

An Enhanced Feature Selection Method for Breast Cancer Prediction Using a Hybrid Framework Based on Logistic Regression with Noise Filtering and Data Balance

Kamal Bashir^{*†}, Saied Pirasteh^{*}, Hassan Abdelrhman[†], Mohamed Mosadag[†], Ashraf Mohammed[†]

^{*}Institute of Artificial Intelligence, School of Mechanical and Electrical Engineering, Shaoxing University
Zhejiang Province 312000, China

[†]College of Computer Science and Information Technology, Karary University, Khartoum, Sudan
E-mail: kamalbashir1@yahoo.com

Abstract. Breast Cancer (BC) is a prevalent disease worldwide, particularly affecting women. Developing accurate classification models is a significant challenge in medical informatics. However, machine learning techniques have effective categorization models and data mining from repositories often involves high dimensionality, incorrect labeling, and class imbalance. These issues can adversely affect the performance of most statistical learning approaches for outcome prediction. This research aims to introduce an innovative method for Feature Selection (FS) that combines Maximum Likelihood Logistic Regression (MLLR) with Noise Filtering (NF) and Data Balance (DB) techniques. Our proposed method involves several steps to improve breast cancer prediction accuracy. Firstly, we utilized multiple iterations of the Iterative Noise Filtering based on the Fusion of Classifiers (INFFC) algorithm to reduce class noise present in cancerous datasets. Next, we addressed the issue of class imbalance by using a modified Synthetic Minority Over-sampling Technique (SMOTE). Subsequently, we applied the Maximum Likelihood Logistic Regression (MLLR) method and determined the statistical significance of each feature's estimated coefficients with the Wald test at a given confidence interval. Additionally, we compared the effectiveness of our proposed method with several well-known feature selection methods. Finally, we utilized five different classifiers namely, Naïve Bayes (NB), Multi-Layer Perceptron (MLP), K-Nearest Neighbors (KNN), AdaBoostM1, and Bagging to assess their performance using the area under the receiver operating characteristic curve (AUC) metric. Our findings demonstrated that our proposed method can effectively select relevant features and improve classifier performance in binary classification problems, especially in MLP which is (0.949). Furthermore, based on the results of five different classifiers, our proposed method outperformed the compared methods used in this study.

Key-terms-*Breast cancer prediction, Machine learning, Feature selection, Noise filtering, Data balance*

I. INTRODUCTION

According to the World Health Organization (WHO), breast cancer (BC) is recognized as the predominant form of cancer affecting women globally and ranks as the second highest contributor to cancer-related mortality [1]. According to Bray et al., BC remains the prevalent form of cancer among women [2]. The likelihood of a woman developing BC at some point in her life is 12%, or one in eight, by the time she reaches 85 years old [3]. Furthermore, approximately 10.9% of all BC patients are under the age of 35 [4]. BC is characterized by the uncontrolled growth of cells within breast tissues, which can be either benign or malignant. While a benign tumor remains localized and does not spread, a malignant tumor has the potential to metastasize, indicating the progression of the disease [5], [6]. Risk factors for BC include obesity, nulliparity, early menarche, lactation duration, unhealthy lifestyle, as well as geographic, racial, and ethnic characteristics [7], [8], [2], [9].

Moreover, comparing Machine Learning (ML) to pathologists has many advantages. ML is regarded as the ideal method for detecting future cancer since it removes manual labor and aids in text and image categorization for doctors. Besides, malignant and benign tumors are two types. Benign tumors do not cause cancer or pose a threat to life. These, though, can make acquiring breast cancer more likely. The threat posed by tumors is greater than that of benign growth tumors. However, essential studies are carried out early. Twenty percent of all women with malignant tumors pass away before their disease is discovered from this illness [10]. Consequently, there is a need for a completely automated approach to detect cancer. Several research studies have explored the application of machine learning methods to evaluate the presence of carcinoma in individuals [11], [12], [13], [14], [15], [16]. They also demonstrated that these algorithms are more successful in diagnosing carcinoma [13]. In order to achieve accurate

patient detection, the involvement of a skilled and experienced doctor is typically necessary [12]. But this skill is developed over years of confirming diagnoses and observing the adverse effects of various patients. There is, however, no guarantee of dependability. Massive data sets, like specialized electronic patient information databases, can now be easily gathered and stored, thanks to processing technology advancements [11]. Health professionals needed a computer to access these sizable databases, mainly when conducting extensive data analysis.

Furthermore, an accurate diagnosis of breast cancer is crucial in determining appropriate treatment, and misdiagnosis can result in patients not receiving the necessary care. Hence, the research focused on precise diagnosis and the classification of breast cancer into benign and malignant groups continues to be a subject of intense debate and discussion. In the previous century, machine learning (ML) techniques were widely used to diagnose breast cancer and identify data patterns. ML is also commonly used to develop models for Breast Cancer Prediction (BCP), which involves identifying patterns and regularities within datasets to generate theories that can uncover new hidden data connections. BCP models are developed using past data to accurately diagnose cancer. However, using this training data, we can estimate and predict unseen cases and classify them as either benign or malignant [17].

Alternatively, the number of samples and the dimensionality of the data available recently have increased across many machine types. Learning programs like text mining, computer vision, and biomedical. Studying how to use these large-scale data is crucial and necessary for knowledge acquisition. High data dimensionality is primarily of interest to us. Due to the enormous volume of high-dimensional data, existing machine learning systems face a sizable challenge. Learning algorithms can become very complex due to noisy, redundant, and irrelevant dimensions. Therefore, performance on learning tasks can be sluggish or even degenerate, making interpreting models challenging. For example, a feature is irrelevant if it provides no information relevant to the classification task, and a feature is redundant if it provides the same information as the currently used features. The predictive performance, execution time, and complexity all suffer when a model based on training data contains redundant and irrelevant features. However, FS has consequently evolved into a crucial component of data mining. FS is locating pertinent features and removing unneeded, redundant, or noisy data [18].

Despite the many FS techniques and advantages, some problems related to data can present difficulties and increase the difficulty of the selected important features. Noisy and unbalanced data primarily characterize these difficulties. The inductive learning of statistical methods can be affected by data noise [19], [20], [21]. Noise can have an impact on the dependent (i.e., whether it is benign or malignant) along with independent variables (i.e., feature noise). Concerning expected syntax, semantics, and values, noise can cause inconsistencies in training data. As a result, making decisions based on erroneous information may result in inaccurate predictions when learning from noisy training data [22], [23].

On the other hand, balanced data and noisy data can affect how statistical methods are learned inductively. When

there are fewer samples than expected in one class (i.e., benign) is considerably more than those in another class (i.e., malignant). Depending on the imbalance ratio, the impact of data skews may change. A high-class imbalance, frequently seen in BC datasets, can be used to identify the minority class using any complicated statistical learning method. Because of the high-class imbalance, the majority class is biased [24]. Therefore, it becomes very challenging for utmost statistical learning techniques to differentiate between the majority and minority classes effectively, creating a task akin to searching for the proverbial needle in a haystack. The skewed data may introduce biased learning by causing the selection of incorrect features or feature subsets that are unable to predict samples from minority classes. Whereas in BCP, the incidence of false positives (i.e., incorrectly projecting a damaged module as benign) is more costly than an erroneous negative (i.e., unfavorable outcomes could result from a learner's prediction that tends toward the majority class).

Thus, because of the challenges of selecting useful features from imbalance and noisy data, we suggest a new method that uses a hybrid framework based on logistic regression using Maximum Likelihood Logistic Regression (MLLR) to select features from enhanced data by applying the Noise Filtering technique to remove noisy and Data Balance to treat the imbalance problem.

The remaining parts of the current study are arranged as follows: Section 2 presents the pertinent work. Section 3 explains the details of the proposed scheme and method. Our findings and analysis are reported in Section 4. Finally, Section 5 provides the conclusion with some suggestions.

II. RELATED WORK

Many studies have been conducted to diagnose BC. Different classification methods have been applied using machine learning, including Logistic Regression (LOR), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RaF), Multi-Layer Perceptron (MLP), K-nearest neighbor (KNN), Linear Regression (LR), and many others [25], [26]. To enhance prediction accuracy and optimize feature selection, various techniques such as filter, wrapper, and embedded approaches were employed. For example, Chougrad et al. suggested a viewing technique to increase breast cancer in its early stages. The author employed a deep convolution technique and computer-aided analysis to effectively screen for breast cancer during its benign stage [27]. Geeitha and Thangamani created a Pearson correlation model to choose the best genes for cervical cancer classification [28]. Identification of the key genes associated with cervical cancer has been accomplished through the utilization of an integrated model called Enhanced Bat Optimization with Hilbert Schmidt Independence Criterion (EBO-HSIC). The classification analysis used two gene samples of normal and abnormal gene data.

In addition, Gopal et al. categorized BC using systematic and unbiased prognostic indicators. To predict BC and make estimates based on the prognosis factor, two learners, including NB and KNN, have been implemented [29]. Dasgupta et al. suggested many models identify different cancer types and

Abbreviations	Phrase	Abbreviations	Phrase
WHO	World Health Organization	NB	Naive Bayes
BC	Breast Cancer	DT	Decision Tree
ML	Machine Learning	RaF	Random Forest
BCP	Breast Cancer Prediction	MLP	Multi-Layer Perceptron
MLLR	Maximum Likelihood Logistic Regression	KNN	K- nearest neighbor
LOR	Logistic Regression	LR	Linear Regression
SVM	Support Vector Machine	MLE	maximum likelihood estimation
INFFC	Iterative Noise Filtering based on the Fusion of Classifiers	CS	Chi-Square
SMOTE	Synthetic Minority Over-Sampling Technique	PCA	Principal Components Analysis
CFS	Correlation Feature Selection	RF	ReliefF
RFW	ReliefF with Weighting	WPBC	Wisconsin Breast Cancer Dataset Prognostic
ROC	Receiver Operating Characteristic curve		

TABLE I: Table of abbreviations

choose features. Machine learning models like NB, Neural Networks, DT, and LOR are used on the BC dataset for FS and BCP [30]. In another study, Vasudev et al. applied three learning methods techniques, including DT, NB, and SVM to diagnose cancer. Furthermore, FS was used to reduce the high dimensionality [31]. Besides, Yue et al. discussed a few techniques of machine learning for parameter selection, including ANN, KNN, and SVM. A new PSO method was used for kernel function parameter optimization in SVM. On the dataset related to cancer, the author performed FS [32]. LDA was used to reduce dimensionality with the Conventional algorithm in Omondigbe et al. hybrid approach for BC diagnosis, including FS and SVM classifiers [33].

Furthermore, the relationship between BC and a few characteristics to lower the mortality rate was investigated by Li Chen. Five distinct classification techniques were employed to apply the model to two datasets [34]. Hajiabadi et al. proposed an objective function using an ANN to diagnose BC. By integrating correntropy, hinge, and cross-entropy loss functions, the author conducted a comprehensive analysis of the dataset at different noise levels. This approach not only enhanced generalization but also facilitated the identification of optimal loss functions [35]. Shravya et al. proposed analyzing BC using kerning methods like LOR, SVM, and KNN learners, demonstrating that SVM is the best learner. Attributes were made irrelevant in the prediction of BC [36]. However, machine learning techniques were studied by Chaurasia et al. to apply risk factors for breast cancer, as well as implemented prediction models for evaluating breast cancer patient's chances of survival. Three models, J48, RBF network, and NB, were used to classify normal and abnormal cells [37]. The models put forth by Aavula and Bhramaramba (2019) addressed BCD, recurrence, and patient survival. For risk calculation and recurrence prognosis analysis, a novel model known as the Extensible BC Prognosis Framework was utilized. The model incorporated Representative FS (RFS) and SVM techniques to enhance prognosis factor performance [38].

Similarly, Nandagopal et al. proposed analyzing breast cancer by choosing important characteristics, categorizing the early and invasive stages for FS, and classifying malignant and benign phases of BC. The authors used maximum likelihood estimation (MLE), LR, and fuzzy logic [39]. Abdar et al.

created a nested two-level ensemble classifier. The evaluation of the model involved K-fold cross-validation, while an ensemble stacking and voting approach was implemented. Through this model, the detection performance of the NB algorithm for breast cancer was significantly enhanced [40]. Wang used X-ray mammography, the gold standard, and the microwave breast imaging technique to predict breast cancer. In his study, breast cancer detection was the primary goal when designing microwave biosensors. The distinction between normal and abnormal tissues was achieved by leveraging the low dielectric property as a discriminative factor [41].

To the best of our experience, over the years, many researchers have used machine learning for BCP and built their model by treating data difficulties such as class imbalance, high dimensionality, and noisy data. Most of the study has been done to understand, treat, and mitigate each problem independently. Thus, the purpose of this study is to develop a novel method used hybrid framework based on MLLR on enhanced data that removes the noise and balances the data. Nevertheless, the following section explains our proposed scheme on how to enhance the feature selection based on maximum likelihood logistic regression for breast cancer prediction.

III. PROPOSED SCHEME OF ENHANCED FEATURE SELECTION METHOD

All supervised, unsupervised, and regression problems require using FS to condense the reduced high dimensionality [42]. The MLLR has a lot of benefits. The first benefit is excellent predictive power, testing and removing the probably zero coefficients and minimizing variance without significantly increasing bias. This is especially useful when dealing with the high-dimensional dataset [43]. This study suggests the FS technique based on MLLR for BCP due to the above benefits. However, it is not used enough in BCP studies. Even though it is practical, noisy data and class imbalance can prevent the MLLR from locating relevant features for classification modeling. So, in this work, we proposed a novel combined FS method based on MLLR, where noisy and imbalanced data treatment based on filtering and oversampling is carried out before the FS.

Initially, the removal of class noise from defect datasets is accomplished through iterative iterations of the Iterative

Noise Filtering based on the Fusion of Classifiers (INFFC) technique. The algorithm Synthetic Minority Over-Sampling Technique (SMOTE) then creates new synthetic malignant instances to make up for the class imbalance using a random linear interpolation between a malignant instance and its nearest neighbor. The final step of the MLLR is to determine the significance of each feature's estimated coefficients at a given confidence interval using the Wald test of statistical significance.

To achieve this objective, three scenarios were carefully examined to comprehensively evaluate and compare the performance of MLLR with other feature selection (FS) techniques. These assessments encompassed various scenarios incorporating the combined utilization of FS, noise filtering, and data balancing techniques: In Case A, we conducted FS on the original data, in Case B, where FS is based on noise-filtered data, and in Case C, where FS is based on noisy filtered and balanced data. The thorough literature review indicates that Case A applies to most of the current FS methodologies. We assumed that an unbalanced class distribution and class mislabels may make it difficult to find useful features and render the FS methods useless. Similar to Case A, Case B's FS techniques could be plagued by class imbalance, producing less-than-ideal outcomes even though the mislabels are filtered [23]. Thus, the best course of action might be to take the more thorough method suggested in Case C, which would balance the class distribution and filter class mislabels before FS. Next, five learners were selected to assess the suggested scheme's effectiveness, which involves conducting FS after noise filtering and balancing the data using NB, MLP, KNN, AdaBoostM1, and Bagging. The fact that these classifiers are frequently employed in BCP studies led to their selection.

The stages utilized in this work are presented in Fig. 1 and are described below in three cases. The MLLR technique is discussed in detail in the following sections, along with a brief introduction to the other feature selection (FS) techniques used for comparison purposes. Additionally, we describe the noise filtering technique and data balancing strategy employed in this study. We also elaborate on the MLLR FS method in this section. By employing the Wald test at a 95% confidence interval, the best subset of features is determined based on the selected significant features. This process involves validating the estimated coefficients to ensure their accuracy and reliability [41], [40]. Five commonly used FS techniques are evaluated in this study: Correlation Feature Selection (CFS), Chi-Square (CS), Principal Components Analysis (PCA), ReliefF (RF), and ReliefF with Weighting (RFW). Furthermore, the MLLR technique is further elaborated on in subsequent sections, along with a brief introduction to the other FS techniques used for comparison.

1) *The Maximum Likelihood Logistic Regression (MLLR) Feature selection:* The log of the odds (LN of the odds) represents a logit transformation when creating the binary LR equation, where the logit is a function of covariates in such a way that.

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \sum_{k=0}^K \beta_k x_{ik}, \quad (1)$$

Where, for any given observation in the i^{th} population, π_i denotes the probability of a malignant module, and $1 - \pi_i$ denotes the probability of a benign module. β is a column vector with a length of $K + 1$. The target of LR is to determine the unknown β of $K + 1$ parameters in Eq. 1. This is accomplished using MLE, which involves identifying the parameters that minimize the error in approximating the observed data. The likelihood function exhibits the values of β as follows for any fixed, known value of y (class variable)

$$f(\beta|y) = \prod_{i=1}^N \frac{n!}{y!(n_i - y_i)} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \quad (2)$$

where N indicates the total number of object, n is a feature vector with elements $n_i (i = 1, \dots, N)$ represents a binomial count in the i^{th} population. stands for the total number of examples in the dataset. $M = \sum_{i=1}^N n_i$. The values for β that maximize the likelihood function in Eq. 2 are known as the maximum likelihood estimates. The likelihood function's first and second derivatives must be calculated to determine the maximum likelihood estimates.

$$\frac{\delta l(\beta)}{\delta \beta_j} = \sum_{i=1}^N y_i x_i - n_i \pi_{ik} \quad (3)$$

As a result, by doing each $K + 1$ equations in Eq. 3., the maximum likelihood estimates for β_j . can be obtained. solving for each β_j and setting in Eq. 3 to be equal to zero. The estimated β_j illustrates influence of corresponding variables, in theory, to predict the outcome variable. Using the generalized linear model (GLM), which simulates the logit in the R statistical program, the β_j parameters for this study are determined. The significance of each β_j in the FS model was evaluated using the Wald test using the the Eq. [44].

$$W = \left(\frac{\beta_j}{S.E(\beta_j)} \right)^2, \quad (4)$$

where $S.E(\beta_j)$ stands for the estimated coefficient's standard error. The alternative to the null hypothesis H_0 alternative H_1 .

$$\underbrace{H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0}_{j=0,1,\dots,K} \quad (5)$$

After that, in order to identify the significant predictors of defective software modules, each Wald statistic is contrasted with a chi-square distribution with one degree of freedom. A large value of chi-squared (with a p value < 0.05) denotes a weak fit, whereas a small chi-squared (with a p -value near to 1) denotes a good fit.

2) *Correlation Feature Selection (CFS):* Based on the heuristic evaluation function [45], CFS might be used to rank features in accordance with a correlation. The evaluation function favors subsets with features that are uncorrelated with one another and strongly correlated with the class. Features are ignored if they are thought to have little relevance to the class. Features that have a strong correlation with one or more of the remaining features are filtered out as they are unnecessary. Therefore, the choice of features is solely based on how well they forecast classes in regions of the instance

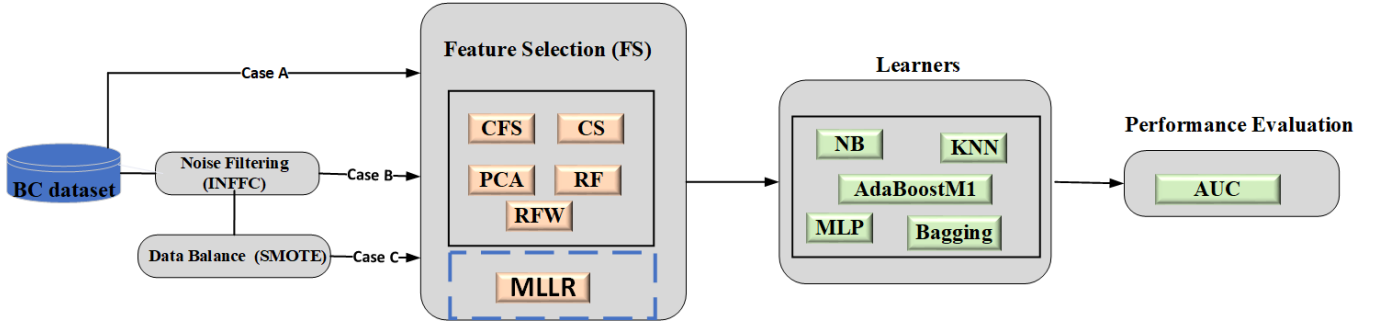


Fig. 1. The proposed scheme of enhanced feature selection method

space where no other features have done so. The evaluation function is depicted by the following;

$$M_S = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}. \quad (6)$$

Here, $\bar{r}_{cf}$ is the sum of all feature-class correlations and $\bar{r}_{ff}$ is the average correlation coefficient between features. The following is a definition of the CFS criterion:

$$CFS = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \dots + r_{cf_k}}{\sqrt{k + 2(r_{f_1f_2} + \dots + r_{f_1f_j} + \dots + r_{f_kf_1})}} \right] \quad (7)$$

Ther $r_{f_1f_j}$ and r_{cf_i} Correlations are used to describe variables. In this study, CFS is combined with the GreedyStepwise search technique

3) *Chi-Square (CS)*: A nonparametric statistical measure called Chi-Square examines the relationship amongbetween the distribution of a feature f and the distribution of the class cited given that feature [46]. Using the following formula, you can calculate it

$$CS = \sum_{i=1}^r \sum_{j=1}^{n_c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \quad (8)$$

Wher er is the number of discrete feature values, n_c indicates the number of classes, $O_{i,j}$ is the observed number of instances with the feature value i in the class j , and $E_{i,j}$ is the predicted number of examples with any value of i in the class j .

A. Principal Components Analysis (PCA)

The dimensionality of the input features can be decreased using the PCA method described in [47] PCA is a technique that generally transforms. n vectors $(x_1, x_2, \dots, x_i, \dots, x_n)$ from a d -dimensional space to n vectors $(x'_1, x'_2, \dots, x'_i, \dots, x'_n)$ in a new, d' -dimensional space.

$$x'_i = \sum_{k=1}^{d'} a_{k,i} e_k, d' \leq d, \quad (9)$$

where e_k is the eigenvector S denotes to the greatest eigenvalue of the scattering matrix d' and $a_{k,i}$ is the projection of the original vector x_i in the object vector e_k . These predictions are named principal components of the original data set. d and d' are positive integers, and the dimension of d' cannot exceed d . We use PCA in conjunction with a ranker search method in this work.

1) *ReliefF (RF)*: RF is a case-based feature ranking technique. Its advantages are that it does not rely on heuristic algorithms, runs in time with low-order polynomials, is robust to noise and interactions, and is suitable for a binary system. or continuous data [48]. RF can be calculated using the following equation.

$$RF(X) = \frac{1}{m} \sum_{i=1}^m [diff(x_i, x_M) - diff(x_i, x_H)], \quad (10)$$

Where X is a random variable that denotes an object, m is the number of samples in total, and x_i is the object value of the th example. This formula defines the $diff()$ function;

$$diff(X_1, X_2) = \frac{|x_1 - x_2|}{x_{max} - x_{min}} \quad (11)$$

where x_{min} and x_{max} stand for the minimum and maximum values of X , respectively. in this study, if we set the WeightByDistance parameter to the default value (false), it call RF algorithm or, if set to True, it refers to Relieff-w (RFW) algorithm.

B. Iterative Noise Filtering based on the Fusion of Classifiers (INFFC)

INFFC was created in a study by Saez et al. [49], considering filtering methods using several classifiers. Every iteration involves three steps. Firstly, Initial filtration is carried out using an FC-based filter that reflects on the three learners (3-NN, C4.5, and LR)). To have less of an impact on subsequent steps, this initial step reduces some of the noise already present in the current iteration. More specifically, it is anticipated that in this step, noise instances that have a high degree of confidence in their identification will be removed.

C. Synthetic malignant Oversampling Based Data Sampling

This work used the SMOTE, which creates new synthetic instances in minority classes. Instead of using data space, artificial instances are created in the feature space. The SMOTE instances are defined as $S = S + u \times (X' - X)$ with $0 \leq u \leq 1$. X' . They are linear aggregates of two comparable instances from the minority class X and X' . Among the K minority class members closest to X , X' is randomly selected. The newly constructed examples increase the fullness and generality of the minority while reducing its rarity [50].

D. Learners

In this study, we use the five most popular supervised learning techniques. NB, MLP, KNN, AdaBoostM1, and Bagging are a few of these. The different classifiers are launched using the WEKA tool. Different learning algorithms have parameter settings that guarantee the best performance. Accordingly, The following configurations are used in this paper with the classifier according to its many research. we use a kernel estimator for the NB learner , we set the useKernelEstimator parameter to true. the kNN parameter is set to 5, and the crossValidate parameter is set to true. The distanceWeighting parameter for KNN is set to Weight by 1/distance. Also modified to 10 is the validationSetSize parameter. In this way, the MLP sets aside 10% of the training data to be used as a validation set to decide the right time to end an iterative training process.

E. Performance Evaluation

Receiver Operating Characteristic (ROC) curve. The curve shows the trade-off between the false alarm rate and the detection rate. As a result, this curve displays a classifier's performance across the full spectrum of potential decision thresholds without making any assumptions about specific misclassification expenses or class previous possibilities. The AUC offers a single numerical metric for assessing model performances and is calibrated over the range of 0 to 1. Indicators with higher values indicate better model performance and vice versa. In this study, we used the Wisconsin Breast Cancer Dataset Prognostic (WPBC) [51], which is mined from the UCI machine learning repository. (151 benign tumors and 47 malignant tumors) The WPBC contains 198 records. It contains 35 variables—35 predictor variables, 34 result variables, and one result variable that indicates whether the tumor is benign or malignant. The range of the predictive attributes is M and B. The most normal state is represented by the B, while the M represents the abnormal.

IV. RESULT AND DISCUSSION

In this research, ten-fold cross-validation was conducted, involving the random division of each training dataset into approximately equal subsets (known as folds). Nine out of the ten folds were utilized for training the models, while the remaining fold was reserved for testing. This process was repeated iteratively, ensuring that each fold was eventually used for both training and testing. Weka 3.8, KEEL, MATLAB R2016a, and R statistical software were used to create and analyze the empirical results. To comprehensively assess and compare the performance of MLLR with other feature selection (FS) techniques in various scenarios, three cases were examined. These cases, namely Case A, Case B, and Case C, involved the combined utilization of FS, noise filtering, and data balancing techniques. In Case A, FS was performed on the original data, while in Case B, FS was conducted on noise-filtered data. Finally, in Case C, FS was implemented on data that underwent both noise filtering and data balancing. The thorough literature review indicates that Case A applies to most of the current FS methodologies. Five classifiers were

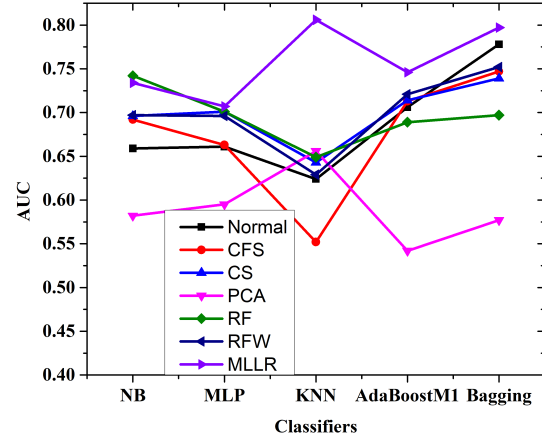


Fig. 2. Classification result for Case A using AUC

selected to assess the suggested scheme's effectiveness, where FS was carried out after noise filtering and data balance, including NB, MLP, KNN, AdaBoostM1, and Bagging. The fact that these classifiers are frequently employed in BCP studies led to their selection. AUC records the developed models classification performance to conduct a comparative analysis. However, the following sub-sections are the findings of this study.

A. Classification Performance of Case A

FS is carried out with the original, unfiltered datasets in this case. The pertinent features are then used to independently develop the models of BCP's NB, MLP, KNN, AdaBoostM1, and Bagging classification. Table II and Fig. 2 show the model's performance using the AUC metric. The MLLR records the

	NB	MLP	KNN	AdaBoostM1	Bagging
Normal	0.659	0.661	0.624	0.706	0.778
CFS	0.692	0.663	0.552	0.713	0.747
CS	0.696	0.701	0.643	0.714	0.739
PCA	0.582	0.595	0.656	0.542	0.577
RF	0.742	0.701	0.649	0.689	0.697
RFW	0.697	0.696	0.629	0.721	0.752
MLLR	0.734	0.707	0.806	0.746	0.797

TABLE II: Classification performance of Case A using AUC

second AUC with RF having a slight difference, followed by RFW, CS, and CFS, according to NB classification. The PCA is the least efficient FS technique. In the case of MLP, MLLR performs best among the FS methods, while RF and CS come in second, followed by RFW and CFS. PCA remains the least efficient FS method. PCA, RF, and CS follow MLLR in terms of AUC measurement, according to the KNN's findings. AdaBoostM1's reported performance is comparable to MLP and KNN classification, where MLLR records the highest AUC values. The least effective FS method is PCA, while RFW and CS outperform the others. The bagging classifier follows the same pattern, where MLLR performs better than

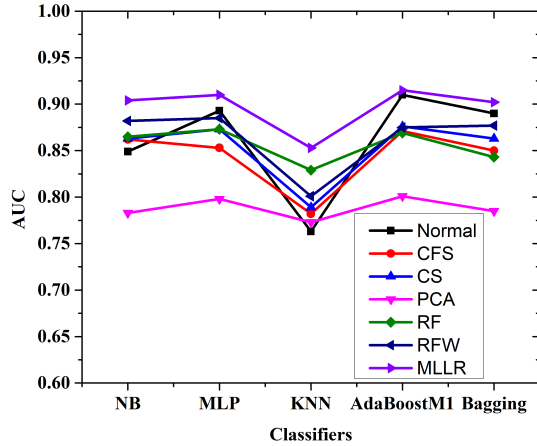


Fig. 3. Classification result for Case B using AUC

the other FS methods, then RFW, CFS, and CS. PCA remains the least effective FS method. Notably, classifiers built on the MLLR consistently outperform those built based on the original data. We also note that the feature subsets produced by the RF technique produce the best AUC results. Another thing to note is that the MLLR with KNN classifiers performs best in Case A.

B. Classification Performance of Case B

Following the noise filtering process, the FS outcomes were obtained. The classification performance of the NB, MLP, KNN, AdaBoostM1, and Bagging classifiers, developed based on the selected feature subsets, is summarized in Table III 2 and Fig. 3, the classification performance of the NB, MLP, KNN, AdaBoostM1, and Bagging classifiers developed on the chosen feature subsets is summarized (in terms of AUC) The evaluation metric used for summarizing the performance is the area under the curve (AUC). RFW, RF, CS, and CFS are the following best FS methods concerning NB, while MLLR consistently outperforms the other FS methods. The

	NB	MLP	KNN	AdaBoostM1	Bagging
Normal	0.849	0.893	0.763	0.910	0.890
CFS	0.862	0.853	0.782	0.871	0.850
CS	0.863	0.873	0.789	0.876	0.863
PCA	0.783	0.798	0.773	0.801	0.785
RF	0.865	0.873	0.829	0.869	0.843
RFW	0.882	0.885	0.801	0.875	0.877
MLLR	0.904	0.910	0.853	0.915	0.902

TABLE III: Classification performance of Case B using AUC

MLP classification results demonstrate MLLR's superior performance because it consistently outperforms the other FS methods. Compared to the other FS methods, RFW outperforms the competition, while PCA shows the worst results. KNN states that the MLLR approach's selection subset reports the highest AUC value, followed by RF, RFW, and CS. PCA is the FS technique that performs the worst. The MLLR performs

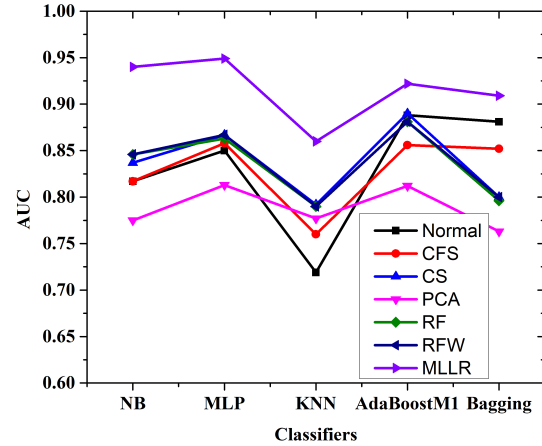


Fig. 4. Classification result for Case C using AUC

the best, followed by RF, RFW, and CS, according to the results of the AdaBoostM1 classifier, which are identical to those of KNN. PCA is another FS method that performs poorly. The best AUC was obtained using the MLLR with the Bagging classifier, followed by RFW, CS, and CFS. The MLLR technique, when compared to all other FS methods, records the best AUC measurements, which is noteworthy to note. Additionally, all MLLR selection subset-based classifiers outperform those developed using the entire feature set.

C. Classification Performance of Case C

After applying noise filtering and data balancing techniques, the six FS methods, including MLLR, were implemented. Table IV and Fig. 4 provide a comparative analysis of the classification performances (measured in terms of AUC) of the NB, MLP, KNN, AdaBoostM1, and Bagging classifiers. The AUC measure appears to perform best in MLLR, followed by RF, RFW, and CS, according to NB. PCA is the FS technique that performs the worst. The MLP results also demonstrate the best MLLR performance (RFW, CS, RF, CFS, and PCA) are the other FS in the following order.

	NB	MLP	KNN	AdaBoostM1	Bagging
Normal	0.817	0.850	0.719	0.888	0.881
CFS	0.817	0.858	0.760	0.856	0.852
CS	0.837	0.867	0.792	0.890	0.800
PCA	0.775	0.813	0.777	0.812	0.763
RF	0.846	0.863	0.790	0.881	0.796
RFW	0.846	0.867	0.790	0.881	0.800
MLLR	0.940	0.949	0.860	0.922	0.909

TABLE IV: Classification performance of Case C using AUC

Similar trends are observed in the KNN classifier, where MLLR exhibits the highest performance, followed by Chi-Square (CS), ReliefF (RF), and ReliefF with Weighting (RFW). Principal Components Analysis (PCA) performs the worst among all feature selection (FS) methods. CS, RF, and RFW show relatively better performance compared to other

Author/yea	FS method/reference	Classifier	Existing method	Proposed method
Chunekar and Ambulgekar 2009	Jordan Elman neural network [52]	SVM	70.725	99
Zhang et al. 2009	Rough set-based multiple Criteria [53]	MLP	65	98.1
Lavanya and Rani 2011	CART with feature selection (Chi-square) [54]	SVM	73.3	99
Stoean 2014	Support vector machine and evolutionary algorithm [55]	SVM	79	98.1
Sridevi and Murugan 2014	Modified Correlation Rough Set Feature Selection (MCRSFS) [56]	SVM	85	99
Aalaei et al. 2016	GA-based on feature selection and PS-classifier [57]	MLP	78.1	98.1
Lee et al 2017	Univariate Selection [34]	SVM	93	99
Naik and reddy 2018	Wrapper (Random Forest) [58]	RaF	96	97
Dua et al. 2019	Lasso Regression [59]	SVM	94	99
Sharma et al. 2020	Hybrid (Information Gain + Genetic Algorithm) [60]	SVM	97	99
Zhang et al. 2021	Improved Genetic Algorithm [61]	SVM	98	99
Abbas et al. 2018	Chi-Square [62]	SVM	92	99
Smith et al. 2020	Ridge Regression [63]	NB	95	97.3
Ji et al. 2020	IBPSO with k-NN [64]	SVM	96.14	99
Salima et al. 2020	ECSCA with k-NN [65]	SVM	95.76	99
Patel and mehta 2021	Wrapper (Bagging, Boosting) [66]	Bagging, Boosting	96,97	97.5, 97.6
Sahu et al. 2022	L1 regularization with SVM [67]	SVM	97	99
Johnson and kim 2022	Hybrid (CFS + Genetic Algorithm) [68]	SVM	97	99
Christo et al. 2022	Cooperative coevolution and RF [69]	SVM	97.1	99

TABLE V: Accuracy performance comparison of our method with other approaches from existing research on WPBC dataset

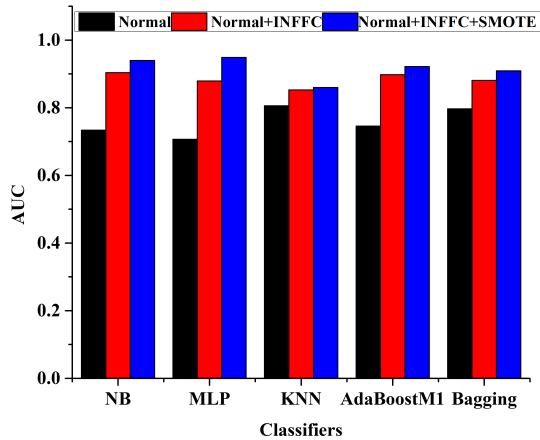


Fig. 5. Classification result for MLLR over all Cases

FS methods, but MLLR still outperforms them all, according to the AdaBoostM1 classifier. PCA continues to exhibit the lowest performance. Similarly, when using the Bagging classifier, MLLR outperforms Correlation Feature Selection (CFS), CS, and RFW. It is worth mentioning that the feature subset selected by MLLR demonstrates superior performance compared to the original dataset where the full feature set was utilized. Among all the FS techniques, MLLR consistently attains the highest AUC scores. More importantly, We can notice that the AUCs achieved by FS with balanced data by SMOTE are greater than the values accomplished when the data is imSignificantly, it is observed that the AUCs achieved through FS with balanced data using SMOTE surpass the values obtained when the data is imbalanced and even outperform the AUC obtained from the original data, as illustrated in Fig.

5. Multiple studies, such as [50], [19], [24], have reported on the effectiveness of SMOTE in achieving data balance. The performance of MLLR further strengthens its applicability for oversampling and FS in the BCP domain, as evident from our findings. Based on the aforementioned results, the treatment of data imbalance directly affects FS performance. The utilization of the MLLR FS method ensures improved classification performance for the BCP model. Notably, among the observed classification models, MLP demonstrates superior performance compared to the others.

D. Comparison of the proposed method with existing methods

To compare our method with existing methods that used the Breast Cancer Wisconsin Prognostic dataset. We used the Accuracy performance metrics and perform another to the classification method, namely Support Vector Machine (SVM) and Random Forest (RaF) in order to consistent with their works. Each study chose a classification method that best complemented their feature selection technique. As presented in Table V, our proposed method outperforms all the existing methods compared in this study.

V. CONCLUSION AND FUTURE WORKS

FS is critical for developing accurate and efficient models for BCP. FS helps to reduce model complexity, minimize overfitting, and enhance the ability of models to detect cancerous modules in out-of-sample data. However, most existing FS methods struggle to identify relevant features for BCP due to noisy samples and unbalanced training data present in breast cancer datasets. To address this challenge, this study proposed a method that performs FS using original data, followed by noise filtering and data balancing to construct a BCP model with the most pertinent features. The study examined three distinct cases involving various FS methods based on original,

filtered, and filtered and balanced data. The NB, MLP, KNN, AdaBoostM1, and Bagging classifiers were trained on the selected feature subsets, and their performance was evaluated using the AUC metric through 10-fold cross-validation. The empirical findings indicated that FS plays a crucial role in identifying valuable features to enhance the performance validation of BCP models after removing noise and balancing the data. Comparatively, the filtered and sampled data demonstrated significantly improved performance compared to FS based on original and noise-filtered data. Moreover, MLLR was identified as the most effective method for BCP, suggesting its potential for enhancing cancer detection and prediction. This study underscores the importance of data quality in selecting appropriate feature selection techniques. Furthermore, our proposed method could apply in the BDP problem domains and any another real-world applications.

In future research, there will be an exploration of additional feature selection (FS) techniques and data sampling methods to further enhance the performance of BCP models within the scope of this study. Furthermore, the combination of deep learning algorithms and optimization techniques has the potential to significantly improve the accuracy of breast cancer diagnosis. By integrating these computational approaches, it becomes possible to capture more intricate relationships within BCP datasets and uncover hidden patterns that may remain undetected by conventional statistical methods. This method could also enable the development of personalized treatment recommendations for patients based on their characteristics and disease profiles. Therefore, future studies should investigate the potential of deep learning and optimization methods for improving the accuracy and efficiency of breast cancer diagnosis and management.

REFERENCES

- [1] "Who breast cancer statistic," (<https://www.who.int/cancer/prevention/diagnosis-screening/breastcancer>, accessed: 3 May 2022).
- [2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [3] R. Kumar, A. Sharma, and R. K. Tiwari, "Application of microarray in breast cancer: An overview," *Journal of Pharmacy and Bioallied Sciences*, vol. 4, no. 1, pp. 21–26, 2012.
- [4] S. Hartmann, T. Reimer, and B. Gerber, "Management of early invasive breast cancer in very young women (< 35 years)," *Clinical breast cancer*, vol. 11, no. 4, pp. 196–203, 2011.
- [5] C. Brekelmans, C. Seynaeve, M. Menke-Pluymers, H. Brüggewirth, M. Tilanus-Linthorst, C. Bartels, M. Kriege, A. van Geel, C. Crepin, J. Blom *et al.*, "Survival and prognostic factors in brca1-associated breast cancer," *Annals of oncology*, vol. 17, no. 3, pp. 391–400, 2006.
- [6] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [7] L. Yang, D. M. Parkin, J. Ferlay, L. Li, and Y. Chen, "Estimates of cancer incidence in china for 2000 and projections for 2005," *Cancer Epidemiology Biomarkers & Prevention*, vol. 14, no. 1, pp. 243–250, 2005.
- [8] L. L. Lipscombe, P. J. Goodwin, B. Zinman, J. R. McLaughlin, and J. E. Hux, "The impact of diabetes on survival following breast cancer," *Breast cancer research and treatment*, vol. 109, pp. 389–395, 2008.
- [9] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.
- [10] T. Subashini, V. Ramalingam, and S. Palanivel, "Breast mass classification based on cytological patterns using rbfn and svm," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5284–5290, 2009.
- [11] S. A. Pavlopoulos and A. N. Delopoulos, "Designing and implementing the transition to a fully digital hospital," *IEEE Transactions on information technology in biomedicine*, vol. 3, no. 1, pp. 6–19, 1999.
- [12] P. Meesad and G. G. Yen, "Combined numerical and linguistic knowledge representation and its application to medical diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 33, no. 2, pp. 206–222, 2003.
- [13] B. Gayathri, C. Sumathi, and T. Santhanam, "Breast cancer diagnosis using machine learning algorithms-a survey," *International Journal of Distributed and Parallel Systems*, vol. 4, no. 3, p. 105, 2013.
- [14] U. Jain, S. Pirasteh, and M. Hussain, "Lightweight, secure, efficient, and dynamic scheme for mutual authentication of devices in internet-of-things-fog environment," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 1, p. e7428, 2023.
- [15] S. Badrloo, M. Varshosaz, S. Pirasteh, and J. Li, "Image-based obstacle detection methods for the safe navigation of unmanned vehicles: A review," *Remote Sensing*, vol. 14, no. 15, p. 3824, 2022.
- [16] V. Mousavi, M. Varshosaz, F. Remondino, S. Pirasteh, and J. Li, "A two-step descriptor-based keypoint filtering algorithm for robust image matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–21, 2022.
- [17] R. Rabiei, S. M. Ayyoubzadeh, S. Sohrabei, M. Esmaeili, and A. Atashi, "Prediction of breast cancer using machine learning approaches," *Journal of biomedical physics & engineering*, vol. 12, no. 3, p. 297, 2022.
- [18] K. K. Gupta, R. Vijay, P. Pahadiya, S. Saxena, and M. Gupta, "Novel feature selection using machine learning algorithm for breast cancer screening of thermography images," *Wireless Personal Communications*, vol. 131, no. 3, pp. 1929–1956, 2023.
- [19] K. Bashir, T. Li, C. W. Yohannese, and Y. Mahama, "Enhancing software defect prediction using supervised-learning based framework," in *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE, 2017, pp. 1–6.
- [20] K. Bashir, T. Li, and C. W. Yohannese, "An empirical study for enhanced software defect prediction using a learning-based framework," *International Journal of Computational Intelligence Systems*, vol. 12, no. 1, pp. 282–298, 2018.
- [21] I. Chakraborty, S. Edirippulige, and P. V. Ilavarasan, "The role of telehealth startups in healthcare service delivery: a systematic review," *International Journal of Medical Informatics*, p. 105048, 2023.
- [22] X. Zhu, X. Wu, and Q. Chen, "Eliminating class noise in large datasets," in *Proceedings of the 20th international conference on machine learning (icml-03)*, 2003, pp. 920–927.
- [23] K. Bashir, T. Li, and M. Yahaya, "A novel feature selection method based on maximum likelihood logistic regression for imbalanced learning in software defect prediction," *Int. Arab J. Inf. Technol.*, vol. 17, no. 5, pp. 721–730, 2020.
- [24] K. Bashir, T. Li, C. W. Yohannese, and M. Yahaya, "Smotefris-inffcc: Handling the challenge of borderline and noisy examples in imbalanced learning for software defect prediction," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 1, pp. 917–933, 2020.
- [25] C. M. Lynch, B. Abdollahi, J. D. Fuqua, A. R. De Carlo, J. A. Bartholomai, R. N. Balgeman, V. H. van Berkel, and H. B. Frieboes, "Prediction of lung cancer patient survival via supervised machine learning classification techniques," *International journal of medical informatics*, vol. 108, pp. 1–8, 2017.
- [26] S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: Deep learning approaches," *International Journal of Medical Informatics*, vol. 148, p. 104371, 2021.
- [27] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.
- [28] S. Geeitha and M. Thangamani, "Incorporating ebo-hsic with svm for gene selection associated with cervical cancer classification," *Journal of medical systems*, vol. 42, no. 11, p. 225, 2018.
- [29] V. N. Gopal, F. Al-Turjman, R. Kumar, L. Anand, and M. Rajesh, "Feature selection and classification in breast cancer prediction using iot and machine learning," *Measurement*, vol. 178, p. 109442, 2021.
- [30] S. Dasgupta, R. Chaudhuri, and S. Purushotham, "Feature selection for breast cancer detection using machine learning algorithms," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9, pp. 2080–2083, 2019.
- [31] M. Vasudev, A. Doegar, and V. Gupta, "A review on automatic classification of breast cancer using supervised learning strategies," in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*. IEEE, 2019, pp. 717–721.

- [32] W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis," *Designs*, vol. 2, no. 2, p. 13, 2018.
- [33] D. A. Omondigbe, S. Veeramani, and A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," in *IOP conference series: materials science and engineering*, vol. 495. IOP Publishing, 2019, p. 012033.
- [34] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl Comput Math*, vol. 7, no. 4, pp. 212–216, 2018.
- [35] H. Hajiabadi, V. Babaiyan, D. Zabihzadeh, and M. Hajiabadi, "Combination of loss functions for robust breast cancer prediction," *Computers & Electrical Engineering*, vol. 84, p. 106624, 2020.
- [36] C. Shravya, K. Pravalika, and S. Subhani, "Prediction of breast cancer using supervised machine learning techniques," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6, pp. 1106–1110, 2019.
- [37] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.
- [38] R. Aavula and R. Bhramaramba, "Xbpfr: an extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 5, pp. 2249–8958, 2019.
- [39] V. Nandagopal, S. Geeitha, K. V. Kumar, and J. Anbarasi, "Feasible analysis of gene expression—a computational based classification for breast cancer," *Measurement*, vol. 140, pp. 120–125, 2019.
- [40] M. Abdar, M. Zomorodi-Moghadam, X. Zhou, R. Gururajan, X. Tao, P. D. Barua, and R. Gururajan, "A new nested ensemble technique for automated diagnosis of breast cancer," *Pattern Recognition Letters*, vol. 132, pp. 123–131, 2020.
- [41] L. Wang, "Microwave sensors for breast cancer detection," *Sensors*, vol. 18, no. 2, p. 655, 2018.
- [42] K. Bashir, T. Ali, M. Yahaya, and A. S. Hussein, "A hybrid data pre-processing technique based on maximum likelihood logistic regression with filtering for enhancing software defect prediction," in *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. IEEE, 2019, pp. 921–927.
- [43] S. A. Czepiel, "Maximum likelihood estimation of logistic regression models: theory and implementation," *Available at czep. net/stat/mlelr.pdf*, vol. 83, 2002.
- [44] A. A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*, 2003.
- [45] M. A. Hall, "Correlation-based feature selection for machine learning, Ph.D. dissertation," 1999.
- [46] H. O. Lancaster and E. Seneta, "Chi-square distribution," *Encyclopedia of biostatistics*, vol. 2, 2005.
- [47] A. Malhi and R. X. Gao, "Pca-based feature selection scheme for machine defect classification," *IEEE transactions on instrumentation and measurement*, vol. 53, no. 6, pp. 1517–1525, 2004.
- [48] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.
- [49] J. A. Sáez, M. Galar, J. Luengo, and F. Herrera, "Infcc: An iterative class noise filter based on the fusion of classifiers with noise sensitivity control," *Information Fusion*, vol. 27, pp. 19–32, 2016.
- [50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [51] S. W. Wolberg, William and O. Mangasarian, "Breast Cancer Wisconsin (Prognostic)," UCI Machine Learning Repository, 1995, DOI: <https://doi.org/10.24432/C5GK50>.
- [52] V. N. Chuneekar and H. P. Ambulgekar, "Approach of neural network to diagnose breast cancer on three different data set," in *2009 International Conference on Advances in Recent Technologies in Communication and Computing*. IEEE, 2009, pp. 893–895.
- [53] Z. Zhang, Y. Shi, and G. Gao, "A rough set-based multiple criteria linear programming approach for the medical diagnosis and prognosis," *Expert Systems with Applications*, vol. 36, no. 5, pp. 8932–8937, 2009.
- [54] D. Lavanya and D. K. U. Rani, "Analysis of feature selection with classification: Breast cancer datasets," *Indian Journal of Computer Science and Engineering (IJCSE)*, vol. 2, no. 5, pp. 756–763, 2011.
- [55] C. Stoean and R. Stoean, "Support vector machines and evolutionary algorithms for classification," *Single or Together*, 2014.
- [56] T. Sridevi and A. Murugan, "A novel feature selection method for effective breast cancer diagnosis and prognosis," *International Journal of Computer Applications*, vol. 88, no. 11, 2014.
- [57] S. Aalaei, H. Shahraki, A. Rowhanimanesh, and S. Eslami, "Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets," *Iranian journal of basic medical sciences*, vol. 19, no. 5, p. 476, 2016.
- [58] R. Naik, "Feature selection and classification of breast cancer data using random forest algorithm," *International Journal of Computer Applications*, vol. 162, no. 1, pp. 1–4, 2018.
- [59] S. D. S. V. T. M. Dua, M., "Comparative study of feature selection methods for breast cancer diagnosis," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5, pp. 662–666, 2019.
- [60] —, "A hybrid feature selection method based on information gain and genetic algorithm for breast cancer diagnosis," *Journal of Computational Science*, vol. 40, no. 10, p. 1079, 2020.
- [61] T. Z. J. G. J. Zhang, H., "An improved feature selection algorithm for breast cancer prognosis," *Journal of Medical Systems*, vol. 45, no. 59, p. 0175, 2021.
- [62] A. M. K. M. A. Abbas, N., "Performance comparison of feature selection methods for breast cancer diagnosis," in *Proceedings of the 2018 6th International Conference on Information and Communication Technology*, 2018, pp. 23–27.
- [63] W. J. C. Y. Smith, A., "Using lasso and ridge regression for breast cancer prognostic classification," *BioMed Research International*, vol. 2020, pp. 1–10, 2020.
- [64] B. Ji, X. Lu, G. Sun, W. Zhang, J. Li, and Y. Xiao, "Bio-inspired feature selection: An improved binary particle swarm optimization approach," *IEEE Access*, vol. 8, pp. 85 989–86 002, 2020.
- [65] S. Ouafeel and M. Abd Elaziz, "Enhanced crow search algorithm for feature selection," *Expert Systems with Applications*, vol. 159, p. 113572, 2020.
- [66] M. P. Patel, H., "A comparative study of ensemble feature selection methods for breast cancer prognosis," *Computers in Biology and Medicine*, vol. 134, p. 104448, 2021.
- [67] P. Sahu, P. Giri, R. Sunkara, and R. Sundararajan, "Extraction of key features and enhanced prediction framework of breast cancer occurrence," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2022, pp. 1679–1685.
- [68] K. S. Johnson, D., "Hybrid feature selection using correlation-based method and genetic algorithm for breast cancer prediction," *Expert Systems with Applications*, vol. 188, no. 3, p. 116009, 2022.
- [69] V. E. Christo, H. K. Nehemiah, J. Brightly, and A. Kannan, "Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest," *IETE Journal of Research*, vol. 68, no. 4, pp. 2508–2521, 2022.



Kamal Bashir is an Assistant Professor in the College of Information Science and Technology, at Karary University he got his B.S. degree in Computer Science from Karary University, Sudan, in 2009. MSc. degree in Computer Science from Khartoum University, Sudan, in 2013. PhD. degree in Computer Science and Technology from Southwest Jiaotong University, China in 2020. His area of research interests includes data Mining, Machine Learning, and Software Quality Assessment.



Saied Pirasteh Saied Pirasteh (Senior Member, IEEE) is a Geomatics and Geospatial Artificial Intelligence professor and the Associate Dean at the Institute of Artificial Intelligence, Shaoxing University. He is also a visiting professor and scientist at the Geospatial Intelligence and Mapping Lab, University of Waterloo. He received his first Ph.D. degree in Remote Sensing and GIS (Geology) in 2004. He also received a Ph.D. degree in geomatics and GeoAI (Geography) from the University of Waterloo in 2018.



Hassan Abdelrhman is an Assistant Professor in the College of Information Science and Technology, at Karary University, got his B.S. degree in Computer Science from Karary University, Sudan, in 2009, MSc. degree in Computer Architecture & Networking from Khartoum University, Sudan, in 2017, PhD. degree in Computer Science and Information Technology from Omdurman Islamic University, Sudan in 2021. His area of research interests includes data Mining, Machine Learning, and Network Engineering, Computer Vision.



Mohamed Mosadag is an Assistant Professor in the College of Information Science and Technology, at Karary University, got his B.S. degree in Computer Science from Karary University, Sudan, in 2010, MSc. degree in Computer Science from Karary University, Sudan, in 2016, PhD. degree in Computer Science and Technology, from Dongla University, Sudan, in 2024. His area of research interests Includes Machine Learning, Data Base.



Ashraf Mohammed Ashraf Mohammed is an a lecturer in College of Computer Science and Information Technology, Karary University, got his B.S. degree in Computer Science from Karary University, Sudan, in 2010, MSc. degree in Information Technology from Elnileen University, Sudan, in 2019. His area of research interests Includes Data Mining, Machine Learning, Data Base.