PREDICTION AND CLASSIFICATION OF BREAST CANCER USING MACHINE LEARNING TECHNIQUES

By

Neha Sunil, BE, Birla Institute of Technology, 2019,

A Major Research Project Report

presented to Ryerson University

in partial fulfilment towards the requirements for the degree of

Master of Science

In the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2022

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

Neha Sunil

PREDICTION AND CLASSIFICATION OF BREAST CANCER USING MACHINE LEARNING TECHNIQUES

Neha Sunil

Master of Science 2022

Data Science and Analytics

Ryerson University

ABSTRACT

Breast cancer, one of the most frequent malignancies among women, is regarded to be nearly one in three cancers diagnosed among women in the United States. For specific care and medical treatment, diagnostic procedures with improved performance are crucial in this domain. In this research, the main aim is to classify whether the breast cancer is malignant or benign and also help predict the reappearance and non-recurrence of cases classified as malignant. The data used was obtained from the public Wisconsin Breast Cancer dataset. An exploratory analysis is conducted using descriptive analytics as well as data visualization to help identify some patterns, or relationships that could help shed light on hidden information in the data available to us followed by the second part of the analysis stemming from training different classifier models that can help classify the tumor in the breast as benign or malignant and predict the reoccurrence of cases classified as malignant by observing the most important features. The performance of the different classifiers was compared using precision, recall as well as Area Under the Curve.

Key words: Breast Cancer, Machine Learning, Feature Selection

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude towards Dr. Farid Shirazi for his constant support and assistance towards the progress and completion of this research project. As my supervisor, Dr. Farid Shirazi's guidance has been monumental throughout the term to guide and direct my research and provide constructive feedback for me to improve upon.

Thank you, Dr. Farid Shirazi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

A. Background

Breast cancer, one of the most frequent malignancies among women, is regarded to be nearly one in three cancers diagnosed among women in the United States. Abnormal cell growth in the breast tissue also referred to as a tumor is one of the main reasons for the occurrence of Breast Cancer. Despite the fact that cancer is curable to some extent especially when in its nascent stages, we still observe a great deal of women diagnosed with late-stage breast cancer. For specific care and medical treatment, diagnostic procedures with improved performance are crucial in this domain

Tumors don't necessarily mean cancer- tumors are sometimes benign as in they're not cancerous, pre-malignant as in the nascent stages of cancer or malignant which means the tumor is cancerous. As few tests such as mammograms, MRIs, ultrasound & biopsy can help diagnose breast cancer.

Machine learning techniques can help contribute to the process of prediction and early diagnosis of breast cancer to a large extent. The prediction falls into two categories (i.e., Malignant or Benign), the reason being the labels in this dataset are discrete making this a classification problem in Machine Learning.

This research paper discusses the following research questions and highlights the machine learning models, evaluating their performance, the dataset utilized and the likelihood of breast cancer classification as well as detection.

B. Research Question

The main aim of this study is to classify whether the breast cancer is malignant or benign and also help predict the reappearance and non-recurrence of cases classified as malignant. Based on the objective and the dataset acquired, the research questions are mapped out below.

1. **Which machine learning models perform the most adequately to detect and classify breast cancer?**

We will compare the different machine learning models based on the performance metrics.

2. **Will the models perform better by adopting a few feature selection techniques to overcome the issue of redundant data?**

   We will try to implement a few feature selection techniques and observe how the models perform differently

3. **Which are the most important features for breast cancer classification that play a vital role with early diagnosis?**

   We will observe the critical features that help in the prediction and classification process.

C. Dataset

The Breast Cancer Wisconsin Dataset, available on the UCI Machine Learning repository defines features that were computed from digitized images of a fine needle aspirant (FNA) of a breast mass. These features report the characteristics of the cell nuclei from the image. Breast mass FNA is considered to play a key role in assessing malignancy in breast cell tissue. This dataset was put together by Dr. William H. Wolberg while he was at the University of Wisconsin-Madison Hospital. This dataset comprises of 569 samples of breast cell tissue which were medically assessed focusing on 32 characteristics of the cell nucleus.

Attribute Information:

1. ID number
2. Diagnosis (M= Malignant, B= Benign)

Ten real-valued features are computed for each cell nucleus (3-32): radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension.

2. LITERATURE REVIEW

The main aim of this study is to classify whether the tumor detected in the breast is benign or malignant and also help predict the reappearance and non-recurrence of cases classified as malignant using machine learning. In cancer research, Machine Learning models have been quite successful in being able to help with not just research but also with the practical implementation in cancer detection [2]. In this section, an overview of relevant work dealing with breast cancer diagnosis leveraging machine learning approaches have been discussed.

To begin, a 2019 paper by Nguyen et al. [3] evaluated the performance of supervised and unsupervised classification models using the Wisconsin Breast Cancer Dataset [1]. To perform feature selection, scaling and Principal Component Analysis were adopted, after which the data was split into training and test sets by a 70:30 ratio. They concluded that the Ensemble Voting approach served as an ideal predictive model for breast cancer. After performing Feature Selection and Principal Component Analysis, various models were trained and tested on the data. The results indicated that only four models among the tested models namely Support Vector Machine, Ensemble-Voting Classifier, Logistic Regression and AdaBoost resulted in approximately 98% accuracy. The performance of the proposed model showed the most potential in classifying the tumor as benign or malignant by analyzing the results of precision & recall, ROC-AUC (receiver operating characteristic curve-area under the curve), F-1 measure and computational time of the models.

A 2013 paper by Ahmad et al. [4] evaluated data using the Iranian Center for Breast Cancer dataset to discover risk factors in breast cancer prediction. A comparative analysis of the performances of decision tree (C4.5), Artificial Neural Network and Support Vector Machine indicated that SVM outperformed both the decision tree and the multilayer perceptron in terms of sensitivity, specificity and accuracy. The paper also describes how due to missing data in their records and exclusion of important variables such as S-phase fraction and DNA index, the performance of the models was affected.

Hasan and Tahir [5] in their research proposed a feature extraction algorithm based on Principal Component Analysis (PCA) using an Artificial Neural Network (ANN) classifier as an optimal tool to help differentiate malignant tumors from benign tumors using the Wisconsin Breast Cancer Dataset [1]. For feature selection, the three rules of rules of thumb of PCA viz. the Kaiser Guttman

rule, the Scree Test, & Cumulative Variance were employed. Using these rules, an ensemble of the reduced dataset was passed as inputs to the ANN classifier with back propagation algorithm resulting in easier classification of normal and breast cancer patients.

A 2019 paper by Omondiagbe et al. [6] suggested an automated method with a primary workflow for the diagnosis of breast cancer using data from the Wisconsin Diagnostic Breast Cancer Dataset (WDBC), analyzing the performance of Support Vector Machine (using radial basis kernel), Artificial Neural Networks and Naïve Bayes. The primary focus was to incorporate these techniques with feature selection/ feature extraction methods to recognize the most fitting approach in the classification of benign and malignant tumors. The paper discusses a hybrid approach by reducing the high dimensionality of features using LDA (Linear Discriminant Analysis) and thereby applying the new reduced dataset to Support Vector Machine producing an accuracy of 98.82%, sensitivity of 98.41%, specificity of 99.07%, and an area under the receiver operating characteristic curve (AUROC) of 0.9994.

A comparison analysis of Decision Tree, Naïve Bayes, Neural Network and Support Vector Machine with three different kernel functions were studied by Nematzadeh et al. [7] to classify original and prognostic breast cancer cases using the Wisconsin Breast Cancer dataset [1]. The motive is to study the impact k in k-fold cross validation and achieve higher accuracy. The study showed that Neural Network with k=10 had the highest accuracy of 98.09% in the diagnostic dataset while Support Vector Machine with Radial Basis Function (RBF) and k=10 had the highest accuracy of 98.32% in the prognostic dataset. What was learnt from this study was that one cannot expect to get an accurate result by having a higher value of k when applying k-fold cross validation.

CFS, a feature selection technique using correlation was employed by Yesuf et al. [8] in his research where a correlation value of 0.7 was set to filter out features with means higher than 0.7. As a subsequent feature selection technique [9], Recursive Feature Elimination (RFE) was also implemented. This technique was implemented by adopting the wrapper approach. All the feature subsets were classified in accordance with their accuracy score and subsets with features having high ranking scores were selected. Their research involved implementing feature extraction techniques such as Principal Component Analysis in addition with Linear Discriminant Analysis along with feature selection techniques namely Correlation-based Feature Selection (CFS) and Recursive Feature Elimination (RFE).

The objective of a 2017 paper by Subrata et al. [10] was to perform feature selection as an approach to obtain the smallest subset of features. By doing so, they were able to obtain an extremely accurate classification of breast cancer as benign or malignant. This was followed by performing a comparative analysis study on Logistic Regression, Naïve Bayes as well as Decision Tree by evaluating the time complexity of each classifier. Their study deduced Logistic Regression classifier to be the best classifier with the highest accuracy in contrast with the other classifiers.

Kumar et al. [11] in his research focused on different classification techniques in data mining to help predict benign and malignant tumors using the Wisconsin Breast Cancer Dataset [1]. Among all the features, clump thickness served as the evaluation class. Performance analysis was carried out using Lazy IBK, Multiclass Classifier, Ada Boost, Naive Bayes, Decision Table, Random Tree, J Rip, Logistics Regression, Multilayer Perceptron and Random Forest.

Lucas et al. [12] in their paper used two machine learning classifiers viz. Bayesian Networks and J48 to discriminate benign from malignant breast tumors. The best results were achieved by Bayesian Networks, resulting in 97.80% accuracy. Bharat et al. [13] assessed the performance of their proposed model with 3 machine learning classifiers to help classify benign and malignant tumors: Naïve Bayes, J48 and Radial Basis Function networks. Their study concluded that the Naïve Bayes model achieved the highest accuracy of 97.3%.

The main objective of a 2017 paper by Ojha and Goel [14] was to determine how accurately data mining algorithms would forecast the likelihood of reoccurrence of breast cancer on the basis of feature importance. After implementing classification algorithms, performing clustering on the dataset and evaluating their performance, the experimental results indicated that classification algorithms were better predictors with decision tree (C5.0) and SVM achieving 81% accuracy and fuzzy c-means achieved the lowest accuracy of 37%.

In their research, Ghosh et al. [15] applied different classification techniques to diagnose and analyze breast cancer with Multilayer Perceptron using Backpropagation Neural Network and Support Vector Machine and by assessing their performance, concluded that SVM was the best classifier.

Osareh & Shadgar [16] in their study to correctly classify data from the Breast Cancer dataset tried combining Support Vector Machine, K-Nearest Neighbor and Probabilistic Neural Network classifiers with signal-to-noise ratio feature ranking, sequential forward selection-based feature

selection and principal component analysis to classify benign tumors from malignant tumors. Their results indicated that SVM-RBF classifier achieved the best overall accuracy of 98.80%.

To compare the models, Bazazeh and Shubair [17] analyzed the performance of 3 classifiers viz. Random Forest (RF), Support Vector Machine (SVM) and Bayesian Networks (BN) using the data from the Wisconsin Breast Cancer dataset [1]. From the experimental results, SVM was highlighted to have performed the best in terms of specificity, accuracy as well as precision. On the other hand, the only classifier to have had the highest probability of classifying benign and malignant tumors correctly was Random Forest.

To easily classify benign and malignant tumors, Azmi and Cob [18] set up a system employing neural networks with Feedforward Back Propagation. Their results revealed that Neural Networks with 7 hidden layers attained the highest accuracy of 96.63%.

In their study to detect breast cancer, Gayathri & Sumathi [19] led a comparative analysis of relevance vector machine (RVM) with a few machine learning algorithms. Linear discriminant analysis (LDA) was implemented for feature reduction. The results indicated that the RVM classifier obtained an accuracy of 96%, sensitivity of 98% and specificity of 94% proving it be better than other machine learning algorithms because of its low computational cost.

Jamal et al. [20] operated on Support Vector Machine and Extreme Gradient Boosting to classify data from the Wisconsin Breast Cancer Dataset into benign and malignant tumors. To classify the data, the number of data attributes were reduced by adopting feature extraction using principal component analysis (PCA) and clustering using k-means. After analyzing the performance of 4 models on the basis of accuracy, specificity and sensitivity, they learnt that k-means showed better results in comparison with PCA in terms of dimensionality reduction.

While working on the Extensible Breast Cancer Prognosis Framework (XBPF), Ravi et al. [21] worked towards trying to predict chances of risk, to forecast the reoccurrence of cancer and analyze the chances of survival. In an attempt to boost the efficiency in prognosis, a Representative Feature for Subset Selection (RFSS) algorithm in addition to Support Vector Machine was implemented. The results indicated noteworthy improvement in performance in comparison with modern methods of cancer diagnosis.

In a study about breast cancer prediction using data mining methods, Haifeng and Won Yoon [22], using clinical records of patients formulated an accurate model to predict the occurrence of breast

cancer. There were four machine learning models that were used to carry out their study: Support Vector Machine (SVM), Naïve Bayes classifier, Artificial Neural Network (ANN), & AdaBoost. These models were tested on data using two datasets: Wisconsin Breast Cancer Database (1991) and Wisconsin Diagnostic Breast Cancer (1995). K-fold cross validation was implemented as well. The reason being, they wanted to evaluate the error of each model on the test set. As a part of their research, they've also discussed about how feature space has a high impact on the efficiency of the learning process and proposed a hybrid between principal component analysis (PCA) and machine learning models.

While evaluating risk of breast cancer at the time of mammography screening, Abdollel et al. [23] in their study evaluated absolute as well as relative breast density-related measures based on area. 392 women who were diagnosed with unilateral breast cancer, had their mammography images examined as well as 817 age-related controls. Measures related to breast density were studied with the help of a completely automated software designed for measuring breast density. Area under the receiver-operating characteristic (AUROC) curve was used as a performance analysis procedure for three cancer risk models once multivariable logistic regression was applied: the first risk model invloved only clinical risk factors, the second model had factors involving measures related to breast density & the third model involved using both clinical risk factors and measures related to breast density. Shravya et al. [24] in her study, proposed and examined the performance of 3 machine learning algorithms in an attempt to improve the performance of predictive models at diagnosing cancer. Out of the 3 classifiers namely Logistic Regression, K-Nearest Neighbors along with Support Vector Machine, used for their experiments, their results indicated that SVM, with an accuracy of 92.7%, was the best for predictive analysis.

To predict breast cancer risks for patients in Nigeria, William et al. [25] used two data mining techniques: Naïve Bayes' and decision tree(J48) algorithms. On evaluation of these techniques, J48 decision tree was the most effective model in terms of predictive analysis with 94.2% accuracy.

Assiri et al. [26], in their research proposed an ensemble classification mechanism on the basis of a majority voting mechanism. The performance of many state-of-the-art machine learning classification models were evaluated on data from the Wisconsin Breast Cancer dataset [1]. Their proposed classifier performed better than the state-of-the-art algorithm, achieving an accuracy of 99.42%.

Darzi et al. [27], in his study for diagnosis of breast cancer demonstrated a process involving genetic algorithm (GA) as well as case-based reasoning (CBR). The former was used in order to look for all likely subsets of features while the latter was used to approximate each subset's evaluation outcome. The results indicated that the proposed model achieved an accuracy value of 97.37%, after performing feature selection.

Marrone et al. [28] in their work proposed to build a completely automated geometrical-based breast-mask extraction method in Dynamic Contrast Enhanced - Magnetic Resonance Imaging using three 2D fuzzy c-means clustering (FCM) with geometrical breast anatomy characterization. To precisely segment breast parenchyma from air and chest-wall, seven well defined key points were taken into account. Their proposed approach has been tested on 30 DCE-MRI studies so far. Rundo et al. [29], in their research utilized the fuzzy c-means algorithm as well for automatic necrosis extraction (*NeXt*) after the GTV segmentation. This unsupervised machine learning method identifies and characterizes the necrotic regions in heterogeneous cancers.

Asri et. al [30] in their study to gauge the accurateness in classifying data with respect to efficiency, conducted a performance -comparison analysis of the following classifier algorithms namely Decision Tree (C4.5), Support Vector Machine (SVM), K Nearest Neighbors (k-NN) and Naive Bayes (NB). This analysis was performed on data from the Wisconsin Breast Cancer dataset [1] with a WEKA data mining tool. From their results, it was indicated that SVM performed the best achieving an overall accuracy of 97.13%.

Most of the literature review has conducted a performance analysis of the classifiers based on accuracy. Nonetheless, measuring the performance based on precision, recall as well as F-score is equally significant since missing out on a condition could have serious implications on patients' health.

3. DESCRIPTIVE ANALYTICS | EXPLORATORY DATA ANALYSIS

This section is a deep dive into the exploratory analysis conducted using descriptive analytics as well as data visualization to help identify some patterns, or relationships that could help shed light on hidden information in the data available to us.

D. Data Acquisition

Data for this research is acquired from the UCI machine learning repository. It defines features that were computed from digitized images of a fine needle aspirant (FNA) of a breast mass. These features report the characteristics of the cell nuclei from the image. This dataset comprises of 569 samples of breast cell tissue which were medically assessed focusing on 32 characteristics of the cell nucleus.

E. Overview of Data Variables

This dataset comprises of characteristics procured from digitized images of fine needle aspirant (FNA) of a breast mass. The ten real-valued attributes of each cell nuclei are described below:

- **radius**
  This attribute illustrates the mean of the distances from the center of the cell nucleus to points on the perimeter.
- **texture**
  This attribute illustrates the standard deviation of gray-scales values of the digitized images
- **perimeter**
- **area**
- **smoothness**
  This attribute illustrates the local variations in radius lengths
- **compactness**
- **concavity**
  This attribute illustrates the severity of the concave parts of the contour
- **concave points**
  This attribute illustrates the number of concave parts of the contour
- **symmetry**

- **fractal dimension**

F. Data Visualization

This segment of the report sheds light on details about the different characteristics of the digitized images that may be present in the dataset. With the help of this information, we can put together experiments to proceed with our analysis.

G. Data Analysis

To begin with, we first load the data from the csv file using the pandas library function into a dataframe and analyze the features before cleaning and preprocessing the data.

H. Data Cleaning and Selecting Relevant Features

Since the aim of our research is primarily to predict patient diagnosis, we will only be focusing on those features that help with the diagnosis. Certain feature columns of the dataset hold null values, and therefore have no contribution in the modeling process. By considering these factors, we can construct an efficient model.

After identifying and dropping the irrelevant features, we work on our exploratory data analysis.

I. Exploratory Analysis

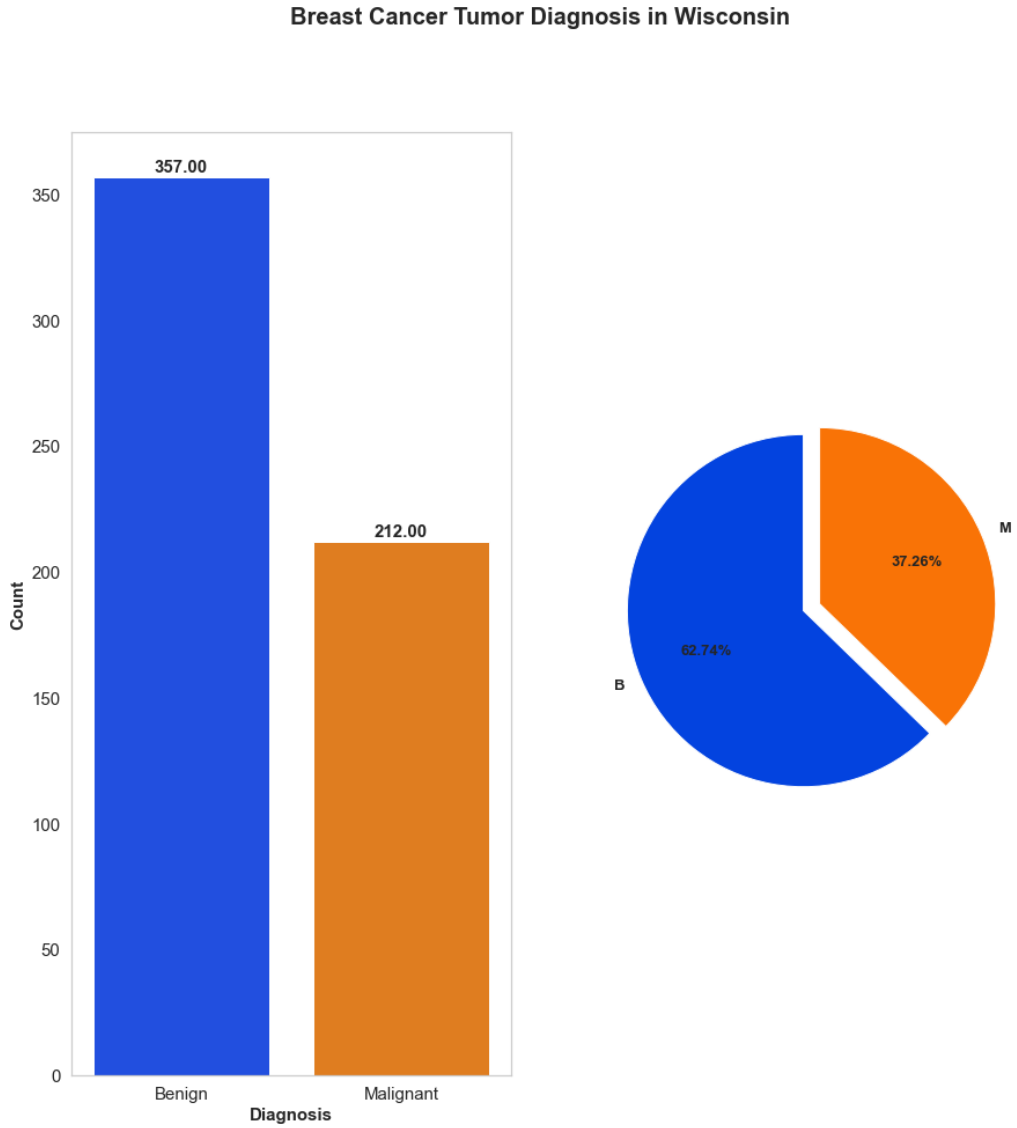

Figure 1: *Breast Cancer Tumor Diagnosis*

From Figure 1, we can see that out of 569 cases, 357 cases i.e., 62.74% of the total cases are Benign while 212 cases i.e., 37.26% are Malignant.

The dataset is slightly imbalanced. If the dataset has an imbalance ratio of 60:40, we may not need to use undersampling or oversampling procedures.

Figures 2-4 highlight the correlation of the features with the 'Diagnosis' attribute.

- **Correlation of Mean Features with Diagnosis**

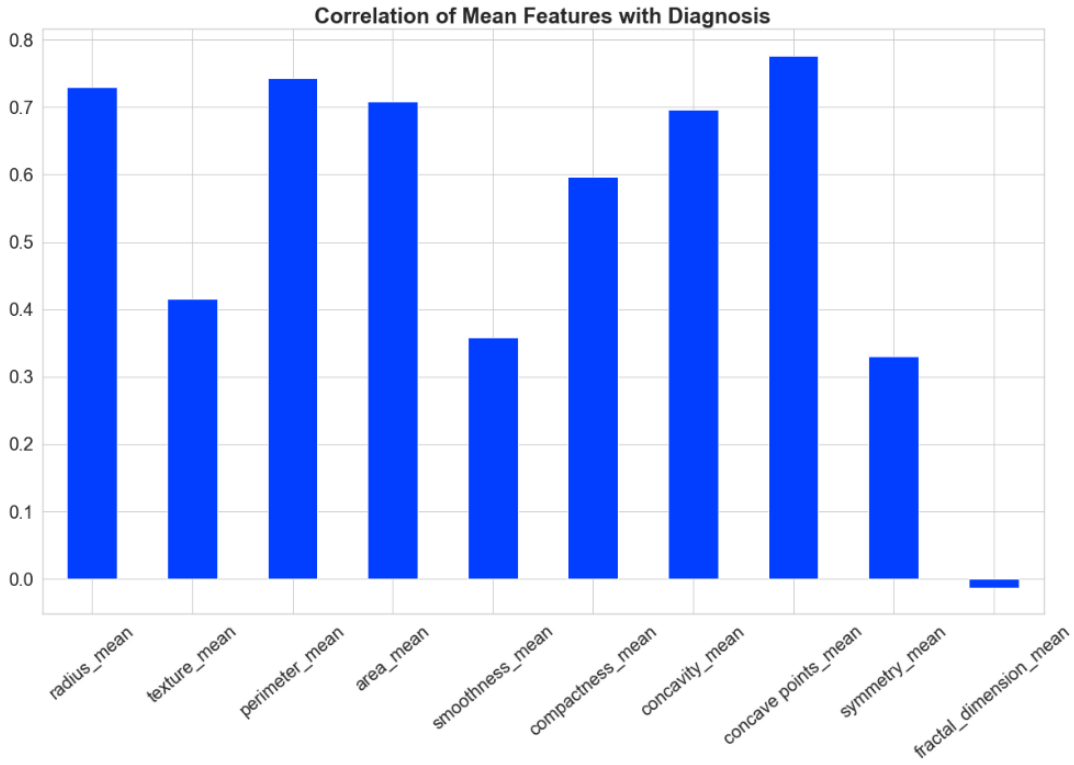

***Figure 2**: Correlation of Mean Features with Diagnosis*

From figure 2, we can see that the fractal_dimension_mean attribute is the weakliest correlated feature with the target variable while all the other features have significant correlation.

- **Correlation of Squared Error Features with Diagnosis**

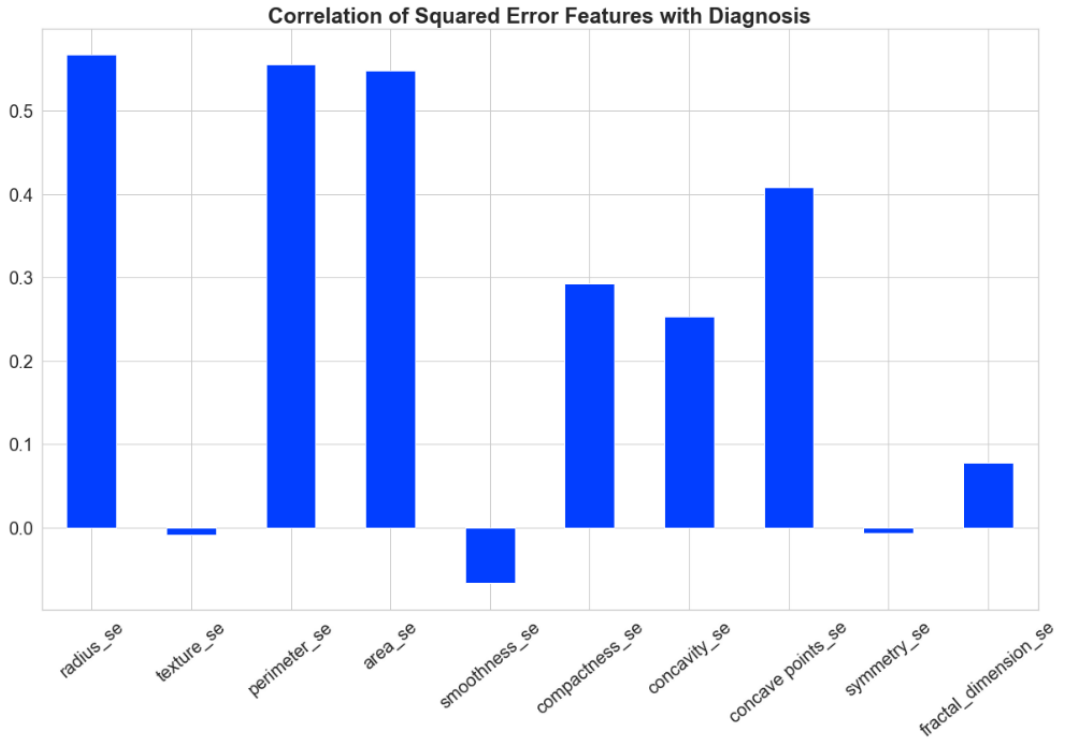

***Figure 3**: Correlation of Squared Error Features with Diagnosis*

From Figure 3, we can see that the features texture_se, smoothness_se and symmetry_se are weakly correlated with the target variable while all the other square errored features have significant correlation with the target variable 'Diagnosis'.

- **Correlation of Worst Features with Diagnosis**

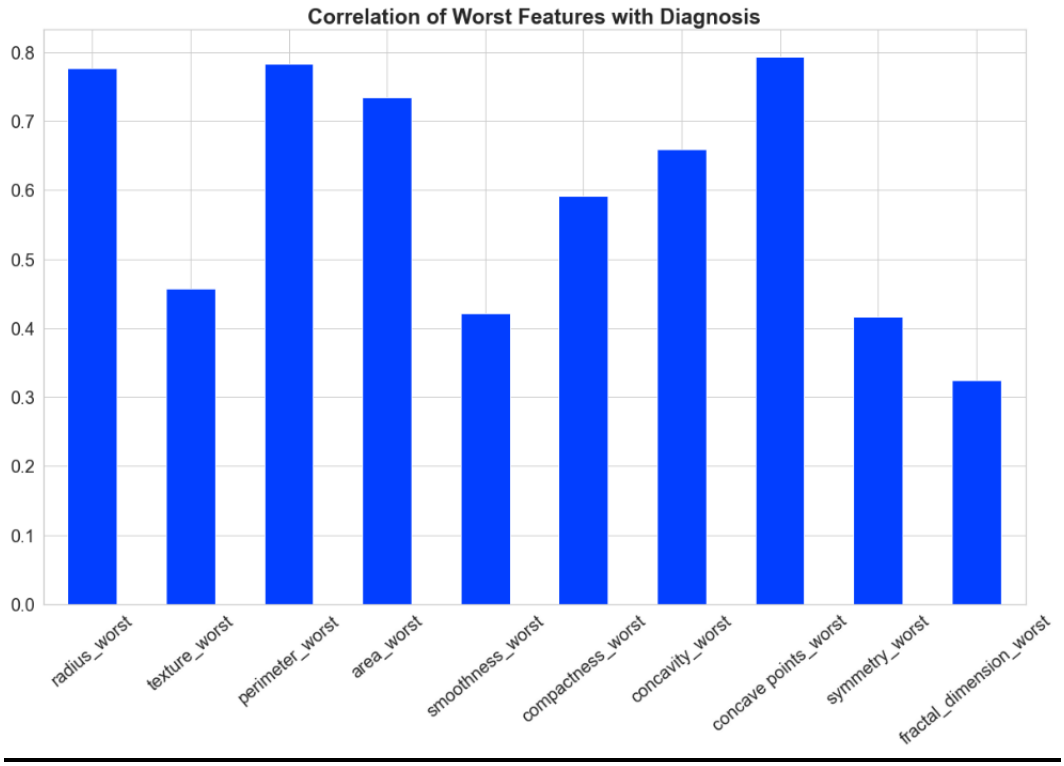

***Figure 4**: Correlation of Worst Features with Diagnosis*

From Figure 4, we can see that all the worst features have significant correlation with the target variable 'Diagnosis'.

Figures 5-7 illustrate violin plots depicting the distribution of the features based on diagnosis.

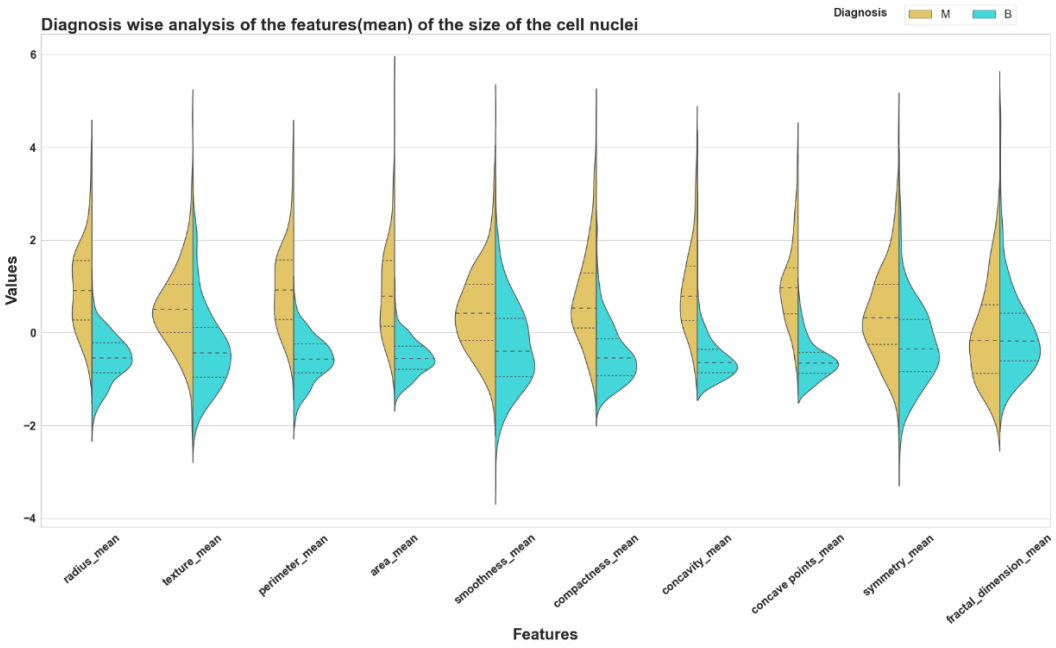

***Figure 5**: Diagnosis wise analysis of the mean features*

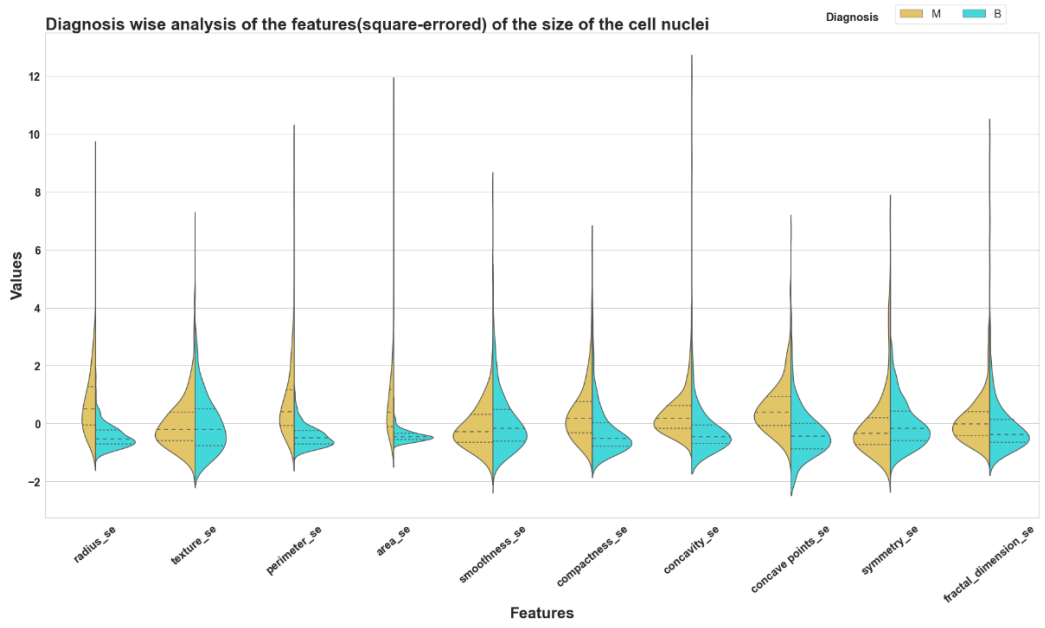

***Figure 6**: Diagnosis wise analysis of the squared error features*

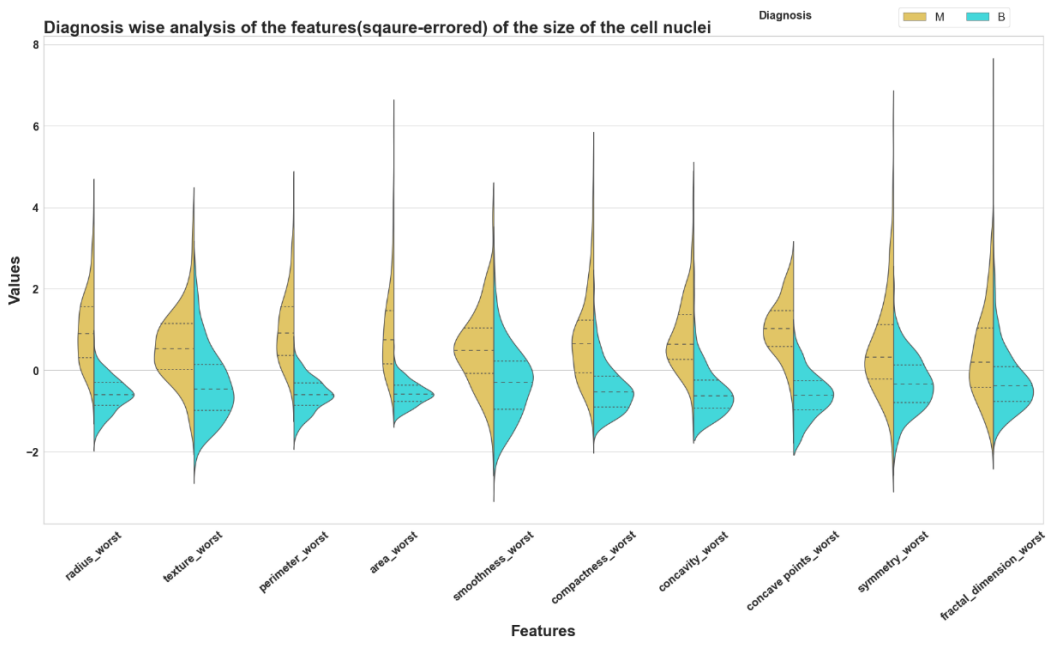

***Figure 7**: Diagnosis wise analysis of the worst features*

From Figure 5, we can see that the features radius_mean, perimeter_mean, area_mean, compactness_mean, concavity_mean, and concave_points_mean are well separated between Malignant and Benign tumors. Large values of these paramters seem to show a correlation with malignant tumors indicating these parameters would be good predicators for the classifer.

From Figure 7, we observe similarities between radius_worst and perimeter_worst. The violin plots for concavity_worst and concave points_worst are also very similar. If two violin plots seem identical, then there is a high correlation between the two features, hence one of the features must be dropped.

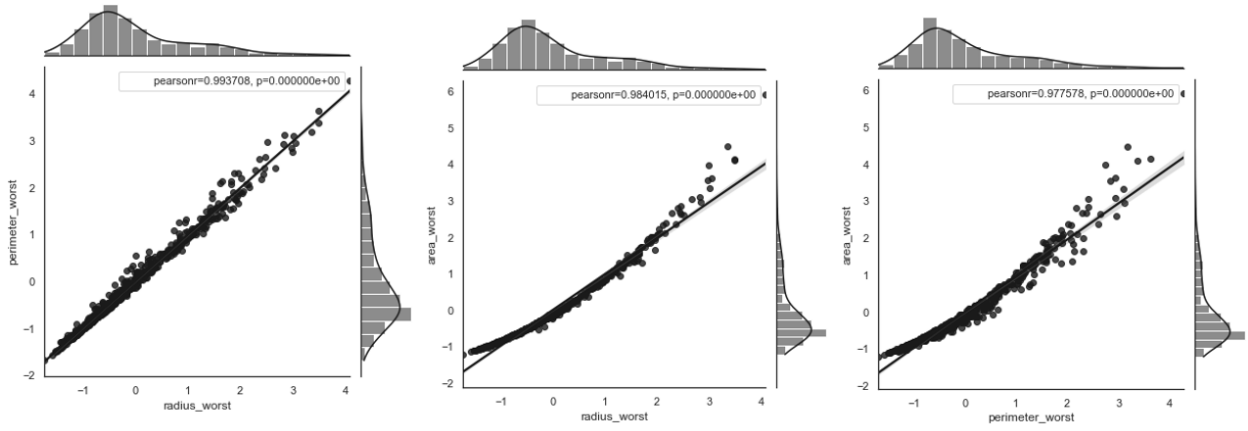

***Figure 8**: Joint plot of features to observe collinearity among features radius_worst, perimeter_worst and area_worst*

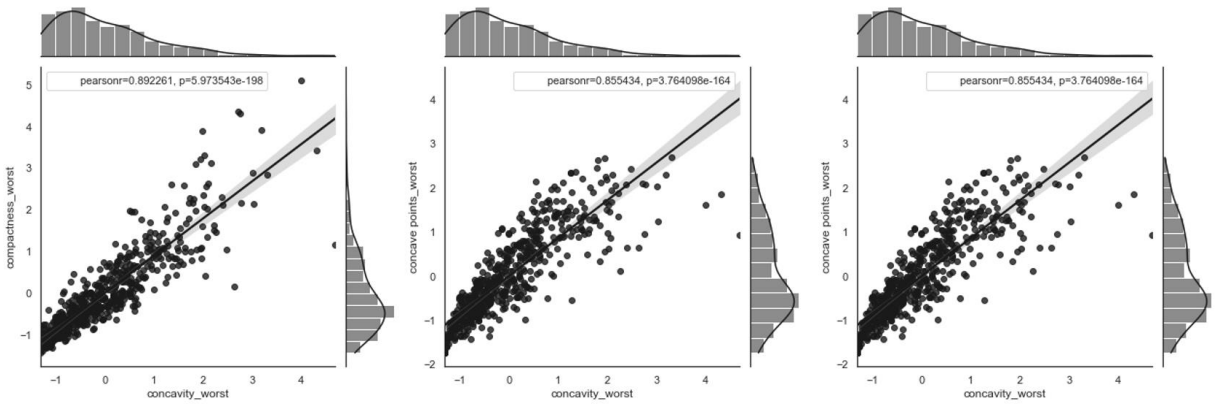

Figure 9*: Joint plot of features to observe collinearity among features compactness_worst, concavity_worst and concave points_worst*

From figures 8 and 9, we can see that radius_worst and perimeter_worst are strongly correlated as expected. The two features have a linear relation owing to the 2pi ratio between radius & perimeter of the cell nuclei. Concavity_worst and concave points_worst also have a strong correlation.

Figure 10 helps us observe the multicollinearity among distinct features of this dataset.

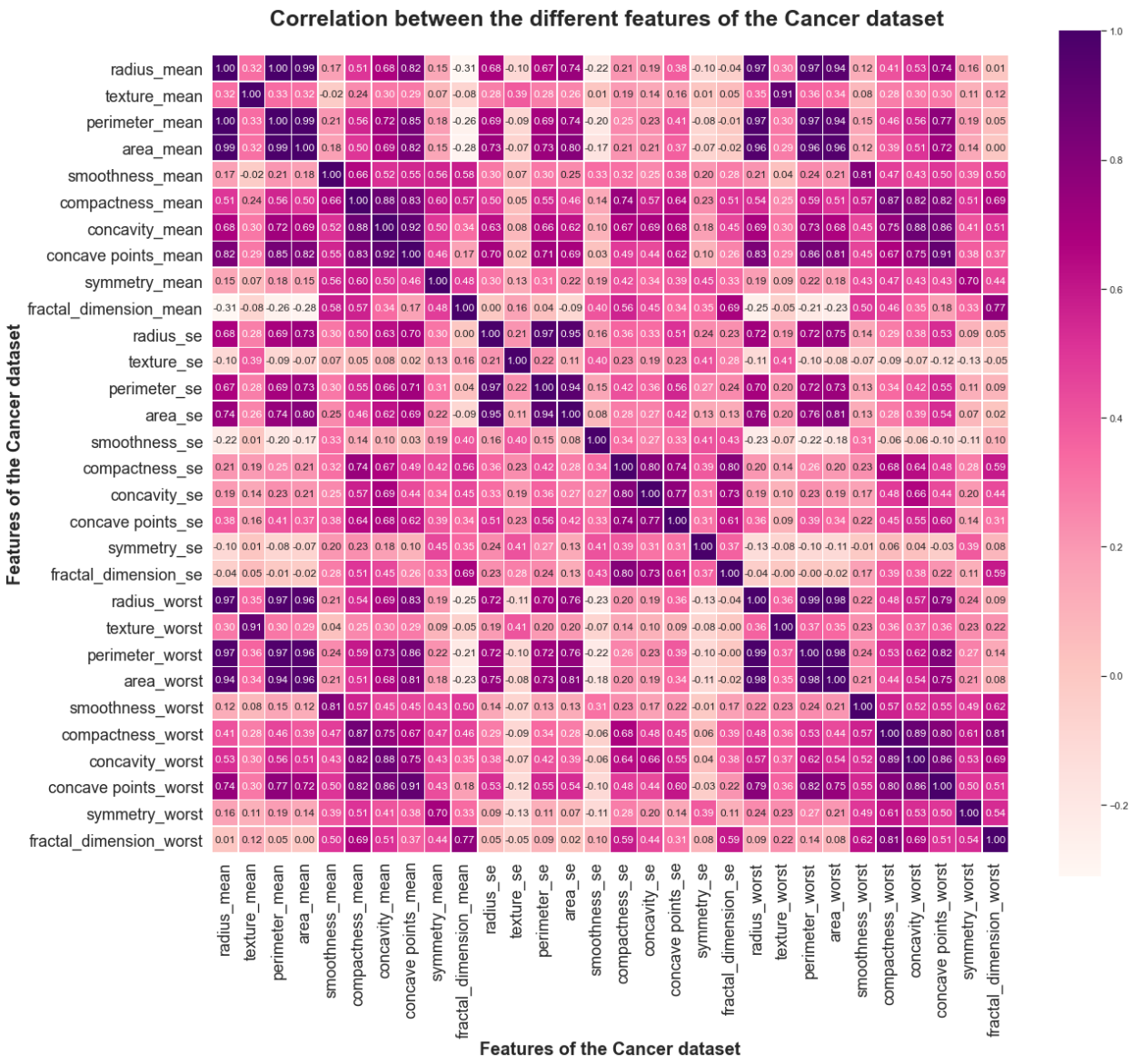

***Figure 10**: Correlation heatmap of the nucleus features*

From figure 10, it is quite clear that the features radius, perimeter and area are highly correlated. This indicates multicollinearity among these variables because all the 3 features describe similar details i.e., the physical size of the nuclei of the cell. Therefore, for going forward with our analysis, we should pick one of the three features. We will use area_mean because the violin plot for this feature indicates significant differences between malignant and benign tumors compared to the rest.

The features compactness, concavity and concave points also show a high correlation. Therefore, for going forward with our analysis, we should pick one of the three features. We will use concavity_mean.

Having redundant features can reduce the generalization capability of a model ultimately affecting the accuracy of the classifier

The same applies for square error and worst features of the dataset.

4. METHODOLOGY AND EXPERIMENTS

J. Aim of Study

The aim of this study is to train different classifier models that can help classify whether the tumor in the breast is benign or malignant and help predict the reoccurrence of cases classified as malignant by observing the most important features. To identify the best trained classifier, 7 different algorithms were applied and their statistic results were analyzed to determine the appropriate approach for this analysis.

K. Response (Dependent) and Independent Variable(s)

In this experiment, the feature **diagnosis** is the **response** or **dependent variable** which is the predicted variable. There are 10 **independent variables** describing characteristics of a breast tumor cell mass that help classify the diagnosis as benign or malignant.

a) radius
b) texture
c) perimeter
d) area
e) smoothness
f) compactness
g) concavity
h) concave points
i) symmetry
j) fractal dimension

Using different feature selection techniques, we identify those features that play a key role in the predictive modeling process.

L. Experimental Design – Setting Up the Experiment

a) Dataset

The Breast Cancer Wisconsin Dataset, available on the UCI Machine Learning repository defines features that were computed from digitized images of a fine needle aspirant (FNA) of a breast mass.

There are 569 rows in this dataset, each with 31 columns.

Attribute Information:

1. ID number 2. Diagnosis (M= Malignant, B= Benign)

b) Data Pre-Processing

To pre-process the data, we have implemented data cleaning by dropping unnecessary features and getting rid of null values, followed by data transformation and standardizing the data since the independent variables used in this process are numerical variables and have values ranging over different scales. If the data is not standardized, the features would have a large impact on the model due to a larger range of values. We have also implemented One Hot Encoding to transform the categorical feature **diagnosis** into individual features of 0 and 1.

c) Randomization (Train/Test Split)

In this experiment, the data of 569 patients was divided randomly into two sets: two-third i.e., 70% of the data was used as the learning set to develop a prediction model while the rest of the one-third i.e., 30% was used as the validation set for validating the developed model.

d) Cross Validation

For this experiment, we trained our models using 5-fold cross-validation, where the data was divided into 5 folds: 4 folds for training the model, while the remaining fold helped evaluate the performance of the model/generalizability.

M. Experiment Performance and Revisions

A few experiments were carried out in order to help the modeling process.

Details of the experiments are provided below.

a) *Experiment 1: Feature Selection*

The purpose of this experiment was to identify features that would contribute the most to our prediction variable. The reason being, having irrelevant features while modeling

can decrease the accuracy of our model and as a result, make our model learn based on irrelevant features. We have implemented the following Feature Selection techniques to go forward with our analysis:

i) Feature Selection using Correlation

After dropping redundant features from the correlation heatmap between different features of the dataset that displayed a high variance, the number of features were reduced from 30 to 16. To check if we have selected the right features, we used the Random Forest Classifier to compute the accuracy of the selected features.

ii) Univariate Feature Selection

By using Univariate feature selection, our aim is to select the best features in accordance with univariate statistical results where each feature will be compared to the target variable to recognize if they are closely related. This is also indicated as Analysis of Variance (ANOVA).

We have used SelectKBest method that selects only the k highest scoring features.

iii) Recursive Feature Elimination with Random Forest

Using Recursive Feature Elimination, we can choose inly those columns that have the most impact on the prediction of the target variable. Recursive Feature Elimination works by initially fitting the chosen machine learning model with all the features of the dataset. It then recursively removes the features, trains the model with the reduced set of attributes, computes the model performance and notes which features are important.

This procedure is repeated until the required number of features are selected.

iv) Recursive Feature Elimination using Cross-Validation with Random Forest

Recursive Feature Elimination, Cross-Validated works by selecting the best subset of attributes for the given estimator by removing 0 to N attributes, N being the number of features by means of Recursive Feature Elimination and then choosing the ideal subset on the basis of the model's cross-validation score.

v) Tree based Feature Selection (ExtraTrees)

Extremely Randomized Trees Classifier also referred to as Extra Trees Classifier is a kind of ensemble learning approach that is centrally based on decision trees. This technique summarizes the outcome of multiple correlated decision trees compiled in a 'forest' to yield the classification result.

vi) Vote based Feature selection

This feature selection technique makes use of a vote-based approach to decrease the number of features in the dataset

b) Experiment 2: Model Building and Validation

In this experiment, we evaluate the seven classification algorithms to classify the breast cancer tumor with the subset of features obtained from each feature selection technique in Experiment 1.

i) Logistic Regression
ii) Random Forest Classifier
iii) Gradient Boosting Classifier
iv) Extra Trees Classifier
v) XGB Classifier
vi) K-Nearest Neighbors (KNN)
vii) Support Vector Machine (SVM)

N. Measuring Classifier Performance

For each model, we have assessed a set of performance measures including precision, recall as well as Area Under the Curve. Although accuracy was also assessed, precision seemed like a better fit. For this experiment we have used traditional performance measures for classification to help classify malignant tumors from benign. These measures are based on four values of the confusion matrix: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). With these

values, we have calculated Positive Predictive Value (PPV) or precision as well as sensitivity or recall.

$$\text{Precision} = \sum TP / (\sum TP + \sum FP)$$

$$\text{Recall} = \sum TP / (\sum TP + \sum FN)$$

Furthermore, the Receiver Operating Characteristic Curve (ROC) was charted as a performance evaluator and the areas under the ROC curve (AUC) were assessed.

An overview of the metrics used:

1. Confusion Matrix

   In Machine Learning, a confusion matrix summarizes the prediction results on a classification problem. In simple words, it describes the manners in which the classification model is confused while forecasting.

2. AUC-ROC

   Area under the Receiver Operating Curve is a valued metric that helps gauge the performance of the classification models. This metric explains how capable the model is in distinguishing the classes. The model is considered to be highly effective in prediction if it has a higher AUC value. The AUC-ROC curve helps measure the performance of classification problems at different thresholds.

O. Algorithm Comparison and Selection

After having selected relevant features that contribute the most to our predicted variable using various feature selection techniques, training and testing different machine learning models for each feature selection technique based on performance evaluating metrics and identifying the best model based on these experiments, the best classifier and its corresponding feature selection technique was identified for further study and as the recommended model to use.

In addition, the accuracy and AUC score were used as metrics to gauge the performance of each model.

Further, hyperparameter tuning can be performed to try and improve the performance of the model on the dataset.

5. RESULTS AND DISCUSSION

P. Exploratory Analysis Results

The initial exploratory data analysis of the Wisconsin Breast Cancer data demonstrated some insightful findings that assisted us in modeling the data in the later phases. We discovered that more than 50% of the cases classified patients as diagnosed with benign tumors while 212 out of 569 patients were identified with malignant tumors.

Further, to identify the features that help in classifying the tumor, a deep dive into visualizing the correlation of all the features with the target variable was performed. We learnt that features such as the radius, perimeter, area, compactness, concavity and concave points of the breast cell nuclei seemed to show a high correlation with malignant tumors indicating these parameters to be good predicators for the classifier.

The exploratory data analysis also confirmed that only a few features were impactful in the decision-making process owing to the multicollinearity among different features of this dataset. The reason being that having redundant features can reduce the generalization capability of a model eventually affecting the model's accuracy.

For the modeling stage, we have implemented different Feature Selection techniques to get rid of the noise in data and incorporate only the selected features that improve the overall model performance.

Q. Machine Learning Experiment Results

A few experiments were conducted to assist in the modeling process and an overview of the results for these experiments are as outlined below:

a) Experiment 1: Feature Selection

   i) Feature Selection using Correlation

      After having dropped the highly correlated features with correlation value greater than 0.9, we obtain a new set of selected features. The number of features were reduced from 30 to 16 and to assess if the right set of features were

selected, the Random Forest classifier was used to compute the accuracy of the model with the selected features.

The Recall is 93.7%. Accuracy and F1 scores look good. This method of performing feature selection with correlation gives us a few incorrect predictions. The selected features using this feature selection approach are shown in Figure 11.

```
The selected features from Feature Selection using Correlation
 ['texture_mean' 'area_mean' 'smoothness_mean' 'concavity_mean'
 'symmetry_mean' 'fractal_dimension_mean' 'texture_se' 'area_se'
 'smoothness_se' 'concavity_se' 'symmetry_se' 'fractal_dimension_se'
 'smoothness_worst' 'concavity_worst' 'symmetry_worst'
 'fractal_dimension_worst']
```

Figure 11: Selected features from Feature Selection using correlation

ii) Univariate Feature Selection

For this approach, we will be basing our analysis by selecting k best features where each feature will be compared to the target variable. We have adopted the SelectKBest method and have opted to employ k=10 to proceed with our analysis.

The Recall is 94.7% while Accuracy and F1 scores are slightly lower compared to Feature Selection using Correlation. The selected features using this feature selection approach are shown in Figure 12.

```
The selected features from Univariate Feature Selection
 ['radius_mean' 'texture_mean' 'perimeter_mean' 'area_mean' 'perimeter_se'
 'area_se' 'radius_worst' 'texture_worst' 'perimeter_worst' 'area_worst']
```

***Figure 12**: Selected features from Univariate Feature Selection*

iii) Recursive Feature Elimination with Random Forest

With Recursive Feature Elimination, we can choose those columns that have a high level of contribution in the decision-making process and have the most impact on predicting the target variable. How this feature selection method works is it initially fits the model with all the features. It then recursively

eliminates features, trains the model with the new set of reduced features, assesses the performance of the model and notes the important features. We have selected n_features_to_select=16 to go ahead with our analysis.

The chosen 16 features by Recursive Feature Elimination are different from the naïve approach as shown in Figure 13, therefore the accuracy and recall for this method will be different as well.

The recall is 96.5% which is slightly better than the previous naïve approach for the same Random Forest classifier with the same random_state value and test set.

```
Chosen best 16 features by rfe:
 Index(['texture_mean', 'perimeter_mean', 'area_mean', 'compactness_mean',
        'concavity_mean', 'concave points_mean', 'area_se', 'radius_worst',
        'texture_worst', 'perimeter_worst', 'area_worst', 'smoothness_worst',
        'compactness_worst', 'concavity_worst', 'concave points_worst',
        'symmetry_worst'],
       dtype='object')
```

***Figure 13**: Selected features from Recursive Feature Elimination*

iv) Recursive Feature Elimination using Cross-Validation with Random Forest

A special attribute of the Random Forest Classifier namely feature_importances allows us to evaluate the importance of the features in each iteration. In this approach, feature importance is computed based on the estimator selected, and eventually a few features will be dropped in each iteration.

The attribute n_features reports the number of features that are important and make a significant impact on predicting the dependent variable. The ranking property specifies the order of importance for each attribute.

With this approach, we are not only able to identify the best features that contribute to the overall performance of the model but also decipher how many

features does a model need to achieve the highest accuracy. The selected features using this feature selection approach are shown in Figure 14.

```
Best features by rfecv: Index(['radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean',
       'compactness_mean', 'concavity_mean', 'concave points_mean',
       'radius_se', 'perimeter_se', 'area_se', 'fractal_dimension_se',
       'radius_worst', 'texture_worst', 'perimeter_worst', 'area_worst',
       'smoothness_worst', 'compactness_worst', 'concavity_worst',
       'concave points_worst', 'symmetry_worst', 'fractal_dimension_worst'],
```

***Figure 14**: Selected features from Recursive Feature Elimination using Cross-Validation*

A graphical representation of the features against their cross-validation scores is shown in Figure 15.

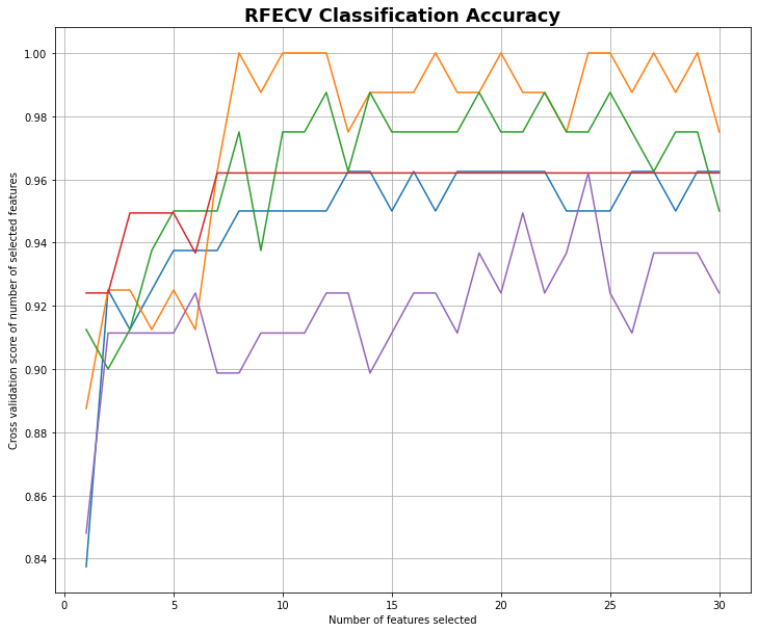

***Figure 15**: RFECV Classification accuracy of the features*

From this plot, we can see that keeping 8, 10, 11, 12, 16, 18, 20, 25, 25 or sometimes 27 features approximately give the same accuracy.

Since the number of optimal features fluctuates a lot for the same classifier, it is quite erroneous to base our understanding of the optimal feature selection on a single occurrence of the cross validation.

v) Tree based Feature Selection

With Extra Tress classifier, we can fit a number of randomized decision trees to the data, which is an approach of ensemble learning. Random splits of the observations assure that the model does not overfit the data.

A further analysis of feature importance helped us understand what the most important features were according to the model. The feature importance results are demonstrated in Figure 16.

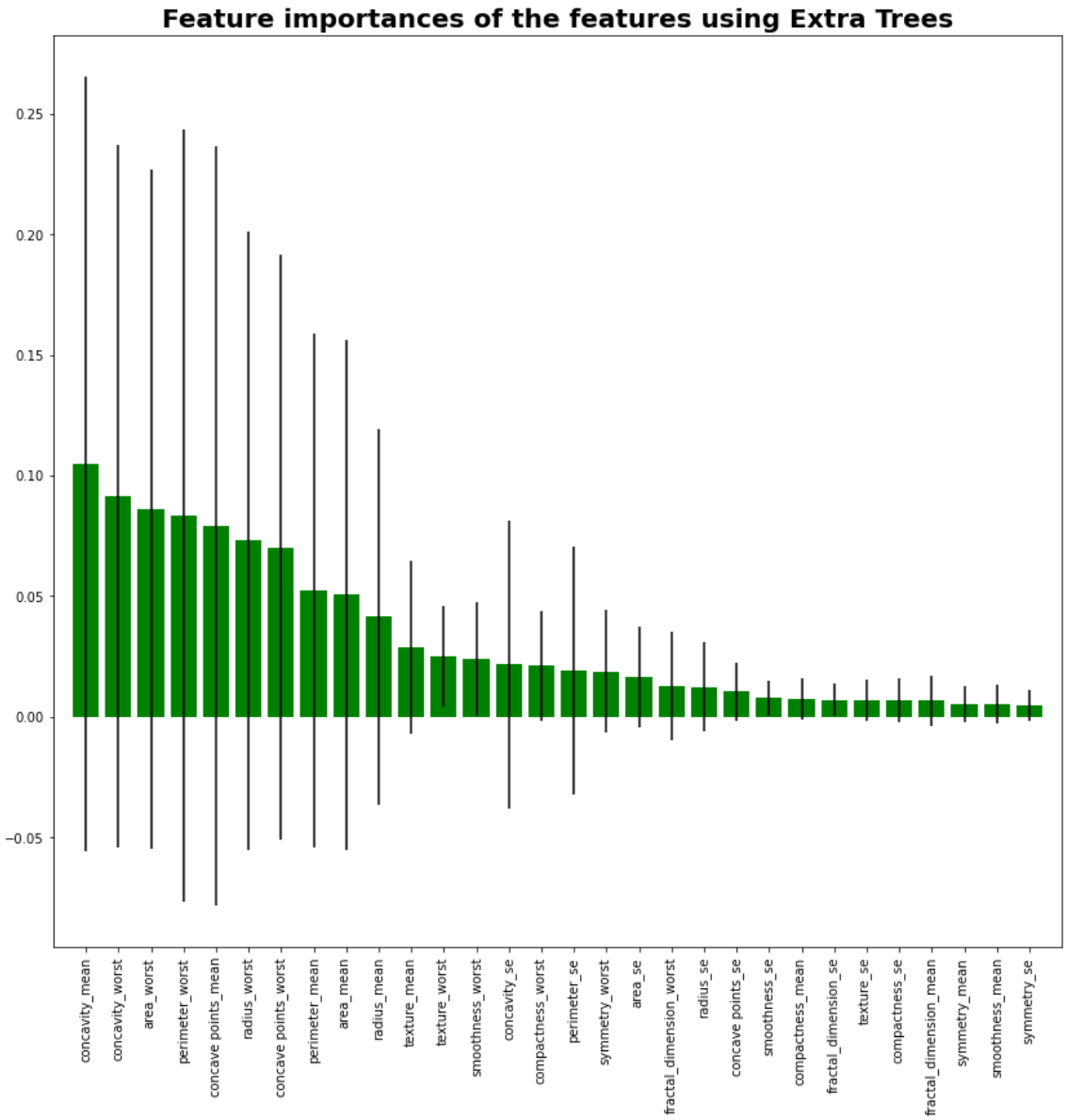

***Figure 16**: Feature Importance Graph*

The selected features using this feature selection approach is shown in Figure 17.

```
Total features: 30
Selected features: 10
Best features by ExtraTrees: Index(['radius_mean', 'perimeter_mean', 'area_mean', 'concavity_mean',
       'concave points_mean', 'radius_worst', 'perimeter_worst', 'area_worst',
       'concavity_worst', 'concave points_worst'],
      dtype='object')
```

Figure 17: *Selected features using Tree based Feature Selection (Extra Trees)*

vi) Vote based Feature Selection

This technique employs a vote-based approach to reduce the number of features in the dataset and selects those features with a final score greater than or equal to 2. The results of this approach are show in Figure 8.

	index	Chi_Square	Extratrees	RFE	RFECV	Final_score
0	area_worst	1	1	1	1	4
1	area_mean	1	1	1	1	4
3	perimeter_worst	1	1	1	1	4
4	perimeter_mean	1	1	1	1	4
5	radius_worst	1	1	1	1	4
9	texture_mean	1	0	1	1	3
14	concave points_worst	0	1	1	1	3
12	concavity_mean	0	1	1	1	3
10	concavity_worst	0	1	1	1	3
15	concave points_mean	0	1	1	1	3
8	texture_worst	1	0	1	1	3
2	area_se	1	0	1	1	3
6	radius_mean	1	1	0	1	3
18	symmetry_worst	0	0	1	1	2
16	compactness_mean	0	0	1	1	2
20	smoothness_worst	0	0	1	1	2
13	compactness_worst	0	0	1	1	2
7	perimeter_se	1	0	0	1	2
11	radius_se	0	0	0	1	1
22	fractal_dimension_worst	0	0	0	1	1
25	fractal_dimension_se	0	0	0	1	1
17	concavity_se	0	0	0	0	0
19	compactness_se	0	0	0	0	0
21	concave points_se	0	0	0	0	0
23	symmetry_mean	0	0	0	0	0
24	smoothness_mean	0	0	0	0	0
26	smoothness_se	0	0	0	0	0
27	texture_se	0	0	0	0	0
28	symmetry_se	0	0	0	0	0
29	fractal_dimension_mean	0	0	0	0	0

***Figure 18**: Vote based Feature Selection*

The selected features using this feature selection approach are shown in Figure 9.

```
['area_worst', 'area_mean', 'area_se', 'perimeter_worst', 'perimeter_mean', 'radius_worst',
'radius_mean', 'perimeter_se', 'texture_worst', 'texture_mean', 'concavity_worst',
'concavity_mean', 'compactness_worst', 'concave points_worst', 'concave points_mean',
'compactness_mean', 'symmetry_worst', 'smoothness_worst']
```

***Figure 19**: Selected features using Vote based approach*

b) *Experiment 2: Model Building and Validation*

Here we evaluate the performance of seven classification algorithms to classify the breast cancer tumor with the subset of features obtained from each feature selection technique in Experiment 1.

Training and testing the performance of the data for each feature selection method

1. Feature Selection technique: Correlation

Table 1: *Training Performance for Feature Selection using Correlation*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.932	0.974	0.917	0.937	0.892	0.981
Gradient Boosting Classifier	0.022	0.59	0.888	0.913	0.866	0.977
KNeighbors Classifier	0.937	0.934	0.912	0.95	0.879	0.959
Logistic Regression	0.935	0.976	0.907	0.949	0.871	0.984
Random Forest Classifier	0.932	0.975	0.913	0.943	0.886	0.979
SVM Classifier	0.942	0.976	0.917	0.951	0.892	0.984

Table 2*: Testing Performance for Feature Selection using Correlation*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.965	0.956	0.983	0.921	0.951
Gradient Boosting Classifier	0.942	0.934	0.934	0.905	0.919
KNeighbors Classifier	0.965	0.959	0.967	0.937	0.952
Logistic Regression	0.965	0.959	0.967	0.937	0.952
Random Forest Classifier	0.959	0.951	0.967	0.921	0.943
SVM Classifier	0.965	0.959	0.967	0.937	0.952
XGB Classifier	0.953	0.946	0.951	0.951	0.935

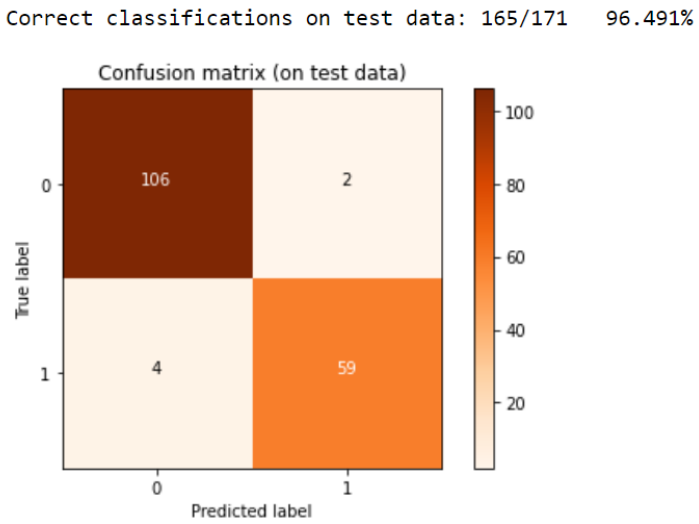

***Figure 20**: Confusion Matrix for Model: SVM Classifier, Feature Selection: Correlation*

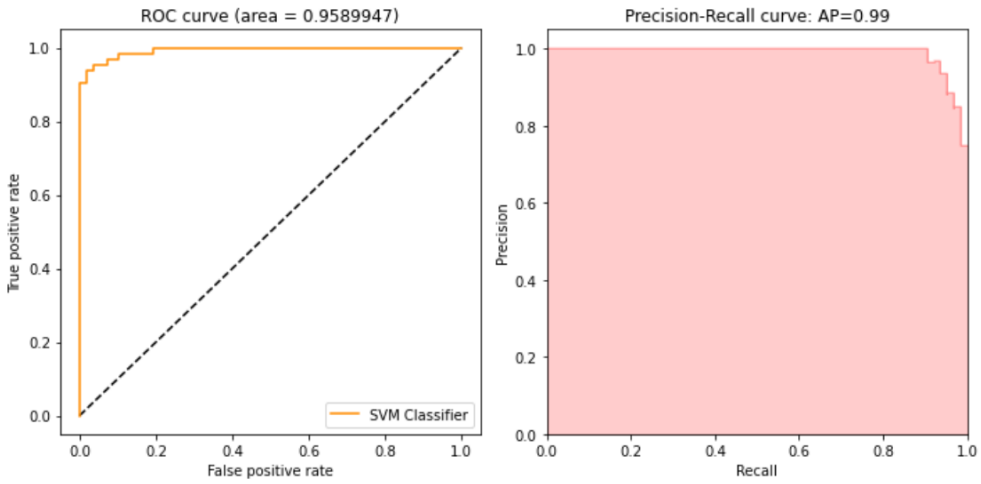

***Figure 21**: ROC Curve for Model: SVM Classifier, Feature Selection: Correlation*

2. Feature Selection technique: Univariate Feature Selection using Chi-Square

***Table 3**: Training Performance for Univariate Feature Selection*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.939	0.975	0.913	0.943	0.899	0.983
Gradient Boosting Classifier	0.922	0.959	0.892	0.914	0.859	0.977

KNeighbors Classifier	0.937	0.934	0.912	0.95	0.879	0.959
Logistic Regression	0.934	0.974	0.907	0.949	0.871	0.984
Random Forest Classifier	0.937	0.974	0.909	0.951	0.892	0.982
SVM Classifier	0.942	0.975	0.92	0.951	0.892	0.983
XGB Classifier	0.927	0.969	0.899	0.928	0.872	0.978

***Table 4**: Testing Performance for Univariate Feature Selection*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.959	0.951	0.967	0.921	0.943
Gradient Boosting Classifier	0.942	0.934	0.934	0.905	0.919
KNeighbors Classifier	0.965	0.959	0.967	0.937	0.952
Logistic Regression	0.965	0.959	0.967	0.937	0.952
Random Forest Classifier	0.965	0.956	0.921	0.921	0.951
SVM Classifier	0.965	0.959	0.937	0.937	0.952
XGB Classifier	0.965	0.962	0.952	0.952	0.952

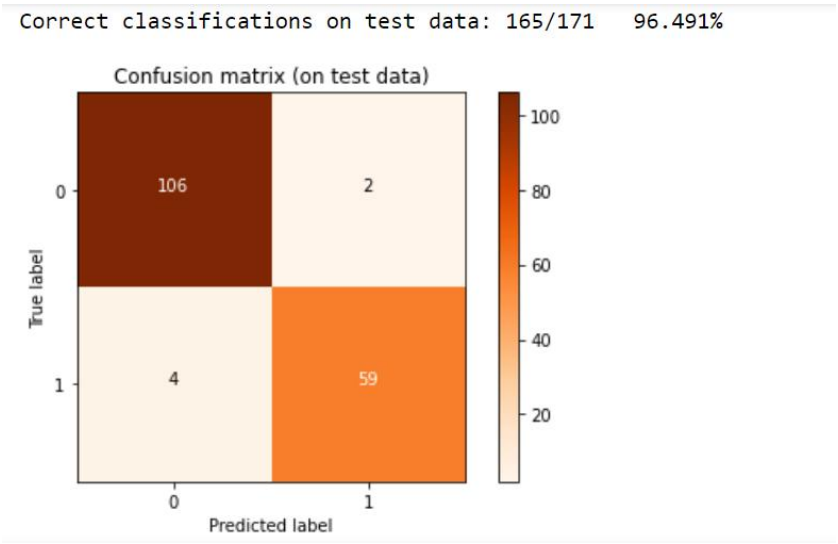

***Figure 22**: Confusion Matrix for Model: SVM Classifier, Feature Selection: Univariate Feature Selection*

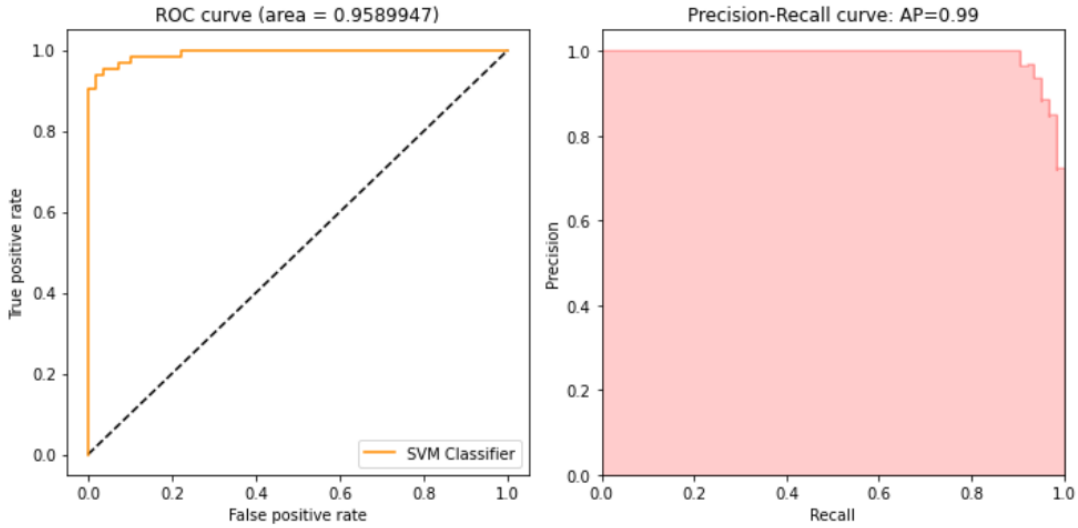

***Figure 23**: ROC curve for Model: SVM Classifier, Feature Selection: Univariate Feature Selection*

3. Feature Selection technique: Recursive Feature Elimination with Random Forest

***Table 5**: Training Performance for Recursive Feature Elimination*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.959	0.989	0.954	0.973	0.932	0.988

Gradient Boosting Classifier	0.942	0.979	0.921	0.939	0.906	0.979
KNeighbors Classifier	0.967	0.976	0.955	0.966	0.946	0.982
Logistic Regression	0/975	0.989	0.965	0.98	0.953	0.989
Random Forest Classifier	0.962	0.983	0.938	0.954	0.919	0.988
SVM Classifier	0.975	0.992	0.965	0.98	0.953	0.993
XGB Classifier	0.949	0.985	0.931	0.941	0.926	0.986

***Table 6**: Testing Performance for Recursive Feature Elimination*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.977	0.972	0.984	0.952	0.968
Gradient Boosting Classifier	0.971	0.964	0.983	0.937	0.959
KNeighbors Classifier	0.965	0.959	0.967	0.937	0.951
Logistic Regression	0.982	0.976	0.98	0.952	0.976
Random Forest Classifier	0.971	0.964	0.983	0.937	0.959
SVM Classifier	0.982	0.979	0.984	0.968	0.976
XGB Classifier	0.977	0.972	0.983	0.952	0.968

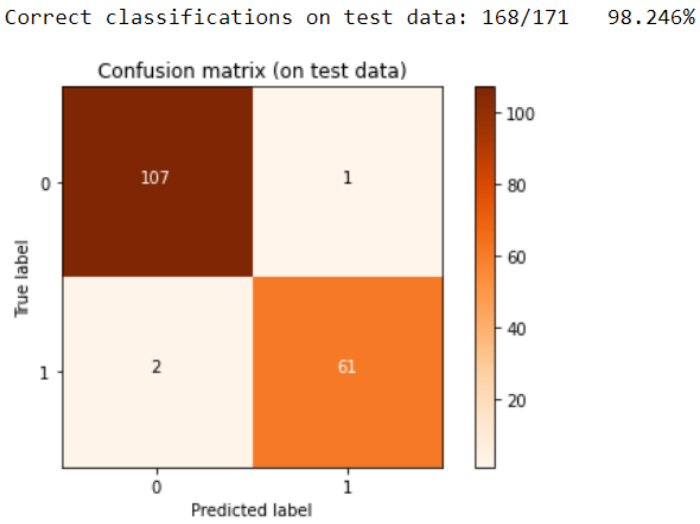

***Figure 24**: Confusion Matrix for Model: SVM Classifier, Feature Selection: RFE*

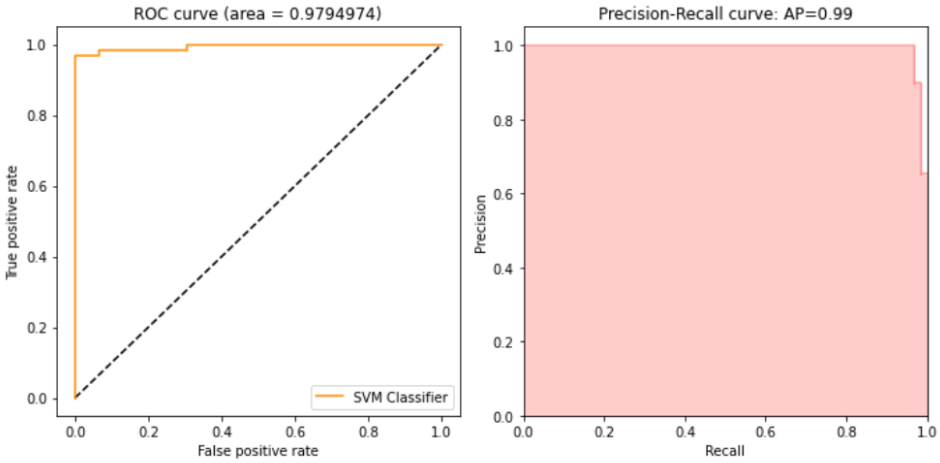

***Figure 25**: ROC curve for Model: SVM Classifier, Feature Selection: RFE*

4. Feature Selection technique: Recursive feature Elimination using Cross -Validation with Random Forest

***Table 7**: Training Performance for Recursive Feature Elimination using Cross Validation*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.957	0.987	0.94	0.973	0.932	0.99
Gradient Boosting Classifier	0.942	0.979	0.929	0.939	0.919	0.979
KNeighbors Classifier	0.97	0.977	0.962	0.979	0.946	0.982
Logistic Regression	0.97	0.99	0.958	0.979	0.939	0.991
Random Forest Classifier	0.955	0.983	0.939	0.953	0.919	0.989
SVM Classifier	0.97	0.993	0.958	0.979	0.952	0.995
XGB Classifier	0.957	0.985	0.941	0.948	0.939	0.986

***Table 8**: Testing Performance for Recursive Feature Elimination using Cross Validation*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.977	0.972	0.984	0.952	0.968
Gradient Boosting Classifier	0.971	0.964	0.983	0.937	0.959
KNeighbors Classifier	0.965	0.959	0.967	0.937	0.952
Logistic Regression	0.982	0.976	1	0.952	0.976
Random Forest Classifier	0.977	0.972	0.984	0.952	0.968

SVM Classifier	0.988	0.984	1	0.968	0.984
XGB Classifier	0.982	0.976	1	0.952	0.976

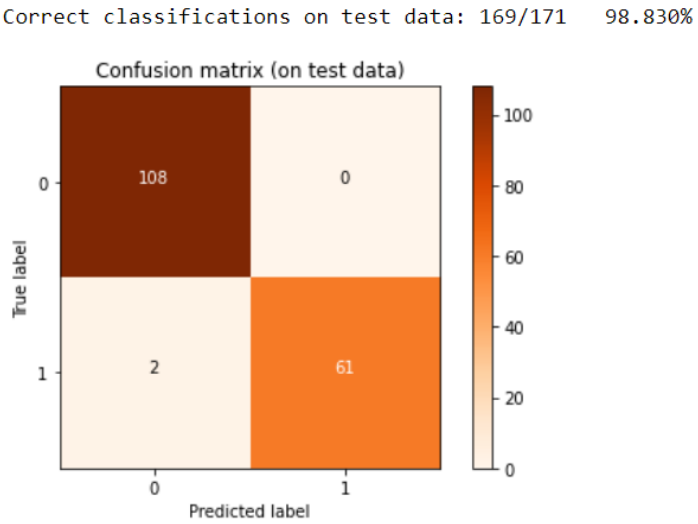

***Figure 26**: Confusion Matrix for Model: SVM Classifier, Feature Selection: RFECV*

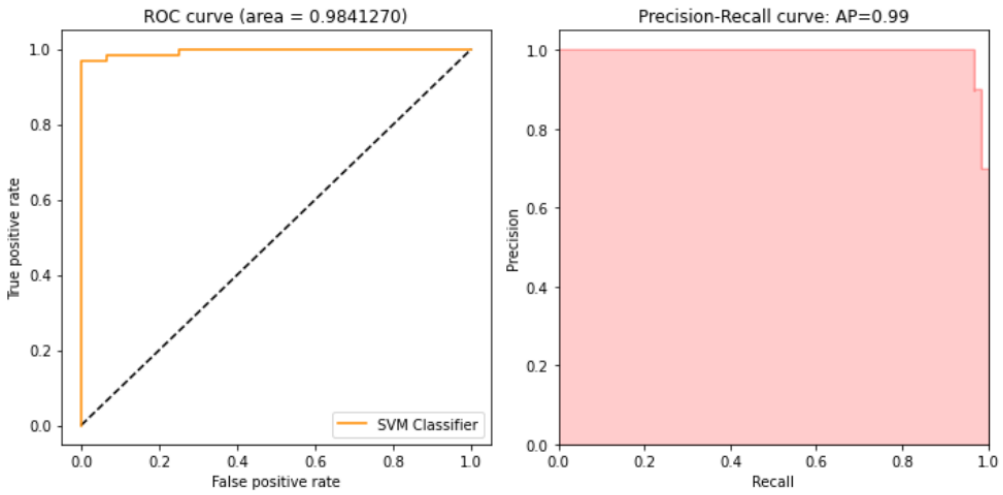

***Figure 27**: ROC curve for Model: SVM Classifier, Feature Selection: RFECV*

5. Feature Selection technique: Tree based Feature Selection (Extra Trees)

***Table 9**: Training Performance for Tree based Feature Selection*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.947	0.975	0.931	0.96	0.912	0.978
Gradient Boosting Classifier	0.939	0.971	0.917	0.945	0.892	0.974
KNeighbors Classifier	0.942	0.947	0.919	0.957	0.886	0.965
Logistic Regression	0.947	0.979	0.927	0.951	0.906	0.983
Random Forest Classifier	0.945	0.976	0.92	0.953	0.912	0.981
SVM Classifier	0.47	0.979	0.925	0.971	0.886	0.98
XGB Classifier	0.939	0.973	0.917	0.932	0.906	0.972

***Table 10**: Testing Performance for Tree based Feature Selection*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.971	0.967	0.968	0.952	0.96
Gradient Boosting Classifier	0.953	0.946	0.951	0.921	0.935
KNeighbors Classifier	0.947	0.945	0.922	0.937	0.929
Logistic Regression	0.965	0.959	0.967	0.937	0.952
Random Forest Classifier	0.965	0.959	0.967	0.937	0.952
SVM Classifier	0.953	0.939	0.982	0.889	0.933
XGB Classifier	0.953	0.946	0.951	0.921	0.935

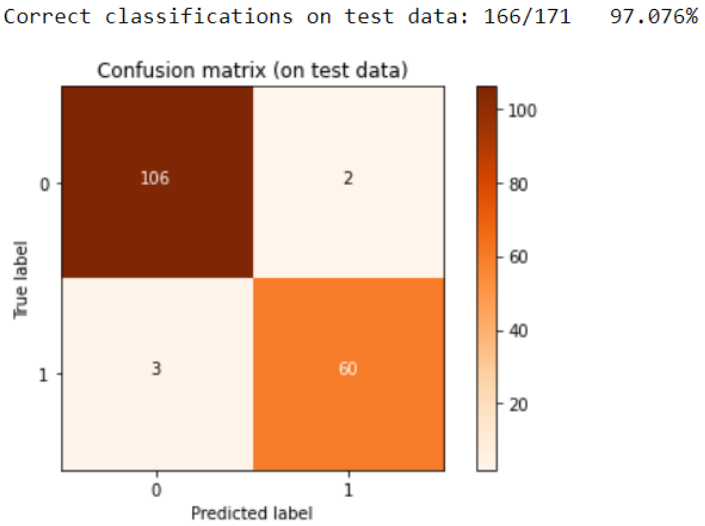

***Figure 28**: Confusion Matrix for Model: Extra Trees Classifier, Feature Selection: Tree based*

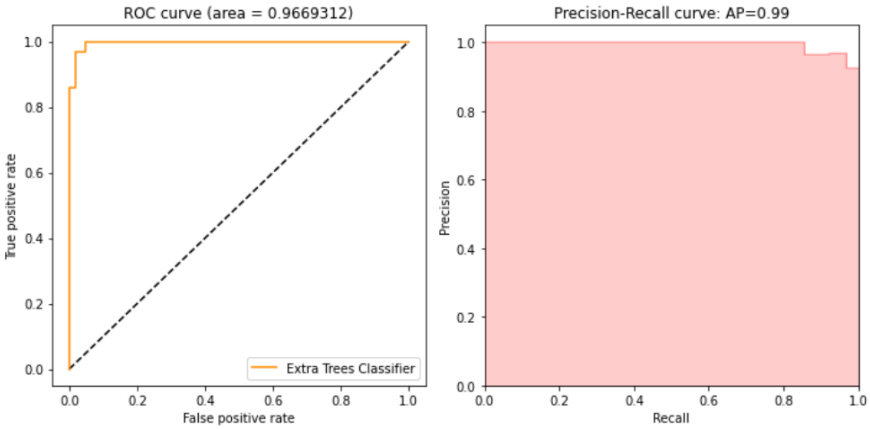

***Figure 29**: ROC curve for Model: Extra Trees Classifier, Feature Selection: Tree based*

6. Feature Selection technique: Vote-based

***Table 11**: Training Performance for Vote based Feature*

Training Performance: 5- fold mean Cross Validation Scores on Training Data						
Algorithm	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
Extra Trees Classifier	0.962	0.986	0.954	0.967	0.926	0.991
Gradient Boosting Classifier	0.942	0.981	0.925	0.939	0.912	0.982

KNeighbors Classifier	0.967	0.976	0.955	0.966	0.946	0.982
Logistic Regression	0.975	0.989	0.965	0.98	0.953	0.989
Random Forest Classifier	0.959	0.983	0.942	0.961	0.926	0.988
SVM Classifier	0.975	0.992	0.965	0.979	0.952	0.993
XGB Classifier	0.955	0.985	0.938	0.948	0.932	0.986

***Table 12**: Testing Performance for Vote based Feature Selection*

Testing Performance					
Algorithm	**Accuracy**	**AUC**	**Precision**	**Recall**	**F1**
Extra Trees Classifier	0.971	0.964	0.983	0.937	0.959
Gradient Boosting Classifier	0.971	0.964	0.983	0.937	0.959
KNeighbors Classifier	0.965	0.958	0.967	0.937	0.952
Logistic Regression	0.982	0.976	0.983	0.952	0.976
Random Forest Classifier	0.977	0.972	1	0.952	0.968
SVM Classifier	0.982	0.979	0.984	0.968	0.976
XGB Classifier	0.977	0.972	0.984	0.952	0.968

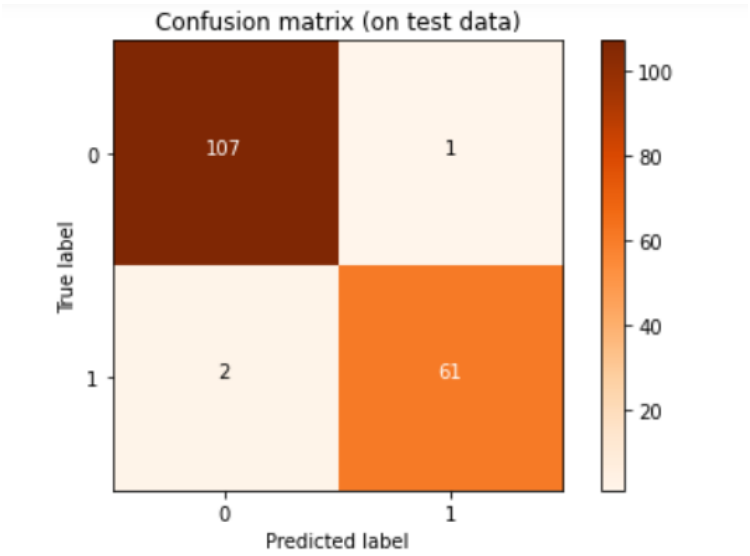

***Figure 30**: Confusion Matrix for Model: SVM Classifier, Feature Selection: Vote based*

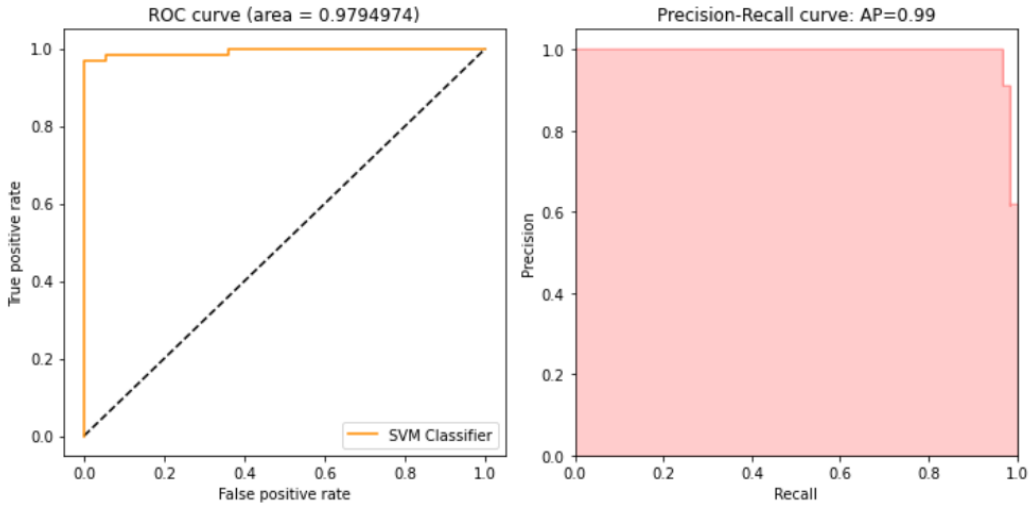

***Figure 31**: ROC curve for Model: SVM Classifier, Feature Selection: Vote based*

Comparing the best selected models from all the feature selection techniques

Based on the above 2 experiments, we have selected the best model from each feature selection technique. A summary of the detailed results has been demonstrated in Table 12 and Table 13.

***Table 13**: Summary of the Training Performance of the best classifiers from each Feature selection method*

Training Performance: 5- fold mean Cross Validation Scores on Training Data							
Algorithm	**Feature Selection**	**Accuracy**	**Average_ Precision**	**F1**	**Precision**	**Recall**	**ROC_AUC**
SVM Classifier	Correlation	0.942	0.976	0.917	0.951	0.892	0.984
SVM Classifier	Chi Square	0.942	0.975	0.92	0.951	0.892	0.983
SVM Classifier	RFE	0.975	0.992	0.965	0.98	0.953	0.993
SVM Classifier	RFECV	0.97	0.993	0.958	0.979	0.952	0.995
Extra Trees Classifier	Extra Trees	0.947	0.975	0.931	0.96	0.912	0.978
SVM Classifier	Vote based	0.975	0.992	0.965	0.979	0.952	0.993

***Table 14**: Summary of the Testing Performance of the best classifiers from each Feature selection method*

Testing Performance

Algorithm	Feature Selection	Accuracy	AUC	Precision	Recall	F1
SVM Classifier	Correlation	0.965	0.959	0.967	0.937	0.952
SVM Classifier	Chi Square	0.965	0.959	0.937	0.937	0.952
SVM Classifier	RFE	0.982	0.979	0.984	0.968	0.976
SVM Classifier	RFECV	0.988	0.984	1	0.968	0.984
Extra Trees Classifier	Extra Trees	0.971	0.967	0.968	0.952	0.96
SVM Classifier	Vote based	0.982	0.979	0.984	0.968	0.976

Applying the different feature selection techniques provided optimal results for this breast cancer dataset overcoming the issue of having redundant data and hence getting rid of the noise. After examining the summary results, it is evident that the SVM classifier generally performed better using all the feature selection techniques; however, we should keep in mind that it's not imperative that the same algorithm must produce optimal results for all binary classifiers.

R. Discussion

This project put to practice a machine learning approach to classify whether the tumor in the breast is benign or malignant and help predict the reappearance of cases classified as malignant by identifying those features that had a significant impact on the prediction of the target variable. We were able to implement different feature selection techniques to minimize redundancy and consequently build different machine learning models with improved accuracy.

According to our findings, Support Vector Machine outperformed all the other models in analysis and prediction of cancer with an accuracy 0.982 and AUC of 0.979 on the test data. Thus, the most significant features that proved to be beneficial in predicting benign and malignant tumors in breast cancer patients are **texture_mean, perimeter_mean, area_mean, compactness_mean, concavity_mean, concave points_mean, area_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst**. This was obtained using Recursive Feature Elimination that generated 16 features.

6. **CONCLUSION AND FUTURE WORK**

Helping women with early diagnosis of breast cancer could lower the risk of prolonged treatment and sometimes death. Since early-stage breast cancer is easier to treat considering the tumor isn't too large and has not spread, it is important that we develop a more efficient, accurate and reliable approach to do so. This can be achieved by dint of machine learning classifiers.

One of the goals of this research was to help women with early diagnosis by building an efficient machine learning model. In the field of medicine, the impact of false positive and false negative results has led to high-priced follow-up tests, being exposed to unneeded danger as well as treatments, but most of the existing studies have based their results mostly on the model's accuracy. For that reason, we have built a model whose performance is evaluated based on precision, recall as well as area under the curve. In this research we have executed different feature selection techniques to identify features that would contribute the most to our prediction variable. The performance of the different classifiers such as Logistic Regression, Random Forest, Support Vector Machine, K-Neighbors, Gradient Boosting, Extra Trees and XGB were compared using precision, recall as well as Area Under the Curve. Our results showed that Support Vector Machine outperformed all the other models in analysis and prediction of cancer with an accuracy 0.982 and AUC of 0.979 on the test data.

One of the future scopes that could be carried out on this research would be to examine how the proposed machine learning algorithms would perform on larger datasets. As a further step, it would be interesting to fine tune the hyperparameters of the existing model along with testing the model with deep learning procedures to classify the type of cancer and improve the model's accuracy.

7. APPENDIX – A | METRICS USED

Mean Absolute Error:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}i - yi)^2$$

Root Mean Square Error:

$$RMSE = \sqrt{MSE}$$

Precision:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F1-Score:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

8. APPENDIX – B | GITHUB LINK

S. Github Link

https://github.com/nehasunil21/Major-Research-Project

9. REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[2] Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 117693510600200030.

[3] Nguyen, Q. H., Do, T. T., Wang, Y., Heng, S. S., Chen, K., Ang, W. H. M., ... & Chua, M. C. (2019, July). Breast cancer prediction using feature selection and ensemble voting. In *2019 International Conference on System Science and Engineering (ICSSE)* (pp. 250-254). IEEE.

[4] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.

[5] H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on Principal Component Analysis," 2010 6th International Colloquium on Signal Processing & its Applications, 2010, pp. 1-4, doi: 10.1109/CSPA.2010.5545298.

[6] Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019, April). Machine learning classification techniques for breast cancer diagnosis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 495, No. 1, p. 012033). IOP Publishing

[7] Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat, “Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques,” Proc. in 2015 10th Asian Control Conf. (ASCC), pp 1-6, IEEE, 2015.

[8] Yesuf, S. H. (2019). BREAST CANCER DETECTION USING MACHINE LEARNING TECHNIQUES. *International Journal of Advanced Research in Computer Science*, *10*(5).

[9] Chen, X. W., & Jeong, J. C. (2007, December). Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)* (pp. 429-435). IEEE.

[10] Mandal, S. K. (2017). Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. *International Journal Of Engineering And Computer Science*, *6*(2), 20388-20391.

[11] Kumar, V., Mishra, B. K., Mazzara, M., Thanh, D. N., & Verma, A. (2020). Advances in Data Science and Management.

[12] Borges, L. R. (1989). Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection. *Group*, *1*(369), 15-19.

[13] Bharat, A., Pooja, N., & Reddy, R. A. (2018, October). Using machine learning algorithms for breast cancer risk prediction and diagnosis. In *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)* (pp. 1-4). IEEE.

[14] Ojha, Uma and Savita Goel. “A study on prediction of breast cancer recurrence using data mining techniques.” *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence* (2017): 527-530.

[15] Ghosh, S., Mondal, S., & Ghosh, B. (2014). A comparative study of breast cancer detection based on SVM and MLP BPN classifier. *2014 First International Conference on Automation, Control, Energy and Systems (ACES)*, 1-4.

[16] Osareh, A., & Shadgar, B. (2010). Machine learning techniques to diagnose breast cancer. *2010 5th International Symposium on Health Informatics and Bioinformatics*, 114-120.

[17] Bazazeh, D., & Shubair, R.M. (2016). Comparative study of machine learning algorithms for breast cancer detection and diagnosis. *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 1-4.

[18] Muhammad Sufyian Bin Mohd Azmi and Z. C. Cob, "Breast cancer prediction based on backpropagation algorithm," Proc. of 2010 IEEE Student Conf. on Research and Development (SCOReD 2010), pp 164-168

[19] B. M. Gayathri and C. P. Sumathi, "Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer," 2016 IEEE Int. Conf. on Computational Intelligence and Computing Research (ICCIC), pp 1-5, IEEE, 2016.

[20] Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality reduction using pca and k-means clustering for breast cancer prediction. *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi*, 192-201.

[21] Aavula, R., & Bhramaramba, R. (2019). XBPF: an extensible breast cancer prognosis framework for predicting susceptibility, recurrence and survivability. *Int. J. Eng. Adv. Technol*, *8*(5), 2249-8958.

[22] Wang, H., & Yoon, S. W. (2015). Breast cancer prediction using data mining method. In *IIE Annual Conference. Proceedings* (p. 818). Institute of Industrial and Systems Engineers (IISE).

[23] Abdolell, M., Tsuruda, K. M., Lightfoot, C. B., Payne, J. I., Caines, J. S., & Iles, S. E. (2016). Utility of relative and absolute measures of mammographic density vs clinical risk factors in evaluating breast cancer risk at time of screening mammography. *The British journal of radiology*, *89*(1059), 20150522.

[24] Shravya, C., Pravalika, K., & Subhani, S. (2019). Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(6), 1106-1110.

[25] Williams, K., Idowu, P. A., Balogun, J. A., & Oluwaranti, A. I. (2015). Breast cancer risk prediction using data mining classification techniques. *Transactions on Networks and Communications*, *3*(2), 01.

[26] Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, *6*(6), 39.

[27] Darzi, M., AsgharLiaei, A., & Hosseini, M. (2011). Feature selection for breast cancer diagnosis: a case-based wrapper approach. *International Journal of Biomedical and Biological Engineering*, *5*(5), 220-223.

[28] Marrone, S., Piantadosi, G., Fusco, R., Petrillo, A., Sansone, M., & Sansone, C. (2016, December). Breast segmentation using Fuzzy C-Means and anatomical priors in DCE-MRI. In *2016 23rd International Conference on Pattern Recognition (ICPR)* (pp. 1472-1477). IEEE.

[29] Rundo, L., Militello, C., Tangherloni, A., Russo, G., Vitabile, S., Gilardi, M. C., & Mauri, G. (2018). NeXt for neuro-radiosurgery: a fully automatic approach for necrosis extraction in brain tumor MRI using an unsupervised machine learning technique. *International Journal of Imaging Systems and Technology*, *28*(1), 21-37.

[30] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.