer in-depth and its classification could help improve inconclusive treatment. Classification stands a supervised learning that can help the system learn from the data and classify the new data based on that learning. There are various machine classification algorithms (Omondiagbe et al., 2019) in literature like neural networks, random forest, support vector machine (SVM), etc. Earlier different studies have used classification approaches for breast

cancer-related problems mentioned in Table 2. All these approaches have shown promising results to some extent.

Table 2 Accuracy obtained using various machine learning method for breast cancer

Year	*Methods*	*Accuracy*
1996	Decision tree (C4.5)	97.80%
	10-cross validation (Akay, 2009)	
1996	Method RIAC – rule induction algorithm depending on classification (Akay, 2009)	96%
1999	Neuro fuzzy techniques (Akay, 2009)	95.06%
1999	Fuzzy genetic algorithm (Akay, 2009)	97.36%
2000	Neuro-rule method (Akay, 2009)	98.1%
2002	LVQ (optimised learning vector quantisation) (Akay, 2009)	96.7%
	Big LVQ	96.8%
	AIRS – artificial immune system (Akay, 2009)	97.2%
2003	Supervised fuzzy clustering (Akay, 2009)	95.57%
		96.8%
2007	SVM robustness (Polat and Güneş, 2007)	98.53%
2007	SVM (Übeyli, 2007)	99.54%
	Training – 37%	
	Testing – 63%	
2009	Collective SVM with feature obtained selecting five features (Akay, 2009)	99.51%
2014	PSO (particle swarm optimised wavelet neural network) (Dheeba et al., 2014)	93.67%
2015	Back propagation neural network with rough set relation (RS-BPNN) (Nahato et al., 2015)	98.6%
2016	Deep belief neural network (Abdel-Zaher and Eldeib, 2016)	99.68%

Since medical data generally suffers from class imbalance problem, thus it leads to error in the classification task (Zhang et al., 2019). Researchers have focused on the pre-processing approach like making the original data balanced using under-sampling or over-sampling. One such method is synthetic minority over-sampling technique (SMOTE) which is used with many existing methods (Bunkhumpornpat et al., 2009). Verbiest et al. (2014) make use of fuzzy as a selection algorithm to minimise the noise produced by SMOTE. Zeng et al. (2009) combined SMOTE with kernel into SVM mainly for imbalanced data issue. Gao et al. (2011) makes use of particle swarm optimisation to strengthen the under-sampling method of SMOTE and RBF categorisation is introduced to minimise the misclassification cases. Jeatrakul et al. (2010) introduced SMOTE with a neural network to derive the performance. Low dimensional data tend to outperform smote in multiple instances, however the same can be improved using SMOTE with SVM as a classifier-base (Lusa, 2013). Considering the importance of SMOTE in handling imbalanced dataset, current work has also taken into advantage.

After pre-processing, relevant features are extracted using a feature selection approach and after that these features are used in classification task. Traditional classification methods of breast cancer use arrangement/ morphology to separate tumours in various categories based on different prognosis and behaviour (Eliyatkın et al., 2015). In the last 11 years, the molecular subcategory of breast cancer is studied rigorously. It is observed that several classification methods suffer from overfitting, the training process takes a lot of time and on an average computation is too expensive (Tomar and Agarwal, 2013). Deep neural network (DNN) is a relatively a new classifier and has been applied successfully in many research areas related to genomics (Dong et al., 2019; Arisdakessian et al., 2019; Abdel-Zaher and Eldeib, 2016). In recent studies, it has been applied to microarray gene expression of breast cancer using denoising autoencoder (Kumar and Misra, 2016). Deep learning has successfully performed linear regression for extracting relevant genes (Mendez et al., 2019). It shows higher accuracy than the other shallow learning methods in classifying ER_positive and ER_negative samples (Alakwaa et al., 2018). Considering various advantages of DNN-based model, we have preferred them in a supervised phase to build our current model.

For bioinformatics problems, DNN performance may improve by selecting appropriate features (Chen et al., 2020; Aggarwal and Singh, 2019; Latkowski and Osowski, 2015; Lamba et al., 2020). Therefore, in our proposed model, we have used a hybrid approach of best-first search and correlation feature selection for identifying relevant features.

2.1 Datasets

The experiments are done on six datasets that are collected from National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.go) Advances. Microarray data experience curse of dimensionality issue, where the count of samples is very less as compared to numbers of genes. As shown in Table 3, the numbers of genes are very high, so it is very important to find only relevant genes followed by classification tasks. Table 4 depicts the distribution of samples in various molecular subtypes/classes which are luminal B, basal-like, claudin, luminal A, HER2 and normal. The sample count in claudin, normal and HER2 are very less as compared to all other classes thus depicting another challenge in the classification task.

Table 3 Detailed description of the various datasets used

Datasets accession no.	*No. of genes*	*Numbers of samples*
GSE25055	13,497	330
GSE10886	16,381	121
GSE18229	12,612	212
GSE20624	13,342	174
GSE21997	16,382	31
GSE34138	16,382	178

Table 4 Molecular types of various datasets with number of attributes

Molecular_types	*GSE25055*	*GSE10886*	*GSE18229*	*GSE20624*	*GSE21997*	*GSE34138*
Basal	122	12	32	24	5	46
Claudin	0	10	19	13	7	0
LumA	99	52	70	67	4	68
LumB	44	26	37	45	4	44
Normal	25	9	32	6	6	7
HER2	40	12	22	19	5	13
Total	330	121	212	174	31	178

3 Proposed model

In this research, we have presented a novel model for categorising the breast cancer into molecular subtype using hybrid feature selection supported by series of pre-processing steps. The whole process is:

1 Data pre-processing include mapping probe-ids to gene id's, normalise the data using the min-max function, discretisation, followed by SMOTE for balancing the minority class using k-nearest neighbour.

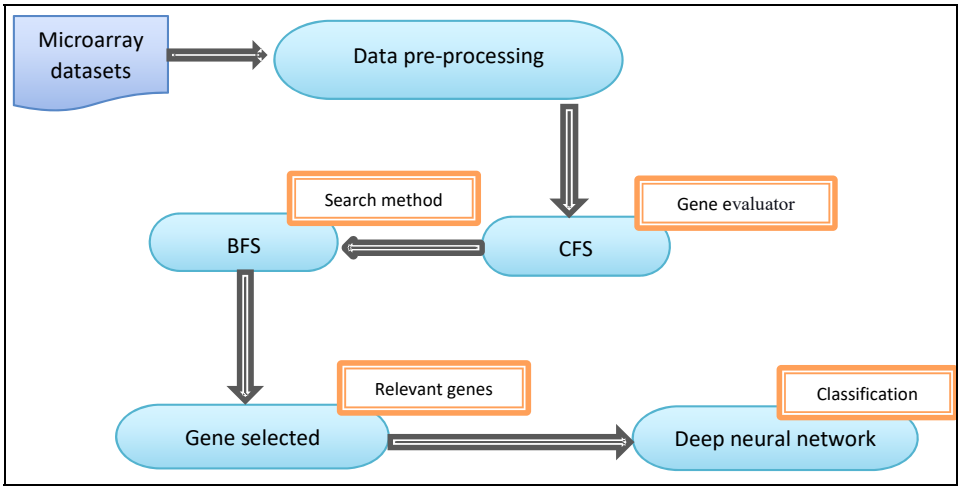
2 Applying feature selection technique using best-first search as searching and correlation-based searching as an evaluator.

3 DNN using soft-max activation functions for the classification task.

4 Evaluate the performance of selected genes on various shallow learning methods along with the proposed DNN on multiple datasets.

The proposed approach is shown in Figure 1, initially, the dataset described in Table 3, which belongs to a different platform (GPL).

Figure 1 Flowchart of proposed method (see online version for colours)



3.1 Data pre-processing

The steps involved in pre-processing are the integration of gene mapping, normalisation using min-max, discretisation, and class balancing using SMOTE. Data is mapped to their gene names using the gene mapping library GEOquery in R Studio (Allaire, 2012). The probe-ids are replaced with gene names where each gene is having an intensity corresponding to every sample. The gene intensity is normalised in range from 0 to 1 using a min-max function. Next step is discretisation, defined as the process where continuous value is converted into discrete values. Liu and Setiono (1995) presented a statically explained heuristic method for discretisation known as chi-square (Chi2). Last step of pre-processing is class balancing which is used when the number of samples is not distributed uniformly in the different categories of molecular subtype in breast cancer as described in Table 4. The problem arises with the datasets is regarding imbalance. SMOTE supports in addressing the imbalance/disproportion dataset problem using oversampling, here synthetic samples are generated for the minority class using the k-nearest neighbour technique. It is also observed that the combination of discretisation and SMOTE can help in improving the results (Jishan et al., 2015). To approach the oversampling task the following steps are followed:

Step 1 Locating the minority class set M, for each $y \in M$, k-nearest neighbour of y is generated by measuring the Euclidean distance among y and every sample present in M.

Step 2 For each $y \in M$, the sampling rate S is decided depending on the imbalanced percentage.

S examples $a_1, a_2, \ldots, a_s$ ($S \leq k$) are chosen aimlessly among k-nearest neighbour, hence generate the set M_1.

Step 3 For each and every example $a_k \in M_1$ ($k = 1, 2, 3, \ldots\ldots, S$), the mentioned formula is used to create the new examples/samples

$$a_{new} = a + rand(0,1) * \left\| (a - a_k) \right\| \tag{1}$$

where a_{new} is new example and $rand(0,1)$ will generate a number lies between 0 and 1.

3.2 Feature selection method

Once the data has been pre-processed, selecting the relevant genes that can contribute to the classification process is required. For selecting these related genes, a hybrid approach is used, we have applied the best first search (BFS) as a searching method for attribute subset space along with correlation-based feature selection (CFS) as feature evaluator. The BFS is a wrapper feature selection process under the category of supervised attribute selection. Selecting relevant genes by choosing the algorithm that can best fit the data plays a vital role. While learning, algorithms face many problems to select the optimal feature subset such as which gene to select and which one to eliminate/reject. So, the objective is to find the best functioning of the learning algorithm. It is very important to discover the relation between feature subset selection and feature to feature associativity. For this filter method helps in searching the optimal feature tailored to a machine learning

algorithm. One of the most reliable is the CFS, which ranks genes subset as per correlation-based heuristic evaluation function. It first calculates a matrix of gene-gene and gene-class correlation from the microarray data. The base of the evaluation function is facing the subsets that consist of genes deeply comparable with class and most incomparable with each other. Genes with low correlation with the class are called irrelevant genes and are ignored. Redundant genes are reserved out as they are hugely comparable by one or more with the left-over genes (Wosiak and Zakrzewska, 2018). The approval of a gene is determined by the extent, and that will predict classes in an area of the illustration space that has not yet been concluded through other genes. However, in cases where some highly predictive genes were eliminated may degrade the performance of machine learning. N_s represent CFS's gene subset evaluation function in equation (2):

$$N_s = \frac{k_{\overline{rc_f}}}{\sqrt{k + k(k-1)_{\overline{r_{ff}}}}} \tag{2}$$

N_s is heuristic 'merit' of a gene subgroup S, consisting of k genes, $\overline{rc_f}$ as mean of gene-class correlation and $\overline{r_{ff}}$ as mean gene-gene intercorrelation. Studies shows CFS give comparable results to the wrapper, which outperformed well on small datasets (Li et al., 2017). Moreover, CFS executes numerous times faster than wrapper thus CFS is used to select final relevant features of all datasets as described in Table 5.

Table 5 Number of features selected in six datasets

Dataset	*Gene selected*
GSE25055	96
GSE10886	71
GSE18229	122
GSE20624	102
GSE21997	105
GSE34138	191

The reason of associating BFS with CFS as feature evaluator is that it helps in identifying the most useful genes. It also helps in eradicating noisy, redundant, and irrelevant features when their importance does not heavily dependent on other genes. This combination of CFS and BFS helps in often eliminating half of the genes. Generally, in most of the cases, classification accuracy is equal or better using the minimised set of genes in comparison to full set of genes. This algorithm initiates with a null set of genes and performs forward searching with complete set of genes. Then look backward and begin at any point and examine in both the direction, thus adding or deleting genes. It searches the possible subset of genes by greedy hill climbing approach boosted with a backtracking advantage. After identifying relevant and minimised genes/features, the next task is to classify the samples.

4 Classification methods

Various studies have shown great performance potential of DNN for breast cancer categorisation.

DNN stands beneficial over other machine learning methods due to following reasons:

a DNN have the capability to understand themselves and generate output that is not restricted to an input provided.

b Data labelling has become obsolete, it is costly and time consuming but DNN as it excels learning with any recommendation.

c In case of large data size, the performance of DNN is tremendous and make the highest possible utilisation of unstructured information.

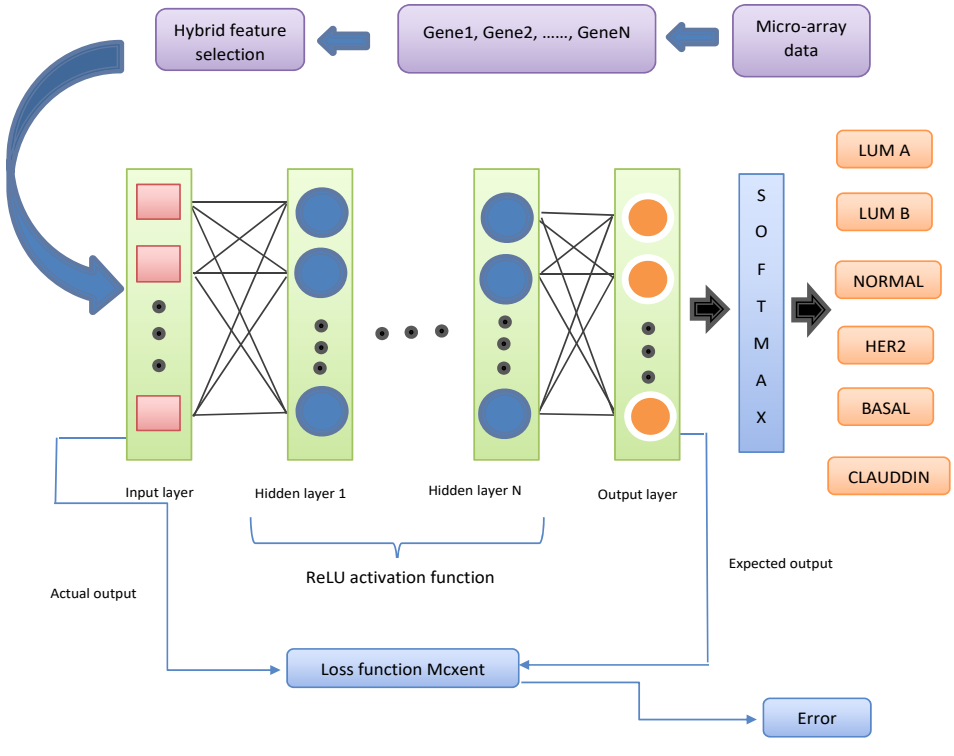
d DNN has high end infrastructure for training the large data in feasible time.

e Due to deficiency of domain knowledge of feature scrutiny, DNN performance outshines.

f The performance of DNN is better even in case of complex problems.

All these benefits have motivated us to explore DNN for this multi-classification task.

Figure 2 A DNN-based breast cancer classification (see online version for colours)



Note: DNN takes input as gene, feed forward artificial neural network to perform gene learning and establish a DNN classifier.

The architecture of DNN is displayed in Figure 2, determining the total number of neurons and hidden layers is based on trial-and-error rule (Sheela and Deepa, 2013). The

DNN structure is made up of various computational layers. Each layer admits input and use that input to generate the output. This output is in the form of nonlinear function consisting of weighted linear combination of the input layer, regulated weight, and threshold with the support of error that is back propagated.

In the process of forward propagation, each neuron results in an output as a nonlinear calculation of the weighted sum of the preceding layer to that neuron. The formula used is defined as:

$$y = f\left(\sum_i \omega_i x_i + b\right) \tag{3}$$

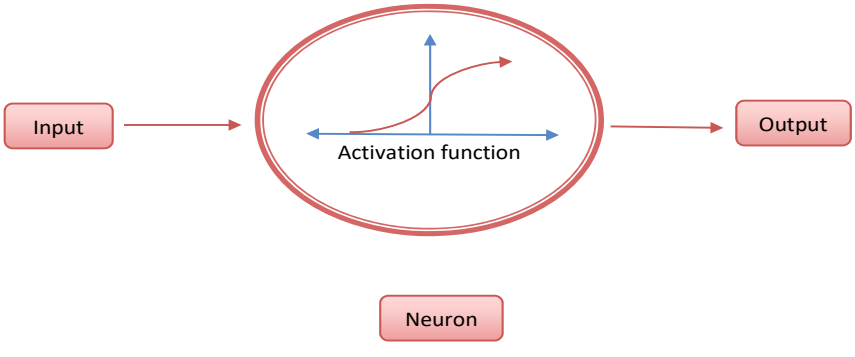
where ω_i represent the weight, x_i represent the input (genes) of the activation, b is the bias and y show the output of activation.

Performance of DNN is highly driven by activation function, it acts as statistical 'gate' sandwiched in the middle of an input feeding the present neuron as well as its output moving on to the following layer. It could be as straightforward as a step function which transforms the neuron output off and on, varying upon a threshold or law. It plays a major role due to amalgamation of arbitrary linear model. When it comes to design solution of complex problem, activation function shown in Figure 3 is changed to nonlinear. Numerous hidden levels of neurons are necessary to discover complicated datasets along with high-level of accuracy. Multiple activation function exists like rectified linear unit (ReLU), tanh, sigmoidal, etc., but ReLU is fragile and needed less computation time along with agile convergence speed. Important benefit of using ReLU is that it helps in minimising the interdependency of criteria which results in defeating the presence of overfitting and cause sparsity of the network. ReLU formula is defined as:

$$f_{ReLu} = \max(x, 0) \tag{4}$$

ReLU does not trigger all the neurons at the very same time. This implies that the neurons shall only be deactivated if the output of the linear transformation is less than 0. Intended for the negative input values, the outcome is zero, which means that the neuron is not getting activated. Because only a certain number of neurons are activated, the ReLU functionality is far more computationally effective when compared to the tanh and sigmoid function.

Figure 3 Activation function that maps the input into output needed for neural network to function (see online version for colours)



Once the current propagation gets completed, loss function named loss multi-class cross-entropy (Mcxent) is used to find the difference among target and predicted values to evaluate the model efficacy. The procedure of decreasing the loss function is the procedure of model training. Afterwards, the concealed (hidden) layer, the result of output from the hidden layer is a probability dispersal using SoftMax function, used to produce output as range of probabilities. In multi-class case, the result gives the probabilities of each class and highest probability is in case of target class. SoftMax function is frequently described as a mixture of several sigmoid. The SoftMax is especially useful as it changes the output layer as probability distribution (Chung et al., 2016). SoftMax is defined in equation (5), where e is a mathematical constant and y_i refers to value of each element in logits (logarithm of the odds). The total of component of output $S(y_i)$ is 1.

$$S\left(y_i\right) = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{5}$$

The experimental results are validated using ten-fold cross-validation (Arlot and Celisse, 2010) where datasets are segregated into testing and training to estimate the exactness without over fitting. Out of ten-fold, model trained on nine splits of data and testing on the remaining split, ensuring training and testing are performed on non-overlapping subsets.

5 Performance measure

Various metrics based on confusion matrix are used to evaluate the proposed model. It helps in understanding the usefulness of our model in case of sensitivity, recall, specificity, Mathews correlation coefficient (MCC), f-measure and precision. The components of confusion matrix are true positive (TP) seems to be correct and it is true. True negative (TN) is incorrect and that is true. False positive (FP) expected correctly and it is incorrect. False negative (FN) seems to be forecasted incorrect and it is false. Recall is count of truly classified correct by total count of positive. F-measure helps to evaluate recall and precision. MCC is known as Mathews correlation coefficient, overcomes the class imbalance problem. Accuracy (ACC) is defined as measure of correct and precise indicator of the classifier, and it gives basic detail like how many genes are misclassified. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

The mean of balanced accuracy per class calculated as per formula:

$$Balanced\ accuracy = \frac{Specificity + Sensitivity}{2} \tag{7}$$

The other metrics generated from a confusion matrix are as follows:

$$Recall\ /\ Sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P} \tag{8}$$

$$Specificity = \frac{TN}{N} \tag{9}$$

$$F\ score = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \tag{10}$$

$$MCC = \frac{\big((TN * TP) - (FN * FP)\big)}{\big((FP + TP)(TP + FN)(TN + FP)(TN + FN)\big)^{0.5}}. \tag{11}$$

6 Experimental outcomes

The suggested DNN paradigm has achieved the best results in terms of sensitivity/TP-rate, precision, FP-rate/fall out, recall, MCC and f-measure. DNN results are highlighted italics in Tables 6–11. We have achieved good results on all six microarray datasets with the highest accuracy of 100% on GSM21997. Classification for the performance of DNNs over the rest of the machine learning algorithm was judged by overall accuracy, balanced mean accuracy, and several unclassified samples shown in Figure 4. As the number of samples per class is not balanced in each dataset, so for the benefit we have evaluated classifier based on balanced accuracy described in Figure 4(b). The result achieved by various machine learning algorithm includes maximum likelihood-based Bayesian network, entropy-based tree classifiers, SVM utilised maximum margin concept, RBF network produce combined result of neuron parameters and input using radial basis function as activation function. In deep learning along with input and output layers, there are multiple hidden layers having activation function at last hidden layer is the key benefit of DNN.

Table 6 Results of various classifiers on dataset – GSM10886

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	0.975	0.004	0.976	0.975	0.975	0.97
Naïve Bayes	0.975	0.004	0.976	0.975	0.975	0.97
LibSVM	0.959	0.011	0.961	0.959	0.959	0.949
SMO	0.967	0.007	0.97	0.967	0.967	0.955
RBF_Network	0.967	0.004	0.971	0.967	0.967	0.961
RandomForest	0.926	0.018	0.929	0.926	0.926	0.907
J48	0.826	0.065	0.826	0.826	0.826	0.761
Filtered classifier	0.826	0.065	0.826	0.826	0.826	0.761
PART	0.826	0.058	0.829	0.826	0.825	0.765
MultiClass classifier	0.95	0.017	0.953	0.95	0.95	0.939
Deep neural network	*0.983*	*0.003*	*0.984*	*0.983*	*0.984*	*0.978*

Table 7 Results of various classifiers on dataset – GSM25055

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	0.933	0.016	0.936	0.933	0.934	0.915
Naïve Bayes	0.93	0.015	0.934	0.93	0.932	0.912
LibSVM	0.921	0.028	0.927	0.921	0.919	0.9
SMO	0.933	0.017	0.934	0.933	0.933	0.915
RBF_Network	0.927	0.023	0.927	0.927	0.927	0.905
RandomForest	0.909	0.035	0.913	0.909	0.904	0.881
J48	0.791	0.068	0.786	0.791	0.788	0.725
Filtered classifier	0.791	0.068	0.786	0.791	0.788	0.725
PART	0.806	0.061	0.805	0.806	0.803	0.747
MultiClass classifier	0.803	0.049	0.815	0.803	0.807	0.753
Deep neural network	*0.948*	*0.014*	*0.949*	*0.948*	*0.948*	*0.935*

Table 8 Results of various classifiers on dataset – GSM18229

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	0.92	0.021	0.923	0.92	0.921	0.9
Naïve Bayes	0.915	0.023	0.917	0.915	0.916	0.893
LibSVM	0.92	0.03	0.923	0.92	0.918	0.897
SMO	0.948	0.019	0.949	0.948	0.948	0.932
RBF_Network	0.958	0.012	0.958	0.958	0.958	0.947
RandomForest	0.892	0.036	0.896	0.892	0.889	0.865
J48	0.854	0.034	0.853	0.854	0.853	0.821
Filtered classifier	0.854	0.034	0.853	0.854	0.853	0.821
PART	0.849	0.036	0.847	0.849	0.847	0.814
MultiClass classifier	0.741	0.08	0.75	0.741	0.742	0.672
Deep neural network	*0.948*	*0.015*	*0.948*	*0.948*	*0.948*	*0.934*

Table 9 Results of various classifiers on dataset – GSM20264

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	0.937	0.02	0.937	0.937	0.936	0.91
Naïve Bayes	0.931	0.023	0.932	0.931	0.931	0.903
LibSVM	0.914	0.031	0.917	0.914	0.914	0.882
SMO	0.948	0.022	0.948	0.948	0.948	0.926
RBF_Network	0.937	0.025	0.936	0.937	0.936	0.912
RandomForest	0.908	0.033	0.909	0.908	0.907	0.877
J48	0.753	0.084	0.753	0.753	0.753	0.668
Filtered classifier	0.753	0.084	0.751	0.751	0.753	0.667
PART	0.776	0.076	0.786	0.776	0.778	0.704
MultiClass classifier	0.828	0.065	0.839	0.828	0.83	0.771
Deep neural network	*0.937*	*0.02*	*0.937*	*0.937*	*0.937*	*0.914*

Table 10 Results of various classifiers on dataset – GSM21997

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	1	0	1	1	1	1
Naïve Bayes	1	0	1	1	1	1
LibSVM	0.935	0.012	0.954	0.935	0.93	0.929
SMO	1	0	1	1	1	1
RBF_Network	1	0	1	1	1	1
RandomForest	1	0	1	1	1	1
J48	0.645	0.072	0.623	0.645	0.625	0.563
Filtered classifier	0.645	0.072	0.623	0.645	0.625	0.563
PART	0.677	0.067	0.692	0.677	0.668	0.614
MultiClass classifier	1	0	1	1	1	1
Deep neural network	*1*	*0*	*1*	*1*	*1*	*1*

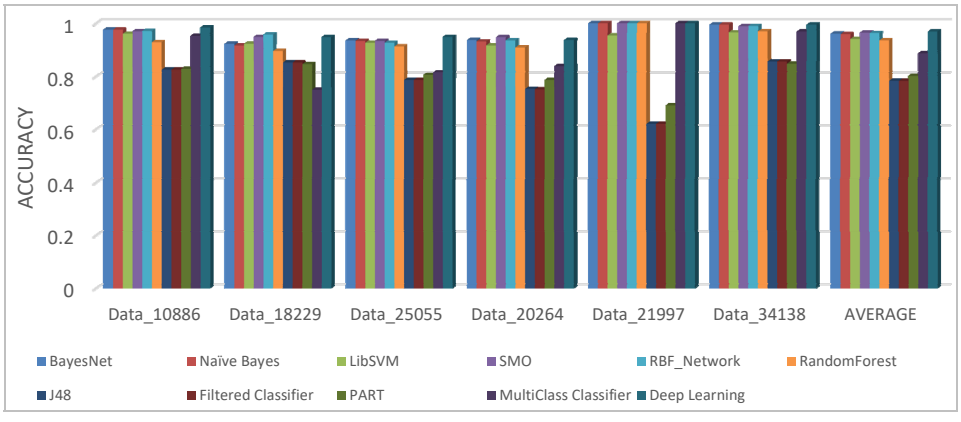
Table 11 Results of various classifiers on dataset – GSM34138

Method name	*Sensitivity/TP_Rate*	*Fall out*	*Precision*	*Recall*	*F score*	*MCC*
BayesNet	0.995	0.002	0.995	0.995	0.995	0.992
Naïve Bayes	0.995	0.002	0.995	0.995	0.995	0.992
LibSVM	0.962	0.022	0.966	0.962	0.956	0.948
SMO	0.989	0.005	0.989	0.989	0.989	0.984
RBF_Network	0.989	0.004	0.99	0.989	0.989	0.986
RandomForest	0.968	0.019	0.97	0.968	0.966	0.957
J48	0.849	0.046	0.857	0.849	0.851	0.805
Filtered classifier	0.849	0.046	0.857	0.849	0.851	0.805
PART	0.843	0.052	0.848	0.843	0.845	0.792
MultiClass classifier	0.968	0.011	0.969	0.968	0.968	0.957
Deep neural network	*0.994*	*0*	*0.995*	*0.994*	*0.994*	*0.993*

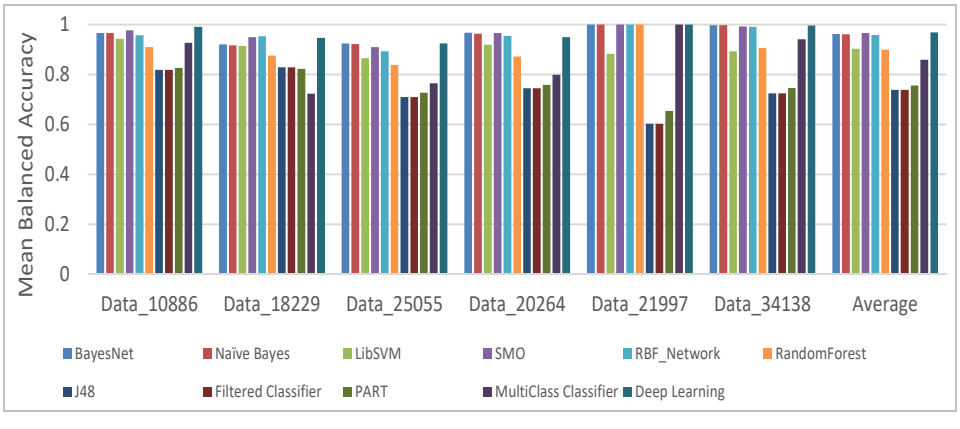
The number of unclassified samples in the case of GSE21997 is zero with classifiers namely RBF network, Bayes net, SMO, naïve Bayes, random forest, multiclass classifier and DNN. Among all the classifiers, DNN has shown the lowest number of unclassified samples, i.e., 42 samples are incorrectly classified from 687 samples obtained after feature selection among six datasets.

Out of 11 classifiers, DNN, SMO, and RBF network have performed well with misclassification of samples. The model has also given satisfactory results considering specificity, recall, f-measure, MCC and sensitivity. It is analysed in results that the proposed model has given excellent results with 0% error in some datasets. The proposed model has given good promising results as compared to other feature selection methods. All the experiment results were implemented on R studio 1.2.5019 (Allaire, 2012) and Weka 3.9.4 (Hall et al., 2009).

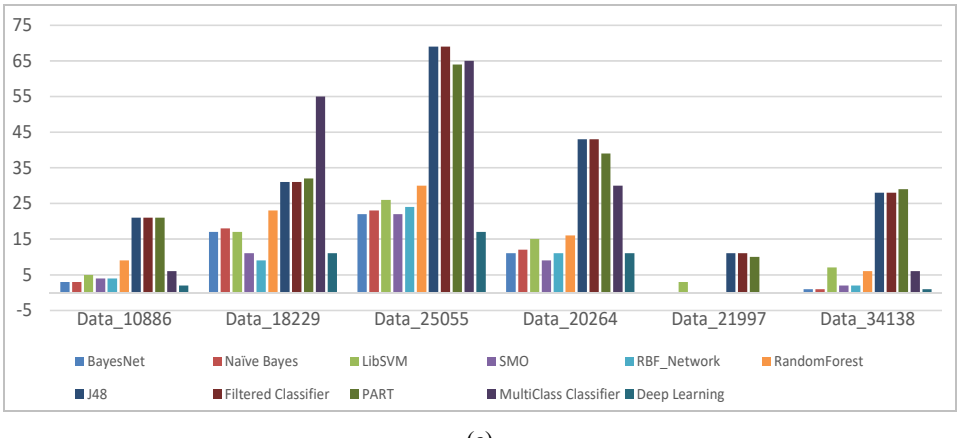
Figure 4 Bar plots of classification of deep learning compared to other machine learning techniques on six-different datasets and average results, (a) balanced accuracy evaluated by mean/average of balanced accuracy per class (b) overall accuracy (c) unclassified samples per class (see online version for colours)



(a)



(b)



(c)

Table 12 Summary of statistical t-test and p-values t-test based on MCC

MCC	*One tailed*			*Two tailed*		
	p = 0.05			*p = 0.05*		
Classification algorithms	*T-test*	*p-value*	*Hypothesis rejected (R)/accepted (A)*	*T-test*	*p-value*	*Hypothesis rejected (R)/accepted (A)*
BayesNet vs. DNN	–0.47782	0.321526	R	–0.47782	0.643052	R
Naïve Bayes vs. DNN	–0.57411	0.28929	R	–0.57411	0.578581	R
LibSVM vs. DNN	*–2.22192*	*0.025265*	*A*	–2.22192	0.05053	R
SMO vs. DNN	–0.34703	0.367879	R	–0.34703	0.735758	R
RBF_Network vs. DNN	–0.33416	0.37258	R	–0.33416	0.74516	R
RandomForest vs. DNN	–1.6992	0.060061	R	–1.6992	0.120123	R
J48 vs. DNN	*–5.60446*	*0.000113*	*A*	*–5.60446*	*0.000226*	*A*
Filtered classifier vs. DNN	*–5.60247*	*0.000113*	*A*	*–5.60247*	*.000227*	*A*
PART vs. DNN	*–6.67937*	*0.000028*	*A*	*–6.67937*	*0.000055*	*A*
MultiClass classifier vs. DNN	*–1.95434*	*0.039589*	*A*	–1.95434	0.079177	R

Table 13 Summary of statistical t-test and p-values t-test based on sensitivity

Sensitivity	*One tailed*			*Two tailed*		
	p = .05			*p = .05*		
Classification algorithms	*T-test*	*p-value*	*Hypothesis rejected (R)/accepted (A)*	*T-test*	*p-value*	*Hypothesis rejected (R)/accepted(A)*
BayesNet vs. DNN	–0.46601	0.325597	R	–0.466	0.65119	R
Naïve Bayes vs. DNN	–0.57101	0.290299	R	–0.57101	0.5806	R
LibSVM vs. DNN	*–2.37468*	*0.019483*	*A*	–2.37468	0.038965	A
SMO vs. DNN	–0.27107	0.395925	R	–0.2711	0.79185	R
RBF_Network vs. DNN	–0.33197	0.373383	R	–0.332	0.746767	R
RandomForest vs. DNN	–1.70202	0.05979	R	–1.70202	0.11958	R
J48 vs. DNN	*–5.33813*	*0.000165*	*A*	*–5.3381*	*0.000329*	*A*
Filtered classifier vs. DNN	*–5.33813*	*0.000165*	*A*	*–5.3381*	*0.000329*	*A*
PART vs. DNN	*–6.05379*	*0.000061*	*A*	*–6.05379*	*0 .000123*	*A*
MultiClass classifier vs. DNN	*–1.95982*	*0.03923*	*A*	–1.9598	0 .078459	R

7 Discussion

The results achieved signify that DNN performance outshines in comparison to various machine learning methods because:

a deep learning works well with large amount of data

b the activation function used in the architecture has major role in achieving reliable performance

c multiple hidden layers are preferred grounded on the number of genes.

Machine learning is extremely vulnerable to inaccuracies. Moreover, the model performance is highly driven by the number of genes selected by feature selection method. Whereas DNN desires plentiful time to allow the algorithms to learn and mature adequate in the direction to accomplish their determination with a substantial amount of precision.

In GSE21997 datasets, we have achieved 100% precision, recall, MCC, sensitivity and f-score. Utilising one and two-tailed student t-test to see whether the means of the measure received from the two different classification algorithms have been different in case of MCC and sensitivity. The following t and p-values reported in Tables 12–13. For statistical significance, we consider the p-value of 0.05, rectified for multiple comparison (Lahmiri et al., 2018). Italic font indicates the significant results.

8 Conclusions and future scope

The manuscript primarily focuses on a hybrid gene selection and DNN-based model used for breast cancer diagnosis based on molecular subtypes. The model highlights the importance of DNNs in classification tasks which can further help in cancer diagnosis. Other important things introduced to readers are pre-processing and hybrid feature selection models that integrate the advantage of CFS and BFS method.

With the selected features from the hybrid model, the entire machine learning models have given satisfactory results, however, it is observed that DNN classifier is giving best results. The DNN-based model is also highly scalable; however, the learning time taken by DNN is very high.

The future effort will be to explore more insights into breast cancer diagnosis using DNN techniques. Due to the flexible architecture of DNN, it might be used to recognise heterogeneity in breast cancer and other cancer types. As DNN architectures are adaptable to huge data thus, they can help in analysing integrated microarray data and other similar data as well, which will give new insight in understanding this complex disease.

References

Abdel-Zaher, A.M. and Eldeib, A.M. (2016) 'Breast cancer classification using deep belief networks', *Expert Systems with Applications*, 15 March, Vol. 46, pp.139–144.

Aggarwal, G. and Singh, L. (2019) 'Age classification with LPCC features using SVM and ANN', in *Information and Communication Technology for Competitive Strategies*, pp.399–408, Springer, Singapore.

Akay, M.F. (2009) 'Support vector machines combined with feature selection for breast cancer diagnosis', *Expert Systems with Applications*, 1 March, Vol. 36, No. 2, pp.3240–3247.

Alakwaa, F.M., Chaudhary, K. et al. (2018) 'Deep learning accurately predicts estrogen receptor status in breast cancer metabolomics data', *Journal of Proteome Research*, 5 January, Vol. 17, No. 1, pp.337–347.

Allaire, J. (2012) *RStudio: Integrated Development Environment for R*, Vol. 537, p.538, Boston, MA.

American Cancer Society [online] https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html (accessed 1 April 2020).

Arisdakessian, C., Poirion, O. et al. (2019) 'DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data', *Genome Biology*, December, Vol. 20, No. 1, pp.1–4.

Arlot, S. and Celisse, A. (2010) 'A survey of cross-validation procedures for model selection', *Statistics Surveys*, Vol. 4, pp.40–79.

Arvold, N.D., Taghian, A.G. et al. (2011) 'Age, breast cancer subtype approximation, and local recurrence after breast-conserving therapy', *Journal of Clinical Oncology*, 10 October, Vol. 29, No. 29, p.3885.

BreastCancer.org [online] https://www.breastcancer.org/symptoms/types/molecular-subtypes (accessed 1 April 2020).

Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. (2009) 'Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem', in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg, 27 April, pp.475–482.

Centers for Disease Control and Prevention [online] https://www.cdc.gov/cancer/breast/young_women/index.htm (accessed 1 April 2020).

Chen, Z. et al. (2020) 'Feature selection may improve deep neural networks for the bioinformatics problems', *Bioinformatics*, March, Vol. 36, No. 5, pp.1542–1552.

Chung, H., Lee, S.J. and Park, J.G. (2016) 'Deep neural network using trainable activation functions', in *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 24 June, pp.348–352.

Clark, S.E., Warwick, J. et al. (2011) 'Molecular subtyping of DCIS: heterogeneity of breast cancer reflected in pre-invasive disease', *British Journal of Cancer*, January, Vol. 104, No. 1, pp.120–127.

Dheeba, V., Singh, N.A. et al. (2014) 'Breast cancer diagnosis: an intelligent detection system using wavelet neural network', in *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, Springer, Cham, pp.111–118.

Doi, K. (2007) 'Computer-aided diagnosis in medical imaging: historical review, current status and future potential', *Computerized Medical Imaging and Graphics*, 1 June, Vol. 31, Nos. 4–5, pp.198–211.

Dong, Y., Yang, W. et al. (2019) 'MLW-gcForest: a multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data', *BMC Bioinformatics*, 1 December, Vol. 20, No. 1, p.578.

Eliyatkın, N., Yalçın, E. et al. (2015) 'Molecular classification of breast carcinoma: from traditional, old-fashioned way to a new age, and a new way', *The Journal of Breast Health*, April, Vol. 11, No. 2, p.59.

Foukakis, T. and Bergh, J. (2016) *Prognostic and Predictive Factors in Early, Non-metastatic Breast Cance*r, in Dizon, D.S. (Ed.), September, UpToDate.

Gao, M., Hong, X. et al. (2011) 'A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems', *Neurocomputing*, 1 October, Vol. 74, No. 17, pp.3456–3466.

Gilbert, F.J., Astley, S.M., Gillan, M.G. et al. (2008) 'Single reading with computer-aided detection for screening mammography', *New England Journal of Medicine*, 16 October, Vol. 359, No. 16, pp.1675–1684.

Gromet, M. (2008) 'Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms', *American Journal of Roentgenology*, April, Vol. 190, No. 4, pp.854–859.

Hall, M., Frank, E. et al. (2009) 'The WEKA data mining software: an update', *ACM SIGKDD Explorations Newsletter*, 16 November, Vol. 11, No. 1, pp.10–18.

Harris, J.R., Lippman, M.E. et al. (2012) *Diseases of the Breast*, 28 March, Lippincott Williams & Wilkins, Philadelphia, PA.

Jeatrakul, P., Wong, K.W. et al. (2010) 'Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm', in *International Conference on Neural Information Processing*, Springer, Berlin, Heidelberg, 22 November, pp.152–159.

Jishan, S.T., Rashu, R.I. et al. (2015) 'Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique', *Decision Analytics*, 1 December, Vol. 2, No. 1, p.1.

Jung, N.Y., Kang, B.J. et al. (2014) 'Who could benefit the most from using a computer-aided detection system in full-field digital mammography?', *World Journal of Surgical Oncology*, December, Vol. 12, No. 1, p.168.

Ko, J.M., Nicholas, M.J. et al. (2006) 'Prospective assessment of computer-aided detection in interpretation of screening mammography', *American Journal of Roentgenology*, December, Vol. 187, No. 6, pp.1483–1491.

Kumar, A. and Misra, B.B. (2019) 'Challenges and opportunities in cancer metabolomics', *Proteomics*, November, Vol. 19, Nos. 21–22, p.1900042.

Lahmiri, S., Dawson, D.A. and Shmuel, A. (2018) 'Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures', *Biomedical Engineering Letters*, Vol. 8, No. 1, pp.29–39.

Lamba, M., Munjal, G. and Gigras, Y. (2020) 'Computational studies on breast cancer analysis', *Journal of Statistics and Management Systems*, Vol. 23, No. 6, pp.999–1009.

Latkowski, T. and Osowski, S. (2015) 'Data mining for feature selection in gene expression autism data', *Expert Systems with Applications*, 1 February, Vol. 42, No. 2, pp.864–872.

Lehman, C.D., Wellman, R.D., Buist, D.S. et al. (2015) 'Diagnostic accuracy of digital screening mammography with and without computer-aided detection', *JAMA Internal Medicine*, 1 November, Vol. 175, No. 11, pp.1828–1837.

Li, J., Cheng, K., Wang, S., Morstatter, F. et al. (2017) 'Feature selection: a data perspective', *ACM Computing Surveys (CSUR)*, 6 December, Vol. 50, No. 6, pp.1–45.

Liu, H. and Setiono, R. (1995) 'Chi2: "feature selection and discretization of numeric attributes"', in *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, 5 November, pp.388–391.

Lusa, L. (2013) 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, 1 December, Vol. 14, No. 1, p.106.

McGuire, A., Lowery, A.J. et al. (2017) 'Locoregional recurrence following breast cancer surgery in the trastuzumab era: a systematic review by subtype', *Annals of Surgical Oncology*, 1 October, Vol. 24, No. 11, pp.3124–3132.

Mendez, K.M., Broadhurst, D.I. et al. (2019) 'The application of artificial neural networks in metabolomics: a historical perspective', *Metabolomics*, 1 November, Vol. 15, No. 11, p.142.

Metzger-Filho, O., Sun, Z., Viale, G. et al. (2013) 'Patterns of recurrence and outcome according to breast cancer subtypes in lymph node-negative disease: results from International Breast Cancer Study Group Trials VIII and IX', *Journal of Clinical Oncology*, 1 September, Vol. 31, No. 25, p.3083.

Nahato, K.B., Harichandran, K.N. et al. (2015) 'Knowledge mining from clinical datasets using rough sets and backpropagation neural network', *Computational and Mathematical Methods in Medicine*, Vol. 2015, 13pp.

National Center for Biotechnology Information (NCBI) [online] https://www.ncbi.nlm.nih.gov (accessed 15 April 2020).

Omondiagbe, D.A., Veeramani, S. and Sidhu, A.S. (2019) 'Machine learning classification techniques for breast cancer diagnosis', *InIOP Conference Series: Materials Science and Engineering*, IOP Publishing, April, Vol. 495, No. 1, p.012033.

Partridge, A.H., Hughes, M.E. et al. (2016) 'Subtype-dependent relationship between young age at diagnosis and breast cancer survival', *Journal of Clinical Oncology*, 20 September, Vol. 34, No. 27, pp.3308–3314.

Polat, K. and Güneş, S. (2007) 'Breast cancer diagnosis using least square support vector machine', *Digital Signal Processing*, 1 July, Vol. 17, No. 4, pp.694–701.

Sheela, K.G. and Deepa, S.N. (2013) 'Review on methods to fix number of hidden neurons in neural networks', *Mathematical Problems in Engineering*, Vol. 2013, 11pp.

Tomar, D. and Agarwal, S. (2013) 'A survey on data mining approaches for healthcare', *International Journal of Bioscience and Biotechnology*, 31 October, Vol. 5, No. 5, pp.241–266.

Übeyli, E.D. (2007) 'ECG beats classification using multiclass support vector machines with error correcting output codes', *Digital Signal Processing*, 1 May, Vol. 17, No. 3, pp.675–684.

Verbiest, N., Ramentol, E. et al. (2014) 'Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection', *Applied Soft Computing*, 1 September, Vol. 22, pp.511–517.

Voduc, K.D., Cheang, M.C. et al. (2010) 'Breast cancer subtypes and the risk of local and regional relapse', *Journal of Clinical Oncology*, 1 April, Vol. 28, No. 10, pp.1684–1691.

Wosiak, A. and Zakrzewska, D. (2018) 'Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis', *Complexity*, Vol. 2018, 11pp.

Zeng, Z.Q., Wu, Q. et al. (2009) 'A classification method for imbalance data set based on kernel SMOTE', *Acta Electronica Sinica*, Vol. 37, No. 11, pp.2489–2495.

Zhang, J., Chen, L. et al. (2019) 'Prediction of breast cancer from imbalance respect using cluster-based undersampling method', *Journal of Healthcare Engineering*, Vol. 2019, 10pp.