TWO-STAGE GENE SELECTION TACTIC FOR IDENTIFYING SIGNIFICANT PROGNOSIS BIOMARKERS IN BREAST CANCER

Monika Lamba

missmonikalamba@gmail.com

Geetika Munjal

Munjal.geetika@gmail.com

Yogita Gigras

yogitagigras@ncuindia.edu

Abstract:

The mining of a subset of informative genes from microarray gene expression data is a significant data preparation task in the classification of breast cancer. Out of all the algorithms developed, CFS-BFS and CONSISTENCY-BFS are the two best ones for gene selection. For reliable prognostication of breast cancer subtypes, a ground-breaking 2-Stage Gene Selection algorithm has been developed. Using CFS-BFS in the first stage and CONSISTENCY-BFS in the second, the majority of the distracting, inappropriate, and redundant genes are removed. To improve algorithm efficacy, the 2-Stage GeS strategy gets around the uncertainty problem with CFS-BFS. Surprisingly, using Hidden Weight Naive Bayes to establish the 2-Stage GeS, more accurate and reliable results are obtained. The standings of recall, precision, f-score, and fallout show encouraging results. The top four genes E2F3, PSMC3IP, GINS1 and PLAGL2 were further verified by applying Kaplan-Meier Survival Model. E2F3 and GINS1 are likely targets for precision therapy.

Keywords: CFS-BFS, Consistency-BFS, gene selection, micro-array gene expression dataset, breast cancer, Kaplan Meier Survival.

1. Introduction

Breast cancer (BC) is a wide variety of diseases with highly adaptable medical behaviours, not a single disease. [1-2]. Diagnosticians have long recognised this morphological multiplicity, which is replicated in Histological Grades (HG) with dissimilar microscopic appearances and correlated with medical outcomes [3-4]. HG, which stands for the morphologic assessment of tumour genetic traits, is a well-established prognostic factor that has been successful in generating significant evidence regarding the clinical behaviour of the disease. [5-6]. The HG scheme shown in Table 1 typically takes the patient's severity into account. These anomalies give clinicians tasks to look for likely targets for the best BC detection and diagnosis. [7].

Table 1. Description of histological grade.

Grade Types	**Growth [8]**	**Mitotic Count (per 10 high power fields)**	**Tubular Differentiation (tumour forms glands)**	**Nuclear Pleomorphism**
Grade 1	Slowly, well-differentiated	< 7 mitoses	> 75%	Small nuclei, no nucleoli, and uniform cells.
Grade 2	Moderate	8-15 mitoses	10 - 75 %	Bigger cells using open vesicular nuclei, moderate in shape and size, visible nucleoli.
Grade 3	Faster, poor differentiation	> 16 mitoses	< 10 %	Cells with variation in size, shape, vesicular nuclei, and prominent nucleoli, marked.

A higher grade may develop and quickly blow out, requiring immediate aggressive treatment. A lower grade denotes slow-growing cancer with a better prognosis. It is still impossible to develop an accurate medical indicator that will commit for improving prognosis and grade-related data [7]. In order to express a tumor's antagonistic behaviour, HG aims to combine measurements of cellular differentiation and replicative potential into a composite score.

The Nottingham Grading System (NGS) is the utmost extensively used technique for BC tumour grading. The grading system of tumour cells is grounded on a microscopic estimation of cytologic and morphologic characteristics, which also include nuclear pleomorphism, mitotic count, and degree of tubule formation [7-9].

The summation of the grading scores classified breast tumours into the following grades:

a. G1 - grade 1 (slow-growing, exceptionally differentiated)
b. G2 - grade 2 (slightly differentiated)

c. G3 - grade 3 (inadequately differentiated, highly proliferative) malignancies.

HG acts as an imperative part in the prognosis, diagnosis, and survival of BC patients. It is becoming a key area to categorize the patients into the correct category and stage of BC. The Genetic Grade (GG) was consistently conceived in multivariate analyses to be a self-determining prognostic symbol of disease reappearance proportionate to lymph node and tumour size status [10-13]. When combined with the Nottingham Prognostic Index (NPI), GG improved the identification of patients with less damaging and destructive tumours who would benefit sufficiently from adjuvant treatment. The findings of Anna et al. show that a GG signature can advance, improve, and facilitate prognosis planning for BC patients, as well as provide comfort that high-grade and low-grade ailment, as stated genetically, replicate separate pathobiological entities rather than a continuation of cancer development [10].

In BC, Micro-Array Gene Expression (MAGE) has the potential to judge thousands of genes simultaneously. Machine Learning (ML) technique has optimized this analysis task. According to research, MAGE-based profiling can provide better and self-determining prognostic information for patients with BC. MAGE data contains many genes, the majority of which are irrelevant or unimportant in the diagnosis of BC. Gene selection will aid in the discovery of relevant genes, and it is useful in a variety of real-world applications, such as identifying relevant genes for a specific disease in microarray data [14-15]. The Best-First Search (BFS) method produces excellent results [13], even when accuracy rankings are average. It also has the greatest influence on the prognostication model. The CFS built on BFS selects the fewest possible features on its own. [16-19, 46-49]. To reduce the genes further with a motive to find biomarker genes, Consistency-BFS is beneficial. Integrating the Hidden Naïve Bayes with 2-Stage GeS has been discussed in detail to predict BC accurately.

This study aims to identify prognostic biomarkers on microarray datasets to forecast the diagnosis and prognosis of breast cancer based on histologic grade subtypes. In future cancer research, the proposed novel architecture demonstrates a cost-effective and powerful predictive tool.

The literature review is covered in Section 2, the GeS method, Hidden Weight Nave Bayes, and the GeS method in detail are covered in Section 3, and the proposed model is highlighted in Section 4. Datasets and experimentation analysis are covered in Section 5. The conclusion and discussion are presented in the last section.

2. Literature Survey

Sankara et al. [9] presented a consolidative approach to recognize Grade-specific biomarkers for BC and constructed networks using grade-specific molecular interactions of cancer Grades 1, 2, and 3 through DEGs (Differentially Expressed Genes). The author discovered a Grade 3 molecular network that is primarily associated with cancer-related procedures. Amongst the top ten associated DEGs in Grade 3, the increment in the expression of the CCNB2 and UBE2C genes was analytically noteworthy among dissimilar grades. Additionally, the expression of the genes CCNB2 and UBE2C, CDK1, KIF2C, CCNB2, and NDC80 is highly pronounced in various Grades and lowers the patient survival rate. Together, the recognised genes can serve as biomarkers for BC diagnosis and prognosis.

Cases were rediverted as one of the following molecular subcategories: LumA; LumB (HER2-); LumB (HER2?); Basal; HER2 subcategory; and five negative phenotypes, as discussed by Engstrm et al. in their discussion of immunohistochemistry and in situ hybridization as alternatives for analysing gene expression. The studies made use of Kaplan-Meier Survival (KMS) models and Cox proportional hazards models. HER2 had the worst prognosis and diagnosis based on molecular subcategory, while LumA had the best prognosis and diagnosis along with five negative phenotypes in the first five years following investigation. Only Grade2 tumours exhibit subcategory-related changes in BC survival. According to histopathological grade or molecular subcategory, there was no difference in the survival of BC after the time of diagnosis. Lymph node, GG, and tumor size status are robust prognostic factors. The high grade related to the non-luminal subcategory [20-21]. The information on prognosis and prediction for the various factors can change after diagnosis [22]. An additional source of prognostic and predictive data regarding the patient's outcome may come from BC's molecular subtyping. Nevertheless, its clinical and medical ramifications have not yet been fully appreciated. The primary goals of the study were to ascertain whether classifying BC into molecular subtypes (Non-Luminal and Luminal) provides more precise and in-depth information regarding outcomes related to traditional HG and to examine BC-specialized survival in the various molecular and grade subcategories. The prognosis varied significantly by molecular subtype during the first five years following diagnosis, with the Luminal A subtype having the best prognosis and HER2 and the five negatives' phenotypes having the worst [20]. While Grade 1 tumours are linked to the best diagnosis, Grade 3 tumours are linked to the worst diagnosis. Although some cases of Grade 2 tumours may resemble Grade 3 and 1 and are more heterogeneous in nature, most cases have a transitional prognosis. [23-24]. Furthermore, examining patients' Gene Expression Profiles (GEP) across different grades and molecular subtypes of BC, aids in determining thoughtful pathogenesis and planning appropriate treatment strategies.

Relating the Histologic Grade with the prognosis of breast cancer along with feature selection [25]. Where the first feature sent is either full or empty. By employing forwarding selection and gradually adding features, it expands the exploration space. Later, it uses backward search to reduce the exploration space by removing one gene at a time.

Utilizing the two steps feature selection, the overall paper comprises of

- The most appropriate genes related to breast cancer will be found through a thorough analysis of correlation and consistency measures with best-first search. Experimental evaluation of identified genes using different classification methods.
- Generating the ranking of important genes by 2-stage GeS tactic.
- Medical validation of identified genes using the Kaplan Meier survival model.

3. GeS Technique

The GeS approach to feature exploration concluded with the finest subgroups of features, and an attempt to discover a subgroup X amongst the challenging 2^X candidate groups. The necessity of this approach is its stopping condition: it avoids comprehensive exploration of subgroups. The GeS technique (shown in Figure 1) primarily involves the following four steps [15]:

1. Creating the succeeding candidate subgroup for the assessment using the *generation technique*
2. Estimating the candidate subgroup utilizing the *estimation function*.
3. When to stop exploring is indicated by the stopping condition, and
4. Validate the subgroups using the *validation technique*.

The generation technique employs an exploratory strategy to generate subgroups of features for evaluation. It begins by employing all or no features or a random subgroup of features. An estimation function helps in the generation of a subgroup, an optimum subgroup is constantly compared to an estimation function like the linear correlation coefficient [27]. In the absence of an appropriate stopping condition, the GeS procedure might run, repeatedly ending up as a liability for the exploration approach. The generation technique and estimation function can affect the judgment or preference aimed at a stopping condition. Instances of stopping conditions grounded based on the generation technique comprise either a predefined count of features or a predefined count of repetitions attained. Instances of a halting condition grounded in an estimation function either facilitate the deletion or addition of any feature, generating an improved subgroup, or an optimum subgroup is attained. The GeS procedure stands still by outputting the chosen subgroup of features. i.e., later authenticated. There are numerous variations to this GeS method, but the vital stages of generation, estimation, and stopping condition are performed in almost every procedure. The authentication practice is not an essential fragment of the GeS method itself. It attempts to examine the genuineness of the chosen subgroup by comparing and verifying the outcomes with earlier established outcomes or with the outcomes of challenging the GeS approach using real-world or artificial datasets.

To deal with dimensionality reduction, Gene Selection [7,15,17,28] is a potent method. GeS is utilized to discover an “optimum” subgroup of significant features, therefore the comprehensive accuracy is amplified although the data size is made smaller, and the comprehensibility is enhanced in the case of classification. GeS approaches comprise two vital characteristics one is the estimation of a candidate feature subgroup and the second is exploration using the feature space.

The GeS is implemented using two techniques named:

a. *Inconsistency* measure corresponds to a feature subgroup i.e. Unpredictable as at least two illustrations through equivalent feature principles through distinctive class markers.

b. *Correlation* measures correspond to correlation either among features or among classes and features.

Contrasting inconsistency measure with correlation measure and studying Best-First Search (BFS) as an inspecting approach.

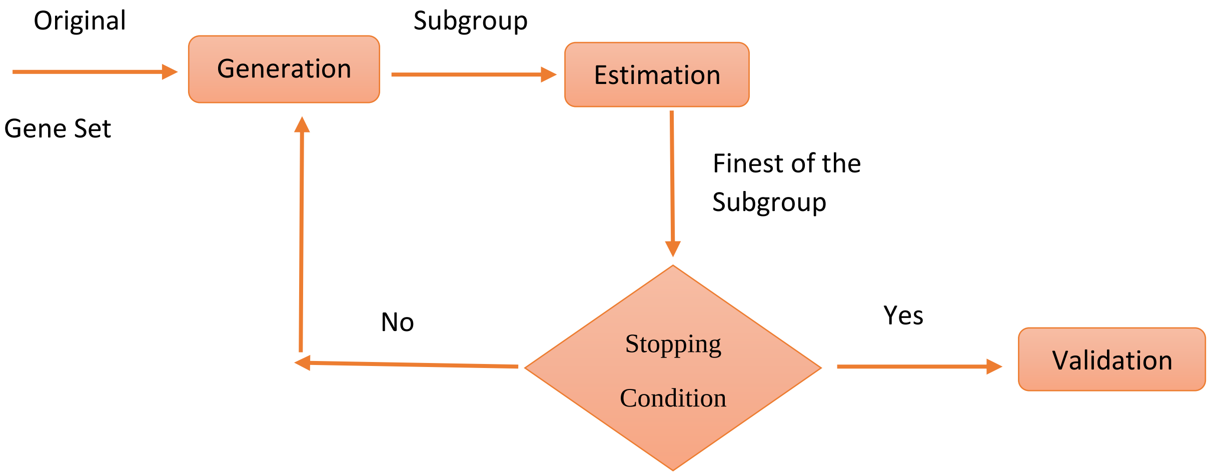

Figure 1. GeS approach

To deal with dimensionality reduction, gene selection [7,15,17,28] is a potent method. GeS is utilized to discover an "optimum" subgroup of significant features, therefore the comprehensive accuracy is amplified although the data size is made smaller, and the comprehensibility is enhanced in the case of classification. GeS approaches comprise two vital characteristics one is the estimation of a candidate feature subgroup and the second is exploration using the feature space.

The GeS is implemented using two techniques named:

a. Inconsistency measure corresponds to a feature subgroup i.e. Unpredictable as at least two illustrations through equivalent feature principles through distinctive class markers.
b. Correlation measures correspond to correlation either among features or among classes and features.

Contrasting inconsistency measure with correlation measure and studying best-first search (BFS) as an inspecting approach.

4. Proposed Model

4.1 Data Pre-processing

In the current study, an innovative 2-GeS model for BC categorization into Histologic Grade subtype is proposed with a Hidden Weight Naïve Bayes (HWNB) classifier shown in Figure 2. In the beginning pre-processing of data is done in the form of Gene Mapping, replacing probe-ids with their corresponding gene IDs utilizing the GEOquery library of R Studio [29], systematizing the gene data employing the min-max method. After mapping, SMOTE and Discretization are performed on the datasets to beat the problem of class unevenness [28,30-31]. The pre-processed data contains thousands of genes, of which only a small number are important. To generate the subgroup of relevant genes, 2-Stage GeS is performed where CFS (Correlation-Based Searching) and Best-First Search (BFS) is applied at the first stage. Consistency is used as an evaluator and best-first search is applied in the second stage to find the final genes after relevant genes have been chosen using CFS-BFS (Correlation Feature Selection and Best-First Search). Further, the classification of BC is carried out using different supervised machine-learning algorithms. Gene produced using 2-stage GeS has enhanced the performance of HWNB over other ML methods.

Since the data is imbalanced, so it creates an extreme repercussion on the performance of the ML algorithms. To resolve this issue, SMOTE is executed after discretization; the inclusion of discretization and SMOTE aided in improving performance results.

The problematic issue is concerning the imbalance in the datasets. In SMOTE, synthetic examples are generated with the k-NN (k-nearest neighbor) tactic for the smaller class to resolve the problem of imbalance data. The following steps are taken for the oversampling task:

Step 1: Identifying the marginal class set Q, for every $b \in Q$, k-NN of b is produced by calculating the distance between b and each instance present in A.
Step 2: For every $b \in Q$, the sampling rate T is calculated as liable on the imbalanced proportion.
T instances $t_1, t_2, \ldots t$ ($T \leq m$) are selected aimlessly amongst k-NN, therefore, producing the set Q_1.
Step 3: For each example $t_m \in Q_1$ (m=1,2,3,……,T), the stated method is utilized to generate the new instances

$$t_{new} = t + rand(0,1) * ||(t - t_m)|| \qquad \text{eq. 1}$$

Where t_{new} is a new instance, and $rand(0,1)$ will produce a number that lies on [0,1].

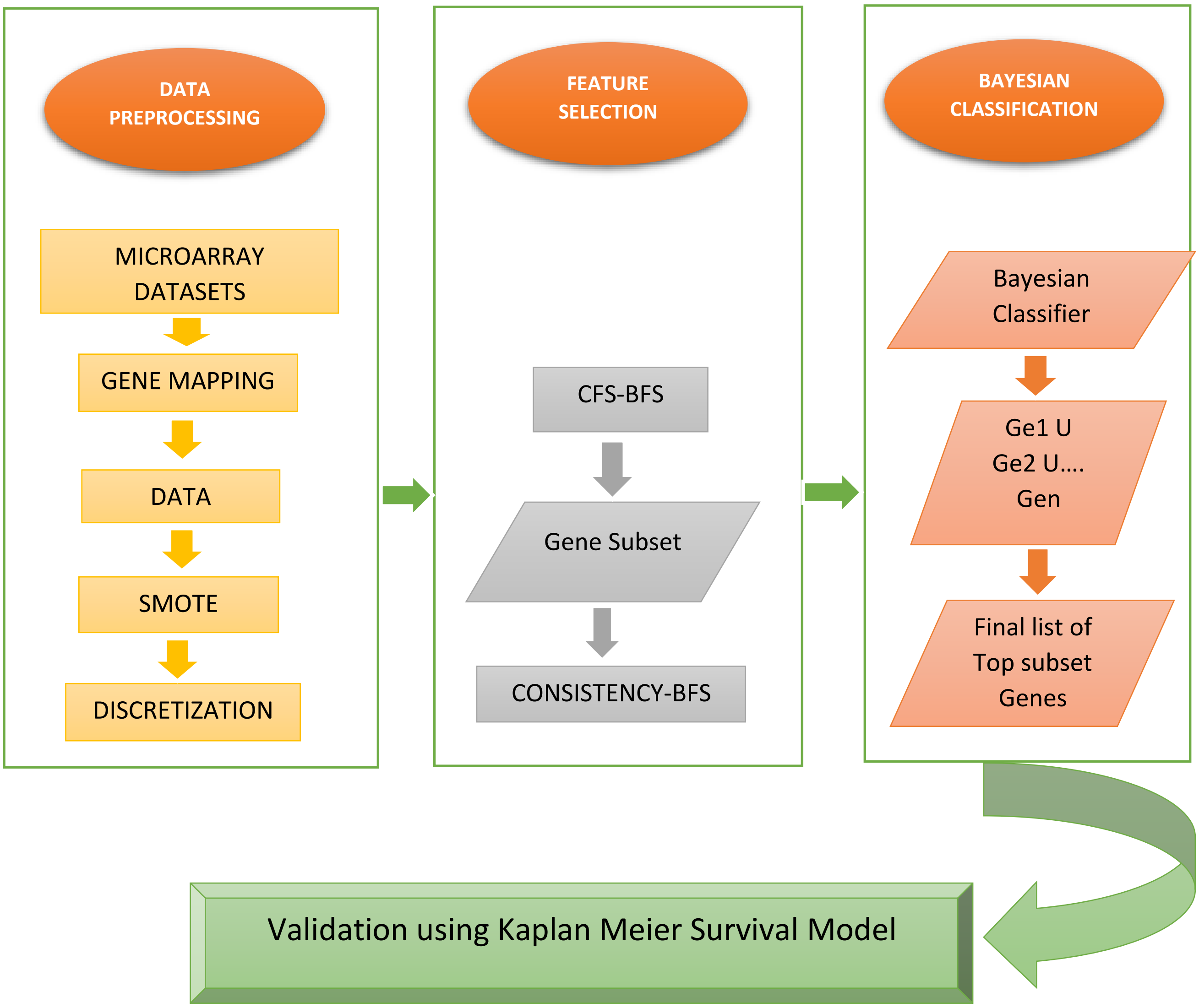

Figure 2. Flowchart of 2 Stage GeS

4.2 GeS Method

To find out the subgroup of relevant genes, a combination tactic is utilized which includes two GeS methods. The first is CFS-BFS at the first stage and Consistency-BFS at the second stage, in which CFS and Consistency act as gene evaluators and BFS acts as an exploration method for gene subgroup. The BFS technique falls under the category of supervised Gene Selection (GeS). By indicating which genes, the algorithm thinks, fit the data the best, chooses the relevant and significant genes. The algorithm encounters a number of difficulties as it learns to determine which genes are relevant and which ones to eliminate. Determining the best genes for the algorithm is therefore GeS's primary goal. The choice of the best gene for the ML technique by filter approach depends on the gene-to-gene correlation and gene subgroup selection, which are important to ascertain. The CFS method is a reliable one because it generates a ranking of genes grounded on associativity determined by the empirical valuation function. By examining each gene's unique ability to predict how much attrition will occur among them, CFS can estimate the value of a subgroup of an attribute. Although there is little association, the subgroup of highly interrelated genes with the class is selected. [17]. Though, a few extremely predictive genes were disregarded which might worsen the performance of ML. *Ac* signifies CFS's gene subgroup assessment function given as:

$$A_C = \frac{f\overline{cr_P}}{\sqrt{f + f(f-1)\overline{C_{PP}}}} \qquad \text{eq. 2}$$

A_c is the experimental 'merit' of a gene subgroup , including of f genes, $\overline{C_{PP}}$ demonstrates gene-gene intercorrelation and $\overline{crp}$ epitomize the gene-class association. According to studies [32], CFS produces results that are comparable to those of the wrapper that outperformed them well on small datasets. In addition, CFS implements much more quickly than wrapper; as a result, CFS is used to select the final appropriate genes.

Despite the fact that training occurrences in the subgroups of qualities are predictable, the Consistency BFS GeS method [33] estimates the value of a subgroup of qualities in the class standards by the level of uniformity. The consistency of the subgroup cannot, under any circumstances, be less than the consistency of the entire set of qualities. As a result, the standard training is to use this subgroup evaluator in aggregate with an exhaustive or random search, which looks for the smallest subgroup with consistency that is equal to the consistency of the entire set of qualities. Consistency measures (CM) are treated differently on the training dataset because of their strong backing and use of min-genes when selecting a subcategory of genes [34]. The goal of min-genes is to define consistency theories over the fewest number of genes possible. It looks for the smallest subgroup size that satisfies the required consistency rate, which is typically set by the user. It is a filter method because it is not dependent on any one classifier that the GeS approach might use to use the output from the carefully selected gene [34–35]. The suggested metric is the dataset's overall inconsistency rate for a particular gene set. A portion of an occurrence known as an outline lacks the class label subset in the explanation that follows. It consists of a gene's subset. Aimed at a given gene subset Z with

a_{g_1}, a_{g_2}, a_{g_3},……….. $a_{g_{|Z|}}$ count of values for genes $g_1, g_2, g_3,$……, $g_{|Z|}$ correspondingly, there are at most m_{g_1}, m_{g_2}, m_{g_3},………, $m_{g_{|Z|}}$ outline.

The inconsistency rate (CM) is determined by performing the calculations described below:

a. For a sample, an inconsistency is obtained by the existences (0 1, 0) and (0 1, 1) where the two genes make the correspondent principles in the two existences even though the character of class fluctuates and the concluding value in the existence. A pattern is hypothetical to be inconsistent uncertain, there occur at least two occurrences like they associate all but with their class markers.

b. The inconsistency count for a gene subgroup's outline is equal to the number of data epochs it examines minus the largest number of inconsistent class labels. For the sake of the sample, let's consider an outline that appears in instances of a gene subgroup where instances have class tags 1, 2, and 3, where $b_1 + b_2 + b_3 = a_y$.

   If b_3 is the largest among the three, then the inconsistency count is ($a{-}b_3$). The sum of entirely a_ys concluded by the different outline y that occur in the data of the gene subgroup X is the overall count of occurrences (P) in the dataset, i.e., $\sum y\ a_y = P$.

c. The sum of all the inconsistent overall designs of the gene subcategory which appears in the data divided by P is the *inconsistency rate (IR)* of a gene subgroup T (*IR(T)*). The following is how CM is still being used for gene selection. CM remains utilized to the gene selection task as follows. Assumed a contender gene subgroup T, inconsistency rate *IR(T)* is calculated. If *IR(T)* $\leq \alpha$ where α is a user-specified IR threshold, the subgroup T is called to be *consistent*. The characteristics of CM are gathered in the description. A gene subgroup may not be able to satisfy the strict condition at that time because real-world data is frequently noisy and uncertainty α is set to 0%. The hashing mechanism makes it possible to compute IR with time complexity. *O(T)* [33]. CM utilises data with discrete value features. In this case, features must first be discretized if the data is continuous [36].

In order to identify the most advantageous genes, it is advantageous to correlate BFS with CFS and Consistency as a gene evaluator. It advocates eliminating unnecessary, obtrusive, and redundant genes once their significance is not largely dependent on other genes. Using greedy hill-climbing techniques that are aided by the ability to go back, BFS investigates the space of attribute subgroups. By combining BFS, CFS, and consistency, fifty percent of the genes are eliminated.

The accuracy of classification is typically superior to or equal to the minimal set of genes in judgment to the complete set of genes in the vast majority of cases. BFS starts with a null group of genes and uses the entire set of genes to accomplish forward searching. Later, it initiates at any point, looks backward, and examines both ways, subsequently removing or including genes. Subsequently identifying suitable, minimised, and pertinent genes, the next step is to classify the samples in order to assess the significance of a smaller subset of important genes, independent of the entire gene cluster present in the datasets. By addressing some noise that is modelled as a proportion of data inconsistencies, CM helps to eliminate redundant and inappropriate genes. A subgroup of genes is continually being checked by this multi-variate measure. In light of this, CM is quick, multi-variate, monotonic, capable of handling data noise, and multi-variate before removing inappropriate genes. CM appears to be more expensive than CFS.

4.3 Hidden Weight Naïve Bayes Classifiers

Classification is an important task in pattern recognition and data mining [37]. Due to its easiness of construction but amazing effectiveness, Naïve Bayes (NB) seems to be the top machine learning tactic [38]. It provides pure semantics utilizing the knowledge of probability. The tactic is used in supervised initiation tasks which helps to achieve good accuracy with predicted class for testing and training data including class information [39]. This classifier is termed as naïve due to the postulation that foretold features are conditionally sovereign in each class and it concludes that no secluded (hidden) features influence the forecast method. These postulates reinforce efficient algorithms for learning as well as classification.

Let A be the arbitrary variable symbolizing the class of an example like gene name, B be a vector of arbitrary genes symbolizing the experimental attribute values, a symbolize a specific class label like types of Grades and b signify the precise detected value vector. Assuming a test case b to categorize, one uses Bayes' rule to figure out the likelihood of each class given the vector of detected values for the foretold genes and then forecasts the utmost probable class.

$$P(A = a \,/\, B = b) = \frac{P(A=a)P(B=b/A=a)}{P(B=b)} \qquad \text{eq. 3}$$

Now $B = b$ signify the event that $B_1 = b \wedge B_2 = b_2 \Lambda \dots \dots B_k = b_k$. Since the occurrence is a combination of gene value assignments, and because these genes are expected to be conditionally sovereign, one attains

$$P(B = b \,/\, A = a) = P(\wedge B_i = b_i \,/\, A = a), \qquad \text{eq. 4}$$

$$= \pi P(B_i = b_i \,/\, A = a)$$

i.e., is modest to calculate for test cases and to guess from training information. Usually, one does not evaluate the distribution in the denominator of Equation 3, as it is just a standardizing factor; as a substitute, one disregards the denominator and then standardizes so that the summation of $P(A = a \,/\, B = b)$ over all classes is one. For discrete features, $P(A = a \,/\, B = b)$ is demonstrated by a number amongst 0 and 1 that signifies the likelihood that the gene B will take on the value b when the class is a. In opposition to each numeric gene is demonstrated by some continuous likelihood distribution over the range of that gene's value. A mutual belief is that values of numeric genes are normally distributed, and can be characterized in terms of standard deviation and mean. For continuous attributes, equations 5 and 6 are framed, where d signifies the probability density function for a gaussian distribution.

$$P(B = b \,/\, A = a) = d(C; \mu_C, \sigma_C), \ \ where \qquad \text{eq. 5}$$

$$d(C; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(C-\mu)^2}{2\sigma^2}} \qquad \text{eq. 6}$$

NB disregards the attribute dependencies. A method for learning an optimal Bayesian network that can avoid computational complications and take the inspirations from all the genes into account. The concept of creating a hidden parent for each gene that trusts the inspirations from all the genes is termed as Weight Hidden Naïve Bayes [40].

Assume Z is a class node i.e., Histologic Grade and parent of all the attribute nodes. Figure 3 defines the structure of NB and HWNB. Each attribute Y_j has hidden parent $Y_{h_{P_j}}, j = 1,2,3,\dots\dots, m$, signified by a dashed circle. The arc from the hidden parent $Y_{h_{P_j}}$ to Y_j is signified by a dashed line, to differentiate it from systematic arcs.

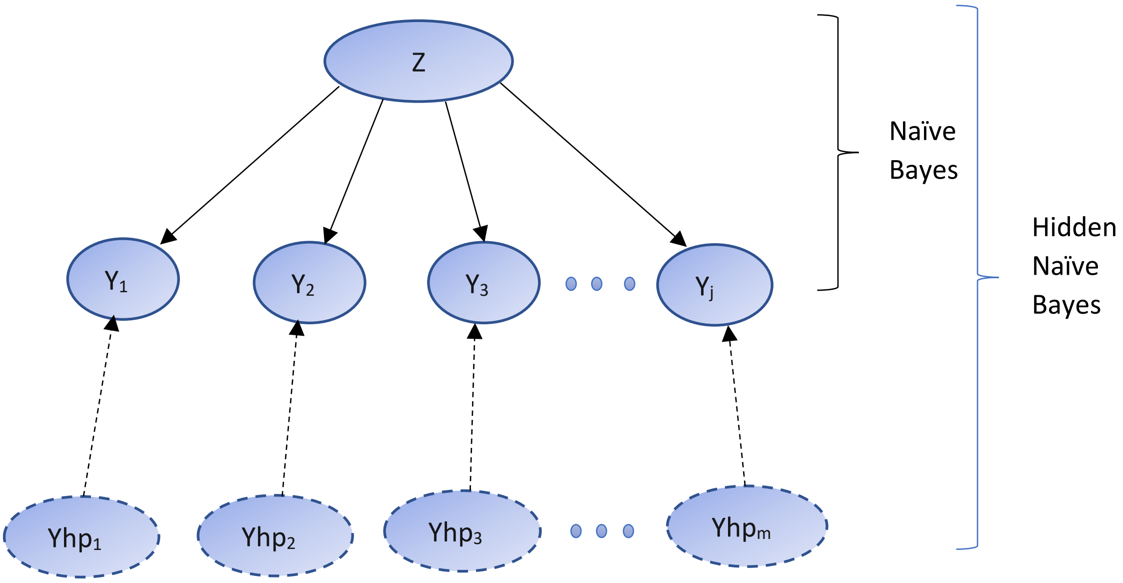

Figure 3. Structural representation of Naïve Bayes and Hidden Naïve Bayes

The joint distribution signified by HNB is defined as follows:

$$P(Y, \dots, Y_m, Z) = P(Z) \prod_{j=1}^{m} P\left(Y_j \,/\, Y_{h_{P_j}}, Z\right) \qquad \text{eq. 7}$$

where,

$$P\left(Y_j \,/\, Y_{h_{P_j}}, z\right) = \sum_{a=1, a\neq b} T_{ba} * P(Y_b / Y_a, Z) \qquad \text{eq. 8}$$

and $\sum_{a=1,a\neq b} T_{ba} = 1$.

The hidden parent $Y_{h_{Pj}}$ for Y_j is fundamentally a combination of the weighted impacts from all other attributes.

Considering the attributes $Y_1, \dots \dots, Y_m$, $P\left(Y_j \,/\, Y_{h_{Pj}}, Z\right)$ can be thought of approximation of $P(Y_1, \dots \dots, Y_m)$. In Equation 6, an approximation is depending on single estimators. Through the principle, arbitrary e-dependence estimators can be utilized to state hidden parents. If $e = m - 1$, any Bayesian network is signified by HNB. HNB is considered equivalent to a Bayesian network in standings of expressive power. It is favoured to outline hidden parents in demand to make the learning procedure well-organized, efficient, and simple.

From equations 7 and 8, the method to regulate weights T_{ba} , $b, a = 1, \dots\dots, m$ and $a \neq b$, is decisive for learning HNB. There are two tactics to find it: one is executing a cross-validation grounded search, or second directly executing the estimated values from data. Adopted the latter, and made use of conditional mutual information amongst attributes Y_a and Y_b as the weight of the $P(Y_a; Y_b|Z)$. More precisely, the weight is defined in eq. 9

$$T_{ba} = \frac{M_p(Y_b;Y_a|z)}{\sum_{a=1,a\neq b} Mp(Y_b;Y_a|z)} \qquad \text{eq. 9}$$

Where $M_p(Y_b; Y_a|z)$ is a conditional mutual information defined as:

$$M_p(A; B|C) = \sum_{a,b,C} P(a, b, c) \, log \frac{P(a,b|c)}{P(^a/_c)P(^b/_c)} \qquad \text{eq. 10}$$

where a,b and c are values of variables A,B, and C respectively.

5 Datasets

The experimentations are conducted on six microarray gene expression datasets extracted from National Centre for Biotechnology Information (NCBI) and is detailed in Table 2. At the initial stage, the count of genes in the datasets is in the thousands, so subsequently removing irrelevant genes is required to gained insights from data Table 3, shows Grade wise distribution of samples. The number of relevant genes selected in 2-Stage GeS is shown in Table 4. Histologic Grade-wise classification with three classifiers namely Naïve Bayes (NB), Hidden Weight Naïve Bayes (HWNB), and Correlation Weighted Feature Naïve Bayes (CWNB) in terms of precision, recall, f-score and fallout are given in Table 5-8. Out of these three classifiers, HWNB outshines in terms of precision, recall, f-score, and fallout highlighted in bold in Table 6. Eleven classifiers have been used namely, Support Vector Machine (SVC), Deep Learning (DL), Decision Table (DT), Random Forest (RF), Logit Boost (LB), JRip, IBK, OneR, NB, CWNB, and HWNB.

Table 2. Detailed Description of Datasets

Datasets	Genes	Samples
GSE7390	13516	196
GSE10886	16380	74
GSE25055	13515	302

GSE25066	16383	486
GSE29044	16384	98
GSE42568	16384	104

Table 3. Detailed distribution of different grades in each sample

Datasets	Grade 1	Grade 2	Grade 3	Grade 4
GSE7390	30	83	83	0
GSE10886	7	25	42	0
GSE25055	19	117	151	15
GSE25066	32	180	259	15
GSE29044	3	53	42	0
GSE42568	11	40	53	0

5.1 Experimentation Analysis

The proposed model consists of 2-stage GeS techniques and Hidden Weight Naïve Bayes classifier in which the number of appropriate genes is chosen at the first stage utilizing the CFS-BFS method and Consistency-BFS at the second stage. The details of the count of genes chosen are presented in Table 4. The number of genes selected using CONSISTENCY-BFS is very few to the genes chosen by the CFS-BFS method. The genes obtained at the second stage are significantly reduced in comparison to the complete set of genes in the original datasets and genes selected by the CFS-BFS technique. All the genes chosen are relevant and perform a significant role in the analysis and prognosis of BC. The overall results of good f-score, recall, and precision are shown by datasets GSE10886 and GSE29044. The highest precision of 96.4%, recall of 96.3%, and f-score of 96.3% with CWNB have been achieved in GSE10886. The second maximum precision of 96.1%, recall of 96%, and f-score of 96% with HWNB, was obtained in GSE29044. The third highest precision achieved is 95.2%, recall of 95%, and f-score of 95.1% with Naïve Bayes (NB) in GSE9044. The minimum fallout of 1.4% with CWNB in GSE10886, followed by 2.2% with HWNB, and NB is achieved in GSE10886. The graphical description of results achieved by all the classifiers with six datasets is shown in Figure 4-7. Figure 4, shows the performance of various ML classifiers on six datasets in terms of precision. Figure 5, displays the superiority of CWNB classifier on Recall measure. Figure 6 shows the performance of F-score with ML methods. Figure 7 shows the line graph comparing the fallout measure of six datasets with ML methods. The overall results show the superiority of HWNB with the remaining classifiers shown in Table 9.

Considering all the selected genes by the 2-GeS tactic, in each dataset where the correlation coefficient is calculated to find the correlation among the genes. Considering all selected gene's coefficients, a ranking of the genes is generated. Combining all the selected genes of six datasets, the ranking of genes is shown in Table 10. Dataset-wise ranking of the top three selected genes is shown in Table 11. As a result, the top four genes namely E2F3, PSMC31P, GINS1, and PLAGL2 were identified by 2-Stage GeS. Later discovered the serious effects of the top four genes in the existence of patients with BC. KMS Plotter tools were utilized to the existence of patients with BC by using publicly available datasets (2015 version; http://kmplot.com/analysis/index.php? p=service&cancer= breast) [40].

The subcategory of Histologic Grade in MAGE can be distinguished into the category of good and bad prognosis. Patients with lower Histologic Grades typically have better survival rates than those with higher Histologic Grades. KMS Model is used to validate whether the proposed model can distinguish between patients with poor and good prognosis using the Relapse-Free Survival rate (RFS) data from the micro-array datasets. The R-Survival project's package was used to implement the survival scrutiny with the Histologic Grade factor, resulting in the RFS arcs of the proposed model, as shown in Figure 8-11, which shows a clear separation between the groups with good and poor prognoses based on grade. A log-rank test was estimated to determine the p-value, and it suggests that a lower p-value indicates a better separation between grade subtypes. Figure 8-11, shows the probability of survival analysis as high or low in BC patients depending on all Grades, Grade 1, Grade 2, Grade 3, and Grade 4 respectively.

The grade of a BC is a predictor, a prognostic indicator, and a marker of the tumor's "hostile potential." Low-grade cancers tend to be less aggressive than high-grade cancers. Grade appears to be very important, and clinicians use this information to help and direct treatment options for patients. Looking at the prognosis of Histologic Grade, the proposed model has taken into consideration of grade parameters to check the importance of grade in terms of breast cancer prognosis and detection. The result substantiates that the proposed model is efficacious in separating BC patients into two prognosis groups depending upon the RFS rate, which can determine the patient’s expectancy level for an event (relapsed at any site). Accordingly aids in easy credentials of the patient’s group which might demand less or more aggressive medication strategy. The Kaplan-Meier curve and log-rank test scrutinizes discovered that the increased E2F3, PSMC3IP, GINS1, and PLAGL2 mRNA levels were meaningfully associated with the Relapse Free Survival (RFS) of all the patients with BC shown in figure 8-11. The patients with BC with high mRNA levels of the E2F3, PSMC3IP, GINS1 [40], and PLAGL2 genes were predicted to have high RFS in Grade 1 and Grade 2. But the survival analysis is not significant with Grade 3.

The expression levels of E2F3 and GINS1 were higher in BC tissues than in normal breast tissues. Survival analysis using the Kaplan-Meier Plotter database revealed that the high transcription levels of E2F3 were linked with low relapse-free survival (RFS) in all the patients with breast cancer. E2F3 is a potential target of precision therapy for patients with breast cancer [42]. Survival analysis exposed that increased expression levels of GINS1 were associated with poor prognoses in all patients with BC [43]. GINS1 was associated with detrimental relapse-free survival (RFS) [44]. All the experiments are performed using the WEKA software [45] and RStudio [29].

Table 4. Count of features selected in both stages

	First Stage	Second Stage

Datasets	CFS-BFS	CONSISTENCY-BFS
GSE7390	102	16
GSE10886	46	7
GSE25055	193	12
GSE25066	212	13
GSE29044	66	10
GSE42568	91	8

Table 5. Precision wise results of NB, CWNB and HWNB

Precision	NB	CWNB	HWNB
Grade 1	78.58	78.58	**84.53**
Grade 2	81.7	81.7	**83.17**
Grade 3	90.05	90.05	**91.02**
Grade 4	74.2	74.2	71.9

Table 6. Recall wise results of NB, CWNB and HWNB

Recall	NB	CWNB	HWNB
Grade 1	78.82	70.77	**79.73**
Grade 2	83.53	83.62	**86.57**
Grade 3	88.95	90.2	89.48
Grade 4	71.7	50	68.35

Table 7. F-Score wise results of NB, CWNB and HWNB

F-Score	NB	CWNB	HWNB
Grade 1	80.07	76.87	**81.93**
Grade 2	82.48	81.8	**84.77**
Grade 3	89.48	88.93	**90.2**
Grade 4	71.3	52.65	68.45

Table 8. Fallout wise results of NB, CWNB and HWNB

Fall out	NB	CWNB	HWNB
Grade 1	3.63	2.43	**2.42**
Grade 2	10.06	11.6	**9.77**
Grade 3	7.98	10.28	**7.28**
Grade 4	1.95	2.8	**1.9**

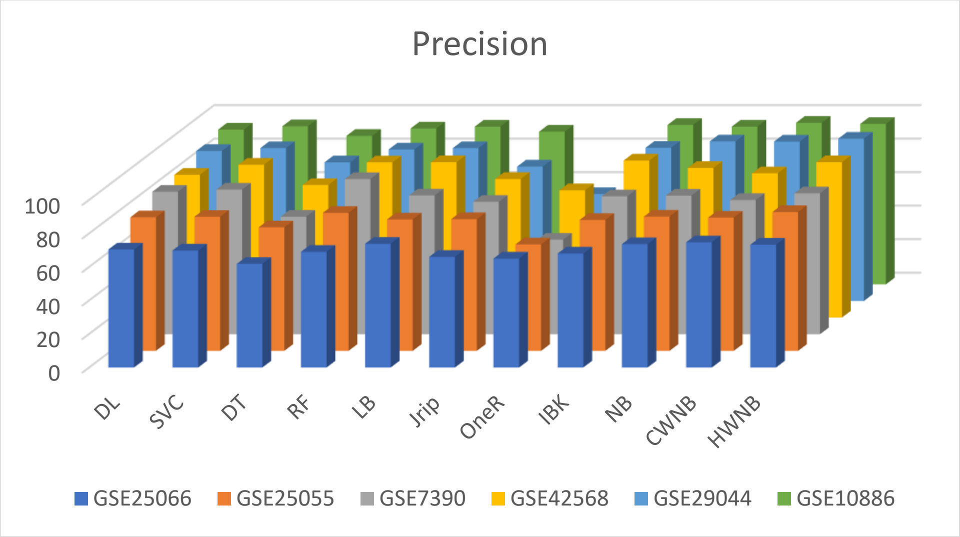

Figure 4. Performance of six datasets based on Precision

Table 9. Performance of Proposed Model in comparison to remaining machine learning classifiers

Classifiers	Precision	Recall	F-score	Fall out
Proposed Model + DL	83.6833	82.5167	82.6167	9.01667
Proposed Model + SVC	85.4667	85.5	85.3167	9.21667
Proposed Model + DT	76	76.35	74.2333	15.9
Proposed Model + RF	86.65	86.5333	86.35	8.28333
Proposed Model + LB	85.5	85.4333	85.2333	9.15
Proposed Model + Jrip	79.6	79.3833	79.2	12.8833
Proposed Model + OneR	61.7333	59.7	59.0833	28.3333
Proposed Model + IBK	85.0333	84.7167	84.6667	9.01667
Proposed Model + NB	85.8333	85.6667	85.6667	8
Proposed Model + CWNB	85.3167	84.95	84.55	9.48333
Proposed Model (2-Stage GeS + HNB)	**87.45**	**87.3667**	**87.3**	**7.45**

6 Conclusion and Discussion

This research proposed a novel 2-stage GeS tactic for BC subtypes prediction based on two methods and Hidden Naïve Bayes classifier, namely CFS-BFS at the first stage, CONSISTENCY-BFS at the second stage utilizing histologic grade, and utilizing the Hidden Weight Naïve Bayes classifier for the classification. CFS-BFS has an O(N^2) complexity, Since Consistency-BFS complexity is linear i.e. (O(N)), it is preferable to CFS-BFS. whereas CFS-BFS is polynomial i.e (O(N^2)), where N is the total number of features. The experiments were performed using six microarray gene expression datasets. The results validate an impressive precision, recall, f-score, and fallout to forecast BC using limited selected appropriate genes in each microarray gene expression dataset. The impressive is achieved by the proposed 2-GeS tactic with Hidden Naïve Bayes classifier. A maximum of the chosen genes is exposed to be correlated to BC grounded on earlier research, although limited are yet to be explored.

This two-stage gene selection strategy can focus research and analysis on a relatively small subset of genes. Most notably, the strategy can be helpful for more sophisticated patient stratification in the future, such as subgroups formed by combining platforms or for groups of patients that have been divided based on treatment response. With the help of this combination tool, it is possible to precisely classify large populations of patients into definite cancer subtypes or treatment groups by regulating the minimum number of genes that must be screened.

Table 10. Ranking of relevant Genes of six datasets after 2-Stage GeS

Genes	Rank	Genes	Rank
E2F3	1	IL1R2	34
PSMC3IP	2	NM_002691	35
GINS1	3	PDHA1	36
PLAGL2	4	FOSB	37
MELK	5	CLTC-IT1	38
CCNB2	6	BC005884	39
FLJ20224	7	GTF3A	40
NMU	8	BTF3	41
SPTBN2	9	MAB21L1 /// MIR548F5	42
BM545088.1	10	NM_002266	43
TPD52L1	11	SDHA	44
C6	12	MUC5AC	45
ATP7B	13	MRPL40	46
I_1109138	14	V39326	47
MYL7	15	ANKRD7	48
HOXC8	16	ACSM2A /// ACSM2B	49
CIAO1	17	TGFBR3	50
RRM2	18	SNX21	51
PPM1G	19	WDR5B	52
BIRC5/// EPR-1	20	CYTH1	53
VSNL1	21	NM_000266	54
NM_001255	22	BECN1	55
EPB41L2	23	KHDRBS1	56
LOC10192	24	NM_006185	57

NM_006430	25	ZNF253	58
PARP4	26	MERTK	59
RRAS2	27	NM_000168	60
C1S	28	NM_003256	61
MZT2A /// MZT2B /// PHGDH	29	M95929	62
HAX1	30	NM_004694	63
PSMB4	31	PCNXL4	64
BG035989	32	RELA	65
NM_004219	33	FGD6	66

The findings showed that the top two genes E2F3 and GINS1 subunits might be new potential predictive biomarkers for BC. The clinical significance of E2F3 and GINS subunits in BC patients, however, still needs to be demonstrated by additional authentication studies. In conclusion, E2F3 and GINS subunits may serve as novel survival biomarkers or therapeutic targets for BC patients. It is expected that this research will improve the accuracy of prognostication in BC patients.

Table 11. Top three Genes after GeS.

Rank	GSE7390	GSE42568	GSE10886	GSE25055	GSE25066	GSE29044
1	E2F3	CIAO1	FLJ20224	HAX1	NM_001255	EPB41L2
2	PSMC3IP	PPM1G	BM545088.1	CLTC-IT1	NM_006430	PARP4
3	GINS1	BIRC5/// EPR-1	ATP7B	SDHA	BG035989	C1S

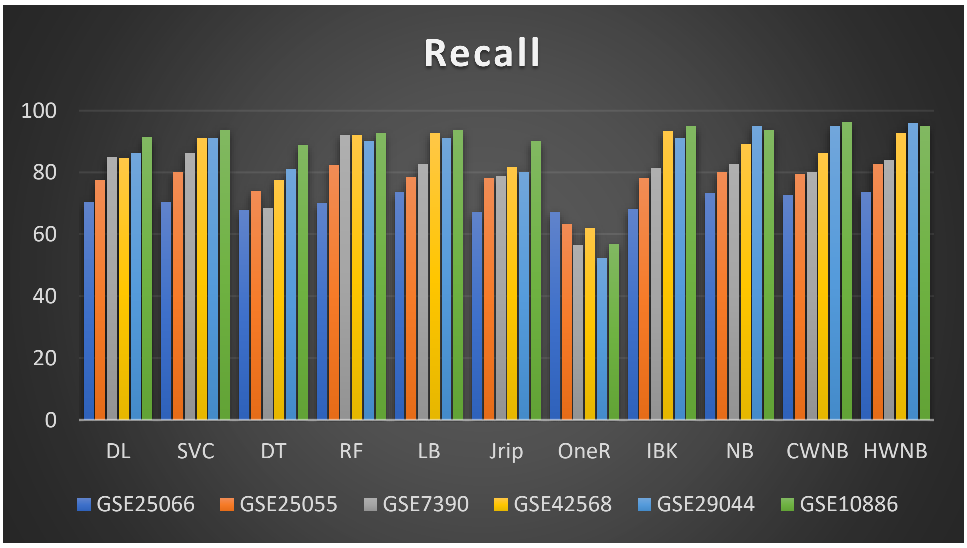

Figure 5. Performance of six datasets depending on Recall parameter

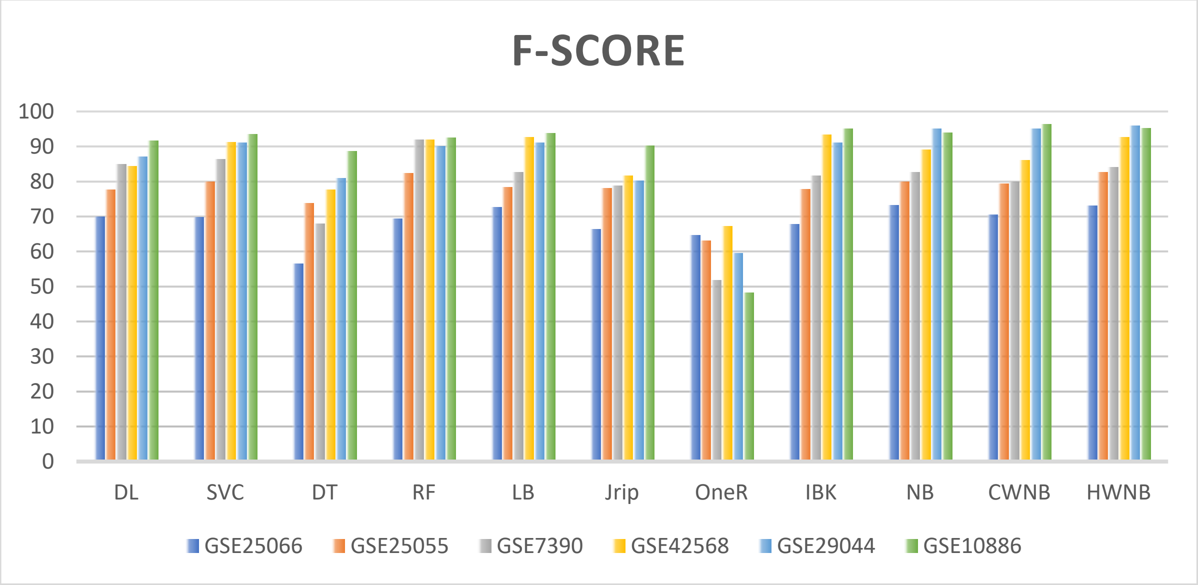

Figure 6. Performance of six datasets based on F-score

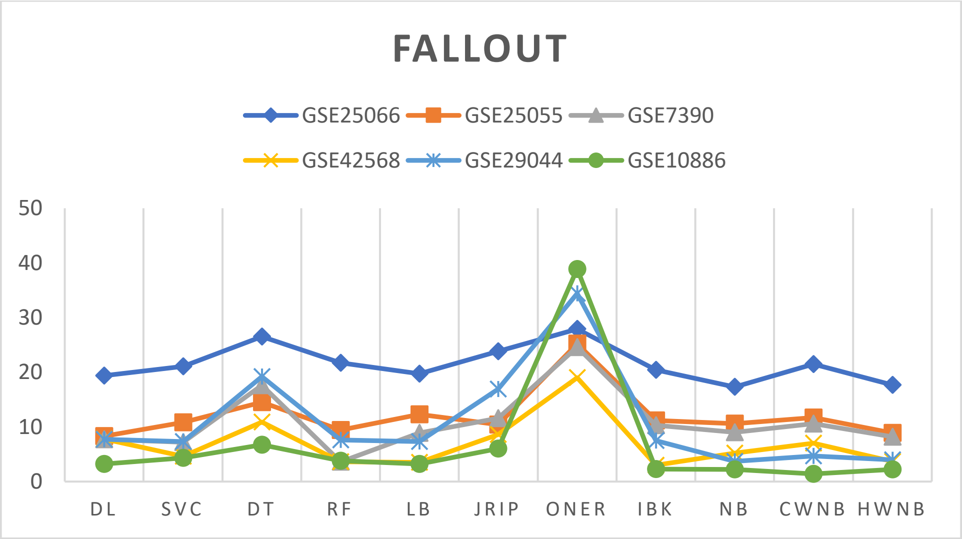

Figure 7. Performance of six datasets depending on Fallout

Table 11. Top three Genes after GeS.

Rank	GSE7390	GSE42568	GSE10886	GSE25055	GSE25066	GSE29044
1	E2F3	CIAO1	FLJ20224	HAX1	NM_001255	EPB41L2
2	PSMC3IP	PPM1G	BM545088.1	CLTC-IT1	NM_006430	PARP4
3	GINS1	BIRC5/// EPR-1	ATP7B	SDHA	BG035989	C1S

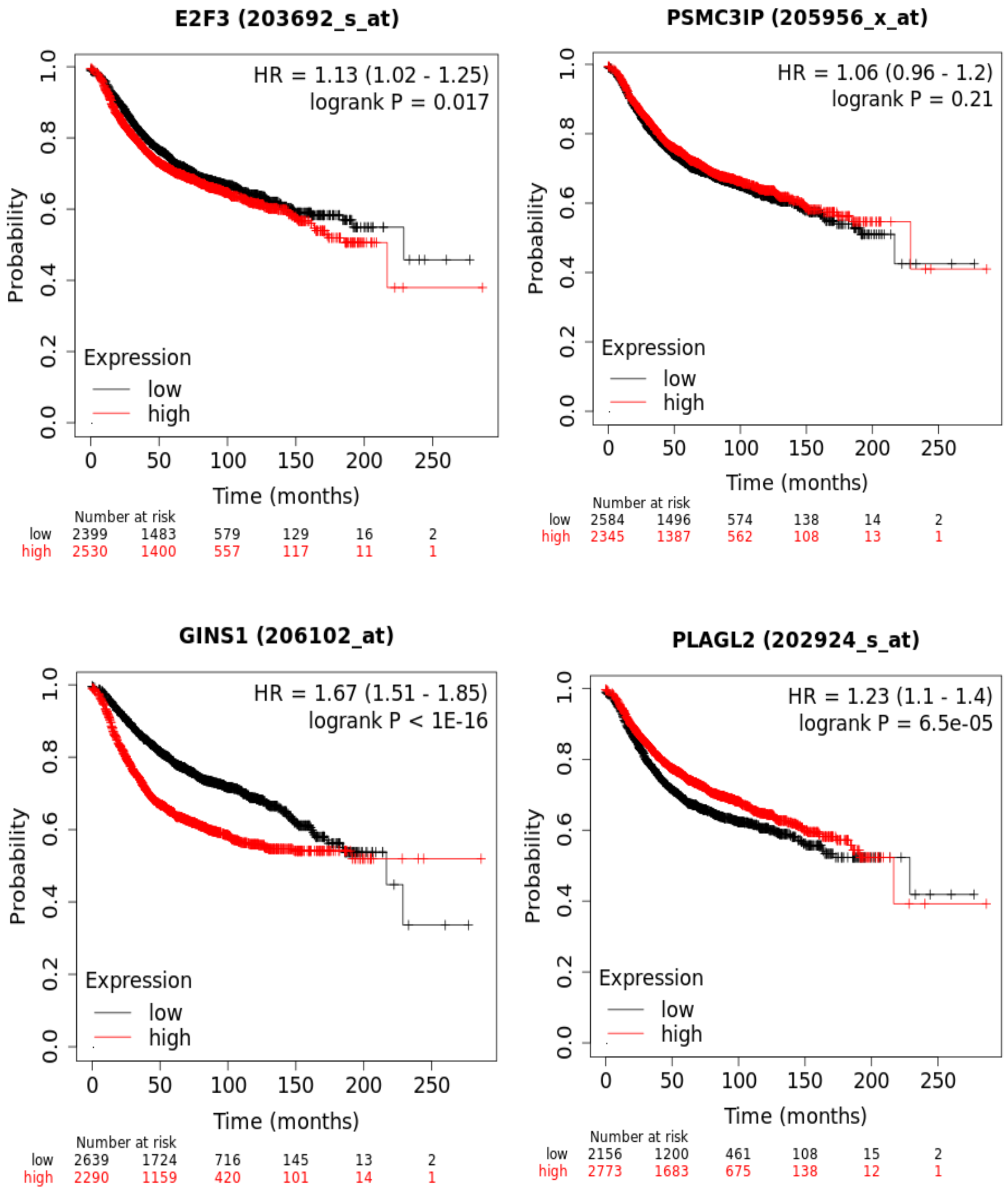

Figure 8. Prognostic Value of mRNA Level of top four genes with RFS with all types of the histologic grade in Breast Cancer Patients (Kaplan-Meier Plotter).

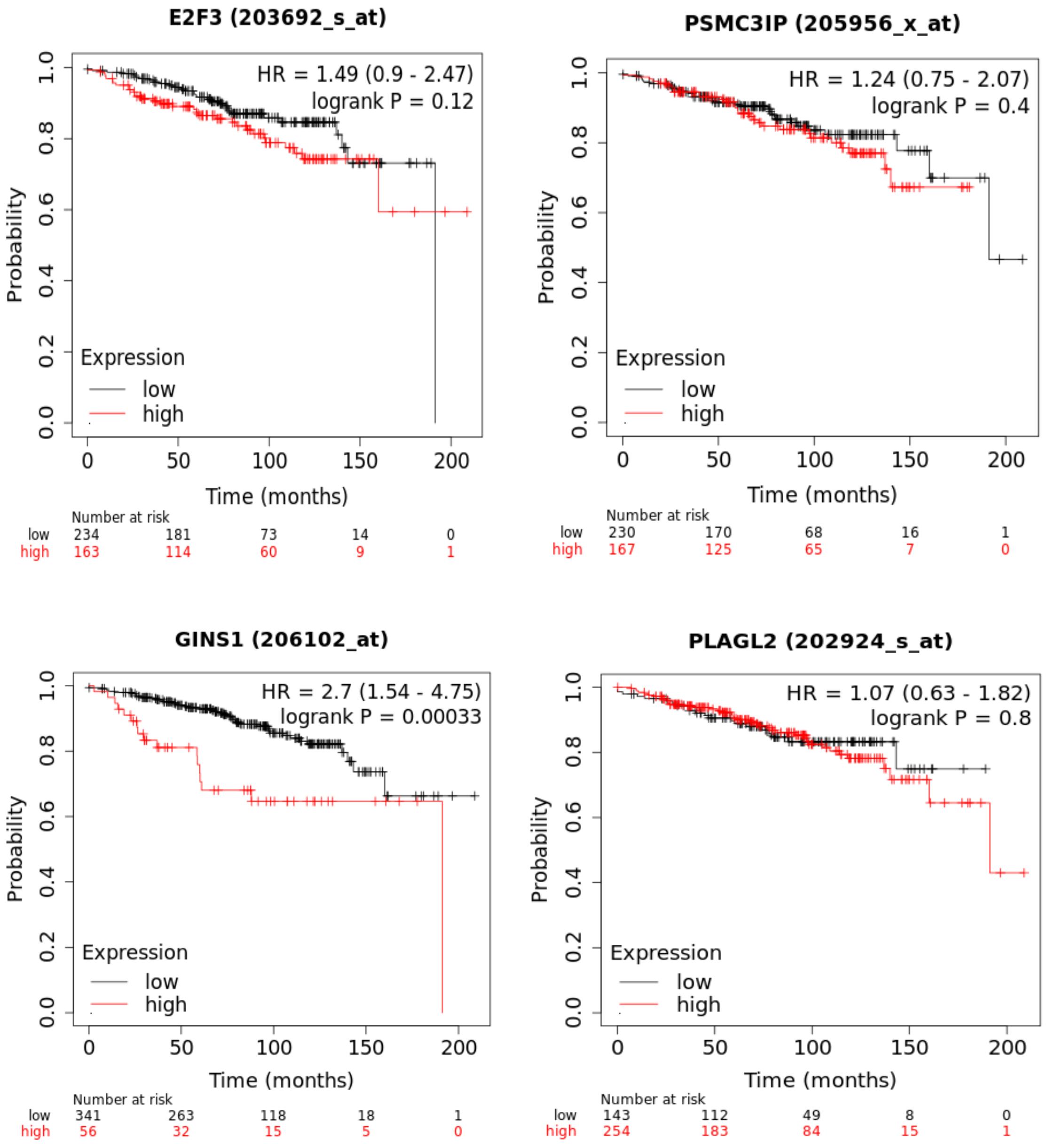

Figure 9. Prognostic Value of mRNA Level of top four genes with RFS with the histologic Grade 1 in Breast Cancer Patients (Kaplan-Meier Plotter).

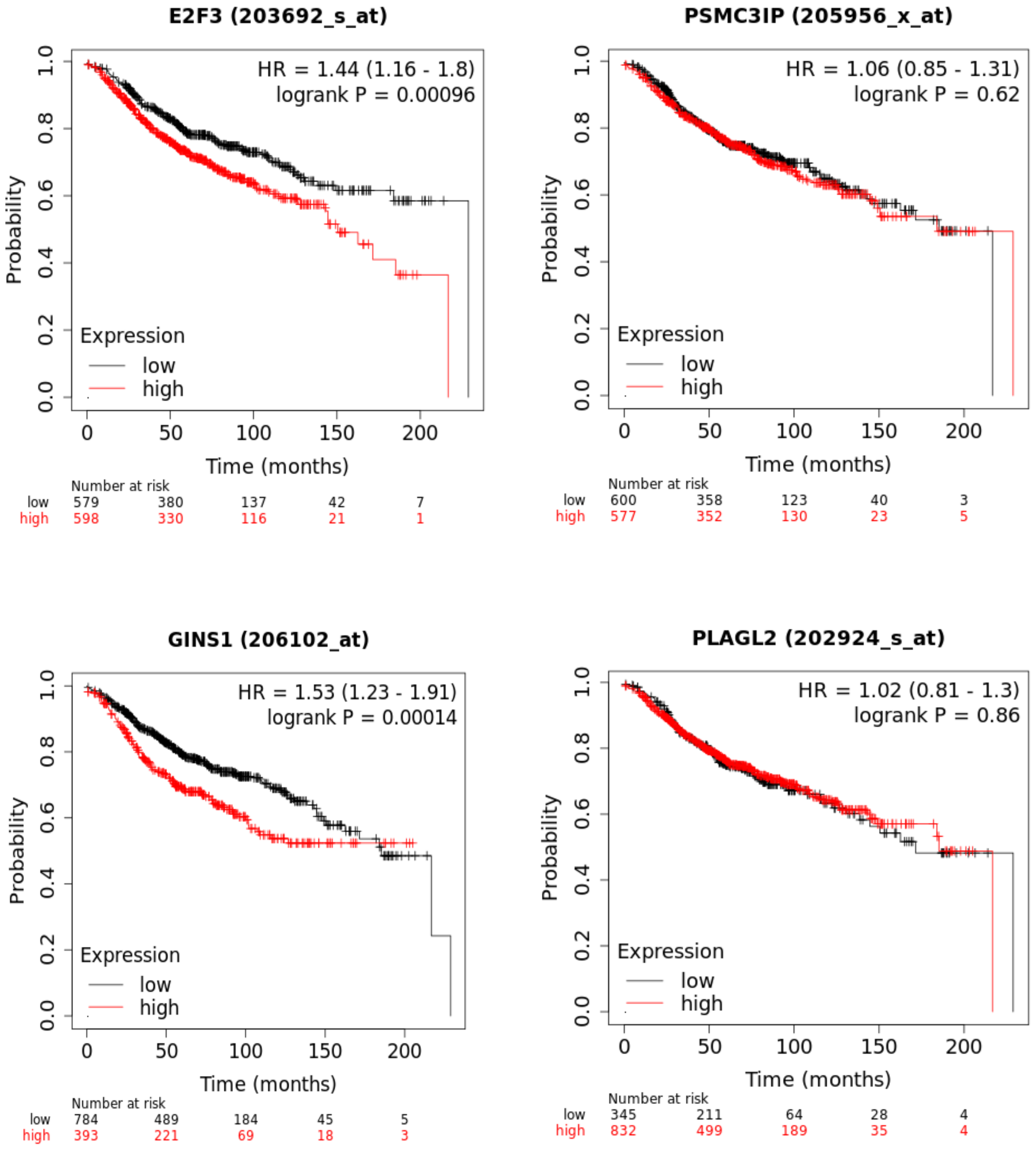

Figure 10. Prognostic Value of mRNA Level of top four genes with RFS with all the histologic Grade 2 in Breast Cancer Patients (Kaplan-Meier Plotter).

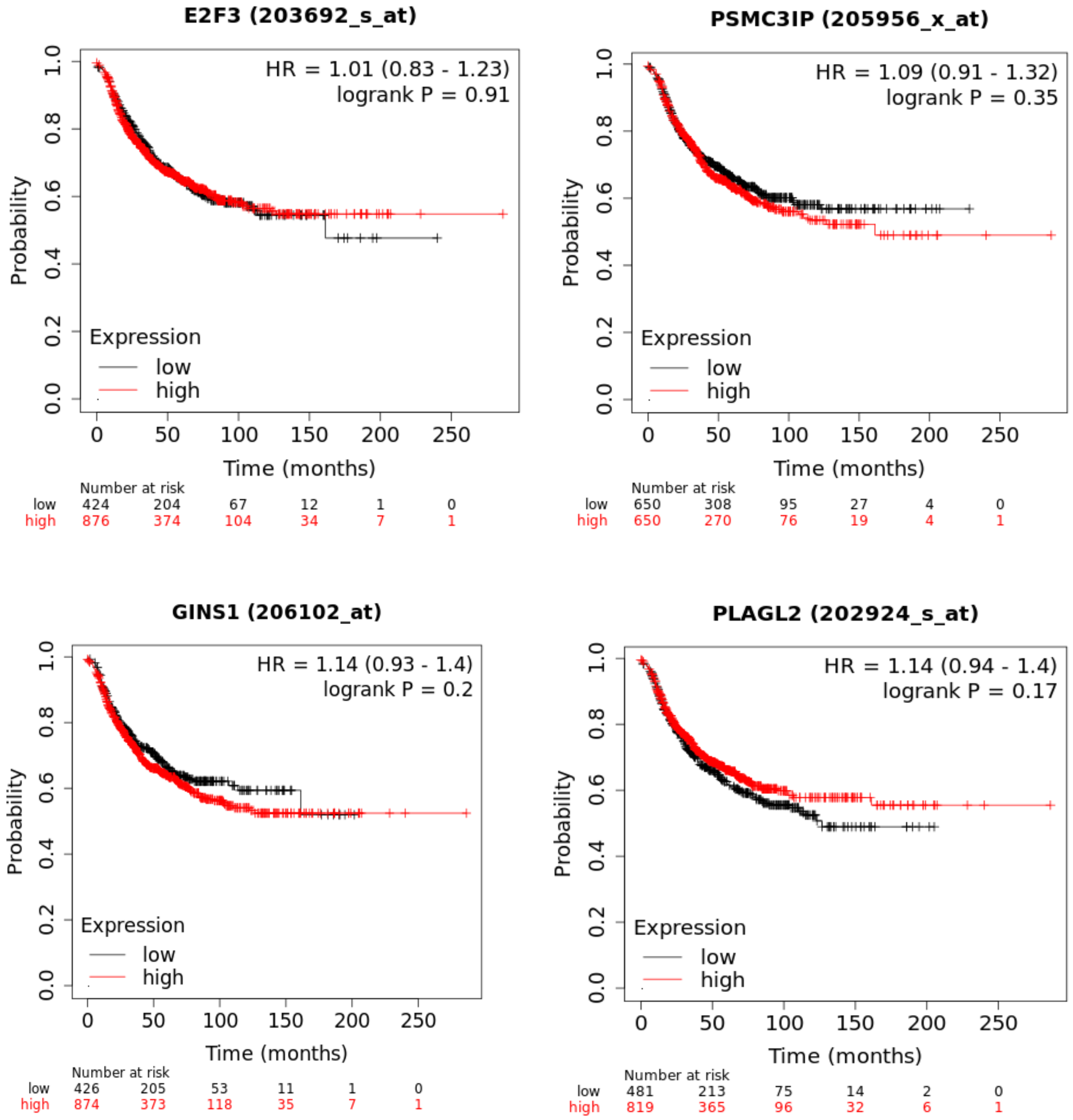

Figure 11. Prognostic Value of mRNA Level of top four genes with RFS with the histologic Grade 3 in Breast Cancer Patients (Kaplan-Meier Plotter).

7. Declaration of Interest

None

References

1. Ali HR, Rueda OM, Chin SF et al (2014) Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology*, *15*(8), 1-14.

2. Lamba, M., Munjal, G., & Gigras, Y. (2021). ECABC: Evaluation of classification algorithms in breast cancer for imbalanced datasets. In *Data Driven Approach Towards Disruptive Technologies: Proceedings of MIDAS 2020* (pp. 379-388). Springer Singapore.
3. Lamba, M., Munjal, G., & Gigras, Y. (2022). Supervising Healthcare Schemes Using Machine Learning in Breast Cancer and Internet of Things (SHSMLIoT). *Internet of Healthcare Things: Machine Learning for Security and Privacy*, 241-263.
4. Lamba, M., Gigras, Y., & Dhull, A. (2021). Classification of plant diseases using machine and deep learning. *Open Computer Science*, *11*(1), 491-508.
5. Lamba, M., Munjal, G., & Gigras, Y. (2021). A MCDM-based performance of classification algorithms in breast cancer prediction for imbalanced datasets. *International Journal of Intelligent Engineering Informatics*, *9*(5), 425-454.
6. Rakha EA, Reis-Filho JS, Baehner F, et al (2010) Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, *12*(4), 1-12.
7. Lamba, M., Munjal, G., Gigras, Y., & Kumar, M. (2023). Breast cancer prediction and categorization in the molecular era of histologic grade. *Multimedia Tools and Applications*, 1-20.
8. Olsson N, Carlsson P, James P et al (2013) Grading breast cancer tissues using molecular portraits. *Molecular & Cellular Proteomics*, *12*(12), 3612-3623.
9. Jayanthi VSA, Das AB, & Saxena U (2020) Grade-specific diagnostic and prognostic biomarkers in breast cancer. *Genomics*, *112*(1), 388-396.
10. Dai X, Li T, Bai Z, et al (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research*, *5*(10), 2929.
11. Souri EA, Chenoweth A, Cheung A et al (2021) Cancer Grade Model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer. *British Journal of Cancer*, 1-11.
12. Rakha EA, & Pareja FG (2021) New advances in molecular breast cancer pathology. In *Seminars in cancer biology* (Vol. 72, pp. 102-113). Academic Press.
13. Jenkins S, Kachur ME, Rechache K et al (2021) Rare Breast Cancer Subtypes. *Current oncology reports*, *23*(5), 1-14.
14. Ang JC, Mirzal A, Haron H, et al (2015) Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics*, *13*(5), 971-989.
15. Lazar C, Taminau J, Meganck S, et al (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM transactions on computational biology and bioinformatics*, *9*(4), 1106-1119.
16. Nagpal, A., & Singh, V. (2019). Feature selection from high dimensional data based on iterative qualitative mutual information. *Journal of Intelligent & Fuzzy Systems*, *36*(6), 5845-5856.
17. Lamba M, Munjal G, & Gigras Y (2021) A hybrid gene selection model for molecular breast cancer classification using a deep neural network. *International Journal of Applied Pattern Recognition*, *6*(3), 195-216.
18. Lamba M, Munjal G, & Gigras Y (2018) Feature Selection of Micro-array expression data (FSM)-A Review. *Procedia computer science*, *132*, 1619-1625.
19. Lamba M, Munjal G, & Gigras Y (2020) Computational studies on breast cancer analysis. *Journal of Statistics and Management Systems*, *23*(6), 999-1009.
20. Engstrøm MJ, Opdahl S, Hagen AI et al (2013) Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients. *Breast cancer research and treatment*, *140*(3), 463-473.
21. Blows FM, Driver KE, Schmidt MK et al (2010) Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short- and long-term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS med*, *7*(5), e1000279.
22. Leong ASY & Zhuang Z (2011) The changing role of pathology in breast cancer diagnosis and treatment. *Pathobiology*, *78*(2), 99-114.
23. Rakha EA, Reis-Filho JS, Baehner F et al (2010) Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, *12*(4), 1-12.

24. Chowdhury N (2011) Histopathological and genomic grading provide complementary prognostic information in breast cancer: a study on publicly available datasets. *Pathology research international*, *2011*.
25. Srivastava, K. R., & Girdhar, N. (2022, January). Retinal Image Segmentation based on Machine Learning Techniques. In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 252-257). IEEE.
26. Dong YN, Zhao JJ, & Jin J (2017) Novel feature selection and classification of Internet video traffic based on a hierarchical scheme. *Computer Networks*, *119*, 102-111.
27. Blessie, E. C., & Karthikeyan, E. (2012). Sigmis: a feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology*, *6*(3), 385-394.
28. Lamba, M., Munjal, G., & Gigras, Y. (2023). Ranking of classification algorithm in breast Cancer based on estrogen receptor using MCDM technique. *International Journal of Information Technology & Decision Making (IJITDM)*, *22*(02), 803-827.
29. Allaire J (2012) RStudio: integrated development environment for R. *Boston, MA*, *770*, 394.
30. Lamba, M., Munjal, G., & Gigras, Y. (2023). Computational Studies in Breast Cancer. *Research Anthology on Medical Informatics in Breast and Cervical Cancer*, 434-456.
31. Lamba, M., Munjal, G. & Gigras Y. (2023). Idenifying Breast Cancer Molecular class using integrated feature selection and deep learning model, *International Journal of Bioinformatics Research and Applications.* 10.1504/IJBRA.2023.10054946
32. Li J, Cheng K, Wang S et al (2017) Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 1-45.
33. Chandrashekar G & Sahin F (2014) A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28.
34. Dash CSK, Behera AK, Dehuri S et al (2020) Building a novel classifier based on teaching learning based optimization and radial basis function neural networks for non-imputed database with irrelevant features. *Applied Computing and Informatics*.
35. Zhang T, Ye S, Zhang K, et al (2018) A systematic dnn weight pruning framework using alternating direction method of multipliers. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 184-199).
36. Maslove DM, Podchiyska T, & Lowe HJ (2013) Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association*, *20*(3), 544-553.
37. Han J, Pei J & Kamber M (2011) *Data mining: concepts and techniques*. Elsevier.
38. Wu X, Kumar V, Ross Quinlan J et al (2008) Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.
39. John GH & Langley P (2013) Estimating continuous distributions in Bayesian classifiers. *arXiv preprint arXiv:1302.4964*.
40. Zhang H, Jiang L & Su J (2005). Hidden naive bayes. In *Aaai* (pp. 919-924).
41. Győrffy B (2021) Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Computational and Structural Biotechnology Journal*, *19*, 4101-4109.
42. Sun CC, Li SJ, Hu W et al (2019) Comprehensive analysis of the expression and prognosis for E2Fs in human breast cancer. *Molecular Therapy*, *27*(6), 1153-1165.
43. Li H, Cao Y, Ma J et al (2021) Expression and prognosis analysis of GINS subunits in human breast cancer. *Medicine*, *100*(11).
44. Nieto-Jiménez C, Alcaraz-Sanabria A, Páez R et al (2017) DNA-damage related genes and clinical outcome in hormone receptor positive breast cancer. *Oncotarget*, *8*(38), 62834.
45. Hall M, Frank E, Holmes G et al (2009) The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1),10-18.
46. Aggarwal, G., & Singh, L. (2022). Comparisons of speech parameterisation techniques for classification of intellectual disability using machine learning. In *Research anthology on physical and intellectual disabilities in an inclusive society* (pp. 828-847). IGI Global.

47. Aggarwal, G., & Singh, L. (2018). Evaluation of supervised learning algorithms based on speech features as predictors to the diagnosis of mild to moderate intellectual disability. *3D Research*, *9*(4), 55.
48. Bhardwaj, A., Dagar, V., Khan, M. O., Aggarwal, A., Alvarado, R., Kumar, M., ... & Proshad, R. (2022). Smart IoT and machine learning-based framework for water quality assessment and device component monitoring. *Environmental Science and Pollution Research*, *29*(30), 46018-46036.
49. Alshehri, M., Kumar, M., Bhardwaj, A., Mishra, S., & Gyani, J. (2021). Deep learning based approach to classify saline particles in sea water. *Water*, *13*(9), 1251.

50.