International Journal of Bioinformatics Research and Applications

ISSN online: 1744-5493 - ISSN print: 1744-5485
https://www.inderscience.com/ijbra

Identifying breast cancer molecular class using integrated feature selection and deep learning model

Monika Lamba, Geetika Munjal, Yogita Gigras

DOI: 10.1504/IJBRA.2023.10054946

Article History:

Received:	17 June 2022
Last revised:	20 June 2022
Accepted:	04 January 2023
Published online:	05 June 2023

Identifying breast cancer molecular class using integrated feature selection and deep learning model

Monika Lamba

The Northcap University,
Gurugram, India
Email: missmonikalamba@gmail.com

Geetika Munjal*

Amity University,
Uttar Pradesh, India
Email: munjal.geetika@gmail.com
*Corresponding author

Yogita Gigras

The Northcap University,
Gurugram, India
Email: yogitagigras@ncuindia.edu

Abstract: The extraction of molecular subcategory is one such valuable evidence concerning breast cancer in determining its cure and prognosis. This manuscript has framed a model for molecular subtype-based feature selection known as CFS-BFS followed by classification using deep learning. The proposed model captures significant genes by utilising pre-processing ladder along with the combination of filter and wrapper-based technique CFS-BFS. The obtained genes are assessed via numerous machine learning methodologies where it is remarked that carefully chosen significant genes are more profitable in explaining this molecular problem using deep learning. The study has attained the maximum precision and beats brilliantly in terms of recall, F-score, TP_Rate, fallout, and MCC. Hence, proposed paradigm is recognised as one of the best effective technique determining the outstanding recital with all the chosen micro-array gene expression datasets for significant obtained genes. The genes identified by integrated model are also validated using Kaplan-Meier survival graph to show their credibility in breast cancer prognosis. Survival analysis show that selected genes using integrated approach can separate luminal, non-luminal subcategory utilising various factors including age, disease free survival, and relapse free survival.

Keywords: feature selection; deep learning; breast cancer; molecular subtype; SMOTE; best first search; CFS-BFS.

Reference to this paper should be made as follows: Lamba, M., Munjal, G. and Gigras, Y. (2023) 'Identifying breast cancer molecular class using integrated feature selection and deep learning model', *Int. J. Bioinformatics Research and Applications*, Vol. 19, No. 1, pp.19–42.

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 an Assistant Professor, Visiting Faculty and her current examination works incorporate the use of machine learning.

Geetika Munjal has a teaching experience of more than 16 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 25 BTech projects, 14 MTech theses. She has published 30 research 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 thirteen 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 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

In bioinformatics field, the prediction of breast cancer (BC) is considered as one extremely substantial research areas. By tradition, categorisation of BC is merely relied on clinical testimony and needs diagnostic specialist knowledge for the biological analysis. The foremost task is the precise categorisation of cancers for enhancing medication and prognosis in medical cancer research. As per the details from centres for disease control and prevention (Miller et al., 2012), more than 1.7 million occurrences have initiated among females. With increase in age, there is a threat among women of age 50 years and above are commonly discovered experiencing from the BC, but practically 11% of the BC is now discovered amongst women under age of 45 years of age. In some cases, it has been found that BC is developed at a very early age which is a major concern for research as it results into physical and psychological burden (Dai et al., 2016; Gilbert et al., 2008).

BC categorisation to its accurate subtype is essential for prescribing the finest conceivable medication to the patients. The prediction and prognosis depending on molecular subtype is such a valuable material concerning BC, which is quite vital in the ability to define its treatment strategy. Thus, an automated and accurate way to identify subtype of BC is required that may take advantage of computational intelligence.

Computer-aided diagnosis system (CAD) appears to be very beneficial for breast radiologists for encouraging diagnosis of cancer in terms of time and correctness (Gilbert et al., 2008; Lehman et al., 2015; Doi, 2007; Gromet, 2008; Lamba et al., 2021a; Jung et al., 2014). CAD-based classifier is essential to support health experts in the primary BC detection. An additional and accountable method to categorise BC gene expression differ on molecular genes, concluded by an examination referred to as PAM50 (Smith et al., 2002). In the beginning of 1970, categorisation of BC started with the status of estrogen receptor (ER). Additionally, other clinicopathological constraints such as lymph node metastasis, histological grade, tumour size, and three well-known indicators HER2, ER, and PR competed a crucial role in treatment choice. BC detection varies based on comprehensive study of intrinsic BC subcategory as LumB and LumA. LumA tumours are low-grade, luminal are ER+, HER2–, and PR+ tumours, LumB have high Ki-67, high-grade, PR–, PR+, and HER2– or HER2+ (Clark et al., 2011; Lamba et al., 2021b; Harris et al., 2012; Foukakis and Bergh, 2016; Lamba et al., 2022a; Metzger-Filho et al., 2013; Lamba et al., 2021c) from the last two decades.

Table 1 Facts of molecular sub-categories of BC

Sub-categories	*Molecular subcategory comprehensive explanation*
Luminal A	LumA develops slowly, commonly found among every race and age
	Good prognosis and low recurrence rate
	PR+, ER+, tumour grade 1 or 2, low Ki-67 value and HER2–
	Treatment is hormonal therapy (Lamba et al., 2021b; Harris et al., 2012; Foukakis and Bergh, 2016; Lamba et al., 2022a; Metzger-Filho et al., 2013; Lamba et al., 2021c)
Luminal B	LumB develop and diagnosed at younger age than LumA
	Marginally worse prognosis
	PR+, ER–, poorer tumour grade, HER2+/–, large tumour size, lymph node positive and high Ki-67 value
	Treatment is chemotherapy, hormone therapy focusing HER2 (Clark et al., 2011; Lamba et al., 2021b; Harris et al., 2013; Metzger-Filho et al., 2013; Partridge et al., 2016)
Normal	Normal is equivalent to LumA and have low-level Ki-67
	Good prognosis and somewhat worse than LumA
	PR+ and/ or ER+ and HER2– (Dai et al., 2016)
HER2–	HER2– is diagnosed at young age in comparison to LumA and LumB
	Poor prognosis
	PR–, ER–, lymph node positive and HER2+
	Treatment is combination of chemotherapy, surgery, and radiation therapy (Clark et al., 2011; Lamba et al., 2022a)
Triple negative/basal	Triple negative is found higher in females undergoing from BRCAI gene mutations and develop quicker than luminal cancer
	Worse prognosis
	ER–, HER2– and PR–
	Treatment is chemotherapy radiation therapy focusing non-HER2 (Clark et al., 2011; Lamba et al., 2021b; Harris et al., 2013; Lamba et al., 2021c)

In connection with such subtypes fluctuate in their complexity (genomic), prognosis and key genetic repetitions. The existence rate of LumA is healthier in comparison to the rest groups, as in all tumours, the low grade remains the constant sign. These sub-categories similarly arise in DCIS (ductal carcinoma in situ) (Clark et al., 2011). Detail's explanation of molecular subtypes is described in Table 1.

Exact molecular grouping is a critical phase towards the BC seriousness recognition due to:

a existence of very limited methods like PAM50 for predicting molecular class of BC

b uncertainty and inconsistency about number of molecular classes available for BC

c preventing patients from undesirable high-priced therapies.

Molecular classification is in research from many years however these classification techniques face various issues as:

1 like over-fitting, high computation cost

2 the dynamic nature of micro-array data may lead to dynamic and inconsistency in predicting BC subtype.

As a result, this manuscript recommends an innovative and valuable method for identifying molecular subtype of BC using deep learning (DL).

Classification of BC is a very vital task as it will help the patients to prevent them from going undesirable high-priced therapies. The present study is partitioned into subsequent sections, Section 2 is related work and literature review, Section 3 introduces the datasets, Section 4 introduced the proposed model explaining approach implemented in detail, Section 5 comprise of classification approaches, Section 6 describes the performance measure, Section 7 states experimental outcomes with biological validation using Kaplan Meier survival (KMS) model and discussion is included in Section 8, Section 9 concluded the paper as conclusion and future aspects.

2 Related work

To study about BC and its categorisation, a reasonable and trustworthy approach is necessary to enhance inconclusive medication. Classification continues to be a monitored learning that could support the scheme to study and categorise the new data from the information established on that learning. In literature, numerous classification algorithms like random forest, neural network (Cao et al., 2019; Wang et al., 2020; Lamba et al., 2022b), support vector machine (SVM), etc. exist. Several research have utilised categorisation tactics for BC-associated difficulties stated as in year 1996, two methods C4.5 (Akay, 2009) and RIAC (rule induction algorithm) scored accuracy of 97.8% and 96% respectively. Fuzzy genetic and neuro fuzzy techniques have achieved accuracy of 97.36% and 95.06% respectively in 1999. Neuro-rule method in 2000 scored 98.1% accuracy. AIRS (artificial immune system) and big LVQ (optimised learning vector quantisation) obtained accuracy of 97.2 and 96.8% respectively. Supervised fuzzy clustering, SVM robustness (Polat and Güneş, 2007), SVM (Übeyli, 2007), and SVM along with feature selection (Geetika, 2012) have obtained good accuracy of 96.8%, 98.53%, 99.54% and 99.51% respectively. Particle swarm optimisation (PSO) (Dheeba

et al., 2014), RS-BPNN (rough set relation) (Nahato et al., 2015) and deep belief NN (DBNN) (Abdel-Zaher and Eldeib, 2016) secured accuracy of 93.67%, 98.6%, and 99.68% respectively. All these tactics have demonstrated encouraging results to a certain extent.

Commonly medical data has suffered from the class imbalance dilemma; therefore, it leads to miscalculation (Zhang et al., 2019). Investigators have concentrated on the pre-processing tactic to make the innovative information balanced by means of over-sampling or under-sampling. SMOTE is one such technique that is being used along with numerous existing methods (Bunkhumpornpat et al., 2009). To reduce the noise generated by SMOTE, Verbiest et al. (2014) use fuzzy as a choice algorithm. For resolving the issue of imbalance data, Zeng integrate SMOTE with kernel into SVM (Zeng et al., 2009; Gandhi and Dhanasekaran, 2013; Jayachandran and Dhanasekaran, 2014; Mahendran and Dhanasekaran, 2015). In 2011, to enhance SMOTE, Geo uses PSO and RBF to minimise the misclassification (Gao et al., 2011). To derive performance, Jeatrakul created SMOTE with a neural network (Jeatrakul et al., 2010). SMOTE integrated with SVM to outperforms low dimensional data in various instances (Rok and Lusa, 2013). Reflecting the significance of SMOTE in managing imbalanced dataset, present work has used SMOTE considering its benefits.

The first step after the pre-processing is the extraction of relevant genes using an approach known as feature selection. After those extracted genes are utilised for categorisation. BC traditional approaches of classification use morphology to distinguish tumours in different category depending on behaviour and prognosis (Eliyatkın et al., 2015). Molecular classification of BC is in continuous study from past 11 years. It has been noticed that multiple classification technique goes through overfitting, a lot of time is taken by training process and its computation seems to be too expensive (Tomar and Agarwal, 2013). In multiple areas of research concerned with genomics, DNN has been implemented magnificently (Dong et al., 2019; Arisdakessian et al., 2019; Abdel-Zaher and Eldeib, 2016). In modern studies, DNN has been utilised using denoising autoencoder with micro-array gene expression datasets of BC (Kumar and Misra, 2019). In Mendez et al. (2019), DL successfully carried out linear regression for obtaining important genes and achieved higher accuracy in comparison to various shallow learning technique to the categories of ER– and ER+ (Alakwaa et al., 2018). Taking into consideration the benefits of deep neural network-based method, this manuscript used DL in a supervised stage to construct the proposed model. In most bioinformatics issues, DL has performed well by choosing the suitable genes (Chen et al., 2020; Lamba et al., 2018). Hence, utilising the tactic of best-first and correlation gene selection in the proposed model to obtain the most important and significant relevant genes.

3 Datasets

The experimentations are implemented on micro-array datasets given in Table 2, gathered from National Center for Biotechnology Information (NCBI) advances (Sayers et al., 2021). Micro-array datasets consist of complex and high-number of genes. The mentioned datasets experience dimensionality curse as the samples are too less in comparison to number of genes. It also suffers from data imbalance problem. Molecular class wise distribution of experimental dataset is detailed in Table 3, where molecular

subcategories like normal, claudin and HER2 are having very less count of samples as depicted in figure 1, hence giving an additional challenge in the task of classification.

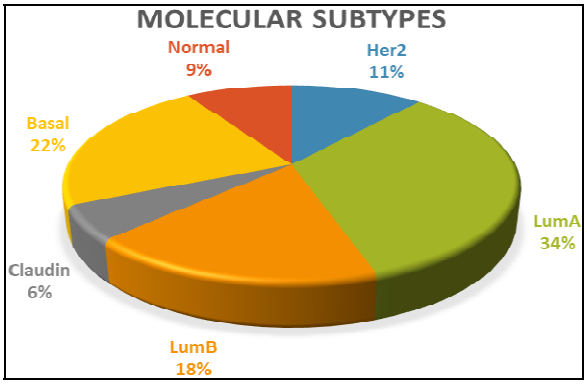
Table 2 Description of experimental datasets

Datasets	*GSE10886*	*GSE20262*	*GSE21997*	*GSE25055*	*GSE18229*	*Total*
No. of genes	16,381	13,342	16,381	13,497	12,612	
No. of samples	120	174	31	329	212	866

Table 3 Number of samples in molecular types of different datasets

Molecular category	*GSE10886*	*GSE20624*	*GSE21997*	*GSE25055*	*GSE18229*	*Total*
HER2	12	19	5	40	22	98
LumA	51	67	4	99	70	291
LumB	26	45	4	43	37	155
Claudin	10	13	7	0	19	49
Basal	12	24	5	122	32	195
Normal	9	6	6	25	32	78
Total	120	174	31	329	212	866

Figure 1 Description of molecular subcategory distribution in dataset (see online version for colours)



4 Proposed model

In this study, CFS-BFS is used to select genes according to molecular subtype of BC and followed by DL for classification task. Detailed steps are as follows:

a Firstly, probe-ids are mapped with gene names followed by normalisation of data using function known as min-max. Once the normalisation is done, discretisation is followed by SMOTE to balance the minority class with the help of k-nearest neighbour.

b Secondly, for searching and evaluator, best-first search (BFS) and correlation-based searching (CFS) are used respectively in feature selection.

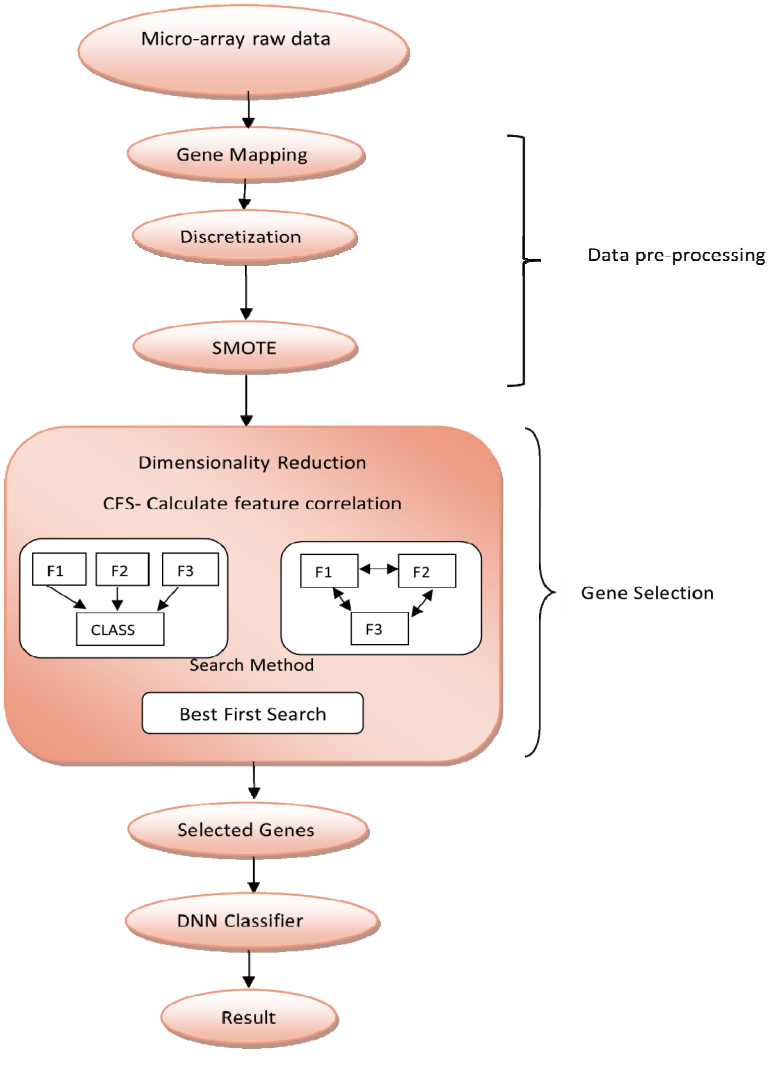
c Thirdly, after obtaining the relevant genes, DL is used with soft-max activation function to perform the classification.

d Performance of newly proposed model with multiple shallow learning techniques using the selected genes on various parameters.

e Using KMS model to evaluate the prognosis of BC patients depending on age, over-all survival (OS), and disease free survival (DFS).

Figure 2 represent the flowchart of proposed model. Datasets given in Table 2 belongs to distinct platform, i.e., GEO platform (GPL)

Figure 2 Flowchart of proposed method (see online version for colours)



4.1 Data normalisation

Gene mapping using GEOquery in R (Allaire, 2012), the probe-ids are replaced with gene names utilising gene mapping. Data is normalised between [0, 1] using min-max function. This function is expressed using equation (1), where a represents the gene intensity of any sample, min(a), max(a) represent the minimum and maximum value of gene intensity, respectively. This is followed by discretisation.

$$Normalise = \frac{a - \min(a)}{\max(a) - \min(a)} \tag{1}$$

4.2 Discretisation

The need of discretisation arises when converting continuous value to discrete values. Discretisation is presented by Setiano and Liu, statically clarified heuristic technique recognised as chi^2 (chi-square) (Tsai and Chen, 2019). It is a statistical characteristic that is predetermined and arranged by introducing each gene cost into its interval as expressed in equation (2) is termed as chi-square to identify whether relative frequencies of the multiple classes in adjacent intervals are adequate to justify merging. Formula for calculating the adjacent interval is given as:

$$x^2 = \sum_{a=1}^{2} \sum_{b=1}^{A} \frac{\left(A_{ab} - E_{ab}\right)^2}{E_{ab}} \tag{2}$$

where A is count of molecular classes, A_{ab} is the count of instances/values in a^{th} interval of b^{th} class. E_{ab} is the expected frequency (probability count) of A_{ab} calculated as:

$$A_{ab} = R_a * \left(\frac{C_b}{N}\right) \tag{3}$$

R_a is the count of instances in i^{th} interval where $\sum_{b=1}^{C} A_{ab}$. C_b is count of instances of class b and N is total count of instances.

4.3 SMOTE

The total samples are distributed not in the uniformity among the dissimilar groups of molecular subtypes in BC as defined in Table 3. Class imbalance is the difficulty associated with data, to resolve SMOTE (synthetic minority over-sampling method) is used.

It generated synthetic samples via k-nearest neighbour to balance the minority class. Integration of discretisation and SMOTE helps in enhancing and improving outcomes (Jishan et al., 2015). The subsequent steps are followed to conclude the oversampling assignment:

1 Discovering the minority class set N, for every $x \in N$, x is calculated by the Euclidean distance for k-nearest neighbour and every single sample present in N.

2 Every $x \in N$, the sampling rate R is agreed varying on the unbalanced ratio. R examples $d_1, d_2, \ldots, d_z$ ($R \leq z$) are preferred purposelessly including k-nearest neighbour, therefore produce the set N_1.

3 In every new example $d_z \in N_1$ (z =1, 2, 3, …, R), the stated formulation is applied to establish the new sample

$$d_{new} = d + rand(0, 1) * \left\| (d - d_z) \right\| \tag{4}$$

where $rand(0, 1)$ will produce a number in the range 0 and 1 and d_{new} is new example generated to balance the minority class.

4.4 Feature selection method

As soon as the data complete the pre-processing steps, choosing most important and relevant genes play an epic role in the performance of classification (Lamba et al., 2020, 2023). For selecting these important genes, a proposed model is taken into consideration consisting of BFS as method of searching and CFS as method of gene evaluator. Wrapper feature selection approach is BFS with the supervision of supervised feature selection. Choosing important genes by indicating the appropriate algorithm which will suits well with the data plays a very crucial role. Algorithm faces multiple issues while learning, to choose the best genes subgroup by selecting and rejecting gene.

Therefore, discovering the finest working of the learning algorithm is the objective. It is utmost significant to determine the association amongst feature-to-feature associativity and feature subset selection. It aids in finding the optimal feature personalised to a machine learning (ML) algorithm.

In proposed feature selection method CFS uses correlation coefficients that search for genes which are the extremely correlated with all their predecessors. It starts with constructing a subgroup of genes with the attribute, i.e., the most associated with the others. Formerly, correlation coefficients amongst the chosen feature and the rest of the parameters are calculated. The attribute with the maximum correlation value is designated as the second feature. The obtained subgroup of two genes is additional extended by adding the attribute of the correlation coefficient with the higher value between the subgroup and remaining parameters. The procedure of appending the genes of the higher correlation values is repeated unless all the correlation coefficients designate statistically significant dependencies (respective values exceed thresholds) or the count of genes in the subgroup is equal to the determined percentage of the total count of attributes.

CFS is one of the most credible that helps in creating ranking of genes subsets according to the association depending on empirical estimation function. CFS computes feature-feature and feature-class association to establish the matrix in the micro-array datasets. The basis of estimation measure is confronting the subgroups which comprise of genes genuinely unmatched in conjunction with each other and similar with class. Inappropriate genes are those having low association with the class, such genes are disregarded. Unwanted genes are retained out as they are enormously one or more with the remaining genes (Wosiak and Zakrzewska, 2018). The consent of a gene typically be subject to the degree it will forecast the correct output in a zone of the illustration space. Though, in circumstances where certain extremely prognostic genes stood excluded

might worsen the performance of ML. P_r stand for CFS's feature subgroup estimation function defined in equation (5):

$$P_r = \frac{l\overline{rcf}}{\sqrt{l + l(l-1)\overline{rff}}} \tag{5}$$

P_r is empirical 'merit' of a feature subcategory r, comprising of l genes, $\overline{rcf}$ signify as average correlation value between feature and class association and $\overline{rff}$ signify average correlation value among two genes.

Equation (5), represent the correlation coefficient. It is shows that the correlation between feature and class variable is a function of the number of genes in the composite and magnitude of the inter-correlation among them, together with the magnitude of the correlations among genes and the class variable. Entering two illustrative values for $\overline{rcf}$ in equation (5), and allowing the values of l and $\overline{rff}$ to vary.

Based on the P_r, few observations are made:

1 Greater the correlations among the genes and the class variable, the greater the correlation between composite and the class variable.

2 The lower the inter-correlations among the components, the higher the correlation between the set of genes and the class variable.

3 As the number of genes in the composite increases (considering the additional components are the same as the original components in terms of their average intercorrelation with the other genes and with the class variable), the correlation between the set of genes and the class variable increases.

Experiments reveals CFS provide similar outcomes to the wrapper, which outpaced properly on small datasets (Li et al., 2017). Furthermore, CFS accomplishes several times more rapidly than wrapper hence, CFS is applied to choose the ultimate appropriate genes of the complete datasets as labelled in Table 4.

Table 4 Count of gene chosen in five micro-array datasets

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

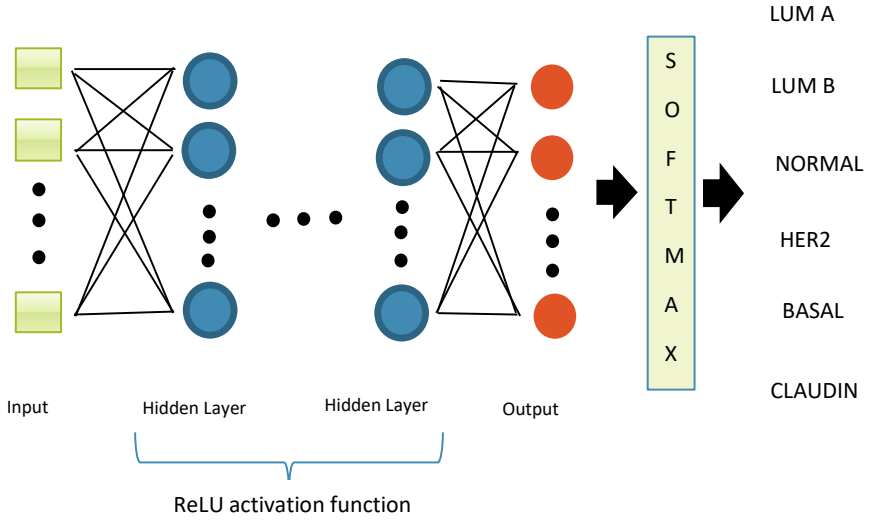
The motivation of linking BFS with CFS as feature assessor is the fact that it facilitates in pinpointing the extremely beneficial genes, facilitates in eliminating noisy, inappropriate, and duplicate genes if their significance does not rely strongly on remaining genes. Integration of BFS and CFS improves eradicating fifty percent of the genes. Usually, classification accuracy is equivalent in manipulating the reduced group of genes in contrast with original set of genes. Hence, integrated approach initiates with an empty set and completes forward searching with the full cluster of genes. Later during backward searching it start investigating in directions at any phase, thereby deleting and adding genes. Algorithm explores the potential subgroup of genes utilising greedy hill climbing method enhanced through backtracking improvement. When recognising minimised and relevant genes, so subsequent duty is to classify the genes.

5 Classification method

Numerous studies show good performance of DL aimed at categorisation task. DL is advantageous across other methods of ML specifically for large datasets. DL performance in complicated problems is much better.

All these advantages have inspired us to discover DL for this multi-classification assignment.

Figure 3 A DL-based BC classification (see online version for colours)



Architectural design of DL is demonstrated in Figure 3, influencing the count of neurons and concealed layers will be grounded on trial-and-fault rule.

Input is the set of selected genes that are fed into the model for the learning process. Weights are provided to those genes which pay more attention towards learning, it is done through scalar multiplication among weight matrix and input layer. Then, transfer function is used to combine multiple inputs of genes into one molecular class output value. Further, processing is done through the hidden layer utilising the activation function, this layer is treated as intermediate layers which helps in doing all the computations. Multiple interconnected hidden layers can be used that follow trail and fault rule, in current model four hidden layer are used. This layer account for searching different hidden genes in the data. Output layer is consisting of molecular classes as LumA, LumB, normal, HER2, basal and claudin.

The DL design has been built upon numerous computational layers. Every single layer acknowledges the input and utilises to produce the result. The result is nonlinear function comprising of linear grouping of regulated weight, threshold, and input layer in conjunction through the assistance of mistake that can propagated back.

In forward propagation, every neuron result as a nonlinear computation of the weighted sum of the previous layer to which the neuron in an output, described in equation (6).

$$z = f\left(\sum_i \omega_i y_i + c\right) \tag{6}$$

where y_i signify the input of the activation, ω_i signify the weight, c represent the bias and z indicate the activation output.

Table 5 CM of classifier DL (see online version for colours)

Confusion matrix	*HER2*	*LumA*	*LumB*	*Claudin*	*Basal*	*Normal*	*Total*
HER2	95	1	2	0	0	0	98
LumA	1	281	6	0	0	2	290
LumB	2	7	143	2	2	0	156
Claudin	1	0	1	46	1	0	49
Basal	1	2	1	0	191	0	195
Normal	2	4	0	1	0	71	78
Total	102	295	153	49	194	73	866

Significant role is played by activation function because of amalgamation of random linear pattern/model. To find the solution of complicated problem, activation function changed to nonlinear. Numerous activation functions are available such as rectified linear unit (ReLu), tanh, sigmoidal, etc. although ReLu required a lesser amount of time for computation and delicate. Crucial advantage of utilising ReLu, it assists in lessening the interconnectedness of principles that results in overcoming the existence of overfitting and is causing the rarity of the network. ReLu equation is described as:

$$f_{ReLu} = \max(x, 0) \tag{7}$$

When the existing propagation gets finalised, Mcxent as loss function used to discover the difference between the objective and the forecast amount to assess the projected model effectiveness. The output of the concealed layer is a likelihood dispersion with SoftMax function. It is used to generate output as variety of chances. The result provides the likelihoods of every class and uppermost likelihood in goal class of multi-class problem.

SoftMax is stated in equation (8), where z_i denotes to the amount of every element in a logit and e is a numerical constant. It helps as it switches the output layer as likelihood dispersal (Chung et al., 2016; Lamba et al., 2021d). The sum of component of output $S(z_i)$ is 1. Ten-cross validation (Arlot and Celisse, 2010) is applied to validate the experimental results.

$$S(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{8}$$

6 Performance measure

Numerous metrics are utilised to assess the proposed model depending on confusion matrix (CM). CM mentioned in Table 5, helps in comprehending the efficacy of the model in term of Mathew's correlation coefficient (MCC), sensitivity, precision, fallouts, and f-score. The elements of CM are true negative (TN) appears incorrect but it is true. True positive (TP) appears correct but it is true. False negative (FN) is predicted incorrect and it is incorrect. False positive (FP) is predicted accurately and seems to be incorrect.

Recall/sensitivity is defined number of truly categorised right/correct by total number of positive.

$$Sensitivity/Recall = \frac{TP}{P} \tag{9}$$

F-score facilitates to assess the twice the product of *recall* (R) and *precision* (P) divided by sum of precision and recall.

$$F\text{-}score = \frac{(2 * P * R)}{(P + R)} \tag{10}$$

MCC helps to overcomes the category imbalance dilemma.

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

Accuracy is measure of precise and correct pointer, and it provide the straightforward detail of the classifier such as how several genes are not classified accurately, and formulated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

The average of balanced accuracy for each class predicted as per formulation:

$$Balanced\ accuracy = \frac{Recall + Specificity}{2} \tag{13}$$

Specificity/fallout is count of TN divided by total count of negatives.

$$Fallout = \frac{TN}{N} \tag{14}$$

7 Experimental outcomes

The result produced by classification methods before feature selection are mentioned in Table 6, where SVM shown better precision results (highlighted in italic). Utilising the advantages of feature selection in reduction of dataset and improving the classification methods results are clearly seen better in case of CFS-BFS mentioned in Table 7.

The proposed model shows a better result with CFS-BFS feature selection method and accuracy of five micro-array datasets with DL is highlighted in Table 7.

The projected DL paradigm takes led to the finest performance in case of molecular subtyping where the results corresponding to various classifiers and performance measures are highlighted in Tables 8–10.

The performance result of sensitivity is approx. 99% with basal and normal subtypes. Fallout is minimum for LumA and normal as 0.0164% and 0.0014% respectively. Highest precision is achieved as 0.9672%, 0.94% and 0.9818% for LumA, LumB and normal respectively. We have attained satisfactory outcomes with 0 percentage fallout on

claudin subtype. In case of basal subtype, recall is 0.9934%. The value of MCC is 0.9918% on normal subtype. The result of F-score is 0.9682%, 0.9374% and 0.9638% for LumA, LumB and normal respectively.

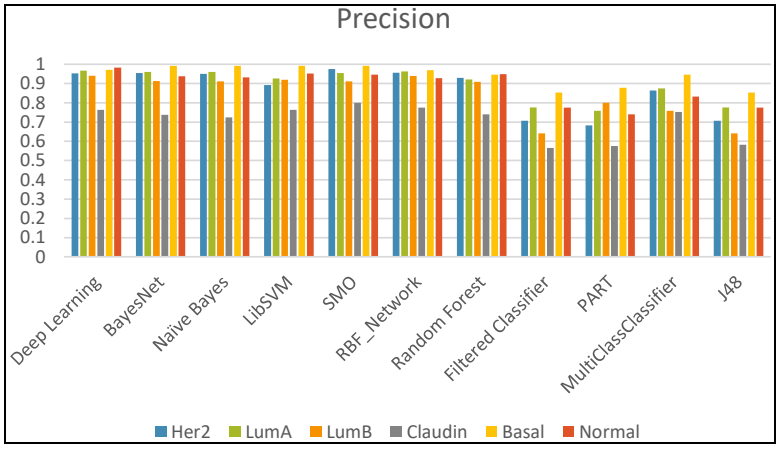
Table 6 Precision performance of classifiers before feature selection

Datasets/classification algorithms	*GSE10886*	*GSE20262*	*GSE21997*	*GSE25055*	*GSE18229*
DL	0.746	0.729	0.729	0.79	0.814
SMO	*0.787*	*0.82*	*0.822*	*0.825*	*0.864*
BayesNet	0.65	0.704	0.704	0.757	0.654
J48	0.671	0.565	0.565	0.684	0.729
Random forest	0.579	0.681	0.69	0.721	0.784
PART	0.624	0.601	0.601	0.691	0.701
Filtered classifier	0.576	0.615	0.615	0.691	0.705

Table 7 Accuracy of feature selection methods using DL classifier

Datasets	*CFS + Subset_Forward*	*Filtered_Attribute + BFS*	*CFS_BFS*
GSE25055	0.832	0.853	*0.949*
GSE18229	0.826	0.885	*0.948*
GSE10886	0.845	0.822	*0.983*
GSE21997	0.736	0.72	*1*
GSE20624	0.7336	0.855	*0.937*

Figure 4 Precision of DL in comparison to other ML methods on molecular subtypes (see online version for colours)



Out of total samples, luminal (LumA and LumB) are 51.5% and non-luminal (basal, normal, claudin and HER2) are 48.5%.

Performance of DL in comparison with shallow ML methods are evaluated using

a overall precision corresponding to molecular subtypes

b balanced overall precision

c fallout corresponding to various molecular subcategory

d unclassified samples.

All the above points are mentioned in Figures 4–7 for various classifiers. Balance precision helps to overcome the problem of imbalanced data. Perfect precision is achieved with zero misclassified samples by random forest, Bayes net, SMO, Bayes net, RBF network, and DL. In comparison to all the mentioned classifiers, DL has shown the minimum count, i.e., 0.0450% of samples are misclassified for 866 samples. Among 11 classifiers, the performance of SMO and DL have shown respectable performance in standings of TP_Rate, fallout, recall, precision, MCC and F-score.

Figure 5 Precision of DL in comparison to other ML methods on five datasets using balanced accuracy estimated by mean of balanced precision per class (see online version for colours)

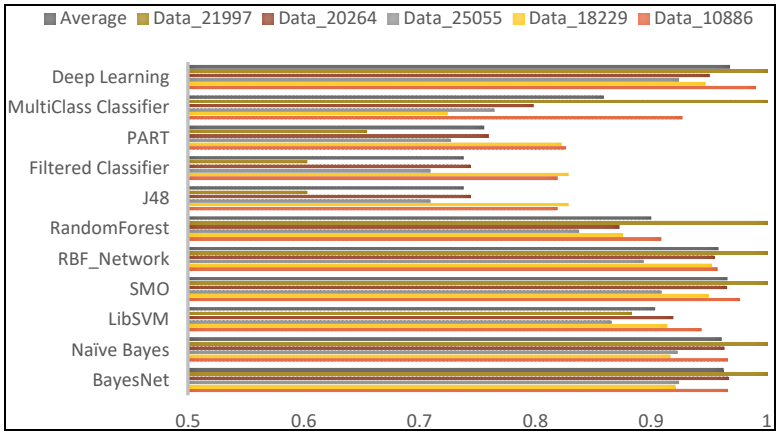


Figure 6 Line plots of fallout of molecular subcategory using 11 ML methods on five datasets (see online version for colours)

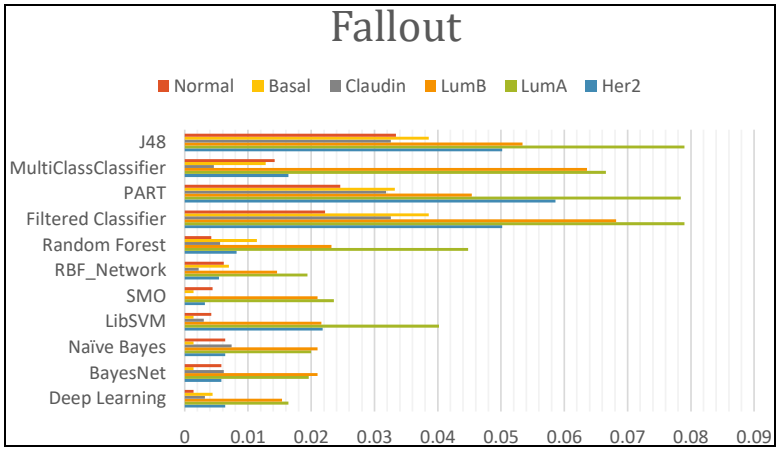


Figure 7 Unclassified samples of each micro-array datasets (see online version for colours)

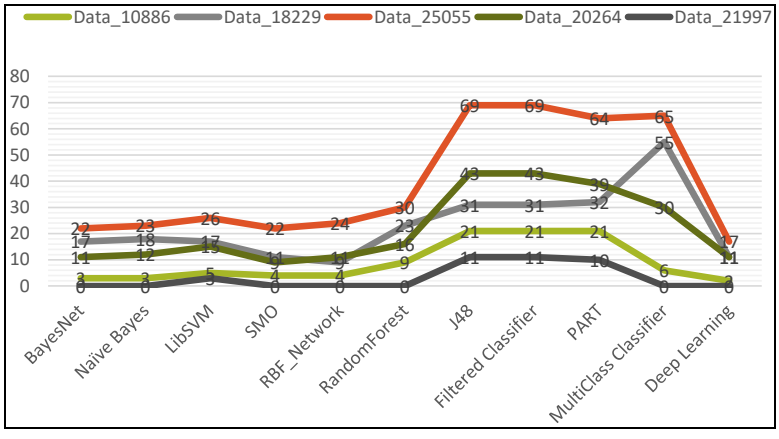


Table 8 Performance of TP_Rate in case of molecular subtypes

TP_Rate/sensitivity	*HER2*	*LumA*	*LumB*	*Claudin*	*Basal*	*Normal*
Deep learning	*0.9774*	*0.9662*	*0.9386*	0.7602	*0.9914*	*0.9918*
Naïve Bayes	0.9556	0.9306	0.902	0.7432	0.9624	0.9528
PART	0.7188	0.7036	0.6322	0.5766	0.907	0.7142
RBF_Network	0.951	0.9502	0.9214	0.7406	0.9774	0.9464
BayesNet	0.9582	0.9312	0.9088	0.7504	0.9636	0.956
LibSVM	0.9432	0.8186	0.8762	0.7464	0.9736	0.9216
Random forest	0.9004	0.921	0.8698	0.7214	0.9296	0.877
Filtered classifier	0.7036	0.6854	0.5718	0.5702	0.8628	0.7618
SMO	0.972	0.9312	0.9138	*0.7942*	0.9724	0.9552
MultiClassClassifier	0.8226	0.8226	0.7332	0.739	0.8646	0.8062
J48	0.7036	0.6854	0.5718	0.5702	0.8628	0.7618

Table 9 Performance of F1-score in case of molecular subtypes

Classifier/molecular types	*HER2*	*LumA*	*LumB*	*Claudin*	*Basal*	*Normal*
Deep learning	*0.9656*	*0.9682*	*0.9374*	0.7636	0.9816	*0.9638*
Naïve Bayes	0.9602	0.9544	0.9168	0.7458	0.975	0.946
PART	0.7182	0.7542	0.709	0.5706	0.8944	0.7244
RBF_Network	0.9562	0.9678	0.9352	0.7598	0.9754	0.93
BayesNet	0.9626	0.955	0.9224	0.7526	0.976	0.949
LibSVM	0.9158	0.8776	0.8964	0.7566	0.984	0.9018
Random forest	0.9146	0.9466	0.8824	0.7356	0.9454	0.853
Filtered classifier	0.7104	0.7498	0.6202	0.574	0.8728	0.7666
SMO	0.975	0.956	0.9184	*0.7946*	*0.9832*	0.9478
MultiClassClassifier	0.8478	0.8778	0.7844	0.7228	0.8978	0.8182
J48	0.7104	0.7498	0.6202	0.574	0.8728	0.7666

Table 11 lists the parameters that are employed in ML methods. Entire experimentation is done on Weka 3.9.4 (Hall et al., 2009) and R studio 1.2.5019 (R Core Team, 2013).

Table 10 Performance of MCC in case of molecular subtypes

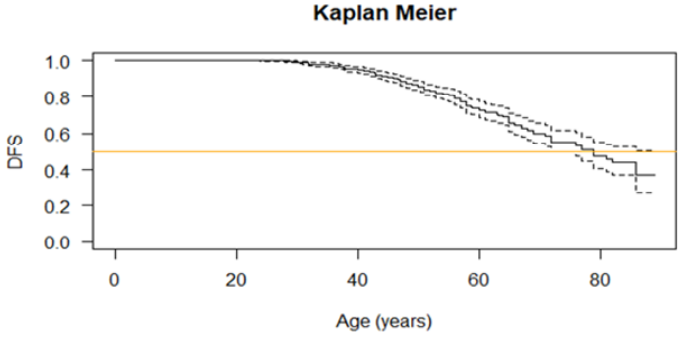
Classifier/molecular types	*HER2*	*LumA*	*LumB*	*Claudin*	*Basal*	*Normal*
Deep learning	*0.9774*	*0.9666*	*0.9364*	0.7602	*0.9826*	*0.9918*
Naïve Bayes	0.9556	0.931	0.8972	0.7414	0.9696	0.9416
PART	0.6736	0.6764	0.6598	0.5394	0.8694	0.71
RBF_Network	0.951	0.9506	0.9192	0.7578	0.9686	0.9256
BayesNet	0.9582	0.9316	0.904	0.7486	0.9708	0.9448
LibSVM	0.9056	0.852	0.874	0.7524	0.9808	0.9008
Random forest	0.9052	0.9184	0.8568	0.7302	0.9344	0.8562
Filtered classifier	0.6638	0.6734	0.5468	0.5426	0.8418	0.7468
SMO	0.972	0.9328	0.9002	*0.7942*	0.9796	0.944
MultiClassClassifier	0.8298	0.816	0.7338	0.72	0.879	0.8044
J48	0.6638	0.6734	0.5468	0.5426	0.8418	0.7468

7.1 KMS model

Patients having luminal type cancer generally have better survival rate compared to non-luminal. Luminal type generally has ER+ whereas non-luminal has ER-status, so ER-group has poor prognosis (Lang et al., 2012; Dunnwald et al., 2003). To regulate if the genes selected using proposed model can distinct the non-luminal (bad prognosis) and luminal (good prognosis) patients, utilising the disease/relapse free survival rate (DFS/RFS) knowledge in the dataset, KMS plots are presented. KMS assessment is done using R-project package called 'survival' (Hall et al., 2009) to perform the survival analysis amongst luminal and non-luminal groups for the micro-array datasets, that generated the DFS curves as shown in Figures 8–11. Implementing KMS analysis for age and DFS survival shown in Figures 8 and 9. Figures 10–11, presents the KMS analysis for DFS/RFS survival luminal and non-luminal patient groups. Figures 12–13, presents survival analysis using Kaplan Meier using overall survival for separating luminal and non-luminal patients. Together survival analysis graphs in Figures 8–13 display great separation among the two prognosis groups. To evaluate the *p*-value, a log-rank test was evaluated that signify that lower p value is the superior separation among luminal and non-luminal subtypes. The log-rank statistical test yielded *p*-value of $8e^{-12}$, that had been statistically noteworthy (i.e., $P < 0.001$) and indicated respectable partition amongst the two groups shown in Figure 4(b). Comparably KMS scrutiny based on DFS, *p*-value is $1e^{-13}$ that is significant statistically to a good difference among luminal and non-luminal groups. In overall survival attained *p*-value is $6e^{-09}$, i.e., significant statistically to give a good separation among luminal and non-luminal subcategory.

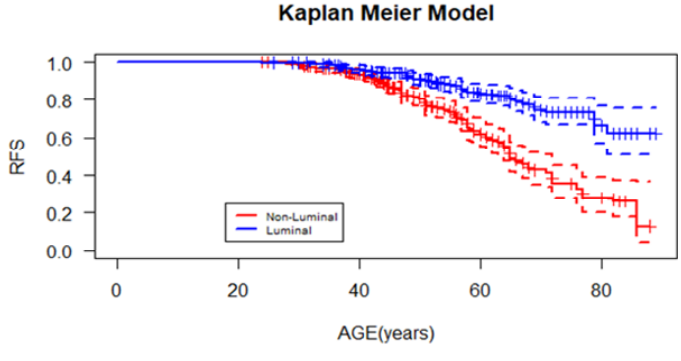
These outcomes authenticate that proposed model is successful in dividing BC patients at the foundation of the DFS rate, into two diagnosis groups which can ascertain the patient's expectation level for an event (relapsed at any site or died of disease). This consequently helps in easy identification of the patient's group which might necessitate less or more hostile medication strategy.

Figure 8 KMS graph for patients of BC in the micro-array datasets for DFS vs. age (see online version for colours)



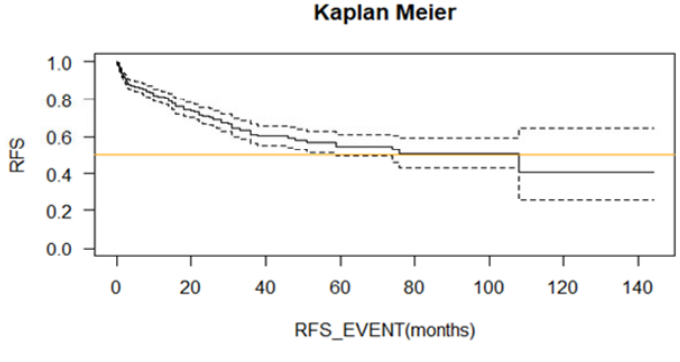
Note: Incorporating the DFS rate with age that distinguishes good and poor prognosis groups.

Figure 9 KMS graph for luminal and non-luminal patient groups in the micro-array datasets (see online version for colours)



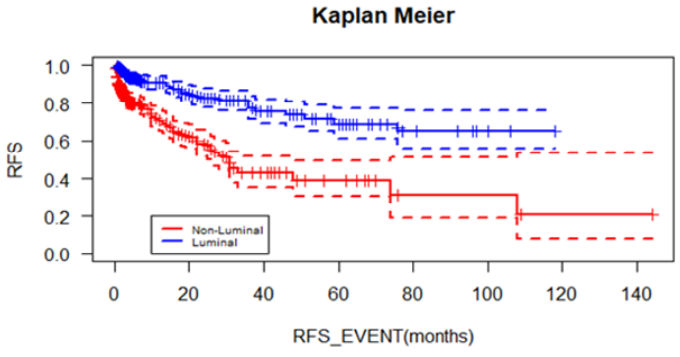
Note: Uniting the RFS rate to differentiate between luminal with ER+/good prognosis and non-luminal with ER–/poor prognosis groups.

Figure 10 KMS graph for patients of BC in the micro-array datasets for relapse free survival vs. RFS_event (see online version for colours)



Note: Incorporating the RFS rate with event that distinguishes good and poor prognosis groups.

Figure 11 KMS graph for luminal and non-luminal patient groups in the micro-array datasets (see online version for colours)



Note: Incorporating the RFS rate to differentiate between luminal with ER+/good prognosis and non-luminal with ER–/poor prognosis groups.

Figure 12 KMS graph for patients of BC in the micro-array datasets for overall survival (see online version for colours)

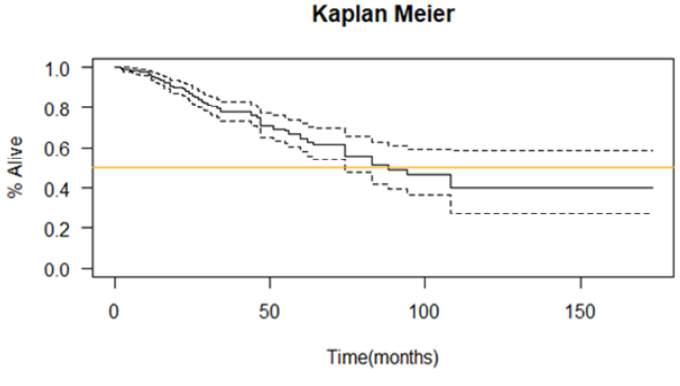
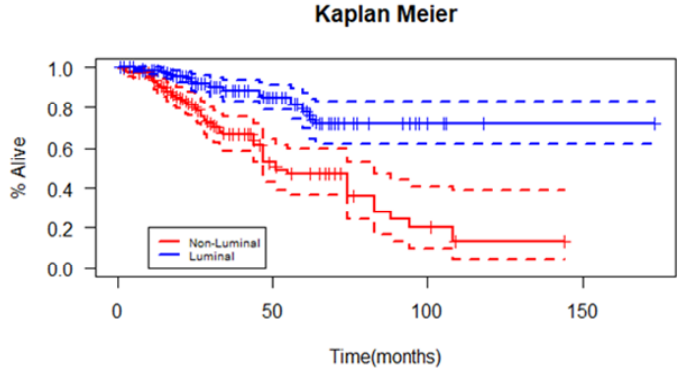


Figure 13 KMS graph for luminal and non-luminal patient groups in the micro-array datasets (see online version for colours)



Note: Incorporating the overall survival rate to differentiate between luminal with ER+/good prognosis and non-luminal with ER–/poor prognosis groups.

Table 11 List of parameters of ML algorithms

Machine learning algorithm	*Parameters*
Deep learning	Num of epochs: 10
	Batch size: 100
	Activation function: ReLu
	Loss function: LossMcxent
Naïve Bayes	Batch size: 100
	Works on Bayes formula
PART	Batch size: 100
	Confidence factor: 0.25
	numFolds: 3
	MDL correction is used for finding splits
RBF network	Batch size: 100
	maxIts: –1
	minStd: 0.1
	numClusters: 2
	Basis function k-means (value is 2) clustering using linear regression and implemented using normalised Gaussian radial basis function network
	Ridge: $1e^{-8}$
Bayes net	Batch size: 100
	Estimator: SimpleEstimator, alpha = 0.5 and search algorithm = K2
LibSVM	SVM type: C-SVC
	Batch size: 100
	Kernel: radial basis function
	Nu: 0.5
Random forest	Batch size: 100
	Count of execution slots is 1 for constructing the ensemble, maximum depth of tree is zero (unlimited) and count of genes obtained as $\log_2$ (number of predictors) + 1
Filtered classifier	Filter used is discretise and classifier is J48
SMO	Batch size: 100
	Calibrator: logistic
	Epsilon: $1e^{-12}$
	Filter type: normalise
	Kernel: polykernel
	Tolerance parameter: 0.001
MultiClassClassifier	Batch size: 100
	Classifier: logistic
J48	Batch size: 100
	Confidence factor of 0.25 for pruning (smaller value incur more pruning) and MDL correction is used for finding splits

8 Discussion

A key role in prognosis and diagnosis of BC is played by selecting important genes from thousands of genes present in the micro-array dataset. Molecular subtypes of BC are used in conjunction with the CFS-BFS approach to first choose key genes. Molecular subtype is discovered after significant genes are provided as input into the DL algorithm. DL has delivered incredibly promising results in a variety of criteria when compared to the outcomes of ML algorithms. Numerous hidden layers are preferred based on the number of genes, and the activation function utilised in the design is crucial for achieving consistent performance. Further the KMS model is used to display the BC prognosis. Depending on the subtype survival rate, such as luminal vs. non-luminal, luminal has a better prognosis than non-luminal.

9 Conclusions and future scope

The work concentrates mainly molecular BC classification where integrated feature selection along with DL has played a crucial role in overall performance. Due to the complexity of the micro-array datasets, it is observed that pre-processing plays an important role. Feature selection models have incorporated superiority of BFS and CFS method and selected very few significant genes and smote has taken care of class imbalance issue. Based on the proposed feature selection approach, all the chosen ML algorithms have achieved adequate results where DL have given excellent results with the short-listed genes. DL model is extremely scalable; though, the learning time needed through DL is very high. Categorising the molecular subtypes into luminal and non-luminal subcategory helps in deciding its prognosis. For prognosis purpose KMS model shows that patients suffering from luminal type of BC have better chances of survival in comparison to non-luminal.

Potential future effort is required to discover more in-depth information about BC diagnosis exploring more feature selection and DL architectures. Because of the resilient architecture of DL, it could be utilised to acknowledge heterogeneity in multiple type of cancer including BC. As feature selection tactic and DL architectures are compliant to enormous data therefore, they may be able to be helpful in scrutinising more complex data thus giving solution in understanding complex disease.

References

Abdel-Zaher, A.M. and Eldeib, A.M. (2016) 'Breast cancer classification using deep belief networks', *Expert Systems with Applications*, Vol. 46, No. C, pp.139–144.

Akay, M.F. (2009) 'Support vector machines combined with feature selection for breast cancer diagnosis', *Expert systems with applications*, Vol. 36, No. 2, pp.3240–347.

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

Allaire, J. (2012) *RStudio: Integrated Development Environment for R*, Vol. 770, No. 394, pp.165–171, Boston, MA.

Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. and Garmire, L.X. (2019) 'DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data', *Genome Biology*, 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.

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, pp.475–482.

Cao, Y., Cao, Y., Wen, S., Huang, T. and Zeng, Z. (2019) 'Passivity analysis of delayed reaction – diffusion memristor-based neural networks', *Neural Networks*, Vol. 109, pp.159–167.

Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., Feng, X., Feng, X., Zhang, Y., Duan, M. and Huang, L. (2020) 'Feature selection may improve deep neural networks for the bioinformatics problems', *Bioinformatics*, 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*, IEEE, pp.348–352.

Clark, S.E., Warwick, J., Carpenter, R., Bowen, R.L., Duffy, S.W. and Jones, J.L. (2011) 'Molecular subtyping of DCIS: heterogeneity of breast cancer reflected in pre-invasive disease', *British Journal of Cancer*, Vol. 104, No. 1, pp.120–127.

Dai, X., Xiang, L., Li, T. and Bai, Z. (2016) 'Cancer hallmarks, biomarkers and breast cancer molecular subtypes', *Journal of Cancer*, Vol. 7, No. 10, p.1281.

Dheeba, V., Singh, N.A. and Singh, J.A.P. (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)*, pp.111–118, Springer, Cham.

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

Dong, Y., Yang, W., Wang, J., Zhao, J., Qiang, Y., Zhao, Z., Kazihise, N.G., Cui, Y., Yang, X. and Liu, S. (2019) 'MLW-gcForest: a multi-weighted gcForest model towards the staging of lung adenocarcinoma based on multi-modal genetic data', *BMC Bioinformatics*, Vol. 20, No. 1, pp.1–4.

Dunnwald, L.K., Rossing, M.A. and Li, C.I. (2007) 'Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients', *Breast Cancer Research*, Vol. 9, No. 1, pp.1–10.

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

Foukakis, T. and Bergh, J. (2016) *Prognostic and Predictive Factors in Early, Non-Metastatic Breast Cancer*, Dizon Ed.

Gandhi, M. and Dhanasekaran, R. (2013) 'Diagnosis of diabetic retinopathy using morphological process and SVM classifier', in *2013 International Conference on Communication and Signal Processing*, IEEE, pp.873–877.

Gao, M., Hong, X., Chen, S. and Harris, C.J. (2011) 'A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems', *Neurocomputing*, Vol. 74, No. 17, pp.3456–3466.

Geetika, G. (2012) 'A survey of classification methods and its applications', *International Journal of Computer Applications*, Vol. 53, No. 17, pp.14–16.

Gilbert, F.J., Astley, S.M., Gillan, M.G., Agbaje, O.F., Wallis, M.G., James, J., Boggis, C.R. and Duffy, S.W. (2008) 'Single reading with computer-aided detection for screening mammography', *New England Journal of Medicine*, Vol. 359, No. 16, pp.1675–84.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) 'The WEKA data mining software: an update', *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp.10–18.

Harris, J.R., Lippman, M.E., Osborne, C.K. and Morrow, M. (2012) *Diseases of the Breast*, Lippincott Williams & Wilkins, Philadelphia, PA.

Jayachandran, A. and Dhanasekaran, R. (2014) 'Severity analysis of brain tumor in MRI images using modified multi-texton structure descriptor and kernel-SVM', *Arabian Journal for Science and Engineering*, Vol. 39, No. 110, pp.7073–7086.

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

Jishan, S.T., Rashu, R.I., Haque, N. and Rahman, R.M. (2015) 'Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique', *Decision Analytics*, Vol. 2, No. 1, pp.1–25.

Jung, N.Y., Kang, B.J., Kim, H.S., Cha, E.S., Lee, J.H., Park, C.S., Whang, I.Y., Kim, S.H., An, Y.Y. and Choi, J.J. (2014) 'Who could benefit the most from using a computer-aided detection system in full-field digital mammography?', *World Journal of Surgical Oncology*, Vol. 12, No. 1, pp.1–9.

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

Lamba, M., Munjal, G. and Gigras, Y. (2018) 'Feature selection of micro-array expression data (FSM) – a review', *Procedia Computer Science*, Vol. 132, pp.1619–1925.

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.

Lamba, M., Munjal, G. and Gigras, Y. (2021a) 'ECABC: Evaluation of classification algorithms in breast cancer for imbalanced datasets', in *Data Driven Approach Towards Disruptive Technologies*, pp.379–388, Springer, Singapore.

Lamba, M., Munjal, G. and Gigras, Y. (2021b) 'A hybrid gene selection model for molecular breast cancer classification using a deep neural network', *International Journal of Applied Pattern Recognition*, Vol. 6, No. 3, pp.195–216.

Lamba, M., Munjal, G. and Gigras, Y. (2021c) 'A MCDM-based performance of classification algorithms in breast cancer prediction for imbalanced datasets', *International Journal of Intelligent Engineering Informatics*, Vol. 9, No. 5, pp.425–454.

Lamba, M., Gigras, Y. and Dhull, A. (2021d) 'Classification of plant diseases using machine and deep learning', *Open Computer Science*, Vol. 11, No. 1, pp.491–508.

Lamba, M., Munjal, G. and Gigras, Y. (2022a) 'Supervising healthcare schemes using machine learning in breast cancer and internet of things (SHSMLIoT)', *Journal: Internet of Healthcare Things*, pp.241–263.

Lamba, M., Munjal, G. and Gigras, Y. (2022b) 'Ranking of classification algorithm in breast cancer based on estrogen receptor using MCDM technique', *International Journal of Information Technology & Decision Making*, Vol. 22, No. 2, pp.1–25.

Lamba, M., Munjal, G. and Gigras, Y. (2023) 'Computational studies in breast cancer', *Research Anthology on Medical Informatics in Breast and Cervical Cancer*, pp.434–456.

Lang, K., Huang, H., Namjoshi, M., Federico, V. and Menzin, J. (2012) 'Initial treatment and survival among elderly breast cancer patients in the United States by estrogen receptor status and cancer stage at diagnosis: an analysis of national registry data 2000–2009', in *Cancer Research*, AMER Assoc. Cancer Research, 615 Chestnut St., 17th Floor, Philadelphia, PA 19106-4404, USA, Vol. 72.

Lehman, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N., Miglioretti, D.L. and Breast Cancer Surveillance Consortium (2015) 'Diagnostic accuracy of digital screening mammography with and without computer-aided detection', *JAMA Internal Medicine*, Vol. 175, No. 11, pp.1828–1837.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H. (2017) 'Feature selection: a data perspective', *ACM Computing Surveys (CSUR)*, Vol. 50, No. 6, pp.1–45.

Mahendran, G. and Dhanasekaran, R. (2015) 'Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms', *Computers & Electrical Engineering*, Vol. 45, pp.312–323.

Mendez, K.M., Broadhurst, D.I. and Reinke, S.N. (2019) 'The application of artificial neural networks in metabolomics: a historical perspective', *Metabolomics*, Vol. 15, No. 11, pp.1–4.

Metzger-Filho, O., Sun, Z., Viale, G., Price, K.N., Crivellari, D., Snyder, R.D., Gelber, R.D., Castiglione-Gertsch, M., Coates, A.S., Goldhirsch, A. and Cardoso, F (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*, Vol. 31, No. 25, p.3083.

Miller, J.W., King, J.B., Joseph, D.A., Richardson, L.C. (2012) 'Centers for disease control and prevention (CDC) breast cancer screening among adult women – behavioral risk factor surveillance system', *MMWR Morb. Mortal Wkly. Rep.*, USA, Vol. 61, No. Suppl., pp.46–50.

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

Partridge, A.H., Hughes, M.E., Warner, E.T., Ottesen, R.A., Wong, Y.N., Edge, S.B., Theriault, R.L., Blayney, D.W., Niland, J.C., Winer, E.P. and Weeks, J.C. (2016) 'Subtype-dependent relationship between young age at diagnosis and breast cancer survival', *Journal of Clinical Oncology*, 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*, Vol. 17, No. 4, pp.694–701.

R Core Team (2013) 'R: A language and environment for statistical computing'.

Rok, B. and Lusa, L. (2013) 'SMOTE for high-dimensional class-imbalanced data', *BMC Bioinformatics*, Vol. 14, No. 1, pp.106–121.

Sayers, E.W., Beck, J., Bolton, E.E., Bourexis, D., Brister, J.R., Canese, K., Comeau, D.C., Funk, K., Kim, S., Klimke, W. and Marchler-Bauer, A. (2021) 'Database resources of the national center for biotechnology information', *Nucleic Acids Research*, Vol. 49, No. D1, p.D10.

Smith, R.A., Cokkinides, V., von Eschenbach, A.C., Levin, B., Cohen, C., Runowicz, C.D., Sener, S., Saslow, D. and Eyre, H.J. (2002) 'American Cancer Society guidelines for the early detection of cancer', *CA: A Cancer Journal for Clinicians*, Vol. 52, No. 1, pp.8–22.

Tomar, D. and Agarwal, S. (2013) 'A survey on data mining approaches for healthcare', *International Journal of Bio-Science and Bio-Technology*, Vol. 5, No. 5, pp.241–266.

Tsai, C.F. and Chen, Y.C. (2019) 'The optimal combination of feature selection and data discretization: an empirical study', *Information Sciences*, Vol. 505, pp.282–293.

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

Verbiest, N., Ramentol, E., Cornelis, C. and Herrera, F. (2014) 'Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection', *Applied Soft Computing*, Vol. 22, pp.511–517.

Wang, S., Cao, Y., Huang, T., Chen, Y. and Wen, S. (2020) 'Event-triggered distributed control for synchronization of multiple memristive neural networks under cyber-physical attacks', *Information Sciences*, Vol. 518, No. 1, pp.361–375.

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

Zeng, Z.Q., Wu, Q., Liao, B.S. and Gao, J. (2009) 'A classfication method for imbalance data set based on kernel SMOTE', *Acta Electronica Sinica*, Vol. 37, No. 11, pp.2489–2495.

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