International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 16, No.1 (Jul-24)

http://dx.doi.org/10.12785/ijcds/160107

Two-Stage Gene Selection Technique For Identifying Significant Prognosis Biomarkers In Breast Cancer

Monika Lamba[1], Geetika Munjal[2] and Yogita Gigras[3]

[1,3]*Department of Computer Science and Engineering,The NorthCap University, Gurugram , India*
[2]*Department of Computer Science and Engineering, Amity University, Uttar Pradesh, India*

Received 20 May 2023, Revised 18 Mar. 2024, Accepted 6 Apr. 2024, Published 1 Jul. 2024

Abstract: One crucial stage in the data preparation procedure for breast cancer classification involves extracting a selection of meaningful genes from microarray gene expression data. This stage is crucial because it discovers genes whose expression patterns can differentiate between different types or stages of breast cancer. Two highly effective algorithms, CONSISTENCY-BFS and CFS-BFS, have been developed for gene selection. These algorithms are designed to identify the genes that are most crucial in distinguishing between different types and stages of breast cancer by analysing large volumes of genetic data. A noteworthy advancement is a refined 2-Stage Gene Selection technique specifically designed for predicting subtypes in breast cancer. The initial phase of the 2-Stage Gene Selection (GeS) approach relies on the CFS-BFS algorithm, which plays a crucial role in effectively eliminating unnecessary, distracting, and redundant genes. The initial filtering process plays a crucial role in simplifying the dataset and identifying the genes that have the highest potential to shed light on the category of breast cancer. The CONSISTENCY-BFS algorithm guarantees that only the most pertinent genes are retained by further refining the gene selection process. This stage is essential for eliminating any remaining uncertainty and enhancing the overall efficiency of the algorithm. This innovative approach represents a significant advancement in the field of bioinformatics as it offers a more accurate and targeted method for selecting genes based on their relevance to breast cancer classification. When the 2-Stage GeS is constructed using Hidden Weight Naive Bayes, remarkably, it yields more precise and dependable outcomes. The indicators that demonstrate positive outcomes encompass recollection, accuracy, f-score, and fallout rankings. The Kaplan-Meier Survival Model was employed to further validate the top four genes, namely E2F3, PSMC3IP, GINS1, and PLAGL2. Presumably, precision therapy will specifically focus on targeting the genes E2F3 and GINS1.

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]. Histological Grade (HG) is a well-described prognostic factor that reflects the morphological characteristics of the tumour as well as the clinical behaviour of the disease. Diagnosticians have long recognised this morphological multiplicity, which is replicated in histological grades 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 I 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].

Patients with grade 3 tumours, for example, require prompt consideration of neoadjuvant or adjuvant chemotherapy during the process of systematic treatment selection. On the other hand, patients with grade 1 tumours may benefit from long-term follow-up. As a result of the fact that grade 2 breast tumours represent an intermediate and highly variable state in terms of morphology, underlying biology, and risk of distant metastasis recurrence, it may be challenging to determine the appropriate treatment for thirty to sixty percent of breast tumours that have been diagnosed. Because of this, patients who have been diagnosed with these tumours run the risk of receiving either insufficient or excessive treatment. It has been proposed that only grades 1 and 3 should be considered when deciding on treatment, and that grades 2 should not be considered informative in the absence of additional metrics. In light of this, the accurate stratification of grade-2 tissues presents many complicated challenges. 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

E-mail address: missmonikalamba@gmail.com, munjal.geetika@gmail.com, yogitagigras@ncuindia.edu

TABLE I. Description of histological grade.

Grade Types	Growth [8]	Mitotic Count	Tubular Differentiation	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.

data [7]. 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] [8] [9].

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

1) G1 - grade 1 (slow-growing, exceptionally differentiated)
2) G2 - grade 2 (slightly differentiated)
3) 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] [11] [12] [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] [17] [18] [19]. 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 topic of motivation is addressed in section 2, the literature review is examined in Section 3, Section 4 provides a detailed analysis of the GeS method, Hidden Weight Nave Bayes, and the GeS method, and Section 5 focuses on the proposed model. Section 6 covers the topics of datasets and experimentation analysis. The final section contains the conclusion and discussion.

2. Motivation

When it comes to advancing our understanding and treatment of this complicated disease, research on breast cancer that makes use of feature selection is an essential component. An improvement in diagnostic accuracy is one of the primary reasons for its significance, and it is only one of many. There is no such thing as uniform breast cancer; rather, it is comprised of numerous subtypes, each of which possesses unique characteristics and behaviors. A more precise and individualized diagnostic approach is made possible through the use of feature selection, which assists in identifying the most pertinent biomarkers and factors associated with these subtypes. Researchers can develop diagnostic tools that can differentiate between the various types of breast cancer by focusing on specific characteristics. This enables clinicians to receive diagnoses that are more accurate and timely. Additionally, the research makes a significant contribution to the understanding of the biology that lies beneath breast cancer, which is a significant contribution. A significant obstacle is presented by the

heterogeneity of the disease; however, feature selection helps determine the primary factors that are responsible for the disease's development, progression, and response to treatment. To gain valuable insights into the complex molecular mechanisms that are at play in breast cancer, it is necessary to identify fundamental characteristics. With this more in-depth understanding, the groundwork has been laid for the development of targeted therapies, which will open up new avenues for treatment strategies that are more precise and individualized.

Furthermore, an efficient and effective method of analysis is required because of the sheer volume of data that is involved in the research on breast cancer. By reducing the dimensionality of the data, feature selection helps researchers address this challenge and enables them to concentrate on the aspects of the data that are most pertinent to their work. This not only makes the process of analysis more efficient but also makes it easier to recognize important characteristics that might otherwise be obscured due to the large amount of information contained in the dataset. To discover new associations and patterns, which in turn helps to advance both clinical practice and scientific knowledge, it is essential to have the ability to navigate and distil this vast amount of information. The incorporation of feature selection into breast cancer research has the potential to revolutionize clinical practice as a result of the advancements that have been made. Improved patient outcomes are a direct result of the development of diagnostic tools that are more accurate and treatment strategies that are more individualized. By basing their decisions on the specific characteristics of an individual's breast cancer, clinicians can make more informed decisions, which ultimately results in interventions that are more tailored and effective. At the same time, the scientific community reaps the benefits of a more in-depth understanding of the complexities of breast cancer, which paves the way for ongoing innovation and the ongoing refinement of treatment approaches. Research on breast cancer that makes use of feature selection is, in essence, a cornerstone in the quest for improved diagnostics, a deeper understanding of the biology of the disease, and the development of treatment modalities that are both personalized and effective. Not only does this multidimensional approach improve clinical outcomes, but it also advances scientific knowledge, which in turn helps to foster a comprehensive and ever-evolving understanding of breast cancer.

3. Literature Survey

The initial prognostic staging system for BC, the Nottingham Prognostic Index (NPI), which was established upon the basis of lymph node stage (1–3), histological grade (1–3), and primary tumour size, continues to be implemented in numerous centres. It remains one of the most economically viable and user-friendly prognostic instruments in BC. The principal purpose of BC staging is to risk stratify patients who warrant therapeutic consideration, rather than to ascertain the precise therapeutic approach. It is noteworthy to mention that tumour behaviour may be altered during therapy, and the initial estimated risk may undergo a revision in light of treatment. Consequently, two predicted estimates comprise the risk classification for BC: the initial therapy-naive risk and the posttreatment risk.

Sankara et al. [9] proposed a comprehensive methodology aimed at identifying grade-specific biomarkers for breast cancer (BC). Differentially Expressed Genes (DEGs) were utilised in their approach, which involved the construction of networks that were based on grade-specific molecular interactions within cancer Grades 1, 2, and 3. A Grade 3 molecular network that is intricately associated with cancer-related processes was discovered as a result of their investigation, which focused on a particular field of study.Through the identification of the top ten differentially expressed genes (DEGs), the research brought attention to the significance of Grade 3 within this network. Particularly noteworthy is the fact that the analysis focused on the remarkable increase in the expression of CCNB2 and UBE2C genes in Grade 3, in comparison to the expression of these genes in other grades. An indication of a possible role for these genes in the distinct molecular landscape that is characteristic of Grade 3 breast cancer was provided by the differential expression shown here. In addition, the research demonstrated that certain genes, such as CCNB2, UBE2C, CDK1, KIF2C, CCNB2, and NDC80, are particularly significant in comparison to others. Researchers discovered that the expressions of these genes were extremely strong in a variety of breast cancer subtypes. It is intriguing to note that the increased expression of these genes was linked to a decrease in the patient survival rate, which suggests that these genes may play a role in the progression of the disease. Based on the findings that Sankara and colleagues came up with, it is clear that the genes that they discovered, particularly CCNB2 and UBE2C, have the potential to be helpful in both the diagnosis and prediction of breast cancer. Their differential expression across grades suggests that they play a role in determining the severity of the disease, and the correlation with patient survival rates highlights the potential utility of these genes in predicting outcomes. This research not only makes a significant contribution to our understanding of the molecular mechanisms that underlie breast cancer, but it also paves the way for the development of targeted diagnostic and prognostic approaches that can be utilised in the clinical management of this complicated disease.

The study by Engström et al. explores the molecular subcategories of breast cancer (BC) and their effects on diagnosis and prognosis. Cases were systematically rediverted into distinct molecular subcategories, including LumA, LumB (HER2-), LumB (HER2+), Basal, HER2 subcategory, and five negative phenotypes [20] [21]. Using immunohistochemistry and in situ hybridization, this categorization was investigated as a potential alternative to the analysis of gene expression. The Kaplan-Meier Survival (KMS) models and the Cox proportional hazards models

were among the analytical tools that were utilised in the research. The findings demonstrated that the prognosis of breast cancer is complex and varied, depending on the molecular subcategories. HER2 was found to have the most unfavourable prognosis and diagnosis, whereas LumA displayed the most favourable outcomes, along with the five negative phenotypes, particularly in the first five years after the investigation. It is important to note that Grade 2 tumours exhibited subcategory-related changes in BC survival, which highlights the significance of tumour grading in comprehending the progression of the disease. The histopathological grade or molecular subcategory did not have a significant impact on the survival rate of breast cancer patients after diagnosis. However, significant prognostic factors such as the involvement of lymph nodes, the grade of the tumour, and the size of the tumour played critical roles. There was a correlation between the non-luminal subcategory and negative outcomes, particularly in high-grade cases [21] [22] [23] [24]. This makes molecular subtyping an important tool for predicting outcomes.

To determine whether or not the division of breast cancer into molecular subtypes (Non-Luminal and Luminal) provides more accurate information than the use of traditional histological grade (HG) alone, the purpose of this study was being investigated. The purpose of this study was to investigate the survival of BC-specialized variants across a wide range of molecular and grade subcategories. In particular, the first five years after diagnosis were marked by significant variations in prognosis based on molecular subtypes. Luminal A displayed the best prognosis and HER2 subcategory, while the phenotypes of the five negatives displayed the worst prognosis. In addition to this, the research acknowledged the ever-changing nature of prognosis and prediction following the diagnosis of breast cancer. Molecular subtyping, in particular the examination of Gene Expression Profiles (GEP), has emerged as an additional source of data that can be used for prognostic assessment and prediction. The clinical and medical implications of molecular subtyping in breast cancer have not been fully appreciated, even though this potential existence exists.

The findings of the study highlighted the fact that Grade 1 tumors were associated with the most accurate diagnosis, whereas Grade 3 tumors displayed the most inaccurate results. Since they were more heterogeneous, tumors of grade 2 demonstrated transitional prognoses, which were similar to those of both grade 1 and grade 3 cases. It was also known that analyzing the Gene Expression Profiles (GEP) of patients across a variety of grades and molecular subtypes could be of assistance in gaining an understanding of how diseases begin and in developing personalized treatment plans. In general, the research contributes to a more nuanced understanding of the prognosis and diagnosis of breast cancer. It highlights the significance of molecular subtyping in addition to the utilization of traditional histological grading.

The molecular classification of breast cancer has emerged as a transformative aspect in the understanding and treatment of breast cancer. It provides a wealth of additional information that paves the way for individualized therapies. This classification not only improves our understanding of the disease, but also reveals new genes that are known to be cancer drivers and potential biomarkers that could be targeted for more effective treatments. There has been some progress made in determining how patients with various types of breast cancer will fare, but the situation is not yet resolved to everyone's satisfaction. Neoadjuvant chemotherapy (NACT) and conventional chemotherapy are two types of chemotherapy that are used to treat triple-negative and basal-like breast cancers (TNBC/BLBC). This is especially true when chemotherapy is administered. Despite the progress that has been made, there are still obstacles to overcome when it comes to transferring potential assays from the laboratory bench to the bedside of the patient. In addition to being time-consuming, the process is fraught with challenges that are associated with the unpredictability of gene alterations that culminate in therapeutic responses. In addition, there are concerns regarding the cost-effectiveness and quality control of these assays, which further complicate the process of implementing them in routine clinical practice. Taking into consideration the varied genetic landscape of breast cancer, there is an urgent need for more specific predictive assays to meet the requirements of personalized therapy. Even though the clinical application of these molecular findings is still in its infancy, there is a significant possibility that they will improve prognostic stratification and, as a result, refine therapeutic interventions. There is a palpable anticipation for therapies that are more precise and targeted, which promises better outcomes for patients suffering from breast cancer. However, it is essential to keep in mind that molecular classification should not be utilised in place of or in addition to routine histopathologic evaluation when it comes to the diagnosis and treatment of breast cancer [25].

Through a meticulous process of feature selection, the study, as indicated by the reference [24], establishes a complex connection between the Histologic Grade of breast cancer and its prognosis. The consideration of an initial feature set is the first step in this process. The initial feature that is sent can either be complete or partial. The methodology makes use of a progressive approach that is known as forwarding selection. This approach involves the methodical addition of features to broaden the scope of the exploration space. The next step is to conduct a backward search, which involves removing one gene at a time in an iterative manner in order to reduce the amount of space designated for exploration.

The two-step feature selection employed in this study serves as the backbone of the entire research endeavor, contributing to the following key components:

1) Identification of Appropriate Genes: The objective of the study is to identify the genes that are most

meaningfully associated with breast cancer using an exhaustive analysis that includes correlation and consistency measures. The use of a best-first search allows for a comprehensive investigation into the correlation between a variety of breast cancer characteristics and the prognosis of the disease, which makes the selection process easier to carry out.

2) Experimental Evaluation of Identified Genes: Experiments are conducted to evaluate the genes that have been identified, going beyond the simple act of identifying them. To thoroughly investigate the significance and relevance of the genes that have been chosen about breast cancer, requires the utilization of a variety of classification techniques. These genes have the potential to be useful in classification scenarios, and this step provides a practical understanding of that potential.
3) Ranking of Important Genes: The development of a ranking system for the genes that have been identified is an important part of the research that is being conducted. Specifically, this is accomplished by employing a GeS (Gene Selection) strategy that consists of two stages, which provides an additional layer of refinement to the selection process. In the context of breast cancer prognosis, this ranking system contributes to a more nuanced understanding of the relative importance of each gene that has been identified.
4) Medical Validation through Kaplan Meier Survival Model: Using the Kaplan Meier survival model, the research utilizes medical validation to determine whether or not the genes that were identified have any clinical significance. A powerful instrument for determining the influence of particular genes on the survival outcomes of patients is this model. The research helps to bridge the gap between the findings of computational analysis and the clinical implications that are being experienced in the real world.

In essence, this all-encompassing method not only identifies genes that are linked to breast cancer, but it also assesses the practical significance of these genes through the use of experimental methods and ranks them according to the importance they exhibit. Bringing together computational discoveries with potential clinical applications is made possible through the incorporation of medical validation, which further strengthens the translational aspect of the research. An intricate interplay between feature selection and gene analysis is the focus of this study, which is a multifaceted investigation to advance the understanding of the prognosis of breast cancer.

4. GeS Technique

The GeS approach to feature exploration concluded with the finest subgroups of features, and an attempt to discover a subgroup 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 [26] [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. There are times when a halting condition based on an estimation function either makes it easier to add or remove any feature, creating a better subgroup, or it achieves the best subgroup. 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 checks to see if the chosen subgroup is real by comparing and verifying the results with results that have already been found or with results from challenging the GeS approach using real-world or fake 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:

1) Correlation measures correspond to correlation either among features or among classes and features.
2) Inconsistency measure corresponds to a feature subgroup i.e. Unpredictable as at least two illustrations through equivalent feature principles through distinc-

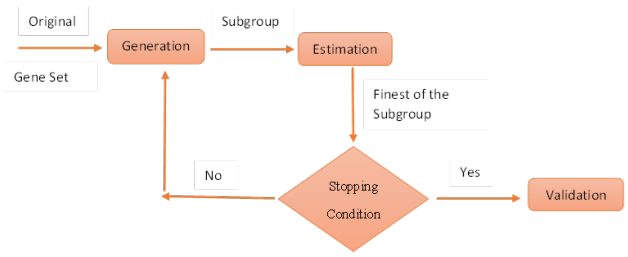

Figure 1. GeS approach

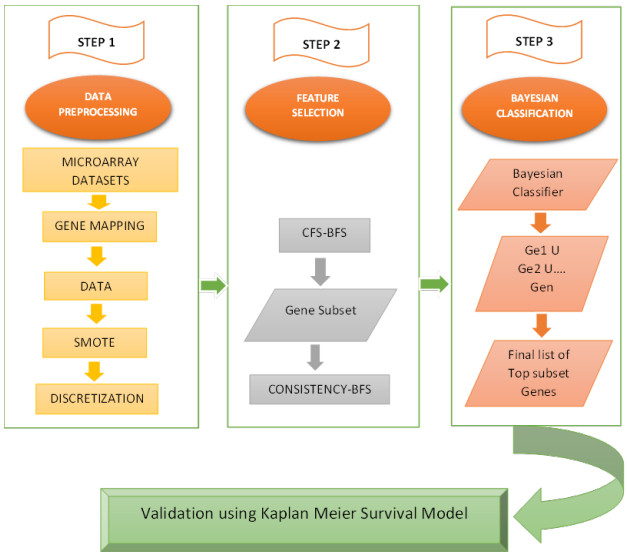

Figure 2. Flowchart of 2 Stage GeS

tive class markers.

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

5. Proposed Model

1) Step 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 concerns 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 imbalanced 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 Q.
Step 2: For every $b \in Q$, the sampling rate T is calculated as liable on the imbalanced proportion. T instances $t_1, t_2, \ldots t$ $(\leq m)$ are selected aimlessly amongst k-NN, therefore, producing the set Q_1.
Step 3: For each example $t_m \in Q_1 (m = 1, 2, 3, \ldots\ldots, T)$,
the stated method is utilized to generate the new instances

$$t_{new} = t + rand(0, 1) * ||(t - t_m)|| \quad (1)$$

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

2) Step 2: Feature Section
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 subgroups. The BFS technique falls under supervised Gene Selection (GeS). By indicating which genes, the algorithm thinks, fit the data the best, and chooses the relevant and significant genes. The algorithm encounters several 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:

$$A_c = \frac{f_{Crp}}{\sqrt{f + f(f-1)c_{pp}}} \quad (2)$$

A_c is the experimental 'merit' of a gene subgroup, including of genes, (c_{pp}) demonstrates genegene intercorrelation and epitomizes the geneclass 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 is implemented much more quickly than wrapper; as a result, CFS is used to select the final appropriate genes.

The Consistency BFS GeS method [33] determines how valuable a subgroup of qualities is in terms of the class standards although training events in the subgroups of qualities are predictable. 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. On the training dataset, consistency measures (CM) are handled uniquely because they have a lot of support and use minimum genes to choose a subset of genes [34]. The goal of min-genes is to define consistency theories over the fewest number of genes possible. It searches for the smallest subgroup size that satisfies the user-specified required consistency rate. 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 as: 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 and uncertain, there occur at least two occurrences like they associate all but with their class markers. 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., $\sum y = P$.

The inconsistency rate (IR) of a gene subgroup T is equal to the sum of all the overall designs that do not match up in the data for that gene subgroup divided by the power (P). The following is how CM is still being used for gene selection. CM remains utilized in 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 utilizes data with discrete value features. In this case, features must first be discretized if the data is continuous [36].

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, minimized, and pertinent genes, the next step is to classify the samples 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 modeled 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.

3) Step 3: Bayesian Classification

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.

$$P(A = a/B = b) = \frac{P((A = a)(P(B = b/A = a))}{P(B = b)} \quad (3)$$

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

$$P(B = b/A = a) = P(\bigwedge B_i = b_i/A = a),$$

$$= \pi P(B_i = b_i/A = a) (4)$$

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. A number between 0 and 1 that represents the likelihood that the gene B will take on the value b when the class is a serves as an example of P(A=a/B=b) for discrete features. In contrast, each numeric gene is demonstrated by some continuous likelihood distribution over the range of that gene's value. A mutual belief is that values

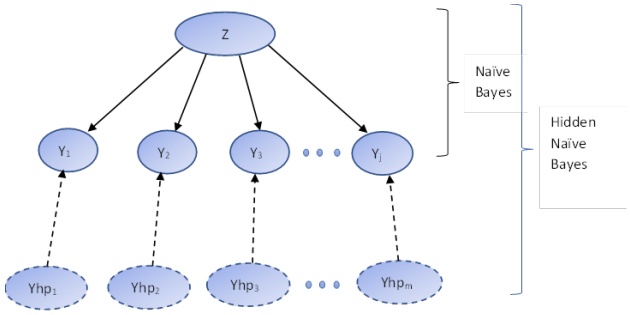

Figure 3. Structural representation of Naïve Bayes and Hidden Naïve Bayes

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.

$$P(B = b/A = a) = d(C : \mu_C, \sigma_C) where \quad (5)$$

$$d(C : \mu, \sigma) = \frac{\surd}{2\pi\sigma} e^{\frac{-(C-\mu)^2}{2\sigma^2}} \quad (6)$$

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 [39] [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 $Y_{h_{Pj}}$, $j = 1, 2, 3, ..., m$, signified by a dashed circle. The arc from the hidden parent $Y_{h_{Pj}}$ to Y_j is signified by a dashed line, to differentiate it from systematic arcs.

The joint distribution that HNB denotes is as follows:

$$P(Y, ..., Y_m, Z) = P(Z) \prod_{j=1.....m} P(Y_j/Y_{h_{Pj},Z}) \quad (7)$$

where

$$P(Y_j/Y_{h_{Pj},Z}) = \sum_{a=1,a\neq=b} T_{ba} * P(Y_b/Y_a, Z) \quad (8)$$

and $\sum\_a = 1, a \neq= bT_{ba} = 1$

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

Considering the attributes $Y_1, ..., Y_m$, $P(Y_j/Y_{h_{Pj}},Z)$ can be thought of approximation of $P(Y_1), ..., Y_m)$. In Equation 6, an approximation depends on single estimators. Through

the principle, arbitrary e-dependence estimators can be utilized to state hidden parents. HNB represents any Bayesian network if e = m-1. 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 $T_{\mathrm{ba}}, b, a = 1, ..., m$ and $a \neq b$, is decisive for learning HNB. There are two tactics to find it: one is executing a cross-validation-based search, or the second is directly executing the estimated values from the data. Adopted the latter and made use of conditional mutual information amongst attributes $\mathrm{Y_a}$ and $\mathrm{Y_b}$ as the weight of the $\mathrm{P(Y_a; } Y_{\mathrm{b}}|Z)$. More precisely, the weight is defined in eq. 9

$$T_{\mathrm{ba}} = \frac{M_{\mathrm{p}}(Y_{\mathrm{b}}; Y_{\mathrm{a}}|Z)}{\sum_{a=1,a\neq b} M_{\mathrm{p}}(Y_{\mathrm{b}} : Y_{\mathrm{a}}|Z)} \tag{9}$$

where $\mathrm{M_p}$ ($\mathrm{Y_b}$: $\mathrm{Y_a}$|Z) is a conditional mutual information defined as:

$$M_{\mathrm{p}}(A; B|C) = \sum_{a,b,c} P(a, b, c) \log \frac{P(a, b|c)}{P(a|c)P(b|c)} \tag{10}$$

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

6. Datasets

The experimentations are conducted on six microarray gene expression datasets extracted from National Centre for Biotechnology Information (NCBI) and is detailed in Table II. 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 III, shows Grade-wise distribution of samples. The number of relevant genes selected in 2-Stage GeS is shown in Table 4. HG 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 V-VIII. Out of these three classifiers, HWNB outshines in terms of precision, recall, f-score, and fallout highlighted in bold in Table VI. 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.

A. 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 IV. 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 IX. Considering all the selected genes by the 2GeS tactic, in each dataset where the correlation coefficient is calculated to find the correlation among the genes. Considering all selected gene's coefficients, the genes are ranked. Combining all the selected genes of six datasets, the ranking of genes is shown in Table X. Dataset-wise ranking of the top three selected genes is shown in Table XI. 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=servicecancer=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

TABLE II. 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 III. 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

TABLE IV. Count of features selected in both stages

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

TABLE V. 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 VI. 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 VII. 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 VIII. 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

TABLE IX. 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

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 shows that the proposed model works to divide BC patients into two groups based on their RFS rate, which can tell them how likely it is that they will have 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. It was thought that people with BC who had a lot of mRNA for the E2F3, PSMC3IP, GINS1, and PLAGL2 genes would have a high RFS in Grades 1 and 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. The Kaplan-Meier Plotter database was used to look at survival rates. It showed that high transcription levels of E2F3 were linked with low relapse-free survival (RFS) in all breast cancer patients. E2F3 is a potential target of precision therapy for patients with breast cancer [41] [42]. Studies of survival showed that higher levels of GINS1 were linked to bad outcomes in all patients with BC [43] [44]. GINS1 was associated with detrimental relapse-free survival (RFS). All the experiments are performed using the WEKA software and RStudio [29].

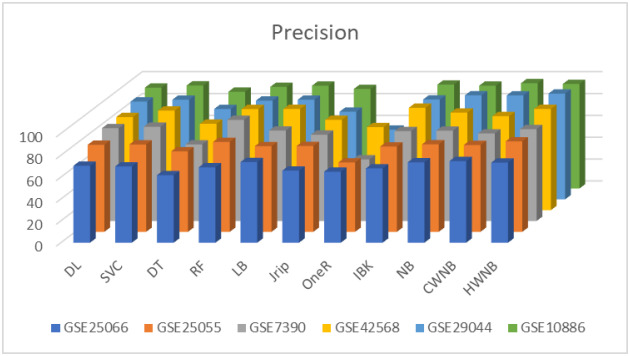

Figure 4. Performance of six datasets based on Precision

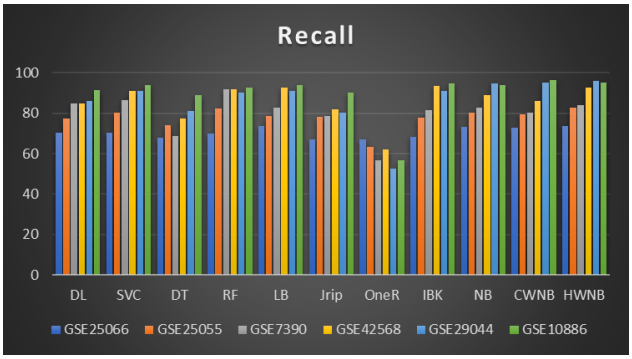

Figure 5. Performance of six datasets depending on Recall parameter

TABLE X. 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

TABLE XI. 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

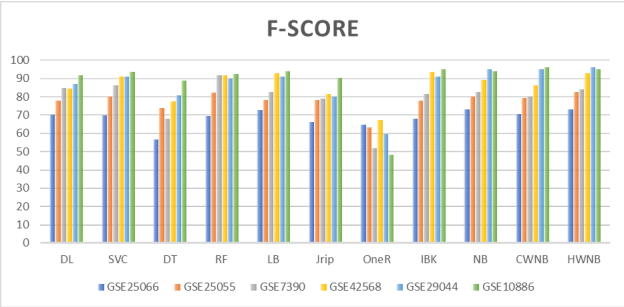

Figure 6. Performance of six datasets based on F-score

7. Conclusion and Discussion

The purpose of this study was to propose a novel two-stage GeS strategy for the prediction of BC subtypes. The strategy was based on two methods and a Hidden Naïve Bayes classifier. The first stage involved the utilization of CFS-BFS, while the second stage involved the utilization of CONSISTENCY-BFS, which utilized histologic grade. Additionally, the classification was carried out with the Hidden Weight Naïve Bayes classifier. It is preferable to use Consistency-BFS because its complexity is linear, which is denoted by the notation (O(N)). On the other hand, CFS-BFS has a polynomial complexity, which is denoted by the notation (O(N^2)), where N is the total number of

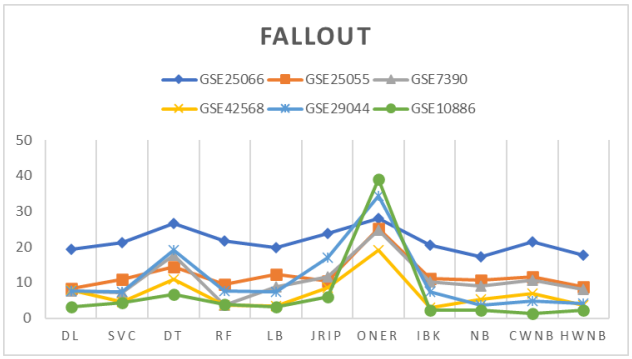

Figure 7. Performance of six datasets depending on Fallout

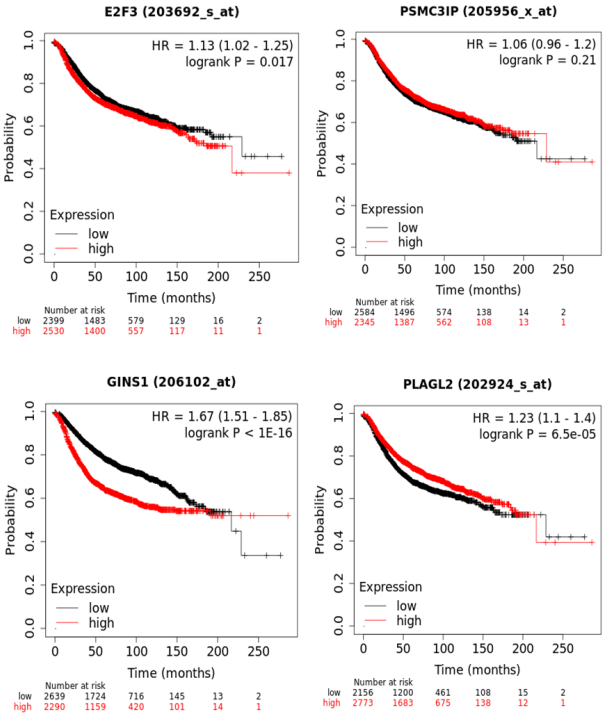

Figure 8. Prognostic Value of mRNA Level of top four genes with RFS with all types of the histologic grade in Breast Cancer Patients (Kaplan-Meier Plotter).

features. Six different microarray gene expression datasets were utilized in the experiments that were carried out. Using a limited number of appropriate genes from each microarray gene expression dataset, the results validate an impressive precision, recall, f-score, and fallout to forecast breast cancer. The proposed 2-GeS tactic and Hidden Naïve Bayes classifier are responsible for the impressive results that have been achieved. The majority of the selected genes have been shown to have a correlation to breast cancer based on previous research; however, only a few of these genes have yet to be investigated.

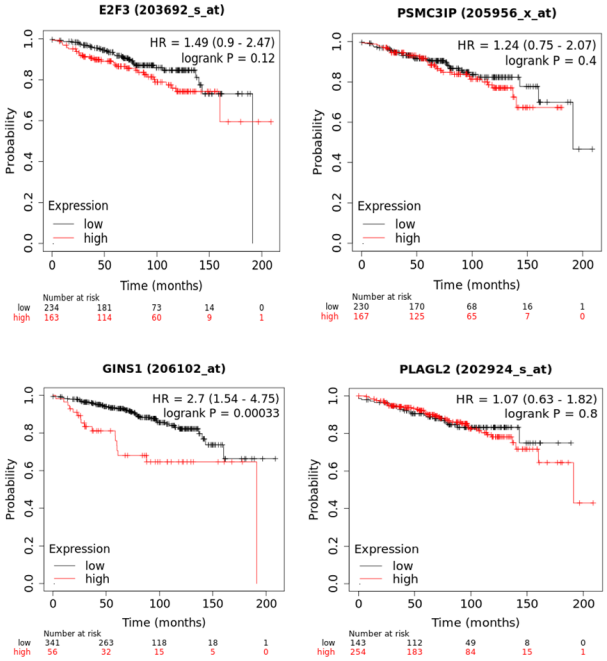

Figure 9. Prognostic Value of mRNA Level of top four genes with RFS with the histologic Grade 1 in Breast Cancer Patients (Kaplan-Meier Plotter).

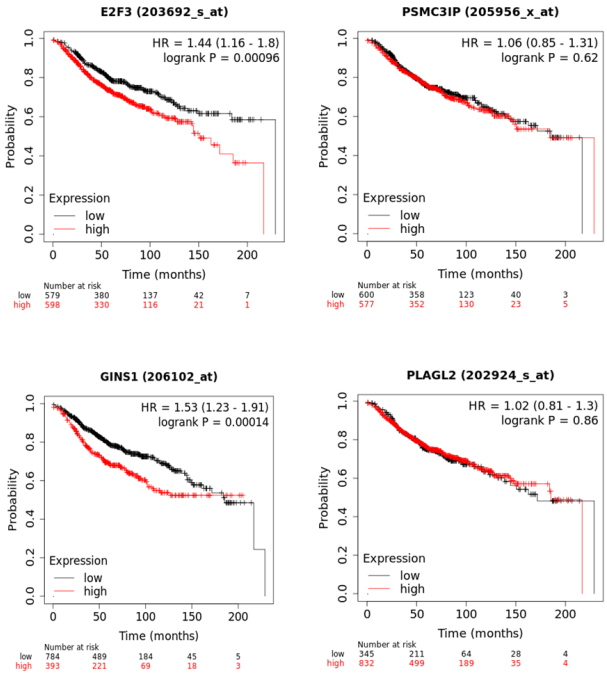

Figure 10. Prognostic Value of mRNA Level of top four genes with RFS with all the histologic Grade 2 in Breast Cancer Patients (Kaplan-Meier Plotter).

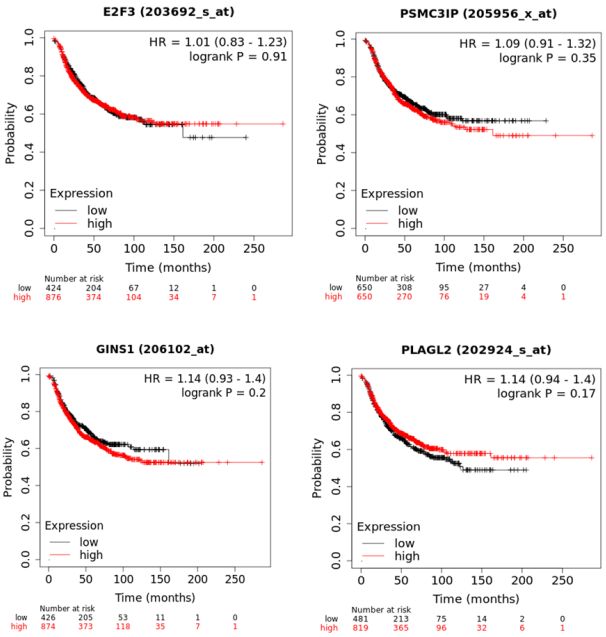

Figure 11. Prognostic Value of mRNA Level of top four genes with RFS with the histologic Grade 3 in Breast Cancer Patients (Kaplan-Meier Plotter).

To concentrate research and analysis on a relatively small subset of genes, this two-stage gene selection strategy can be utilized. Particularly noteworthy is the fact that the strategy has the potential to be useful for more complex patient stratification in the future. For example, it could be useful for subgroups that are formed by combining platforms or for groups of patients that have been divided based on how treatments have been received. It is possible to precisely classify large populations of patients into specific cancer subtypes or treatment groups with the assistance of this combination tool. This is accomplished by regulating the minimum number of genes that need to be screened. The detection of subtle genetic variations is hampered by smaller sample sizes in each stage, which may result in a reduction in statistical power. This is another limitation. Through the two-stage model, cooperative effects may be overlooked due to the isolation of individual genes, which may obscure important interactions within gene networks. The future of cancer research and personalized medicine appears to be greatly promising if the limitations related to the identification of minimum relevant genes in two-stage breast cancer analysis are overcome. Precision medicine may enter a new era if issues like the intrinsic heterogeneity of breast cancer and our incomplete understanding of genetic factors are addressed. Further investigation into gene interactions and larger sample sizes may reveal more reliable biomarkers that improve prognostic accuracy and early detection. Thus, better treatment outcomes could result from the creation of more potent therapeutic targets. Finding new genes and pathways may be facilitated by integrating diverse omics data and resolving data integration obstacles to offer a more comprehensive understanding of breast cancer. Developments in the modeling of the time-dependent nature of gene expression changes and the dynamic nature of cancer progression can guide interventions at different phases of development. Understanding the influence of racial and ethnic diversity can result in more inclusive research, ensuring that conclusions hold true for a range of demographics. While longitudinal studies and real-world data integration provide a comprehensive understanding of the genetic landscape over time, leveraging artificial intelligence and machine learning can reveal subtle patterns within complex datasets. All things considered, getting past these obstacles could completely transform the field of cancer research and open the door to more inclusive, individualized, and focused methods of diagnosing and treating breast cancer.

The findings showed that the top two genes E2F3 and GINS1 subunits might be new potential predictive biomarkers for BC. However, additional authentication studies are still required to demonstrate the clinical significance of E2F3 and GINS subunits in BC patients. 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.

8. Declaration of Interest

None

References

[1] H. R. Ali, O. M. Rueda, S.-F. Chin, C. Curtis, M. J. Dunning, S. A. Aparicio, and C. Caldas, "Genome-driven integrated classification of breast cancer validated in over 7,500 samples," *Genome biology*, vol. 15, pp. 1–14, 2014.

[2] M. Lamba, G. Munjal, and Y. Gigras, "Ecabc: Evaluation of classification algorithms in breast cancer for imbalanced datasets," in *Data Driven Approach Towards Disruptive Technologies: Proceedings of MIDAS 2020*. Springer, 2021, pp. 379–388.

[3] M. lamba, G. Munjal, and Y. Gigras, "Supervising healthcare schemes using machine learning in breast cancer and internet of things (shsmliot)," *Internet of Healthcare Things: Machine Learning for Security and Privacy*, pp. 241–263, 2022.

[4] M. Lamba, Y. Gigras, and A. Dhull, "Classification of plant diseases using machine and deep learning," *Open Computer Science*, vol. 11, no. 1, pp. 491–508, 2021.

[5] M. Lamba, G. Munjal, and Y. Gigras, "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, 2021.

[6] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani *et al.*, "Breast cancer prognostic classification in the molecular era: the role of histological grade," *Breast cancer research*, vol. 12, no. 4, pp. 1–12, 2010.

[7] M. Lamba, G. Munjal, Y. Gigras, and M. Kumar, "Breast cancer prediction and categorization in the molecular era of histologic grade," *Multimedia Tools and Applications*, pp. 1–20, 2023.

[8] N. Olsson, P. Carlsson, P. James, K. Hansson, S. Waldemarson, P. Malmström, M. Fernö, L. Ryden, C. Wingren, and C. A. Borrebaeck, "Grading breast cancer tissues using molecular portraits," *Molecular & Cellular Proteomics*, vol. 12, no. 12, pp. 3612–3623, 2013.

[9] V. S. A. Jayanthi, A. B. Das, and U. Saxena, "Grade-specific diagnostic and prognostic biomarkers in breast cancer," *Genomics*, vol. 112, no. 1, pp. 388–396, 2020.

[10] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, and B. Shi, "Breast cancer intrinsic subtype classification, clinical use and future trends," *American journal of cancer research*, vol. 5, no. 10, p. 2929, 2015.

[11] E. Amiri Souri, A. Chenoweth, A. Cheung, S. N. Karagiannis, and S. Tsoka, "Cancer grade model: a multi-gene machine learning-based risk classification for improving prognosis in breast cancer," *British Journal of Cancer*, vol. 125, no. 5, pp. 748–758, 2021.

[12] E. A. Rakha and F. G. Pareja, "New advances in molecular breast cancer pathology," in *Seminars in cancer biology*, vol. 72. Elsevier, 2021, pp. 102–113.

[13] S. Jenkins, M. E. Kachur, K. Rechache, J. M. Wells, and S. Lipkowitz, "Rare breast cancer subtypes," *Current oncology reports*, vol. 23, pp. 1–14, 2021.

[14] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 13, no. 5, pp. 971–989, 2015.

[15] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012.

[16] A. Nagpal and V. Singh, "Feature selection from high dimensional data based on iterative qualitative mutual information," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 5845–5856, 2019.

[17] M. Lamba, G. Munjal, and Y. Gigras, "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, 2021.

[18] M. lamba, G. Munjal, and Y. Gigras, "Feature selection of microarray expression data (fsm)-a review," *Procedia computer science*, vol. 132, pp. 1619–1625, 2018.

[19] M. Lamba, G. munjal, and Y. Gigras, "Computational studies on breast cancer analysis," *Journal of Statistics and Management Systems*, vol. 23, no. 6, pp. 999–1009, 2020.

[20] F. M. Blows, K. E. Driver, M. K. Schmidt, A. Broeks, F. E. Van Leeuwen, J. Wesseling, M. C. Cheang, K. Gelmon, T. O. Nielsen, C. Blomqvist *et al.*, "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 medicine*, vol. 7, no. 5, p. e1000279, 2010.

[21] A. S.-Y. Leong and Z. Zhuang, "The changing role of pathology in breast cancer diagnosis and treatment," *Pathobiology*, vol. 78, no. 2, pp. 99–114, 2011.

[22] M. J. Engstrøm, S. Opdahl, A. I. Hagen, P. R. Romundstad, L. A. Akslen, O. A. Haugen, L. J. Vatten, and A. Bofin, "Molecular subtypes, histopathological grade and survival in a historic cohort of breast cancer patients," *Breast cancer research and treatment*, vol. 140, pp. 463–473, 2013.

[23] A. Taherian-Fard, S. Srihari, and M. A. Ragan, "Breast cancer classification: linking molecular mechanisms to disease prognosis," *Briefings in bioinformatics*, vol. 16, no. 3, pp. 461–474, 2015.

[24] N. Chowdhury, "Histopathological and genomic grading provide complementary prognostic information in breast cancer: a study on publicly available datasets," *Pathology Research International*, vol. 2011, 2011.

[25] K. R. Srivastava and N. Girdhar, "Retinal image segmentation based on machine learning techniques," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2022, pp. 252–257.

[26] M. Roy, A. M. Fowler, G. A. Ulaner, and A. Mahajan, "Molecular classification of breast cancer," *PET clinics*, 2023.

[27] E. C. Blessie and E. Karthikeyan, "Sigmis: A feature selection algorithm using correlation based method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385–394, 2012.

[28] M. Lamba, G. Munjal, and Y. Gigras, "Ranking of classification algorithm in breast cancer based on estrogen receptor using mcdm technique," *International Journal of Information Technology & Decision Making*, vol. 22, no. 02, pp. 803–827, 2023.

[29] J. Allaire, "Rstudio: integrated development environment for r," *Boston, MA*, vol. 770, no. 394, pp. 165–171, 2012.

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

[31] M. Lamba, G. Munjal, and Y. gigras, "Identifying breast cancer molecular class using integrated feature selection and deep learning model," *International Journal of Bioinformatics Research and Applications*, vol. 19, no. 1, pp. 19–42, 2023.

[32] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[33] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[34] C. S. K. Dash, A. Kumar Behera, S. Dehuri, and S.-B. Cho, "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*, vol. 18, no. 1/2, pp. 151–162, 2022.

[35] T. Zhang, S. Ye, K. Zhang, J. Tang, W. Wen, M. Fardad, and Y. Wang, "A systematic dnn weight pruning framework using alternating direction method of multipliers," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 184–199.

[36] D. M. Maslove, T. Podchiyska, and H. J. Lowe, "Discretization of continuous features in clinical datasets," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 544–553, 2013.

[37] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[38] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, pp. 1–37, 2008.

[39] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," *arXiv preprint arXiv:1302.4964*, 2013.

[40] H. Zhang, L. Jiang, and J. Su, "Hidden naive bayes," in *Aaai*, 2005, pp. 919–924.

[41] B. Győrffy, "Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer," *Computational and structural biotechnology journal*, vol. 19, pp. 4101–4109, 2021.

[42] C.-C. Sun, S.-J. Li, W. Hu, J. Zhang, Q. Zhou, C. Liu, L.-L. Li, Y.-Y. Songyang, F. Zhang, Z.-L. Chen *et al.*, "Retracted: comprehensive analysis of the expression and prognosis for e2fs in human breast cancer," *Molecular Therapy*, vol. 27, no. 6, pp. 1153–1165, 2019.

[43] C. Nieto-Jiménez, A. Alcaraz-Sanabria, R. Páez, J. Pérez-Peña, V. Corrales-Sánchez, A. Pandiella, and A. Ocaña, "Dna-damage related genes and clinical outcome in hormone receptor positive breast cancer," *Oncotarget*, vol. 8, no. 38, p. 62834, 2017.

[44] H. Li, Y. Cao, J. Ma, L. Luo, and B. Ma, "Expression and prognosis analysis of gins subunits in human breast cancer," *Medicine*, vol. 100, no. 11, 2021.

Dr. Monika Lamba is presently employed as an Assistant Professor in the Department of Computer Science and Engineering at The NorthCap University in Gurugram. She has completed her Bachelor of Technology in Computer Science and Engineering from Ansal Institute of Technology (AIT), which is affiliated with IP University in Dwarka, New Delhi. Afterwards, she obtained her M.Tech. degree in Computer Science and Engineering from the University School of Information, Communication and Technology (USICT), located in Dwarka, New Delhi. She passed the GATE exam four times. She has obtained a Ph.D. in Machine Learning from The NorthCap University, Gurugram. She has accumulated approximately 6 years of professional experience in the fields of teaching and research. She, being a dedicated researcher, has delivered academic papers at numerous international conferences. She has published research papers in internationally recognised journals and conferences indexed in SCI/ESCI/Scopus. She has written two chapters for books. She is currently fulfilling the role of a reviewer for numerous academic journals.

Dr. Geetika Munjal is presently employed as an Associate Professor at Amity University. She is a goal-oriented, committed expert with over 17 years of experience in the fields of teaching and research. She possesses an MTech in CSE from Punjab Technical University and a Ph.D. from The NorthCap University, specialising in Machine Learning. She has successfully finished a project sponsored by DST in the field of datamining. She is currently conducting research in the fields of Data Mining, Pattern Recognition, Machine Learning, and Deep Learning. She has authored approximately 20 papers in internationally recognised peer-reviewed journals, which are well-indexed. Additionally, she has contributed to esteemed national and international conference proceedings and book chapters. She has presided over multiple sessions at academic conferences organised by Springer and IEEE in India. She has supervised about 30 B. Tech projects, 10 M.Tech theses, and 1 Ph.D scholar.

Dr. Yogita Gigras is presently employed as an Associate Professor (Selection Grade) in the Department of Computer Science and Engineering and Information Technology at the School of Engineering and Technology, NCU. She possesses over nine years of extensive teaching experience at both the postgraduate and undergraduate levels. She is a dedicated researcher specialising in Soft Computing and has successfully obtained her PhD in the same field. She completed her M.Tech in Computer Science and Engineering with honours from Banasthali University, Rajasthan in 2009. She completed her industrial training at ST Microelectronics in Greater Noida and successfully finished two live projects as part of her M.Tech degree requirements. She is particularly interested in the fields of Algorithm Analysis and Design, Object Oriented Programming, Operating Systems, Computer Networks, Soft Computing, and Cyber Security. She has supervised 14 M.Tech Projects and over 25 B.Tech Projects. She successfully passed the Graduate Aptitude Test in Engineering (GATE) in the year 2007. She has authored over 20 papers that have been published in internationally recognised peer-reviewed journals and presented at IEEE/Springer international conferences. She holds the positions of reviewer and assistant editor for multiple international journals. In addition, she has obtained certifications in Data Scientist Tool, Exploratory data analysis, and Getting and Cleaning Data from Johns Hopkins University. She holds a lifetime membership with ISTE.