International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

IJISAE

ISSN:2147-6799 www.ijisae.org Original Research Paper

Breast Cancer Detection Using Random Forest Supported by Feature Selection

[1]Mustafa Ali Hasan Dalfi, [2]Sihem Chaabouni, [3]Ahmed Fakhfakh

Submitted: 17/07/2023 **Revised**: 07/09/2023 **Accepted**: 24/09/2023

Abstract: Breast cancer has been responsible for the loss of approximately 1.5 million lives over the past 35 years, despite considerable investments in mammography-based detection and treatment. This persistently high death rate underscores the urgency for improved strategies. Research consistently emphasizes the significance of detecting cancer at its early stages, ideally when the tumor size is confined to a modest 5-10 millimeters, thus minimizing the need for invasive procedures such as intensive chemotherapy or radiation. However, the current primary detection methods often fall short in identifying these small, elusive tumors, particularly when they are nestled within dense breast tissue. Consequently, there is a pressing need for more efficient screening techniques. In this study, we propose an innovative machine learning based methodology for Breast Cancer Detection that employs the Feature Selection-Aided Random Forest Algorithm. The research framework incorporates advanced feature selection techniques, such as Variance Inflation Factor (VIF), Model-based Feature Selection, Recursive Feature Elimination, and Univariate Feature Selection, to extract highly relevant features and uncover hidden patterns associated with tumors. Experimental results demonstrate the remarkable effectiveness of this approach, with feature selection facilitated by the Variance Inflation Factor (VIF) algorithm achieving 98.83% accuracy when evaluated on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. This approach effectively identifies the most appropriate features, significantly enhancing the breast cancer detection system's performance.

Index Terms *- Breast Cancer, Early Detection, Feature Selection, Mammography, Random Forests*

1. Introduction

Breast cancer, second only to skin cancers, stands as the recurrent prevalent form of cancer among women in the United States. It presents a notably significant threat to women aged 35 to 54, as evident from the distressing statistic that approximately 1.5 million American women have lost their lives to this disease over the past 35 years. This somber reality translates to a distressing ratio of approximately 1 in 39 women, or about 2.6%, succumbing to breast cancer. [1, 2]. Notwithstanding the substantial financial investments allocated to mammography, which serves as the principal modality for the detection of breast cancer, and the extensive endeavors dedicated to screening and treatment, there has been a dearth of notable reduction in the mortality rate associated with this malignancy over the preceding 15-year period.

This underscores the pressing necessity for the development and implementation of enhanced methodologies for both detection and treatment. Extensive research substantiates that the most optimal opportunity to mitigate these fatalities lies in the realm of early detection, particularly when the tumors manifest as diminutive entities, typically measuring between 5 to 10 millimeters in size [3]. A Dutch study [4] supports this approach, revealing that almost no women died from breast cancer when it was removed at this small size. Furthermore, these smaller cancers usually don't necessitate intense chemotherapy or radiation, as they rarely reach the lymph nodes.

Diagnostics of breast cancer typically entails the utilization of a diverse range of advanced diagnostic instruments, encompassing but not limited to mammography, Breast Ultrasound (BUS), and tomography. Moreover, in instances involving intricate scenarios, the inclusion of Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans is not uncommon, while Positron Emission Tomography (PET) scans are occasionally warranted to facilitate a more comprehensive investigation [5]. Due to the heightened susceptibility of the breast as a complex organ within the human body, the choice of diagnostic techniques is intricately determined by patient-specific circumstances and the characteristics of the tumor. Although mammography is commonly regarded as the preferred method for early-stage breast cancer detection due to its cost-effectiveness and safety profile, it regrettably demonstrates limited efficacy when confronted with the challenge of dense breast tissue, particularly prevalent among younger women.

Nevertheless, when it comes to detecting small-sized tumors, approximately half of them go undetected by these conventional approaches, particularly among women with dense breast tissue. This emphasizes the urgent requirement for more efficient screening modalities, specifically tailored to address the needs of women with dense breasts. Machine learning strategies, purposefully crafted to precisely identify

[1]*ATISP research unit National School of Electronics and Telecommunications of Sfax (ENET'Com) University, Sfax, Tunisia*

[2]*LAB. SM@RTS, CENTRE DE RECHERCHE EN NUM´ERIQUE DE SFAX, TUNISIA ENET'COM, Sfax University, Technopole de Sfax, 3021 Sfax, Tunisia*

[3]*LAB. SM@RTS, CENTRE DE RECHERCHE EN NUM´ERIQUE DE SFAX, TUNISIA*

minute cancerous lesions within dense breast tissue, harbor the potential to revolutionize the early detection and treatment of this ailment, thereby substantially enhancing survival rates. However, despite the evident efficacy of this approach, its integration into routine medical practice has regrettably been met with sluggish acceptance.

Within the confines of this study, we present an inventive methodology for the detection of Breast Cancer, which seamlessly integrates the Random Forest Algorithm with various feature selection techniques aimed at extracting remarkably pertinent features and unveiling concealed patterns intertwined with tumors. Our proposed methodology deftly employs random forests to skillfully train the model, using the meticulously chosen features. In pursuit of enhancing the precision of breast cancer detection, this investigation employs a diverse array of feature selection techniques. Most notably, the Variance inflation factor yields a resounding accuracy rate of 98.83% when employed with a subset of 25 features. Similarly, the Univariate feature selection technique showcases its prowess by attaining an accuracy of 95.32% while utilizing a reduced set of 5 features. Furthermore, Recursive feature elimination emerges as a contender, delivering a commendable accuracy of 95.91% using an identical number of features. The Model-based feature selection methodology, adopting a judicious selection of 9 features, procures an accuracy of 97.08%. Additionally, by skillfully incorporating Principal component analysis (PCA) with a mere 4 principal components, an accuracy rate of 97.08% is successfully attained.

I. Our Contribution

Our contribution focuses on the utilization of feature selection and dimensionality reduction techniques in the context of BC Detection. By exploring the Wisconsin breast cancer dataset, consisting of 569 observations, we aimed to train a machine learning model capable of accurately classifying breast cancer cases as benign or malignant. In this study, we employed the random forest classifier as our chosen model. The techniques employed include Variance Inflation Factor (VIF), Model-based Feature Selection, Recursive Feature Elimination, Univariate Feature Selection and Principal component analysis. We assessed the effectiveness of each technique by evaluating the accuracy of our model's predictions using the confusion matrix. Table I underscores a selection of some of the seminal contributions that have effectively employed contemporary machine learning and deep learning strategies to identify salient features for breast cancer detection.

TABLE I

A Brief Comparison Of Our Work With Previous Works

Ref	Year	Contribution
[6]	2022	The authors have ingeniously formulated a novel ensemble-based machine learning framework titled Meta-Health Stack, aimed at proficiently predicting breast cancer. By employing the Extra Trees classifier to amalgamate features and incorporating a blend of Boosting, Bagging, and Voting techniques, the framework has demonstrated remarkable performance. It has achieved an impressive F1-score of 97% and a precision rate of 98% when evaluated on the Wisconsin Diagnostic Breast Cancer dataset. Encouragingly, the authors advocate for the utilization of their framework in the early-stage diagnosis of breast cancer, supported by its effectiveness demonstrated on three additional medical datasets.
[7]	2021	The study centers its attention on breast cancer, employing a diverse range of feature selection methodologies and implementing Random Forest Classification in the model. The authors present affirmative findings that highlight the immense potential of this approach in the realm

		of breast cancer detection. The primary objective of the research is to ascertain the pivotal features associated with breast cancer, offering valuable insights into the intricate techniques and mathematical foundations underpinning machine learning algorithms for this specific purpose.
[8]	2021	The researchers have innovated a novel technique for the detection of breast cancer, leveraging the power of machine learning algorithms in conjunction with clinical data. The devised model, which combines the Relief algorithm with the Support Vector Machine, surpasses previous methodologies and alternative feature selection approaches, attaining a remarkable accuracy rate of 99.91%. Rigorously validated through k-fold cross-validation, it consistently exhibits superior performance when compared to baseline methods. This breakthrough holds significant promise for enhancing the accuracy and efficacy of breast cancer detection, thereby offering new possibilities for early diagnosis and treatment.
[9]	2021	The researchers have pioneered the development of an advanced model, termed Hierarchical Clustering Random Forest (HCRF), aimed at augmenting the conventional Random Forest technique for the purpose of breast cancer detection. By integrating the HCRF model with an efficient Variable Importance Measure (VIM) approach to select optimal features, the study yielded exceptional results. Notably, the HCRF model demonstrated superior accuracy rates of 97.05% and 97.76% when evaluated on the Wisconsin Diagnosis Breast Cancer and Wisconsin Breast Cancer databases, respectively, surpassing the performance of traditional Decision Tree, Adaboost, and Random Forest methods. This breakthrough presents a significant leap forward in enhancing the precision and reliability of breast cancer detection methodologies.
[10]	2020	This study presents a comprehensive examination of the utilization of machine learning techniques for the prediction of breast cancer, meticulously exploring prior research endeavors that have explored an array of algorithms. Notably, Support Vector Machine (SVM) consistently emerges as a formidable contender, renowned for its remarkable accuracy. Building upon these foundations, the

		authors proceed to introduce their own investigation, which encompasses an intricate analysis and comparison of eight distinct machine learning methods. Employing the esteemed Breast Cancer Wisconsin dataset, this study embarks on a quest to unravel the efficacy and potential of these methodologies in the realm of breast cancer prediction.
Ours	2023	This study sets itself apart by meticulously integrating the Random Forest algorithm with a wide array of feature selection techniques for breast cancer detection. In contrast to earlier studies that employed limited method combinations, this research demonstrates the flexibility and efficiency of combining Random Forest with different feature selection methods. Specifically, the Variance Inflation Factor achieved a 98.83% accuracy using 25 features. The Univariate Feature Selection attained 95.32% accuracy with only five features. Recursive Feature Elimination delivered 95.91% accuracy with five features. Model-based Feature Selection achieved 97.08% accuracy with nine features. Principal Component Analysis reached 97.08% accuracy with only four components. By thoroughly evaluating diverse feature selection techniques, this study contributes a finely tuned and adaptable approach to breast cancer detection. The exhaustive feature selection process uncovers optimized model configurations that enhance precision, such as the Variance Inflation Factor's 98.83% accuracy. This study fills a gap by emphasizing feature selection diversity and adaptability within the Random Forest method for more refined breast cancer detection..

2. Literature Review

The application of machine learning techniques, particularly the utilization of machine and deep learning architectures, has revolutionized the field of life sciences, offering promising outcomes. In the context of breast cancer research, various machine learning approaches, including, decision trees [11], Artificial neural networks [12], K-nearest neighbors (KNN) [13], support vector machines (SVM) [14], and ensemble classifiers etc., have been extensively employed to train and evaluate features for the accurate classification of objects into malignant or benign classes

Building upon the preceding examination of different supervised machine learning algorithms for breast cancer detection, the existing body of literature also offers a wealth of comprehensive information on the utilization of random forest in conjunction with feature selection methods. In this section, we delve into a detailed discussion of significant studies conducted within the realm of breast cancer detection, with a specific focus on the application of random forests and feature selection techniques.

In their breast cancer detection study [15], the authors implemented a comprehensive feature selection approach. In the first stage, they addressed multicollinearity by removing features with a correlation coefficient exceeding 0.8, resulting in a set of 16 features. For the second stage, they utilized Recursive Feature Elimination, Logistic Regression, and Univariate Selection methods to identify the top eight features. To mitigate method biases, features were chosen if they ranked in the top eight in at least two of the methods. Building upon this selection process, the authors employed the Random Forest algorithm as their primary model for classifying breast tumors. The Random Forest models, trained on the two subsets of features, achieved impressive accuracies of 100% and 99.30% respectively. Comparison with other classification algorithms confirmed the superiority of Random Forest for breast cancer diagnosis. Thus, their study highlights the effectiveness of feature selection methods and the superior performance of the Random Forest algorithm in breast cancer detection.

Another study [16] employed a two-phased approach for feature selection in BC detection. In the first phase, they used the learning algorithm RF to select the best features based on Bayesian probability and feature impurity. Backward elimination was implemented to evaluate the influence of each feature, resulting in a set of selected features. In the second phase, only the selected features were used to train the classifier, improving classification accuracy. The proposed method involved a four-step classification algorithm, including n-fold cross validation, estimation of Bayesian probability, feature ranking, and backward elimination. The algorithm aimed to find an optimal feature subset to enhance classification accuracy.

Primary objective of [7] was to examine the application of feature selection techniques in BC detection using the WDBC. The researchers sought to identify the most informative features that could effectively differentiate between benign and malignant cases. To achieve this, they employed various feature selection methods, including correlation analysis, selectKBest, and Recursive Feature Elimination (RFE) with Random Forest. These techniques proved instrumental in identifying the most significant features for BC classification. The study's findings demonstrated the efficacy of feature selection in enhancing breast cancer detection. Utilizing correlation analysis, an impressive accuracy of 95.32% was achieved. Implementing the selectKBest method with a set of five features resulted in an accuracy of 94.15%. Remarkably, the RFE method also identified the same top five features as the selectKBest method, confirming their crucial role in breast cancer classification.

Similarly, [17] aimed to investigate a range of feature selection and machine learning techniques for BC detection. The authors employed a genetic algorithm-based approach to identify the most relevant attributes. Various data mining techniques, including Random Forest (RF) Logistic Regression, Decision Trees, and Rotation Forest, were applied for classification purposes.

The results showcased the superior performance of the Random Forest algorithm, achieving remarkable accuracies of 100% and 99.30% on distinct subsets of the dataset. Comparative analysis with four other classification algorithms consistently demonstrated the superiority of Random Forest in breast cancer diagnosis. Furthermore, correlation analysis, selectKBest, and RFE methods were utilized to select the most informative features. This investigation emphasized the critical role of feature selection in enhancing the accuracy of breast cancer detection.

The results showcased the superior performance of the Random Forest algorithm, achieving remarkable accuracies of 100% and 99.30% on distinct subsets of the dataset. Comparative analysis with four other classification algorithms consistently demonstrated the superiority of Random Forest in breast cancer diagnosis. Furthermore, correlation analysis, selectKBest, and RFE methods were utilized to select the most informative features. This investigation emphasized the critical role of feature selection in enhancing the accuracy of breast cancer detection.

Authors et al. [9] aimed to develop an accurate BC detection model while deploying ML techniques. To address the limitations of decision trees, HCRF model was introduced, which selected representative trees with low similarity and high accuracy. The Variable Importance Measure (VIM) method optimized the selected features for BC prediction. WDBC and Wisconsin Breast Cancer (WBC) databases were utilized for evaluation. The proposed HCRF algorithm with VIM achieved the highest accuracy of 97.05% and 97.76% on the WDBC and WBC datasets, respectively, outperforming AdaBoost, Decision Tree, and Random Forest algorithms. This study provides an effective tool for accurate breast cancer diagnosis. Table II provides a concise overview of the research studies conducted on breast cancer detection using random forests and feature selection techniques, highlighting their key findings and contributions.

The WDBC dataset is more widely used than WBCD for breast cancer classification problems in machine learning. WDBC has 569 samples with 30 real-valued features computed from digitized biopsy images. This helps benchmark algorithms effectively. WBCD has 699 samples but only 9 categorical features derived from WDBC's richer feature set. As a result, WBCD loses important nuanced feature information present in WDBC. While both classify breast cancer biopsies into benign and malignant, WDBC provides a more detailed, authoritative representation.

TABLE II

A CONCISE OVERVIEW OF THE RESEARCH STUDIES CONDUCTED ON BREAST CANCER DETECTION USING RANDOM FORESTS AND FEATURE SELECTION TECHNIQUES

Ref	Methodology	Dataset	Results
[7]	The study employed feature selection techniques, including correlation analysis, selectKBest, and RFE with Random Forest, to identify informative features for BC detection using the WDBC dataset.	WBCD	The study achieved high accuracy in breast cancer detection, with 95.32% accuracy using correlation analysis and 94.15% accuracy with the selectKBest method. The RFE method identified the same top five features as the selectKBest method, further validating their importance in breast cancer classification.
[9]	Study introduced a Hierarchical Clustering Random Forest (HCRF) algorithm to improve BC detection. The model selected representative trees with low similarity and high accuracy. The Variable Importance Measure (VIM) method optimized the selected features.	WDBC and WBC	The proposed HCRF algorithm with VIM achieved the highest accuracy of 97.05% and 97.76% on the WDBC and WBC datasets, respectively.
[17]	The study utilized a genetic algorithm-based approach for feature selection and employed various data mining techniques, including Random Forest (RF) Logistic Regression, Decision Trees, and Rotation Forest, for breast cancer classification.	WBCD	The Random Forest algorithm showcased exceptional performance, achieving accuracies of 100% and 99.30% on different subsets of the dataset.
[15]	The authors employed a comprehensive feature selection approach, consisting of addressing multicollinearity by removing highly correlated features and utilizing Recursive Feature Elimination, Logistic Regression, and Univariate Selection methods to identify the top eight features.	WBCD	The Random Forest models trained on two subsets of features achieved impressive accuracies of 100% and 99.30% respectively
[16]	The study used a two-phase approach for feature selection in breast cancer detection, utilizing the RF learning algorithm with Bayesian probability and feature impurity for selecting the best features. Backward elimination was applied to evaluate feature contribution, resulting in a selected feature set. A four-step classification algorithm involving n-fold cross-validation, Bayesian probability estimation, feature ranking, and backward elimination was employed.	WBCD	Classification Accuracy 100%

Researchers can gain more insights from the refined attributes in WDBC compared to the noisier, compressed WBCD. Caution should be taken before using the two interchangeably, as WDBC is considered the standard benchmark dataset for its ability to capture intricacies better through finely engineered features[18].

3. Research Method

In this study, we have constructed a framework aimed at breast cancer prediction, while harnessing the capabilities of the random forests supervised machine learning algorithm and an assortment of feature selection techniques. Our method is particularly attuned to the F1-score, a metric of paramount importance in this context. This section will highlight in detail research methodology.

I. Dataset Acquisition and Overview

A multitude of publicly accessible repositories, inclusive yet not limited to the Mammographic Image Analysis Society (MIAS), Digital Database for Screening Mammography (DDSM), WBCD, Breast Cancer Digital Repository (BCDR), and the National Biomedical Imaging Archive (NBIA) [19] offer a substantial breadth of breast imaging data. However, to demonstrate the efficacy and accuracy of our proposed model, we harnessed the WDBC dataset procured from the esteemed University of California — Irvine repository [20]. The WBCD dataset comprises 569 breast cancer observations, providing insights into cell nuclei characteristics. Among these observations, 357 pertain to benign tumors, while 212 correspond to malignant tumors as shown in Figure 1. As per the comprehensive elucidation associated with this dataset, it is understood that the columns encapsulate ten quantitatively significant attributes pertaining to each cellular nucleus. These include compactness, perimeter, concave points, texture, symmetry, area, smoothness, concavity, fractal dimension, and radius. Moreover, for each specified group in this dataset, three distinct metrics are meticulously evaluated: the average value, the standard error, and the maximal value. These trio of measurements within each group are acknowledged as distinct features within the dataset. Consequently, the dataset is composed of a total of 30 attributes, enhancing its multidimensionality and comprehensive nature. Thus, we have 30 numerical variables and only 1 categorical variable, which is our target variable (diagnosis). The goal is to label each observation as either benign or malignant. Table III provides description of dataset attributes.

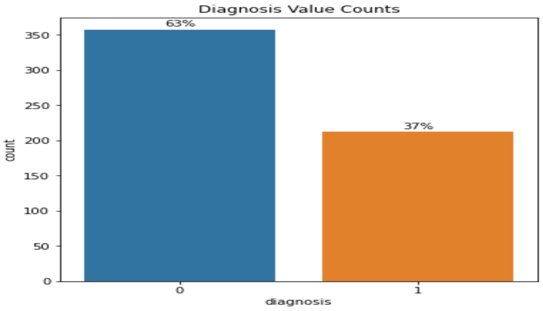

FIG 1

DIAGNOSIS VALUE COUNTS

TABLE III

DATASET DESCRIPTION

Attribute	Description
Radius	Average measurement from the center to the boundary on the perimeter
Texture	Variation in the intensity of pixels in the grayscale representation
Perimeter	The total length defining the outer boundary of the tumor in the depicted image
Area	The spatial extent within the tumor's boundaries depicted in the image
Smoothness	Fluctuation in the length of the radius around the tumor
Compactness	An equation derived from the perimeter squared divided by area, and reduced by 1
Concavity	Measure of the depth and frequency of dips in the contour of the tumor
Concave points	Count of indented or inward segments present in the tumor's contour
Symmetry	The degree to which the two halves of the tumor image mirror each other
Fractal dimension	A mathematical approach that tries to quantify the complexity of the tumor's border, by approximation - 1

II. Data Pre-processing

The initial phase of this extensive process involves the acquisition of image data. This is closely followed by the deployment of a series of meticulous preprocessing and normalization operations to refine the raw data, ensuring its pristine condition before it is propelled into subsequent stages of processing.

During the data refinement process, a noteworthy observation is made regarding a particular column in the data frame, named 'Unnamed', which lacks any data entries. This column, in fact, contains no information whatsoever, thus suggesting that excluding it would be the most prudent approach to uphold the integrity and consistency of the data analysis. Similarly, another column, 'id', is identified as a potential candidate for exclusion. The reasoning behind this decision stems from the fact that despite its presence, the 'id' column does not provide any meaningful insights or contribute valuable context that could enhance the classification of cancer cells. Consequently, removing this column aligns with the overarching objective of streamlining the data set for more effective analysis.

Standardizing, or normalizing, data is a critical preprocessing step in machine learning that adjusts values on different scales to a common one. This process is implemented through Z-score normalization [21], where each value is subtracted from the feature's mean and divided by its standard deviation. As a result, standardized features acquire a mean of 0 and a standard deviation of 1. Standardization ensures equal contribution from all features, preventing those with larger scales from dominating the model and potentially leading to suboptimal performance. Bringing all features to a similar scale through this technique optimizes the functionality of many machine learning algorithms.

Once the data is standardized it is divided into three groups feature_mean, feature_se, feature_worst for mean, standard error, and 'worst' or 'largest mean value' feature analysis.

III. Feature Engineering

In the context of medical scenarios, accurately predicting malignant cases holds significant importance. Feature importance serves as a metric to gauge the effectiveness of attributes in predicting the target variable. By leveraging attributes with the highest importance, valuable insights can be gleaned from the dataset, ultimately enhancing the predictive model. Moreover, this approach aids in achieving our objective of precisely identifying malignant cases.

To identify and eliminate the features that displayed high correlations with each other – multicollinearity - we utilized a heatmap correlation matrix. The color scheme facilitates differentiation of correlation values, where warm colors represent positive correlations and cool colors represent negative correlations. Exploratory data analysis revealed potential multicollinearity issues, particularly among radius, perimeter, and area variables. The heatmap for correlation between predictor variables (figure 2) provides a visual representation of the correlation matrix, making it easier to identify strong positive or negative correlations between predictor variables.

IV. Understanding Predictor-Target Variable Relationship

In figure 3, violin plots visualize the association between predictor variables (features) and the target variable (diagnosis) using the Seaborn library. The goal is to identify pivotal features for distinguishing benign and malignant cancers. Violin plots combine box plots depicting quartiles and medians with kernel density plots showing distribution shape. The hue differentiates feature distributions between diagnosis groups. It is evident malignant cells have higher values across most features except fractal dimension. This suggests promise in classifying cancer cells. Specifically, the x-axis lists features while y-axis shows values. Each feature has two violins representing benign and malignant cases. Violin width correlates to data density - broader sections mean concentrated data points. Side-by-side violins enable comparison of benign and malignant distributions per feature. Quartile markers provide median values and dispersion insights. Thin tails may be outliers and thick sections highlight patterns. By assimilating these attributes, violin plots help recognize influential features for detection and prediction, critical for breast cancer research.

Violin plots visualize the distribution of numeric data between groups. They combine the insights of a box plot showing quartiles and density plots showing frequency distribution. Thicker sections represent higher frequency values. Violin plots contain all data points unlike bar graphs with averages.

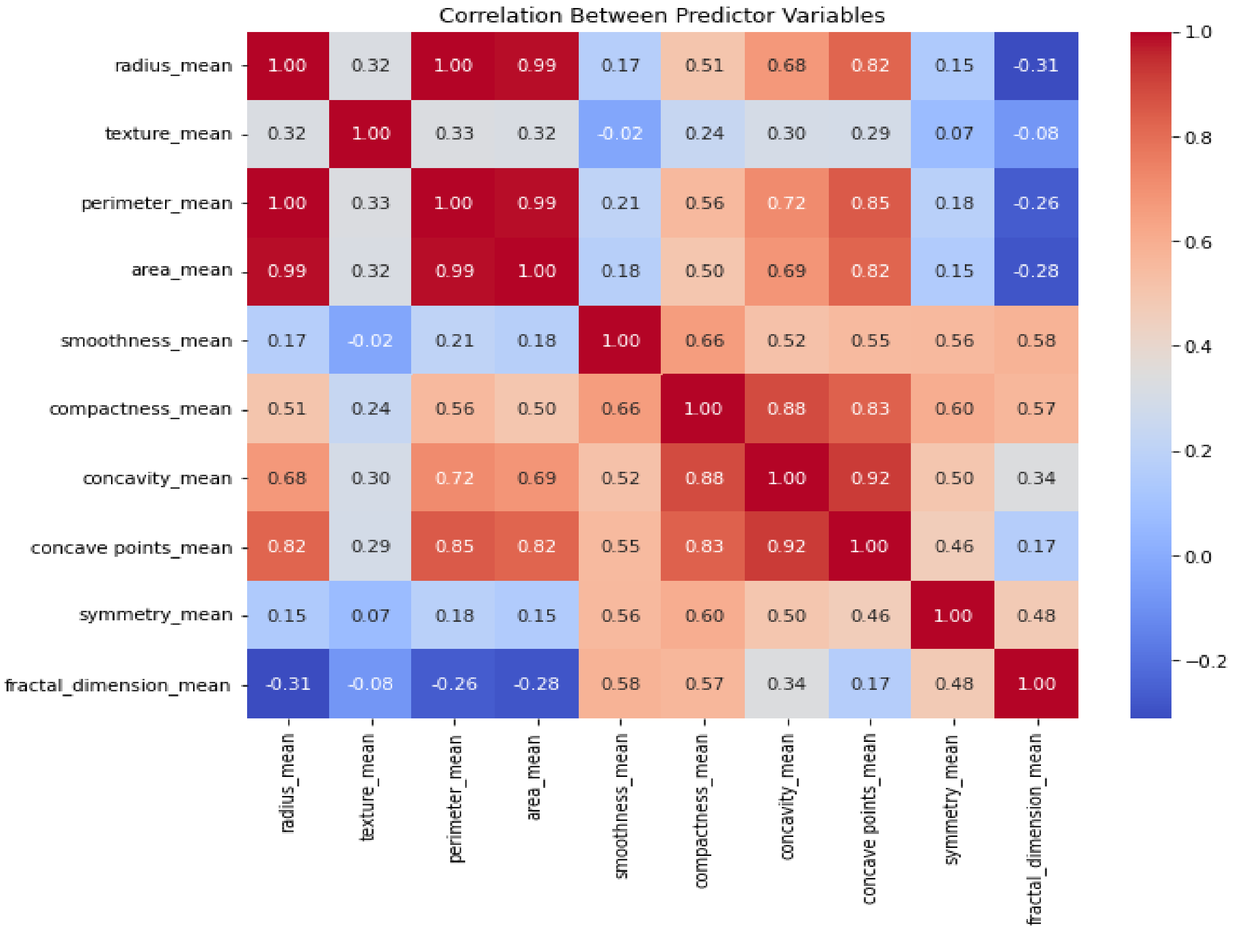

Fig 2 Heatmap For Correlation Between Predictor Variables

This makes them useful for small sample sizes and non-normal data. We created violin plots with Python for 10 features to compare distributions between groups in our dataset. Violin plots help observe differences in center, shape, variance and outliers across groups.

Figure 3 shows that most features show clear separation between the distributions of the benign and malignant groups, indicating their potential for accurately classifying the tumor cells. Specifically, the features radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, and symmetry_mean have distinguishing distributions, with the 75th percentile of the benign group below the 25th percentile of the malignant group. This separation suggests these features can serve as good candidates for predictors in building a classification model, as their value ranges do not overlap much between the two groups. The one exception is fractal_dimension, which shows substantial overlap between the benign and malignant violins, implying it may not be as useful of a predictor.

V. Feature Selection

Feature selection is a crucial step in machine learning for breast cancer detection, enhancing model performance and improving diagnosis methods, especially in the early stages. By selecting relevant features, the model's accuracy and reliability is significantly improved, enabling it to capture essential patterns and characteristics of breast cancer. This process not only enhances classification results but also reduces computational complexity. Consequently, there arises a necessity to ascertain an optimal subset of features that can enhance the precision of classification [22].

With a thorough understanding of our data, we can proceed to feature selection for our model's training set. As a starting point, we will utilize all the features available in the dataset for a base case. Upon training the model using this approach, we achieved an impressive accuracy of 98.25% and an F1 score of 0.98. However, our objective is to explore various techniques for feature selection and dimensionality reduction. We aim to achieve a comparable level of accuracy while reducing the number of features in our training set.

The techniques we will consider include Model-Based Feature Selection, Variance Inflation Factor (VIF)

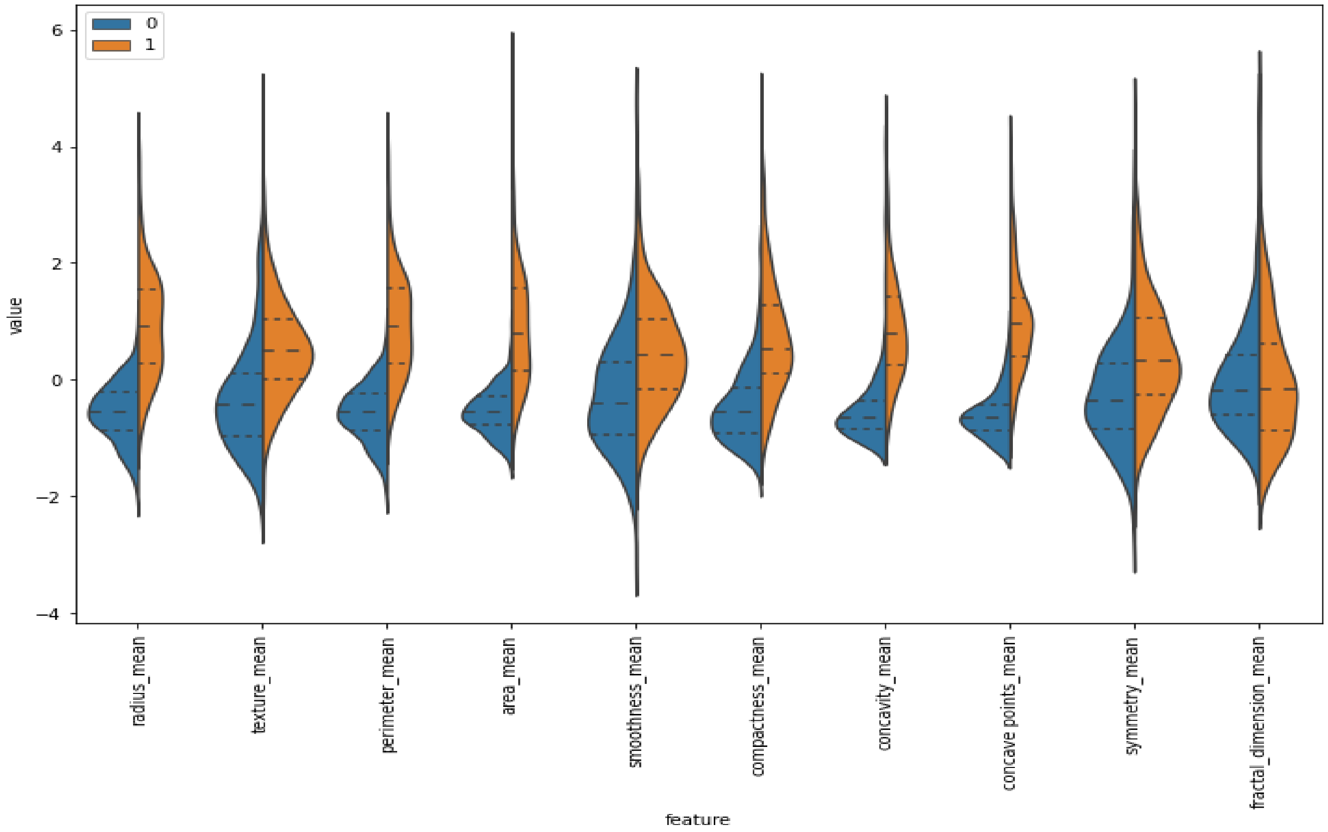

Fig 3 Violin Plot Showing That All Features, Except For Fractal Dimension, Exhibit Promising Potential In Classifying Cancer Cells

[23], Recursive Feature Elimination [24], Univariate Feature Selection [25], and Principal Component Analysis (PCA) [26]. By employing these methods, we strive to identify the most influential features and optimize our model's performance.

VI. Classification Algorithm - Random Forests

We have employed the Random Forest algorithm for the classification of our model. Random forest, an ML by Leo Breiman and Adele Cutler, amalgamates the outputs of several decision trees to generate a single outcome [27]. It's a blend of flexibility and simplicity, capable of managing both classification and regression issues. Building on the bagging method, the random forest algorithm employs feature randomness in conjunction with bagging to create an uncorrelated ensemble of decision trees, often referred to as a "forest". This feature randomness or feature bagging produces a random subset of features, thereby ensuring minimal correlation among the decision trees. Diverging from singular decision trees that contemplate all possible feature splits, random forests select only a subset of these features, thereby reducing overfitting, bias, and overall variance to enhance prediction precision. The random forest algorithm involves setting three primary hyperparameters before training - node size, the number of trees, and the number of features sampled [28]. It utilizes a bootstrap sample, a data sample extracted with replacement from the training set, to construct each tree in the ensemble. An element of randomness is introduced via feature bagging, which enriches dataset diversity and lowers correlation among decision trees. Depending on the problem type, the prediction determination varies - for regression tasks, the individual decision trees are averaged, while for classification tasks, the most frequent categorical variable is chosen as the predicted class through a majority vote. Ultimately, an out-of-bag sample, which constitutes one-third of the training sample, is utilized for cross-validation to finalize the prediction. Academic researchers have previously utilized the random forest algorithm in diagnosing a diverse array of diseases, underscoring its versatile application in the medical field [29-32].

VII. Performance Evaluation Metrics

The performance evaluation of Random Forests using selected feature techniques is assessed based on two key metrics: Accuracy and F1-score, as represented in equations (1) and (2). Accuracy, a widely adopted metric, provides an overall measure of correctness in predictions and must be as close as possible to 1. However, in scenarios where class imbalance exists, accuracy alone might not provide a complete picture of model performance. This is where the F1-score comes into play. The F1-score considers both precision and recall, offering a balanced measure that accounts for both false positives and false negatives. By incorporating the F1-score alongside accuracy, we gain a more comprehensive evaluation of the model's performance, especially in situations where class distribution is uneven or when false positives and false negatives need to be equally weighed.

Accuracy = (TP + TN) / (TP + FP + FN + TN) (1)

Where TP denotes True Positives, signifying instances where both predicted and actual labels are positive. Conversely, TN, or True Negatives, signifies cases where both labels are negative. FP and FN represent False Positives and False Negatives, respectively, reflecting the erroneous classifications: FP where the model erroneously assigns positive labels to negative instances, and FN where positive instances are incorrectly labeled as negative. The quotient's numerator, (TP + TN), aggregates the correct classifications, while the denominator ascertains the total data points classified. The ratio, thus, elucidates the proportion of precise classifications amidst the total predictions, symbolizing the model's fidelity.

F1-score=2(precision*recall)/(precision+recall)*
(2)

With:

precision = TP / (TP + FP)
(3)

The precision metric is a proportion delineating the model's aptitude in accurately discerning positive instances from the sum of instances it designates as positive. Elevated precision is indicative of the model's proficiency in minimizing erroneous positive classifications.

recall = TP / (TP + FN)
(4)

Recall quantifies the efficacy of a model in correctly classifying the positive samples from a dataset, formalized mathematically as the ratio of true positives to the sum of true positives and false negatives, where the former constitutes cases accurately diagnosed as positive while the latter denotes instances where the model erroneously predicts the negative class despite a positive ground truth; consequently, higher recall signifies greater effectiveness in capturing actual positives, and fewer detrimental false negatives - for example, a recall of 0.9 indicates 90% of malignant cases were correctly identified with only 10% mislabeled as benign, exemplifying proficient diagnosis. By distilling model performance on positive identification into a single metric, recall facilitates model comparisons, especially in sensitive domains like cancer prediction where overlooking true positives provokes severe repercussions, thus models maximizing recall are imperative.

VIII. Testing and Training

In line with established research methodology, a partitioning approach is employed to allocate 70% of the dataset as the training set, while reserving the remaining 30% as the test set. This systematic division allows for rigorous evaluation and assessment of the model's accuracy and performance.

4. Results

I. Feature engineering results

This research study leverages Python, a widely adopted open-source programming language, to conduct experimental analysis. As discussed previously, in the realm of feature selection, two distinct approaches were applied on all 30 features. The first approach involves utilizing the Heatmap correlation matrix, while the second approach incorporates various techniques such as Variance Inflation Factor (VIF), Model-based Feature Selection, Recursive Feature Elimination, Univariate Feature Selection and Principal component analysis in conjunction with the random forest algorithm. These feature selection techniques are specifically employed to detect breast cancer within the dataset under investigation.

II. Variance Inflation Factor (VIF)

By identifying and excluding the top five features with the highest Variance Inflation Factor (VIF), the code showcases the potential to refine the model's performance without compromising accuracy. With these features removed, the resulting model achieves an accuracy of 98.83% and an F1 score of 0.98, indicating its strong classification capabilities. A thorough analysis of the confusion matrix heatmap further enhances our understanding of the model's classification outcomes. Figure 4 shows confusion matrix to evaluate model accuracy.

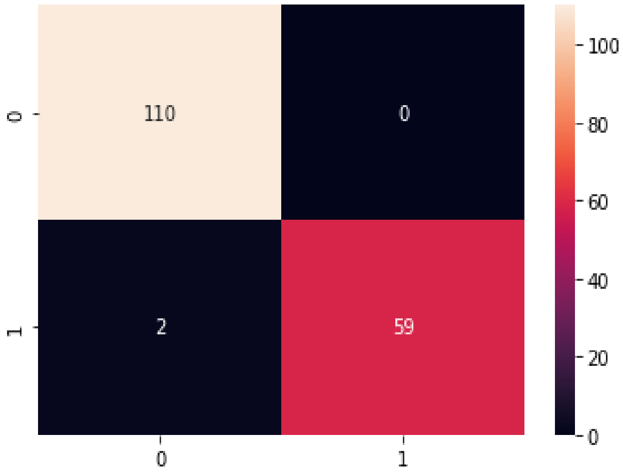

Fig 4 Confusion Matrix For Variance Inflation Factor (Vif)

It shows that with 110 TP, the model accurately identified 110 positive instances. Remarkably, there were no false negatives, indicating that no positive instances were mistakenly classified as negative. However, there were 2 false positives, representing negative instances incorrectly predicted as positive. In contrast, the model correctly identified 59 negative instances as TN. These results demonstrate the model's proficiency in accurately predicting

TP and TN. Nevertheless, the presence of FP suggests a slight vulnerability in classifying negative instances.

II. Univariate feature selection

By applying feature selection using SelectKBest with chi-square scoring, the top five features are identified. These selected features are then utilized to train a Random Forest classifier. Remarkably, even with the reduced feature set, the model achieves an impressive accuracy of 95.32%, showcasing its efficacy in accurately classifying cancer cells. The F1 score provides additional confirmation of the model's precision and recall. Figure 5 shows confusion matrix to evaluate model accuracy.

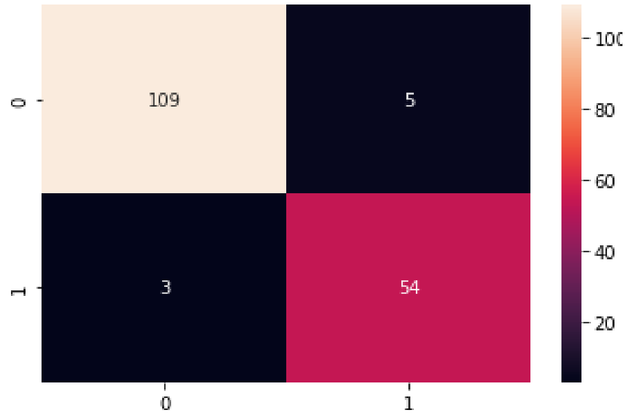

Fig 5 Confusion Matrix For Univariate Feature Selection (Ufs)

It illustrates the performance of a binary classification model utilizing Univariate Feature Selection, with TP of 109, FN of 5, FP of 3, and TN of 54. The model boasts an accuracy of 95.32%, signifying that over 95% of the samples were correctly classified, and an F1 score of 0.93, indicating a strong balance between precision and recall. The high TP coupled with low FN and FP shows the model's efficiency in accurately identifying positive cases and minimizing errors. This result is an analytical snapshot of the model’s capability to make reliable and accurate classifications for the given dataset using Recursive Feature Elimination.

III. Recursive feature elimination (RFE)

RFE technique is utilized to select the most critical features in the dataset. The model's accuracy is evaluated after applying RFE with a Random Forest classifier. The obtained accuracy of 95.91% and F1 score of 0.94 signify the model's exceptional performance in precise classification. Visualizing the confusion matrix heatmap provides valuable insights into the classification outcomes. Importantly, the achieved accuracy closely aligns with the results obtained through the univariate feature selection approach. Figure 6 shows confusion matrix to evaluate model accuracy.

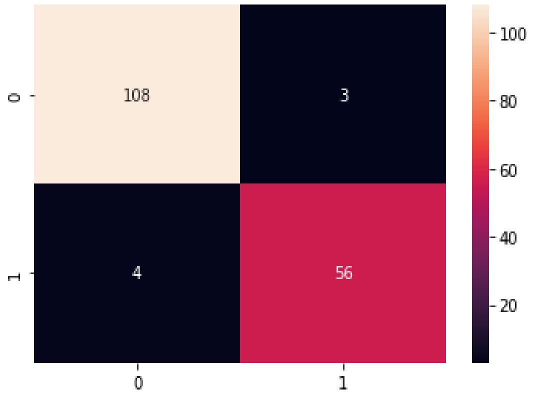

Fig 6 Confusion Matrix For Recursive Feature Elimination (Rfe)

Figure 6 presents a confusion matrix displaying the results of a classification model using Recursive Feature Elimination with the top 5 features: 'concave points_mean', 'radius_worst', 'perimeter_worst', 'area_worst', 'concave points_worst'. The matrix reports TP of 108, FN of 3, FP of 4, and TN of 56. These values reflect the model's precision in correctly identifying cases (TP and TN) and its few misclassifications (FN and FP) when employing these specific features. The superior performance is further affirmed by the model's high accuracy of 95.91% and a robust F1 score of 0.94. This analytical snapshot provides a credible demonstration of the model's effective and reliable capabilities in producing accurate classifications using the selected features.

IV. Model-based feature selection

The analysis begins with a feature selection process using a Random Forest classifier, where each feature's importance is ranked. Features with an importance above 5% are selected to create a refined feature set. A subsequent bar plot visualizes the feature importance, shedding light on the most influential features. In Figure 7, it is evident that concave points_worst, area_worst, perimeter_worst, and radius_worst exhibit considerable deviations in their feature importance values during the prediction process. Particularly, concave points_worst and area_worst demonstrate high deviations.

Consequently, the reliability of these features in accurate prediction is compromised due to their significant variations. With only 9 features in the refined set, our model achieves an impressive accuracy of 97.08% and an F1 score of 0.96. These results demonstrate the model's strong performance, coming very close to the accuracy achieved under the base case scenario (98.25%) where all features were used. Additionally, a confusion matrix heatmap enhances our understanding of the model's classification outcomes, providing valuable insights.

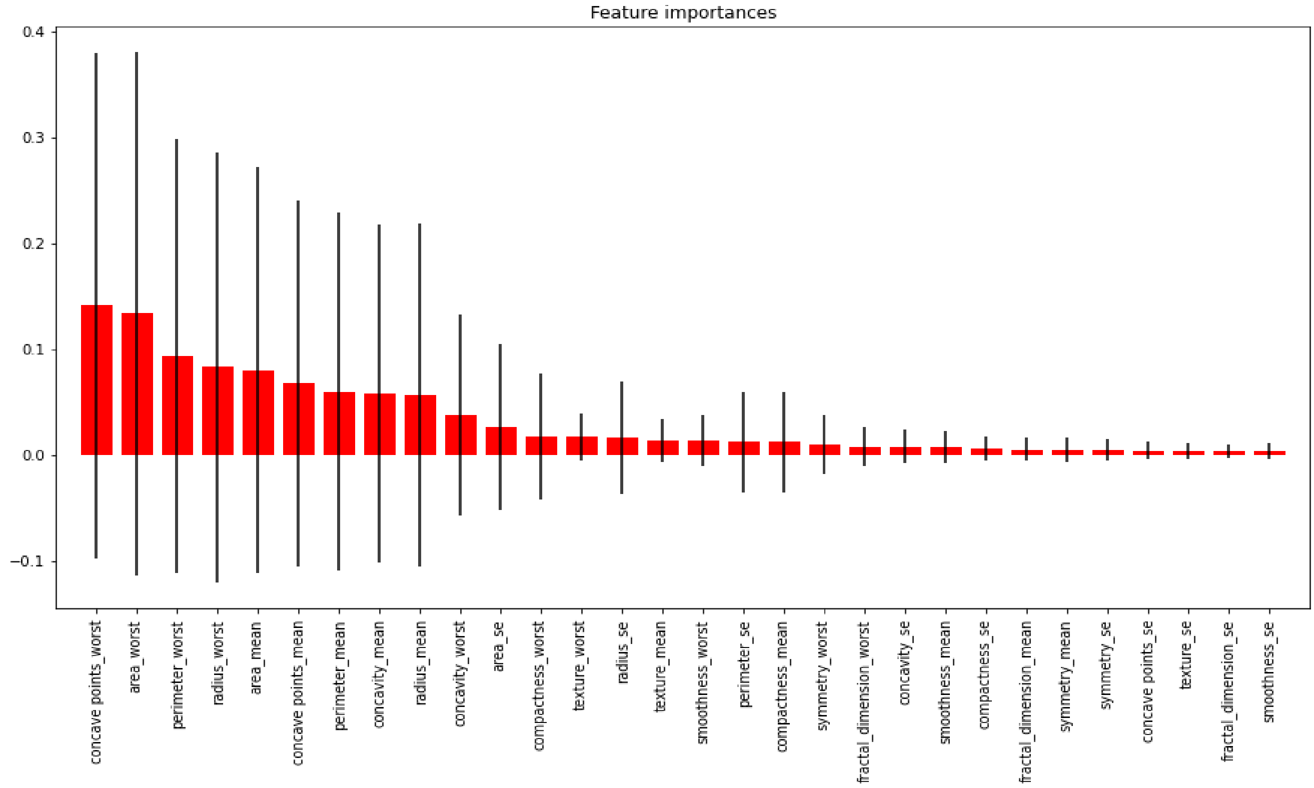

Fig 7 Bar Plot For Variation In Feature Importance

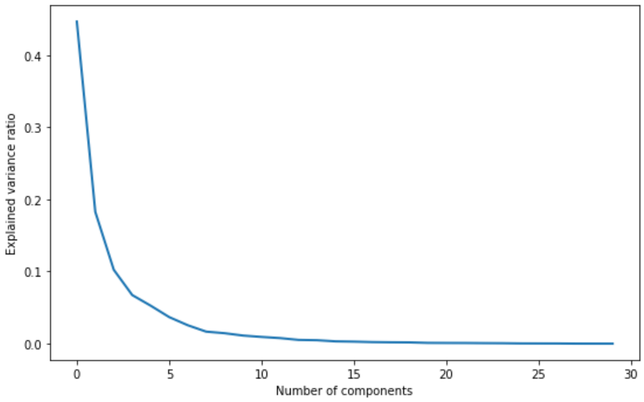

Fig 8 (A) Variance Ratio

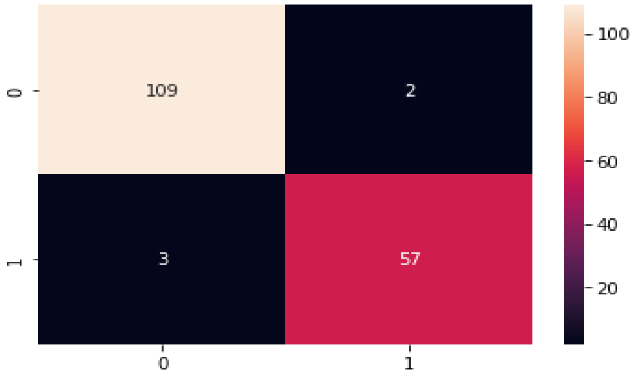

Fig 8 (B) Confusion Matrix For Principal Component Analysis

V. Principal component analysis (PCA)

In this method, the dataset is split into training and test sets using standardized features. Principal Component Analysis (PCA) is then applied to the training set to determine the optimal number of components that capture the variance in the data. Figure 8 (a) shows the elbow method. It is utilized to determine the optimal number of components for Principal Component Analysis (PCA). The plot showcases the explained variance ratio against the number of components, revealing an elbow point at four components. This elbow point is critical as it indicates that increasing the number of components beyond four yields diminishing returns in explained variance. Therefore, selecting four components is computationally efficient and prevents overfitting while still capturing most of the variance in the data.

Subsequently, PCA is performed again on both the training and test sets, this time using 4 components. A Random Forest classifier is trained on the transformed data and used to make predictions on the test set. The model achieves an impressive accuracy of 97.08% and an F1 score of 0.96, demonstrating its strong performance in accurately classifying cancer cells. While PCA reduces dimensionality, it reduces interpretability by transforming features into principal components. However, PCA remains a robust technique for summarizing features and ensuring the model captures important information for accurate predictions. Figure 8(b) illustrates a confusion matrix evaluating the performance of a classification model employing Principal Component Analysis (PCA) with four principal components. The matrix displays 109 True Positives, 2 False Negatives, 3

False Positives, and 57 True Negatives. With an accuracy of 97.08% and an F1 score of 0.96, the model demonstrates exceptional performance in classifications.

Notably, PCA allows the model to achieve such high accuracy by effectively reducing the dimensions, utilizing only four principal components as opposed to nine distinct features in alternative methods. However, it's important to recognize that PCA's combination of original features into principal components limits interpretability, obscuring the individual contribution of each feature. Nevertheless, PCA's robustness in summarizing features into key components ensures the capture of critical information, enabling the model to make highly accurate predictions.

Table IV presents a comprehensive overview of the experimental outcomes achieved through various feature selection algorithms. Each algorithm provided valuable insights by establishing a threshold for feature impact and specifying the number of critical features to consider. For instance, the Variance Inflation Factor algorithm attained an impressive accuracy of 98.83% while utilizing 25 selected features. Similarly, the Univariate Feature Selection technique showcased an accuracy of 95.32% with only 5 selected features. The Recursive Feature Elimination method achieved an accuracy of 95.91% while considering 5 important features. Model-Based Feature Selection delivered an accuracy of 97.08% with a slightly larger feature set of 9 selected features. Additionally, the Principal Component Analysis (PCA) approach demonstrated the ability to achieve an accuracy of 97.08% by utilizing only 4 principal components. These findings underscore the effectiveness of each feature selection algorithm in identifying influential features, enhancing model performance, and optimizing the feature set.

TABLE IV

EVALUATION METRICS USED TO ASSESS THE PERFORMANCE OF FEATURE SELECTION ALGORITHMS.

Feature selection algorithms	F1-Score	Accuracy	# of Features
Variance Inflation Factor	0.98	98.83%	25
Univariate Feature Selection	0.93	95.32%	5
Recursive Feature Elimination	0.94	95.91%	5
Model-Based Feature Selection	0.96	97.08%	9
Principal component analysis	0.96	97.08%	4

5. Discussion

Our work presents a novel framework that integrates feature engineering and random forest classification technique to enhance the accuracy and effectiveness of breast cancer diagnosis. By adopting this framework in clinical practice, it has the potential to significantly reduce breast cancer mortality rates. The focus of our research was on WBCD dataset comprising ten key features relevant to breast cancer. Our proposed model leverages the concept of feature importance to select the most influential attributes as input for our prediction models. This approach not only provides a comprehensive understanding of the model's behavior but also incorporates the interactions between features, resulting in more accurate predictions. By incorporating this framework into clinical decision-making, healthcare professionals can make informed and timely decisions for improved patient outcomes.

6. Conclusion and Future Work

Swift and accurate cancer diagnosis is imperative, especially in the context of breast cancer, a leading cause of mortality among women worldwide. Motivated by this challenge, our study aims to enhance current classification methods for predicting breast cancer. Our comprehensive framework encompasses two crucial components: feature selection and classification. Initially, we employ a heatmap to visualize the correlation matrix, facilitating the identification of robust relationships between predictor variables. Building upon this, we integrate diverse techniques, such as Variance Inflation Factor (VIF), Univariate Feature Selection, Recursive Feature Elimination, Model-Based Feature

Selection, and Principal Component Analysis (PCA), into the random forest algorithm. By applying this framework to the breast cancer dataset, significant enhancements in diagnosis performance are observed. Moving forward, we recommend expanding the evaluation to larger datasets and exploring the integration of optimization techniques like Particle Swarm Optimization (PSO), Genetic Algorithm (GA), or Ant Colony Optimization (ACO) algorithms to achieve precise parameter selection for ensemble algorithms. Additionally, the development of a user-friendly graphical interface holds promise for enabling medical practitioners with limited ML or data science expertise to effortlessly utilize the framework, empowering them to input patient data and obtain accurate classification results.

References

[1] C. D. Runowicz *et al.*, "American cancer society/American society of clinical oncology breast cancer survivorship care guideline," *CA: a cancer journal for clinicians,* vol. 66, no. 1, pp. 43-73, 2016.

[2] M. Ataollahi, J. Sharifi, M. Paknahad, and A. Paknahad, "Breast cancer and associated factors: a review," *Journal of medicine and life,* vol. 8, no. Spec Iss 4, p. 6, 2015.

[3] T. J. Whelan *et al.*, "External beam accelerated partial breast irradiation versus whole breast irradiation after breast conserving surgery in women with ductal carcinoma in situ and node-negative breast cancer (RAPID): a randomised controlled trial," *The Lancet,* vol. 394, no. 10215, pp. 2165-2172, 2019.

[4] A. H. Eijkelboom *et al.*, "Impact of the suspension and restart of the Dutch breast cancer screening program on breast cancer incidence and stage during the COVID-19 pandemic," *Preventive medicine,* vol. 151, p. 106602, 2021.

[5] A. Kumar *et al.*, "Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer," *Information Sciences,* vol. 508, pp. 405-421, 2020.

[6] M. Samieinasab, S. A. Torabzadeh, A. Behnam, A. Aghsami, and F. Jolai, "Meta-Health Stack: A new approach for breast cancer prediction," *Healthcare Analytics,* vol. 2, p. 100010, 2022.

[7] S. Raj, S. Singh, A. Kumar, S. Sarkar, and C. Pradhan, "Feature selection and random forest classification for breast cancer disease," *Data Analytics in Bioinformatics: A Machine Learning Perspective,* pp. 191-210, 2021.

[8] A. U. Haq *et al.*, "Detection of Breast Cancer Through Clinical Data Using Supervised and Unsupervised Feature Selection Techniques," *IEEE Access,* vol. 9, pp. 22090-22105, 2021, doi: 10.1109/ACCESS.2021.3055806.

[9] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," *IEEE Access,* vol. 10, pp. 3284-3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[10] P. H. Prastyo, I. G. Y. Paramartha, M. S. M. Pakpahan, and I. Ardiyanto, "Predicting Breast Cancer: A Comparative Analysis of Machine Learning Algorithms," in *Proceeding International Conference on Science and Engineering*, 2020, vol. 3, pp. 455-459.

[11] B. Padmapriya and T. Velmurugan, "Classification algorithm based analysis of breast cancer data," *International Journal of Data Mining Techniques and Applications,* vol. 5, no. 1, pp. 43-49, 2016.

[12] P. Singhal and S. Pareek, "Artificial neural network for prediction of breast cancer," in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on*, 2018: IEEE, pp. 464-468.

[13] A. A. Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *International Journal of Machine Learning and Computing,* vol. 9, no. 3, pp. 248-254, 2019.

[14] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications,* vol. 41, no. 4, pp. 1476-1482, 2014.

[15] M. Minnoor and V. Baths, "Diagnosis of Breast Cancer Using Random Forests," *Procedia Computer Science,* vol. 218, pp. 429-437, 2023/01/01/ 2023, doi: https://doi.org/10.1016/j.procs.2023.01.025.

[16] C. Nguyen, Y. Wang, and H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," 2013.

[17] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Computing and applications,* vol. 28, pp. 753-763, 2017.

[18] M. H. Alshayeji, H. Ellethy, and R. Gupta, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control,* vol. 71, p. 103141, 2022.

[19] L. Dora, S. Agrawal, R. Panda, and A. Abraham, "Optimal breast cancer classification using Gauss–

Newton representation based algorithm," *Expert Systems with Applications*, vol. 85, pp. 134-145, 2017.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," ed: Irvine, CA, USA, 2007.

[21] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[22] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *The Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

[23] C. G. Thompson, R. S. Kim, A. M. Aloe, and B. J. Becker, "Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results," *Basic and Applied Social Psychology*, vol. 39, no. 2, pp. 81-90, 2017.

[24] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC genetics*, vol. 19, no. 1, pp. 1-6, 2018.

[25] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 25-29 May 2015 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

[26] G. N. Ramadevi, K. U. Rani, and D. Lavanya, "Importance of feature extraction for classification of breast cancer datasets—a study," *International Journal of Scientific and Innovative Mathematical Research*, vol. 3, no. 2, pp. 763-368, 2015.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[28] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, p. e1301, 2019.

[29] J. Gómez-Ramírez, M. Ávila-Villanueva, and M. Á. Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," *Scientific Reports*, vol. 10, no. 1, pp. 1-15, 2020.

[30] V. E. Christo, H. K. Nehemiah, J. Brighty, 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.

[31] D. Paul, R. Su, M. Romain, V. Sébastien, V. Pierre, and G. Isabelle, "Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier," *Computerized Medical Imaging and Graphics*, vol. 60, pp. 42-49, 2017.

[32] S. Wang, Y. Wang, D. Wang, Y. Yin, Y. Wang, and Y. Jin, "An improved random forest-based rule extraction method for breast cancer diagnosis," *Applied Soft Computing*, vol. 86, p. 105941, 2020.

[33] Singh, J. ., Mani, A. ., Singh, H. ., & Rana, D. S. . (2023). Solution of the Multi-objective Economic and Emission Load Dispatch Problem Using Adaptive Real Quantum Inspired Evolutionary Algorithm. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1s), 01–12. https://doi.org/10.17762/ijritcc.v11i1s.5989

[34] Dhabliya, P. D. . (2020). Multispectral Image Analysis Using Feature Extraction with Classification for Agricultural Crop Cultivation Based On 4G Wireless IOT Networks. Research Journal of Computer Systems and Engineering, 1(1), 01–05. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/10