Prediction of Breast Cancer Diagnosis through Machine Learning Techniques

Md Mahmudul Hasan Mamun
Ruthwik Nadam
Ayman Almomany

Table of Contents

Abstract

This project focuses on using machine learning techniques to classify diagnoses between benign and malignant cases using the Breast Cancer Wisconsin Diagnostic Dataset (WDBC). The dataset comprises digitalized measurements from breast lumps via fine needle aspirants to predict tumor characteristics. Through extensive Exploratory Data Analysis (EDA), the study investigates the dataset's structure, feature distribution, correlations, and challenges such as missing values and outliers. The insights gained serve as a basis for developing effective machine learning models to aid in breast cancer diagnosis. The study identifies KNN and Random Forest as the top-performing models across various metrics. Notably, KNN achieves 100% prediction accuracy, but its AUC remains close to 0.5, questioning the practical utility of AUC in assessing machine learning algorithms. Furthermore, the study demonstrates the consistent improvement of prediction accuracy across all models with the Synthetic Minority Over-sampling Technique (SMOTE) balancing technique. While SMOTE was applied to balance the data, its effectiveness suggests potential benefits for enhancing model performance, especially with small datasets.

Introduction

Breast cancer is the most commonly occurring cancer in women globally and represents a significant health challenge that transcends gender, affecting both men and women, albeit it is substantially more common in women [1] The evolution of machine-assisted diagnostic systems marks a pivotal advancement in the early detection of breast cancer, enhancing the diagnostic capabilities of radiologists significantly. The primary modalities utilized in the diagnosis of breast cancer encompass mammography, tomography, Breast Ultrasound (BUS), MRI, CT scans, and, for more in-depth analysis, PET scans are employed [2]. The observed sensitivity for the diagnosis procedures are different for each of them. Even though the sensitivity for FNA varies a lot between 65% - 98% it indicates there might be chances of providing a incorrect diagnosis resulting in life threatening cases [3]. The efficacy of breast cancer diagnosis heavily depends on the robustness of these classification tools.

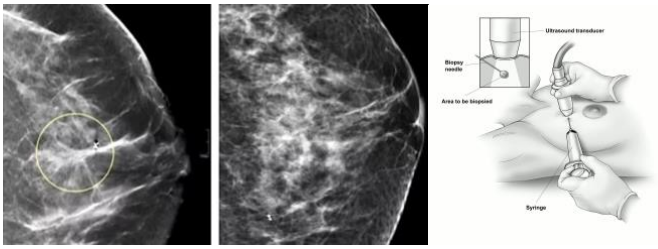

Fig.1 *Different types of screening methods to diagnose breast cancer. The left image shows how the tumor is diagnosed using mammograph, right image is the fine needle aspiration using ultrasound.*

The motivation for this study stems from the appeal to force the perspective of ML to save life expectancy by planning more accurate diagnoses of breast cancer. Present diagnostic methods can be imposing, costly, and, at times, model to mental distress. ML models with rigor and precision, can increase the diagnostic process making it less imposing and more cost-effective. Prior research utilizing machine learning techniques such as SVM, decision trees, random forests, ANNs, and ensemble classifiers for feature training and categorization delineates the objects into malignant or benign classes [4],[5]. These methodologies have demonstrated the capacity of machine learning to assist in the early detection of breast cancer, enhancing the accuracy of diagnoses. The advent of machine learning, particularly the emergence of deep learning architectures, has ushered in promising outcomes, with CNNs garnering significant interest among researchers for breast tumor detection and classification. Notable CNN architectures such as AlexNet, CiFarNet [6], GoogLeNet [7], VGG16, and VGG 19 [8],[9] have been investigated. Despite the advances, the literature reveals a dearth of comparative studies that scrutinize the performance of various machine-learning models in the context of breast cancer diagnosis. Additionally, many studies are limited by the number of datasets, potentially not reflecting the complexity of breast cancer cases in the actual clinical setting [10].

Early recognition and accurate diagnosis are critical for increasing breast cancer survival rates. ML techniques have been proposed to boost the accuracy of breast cancer diagnostics [11]. Nonetheless, gaps remain in the literature: comparative analyses are scarce, limited datasets

constrain existing studies, and advanced feature selection reasonings are needed for identifying leading features for diagnosis.

The prime question in creating an ML model for breast cancer findings is ensuring high sensitivity and specificity. The cost of false negatives (missing a cancer diagnosis) can be life-dangerous, while false positives (wrongly diagnosing cancer) can cause excessive anxiety and medical procedures. Thus, a difficult balance must be uncovered to minimize both anomalies, detect outliers, and recognize feature associations within the WDBC dataset.

The aim of this analysis is multifold:

1. To evaluate several ML models on the dataset and recognize the most favorable ones based on performance metrics such as accuracy, precision, recall, and F1-score.
2. To examine model-specific performance indicators such as ROC curves and confusion matrices, thereby assessing the true positive and false positive rates, which are crucial for clinical applications.
3. To concrete the way for the training of an optimal predictive model that can accurately classify breast tumors as benign or malignant, thus aiding clinicians in diagnosis and treatment forecast.

The goal of the diagnostic aspect of our research is to develop a relatively objective system that diagnoses FNAs with an accuracy that approaches the best achieved visually. (Need to paraphrase this)

Dataset Description

The dataset comprises 569 instances and 32 attributes, with each instance representing features computed from images of cell nuclei. The features present in this dataset were digitally measured using the Xcyt image analysis program. These features include characteristics such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The attributes encompass the ID number, diagnosis (denoted as 'M' for malignant and 'B' for benign), and 30 real-valued input features computed as the mean, standard error, and "worst" (largest) values. The class distribution shows 357 instances labeled as benign and 212 as malignant, indicating an imbalance favoring benign instances. This dataset serves as a valuable resource for tasks related to the diagnosis of cell nuclei as either malignant or benign, providing researchers and practitioners with rich information for analysis and modeling purposes.

Methodology

The project commenced with downloading the WDBC data from UCIML. The next step would be to perform exploratory data analysis to detect any missing values or outliers within the data. This

preliminary analysis revealed several challenges, such as the features varied in measurement scales, the target variable was imbalanced, and there was significant multicollinearity among the predictors.

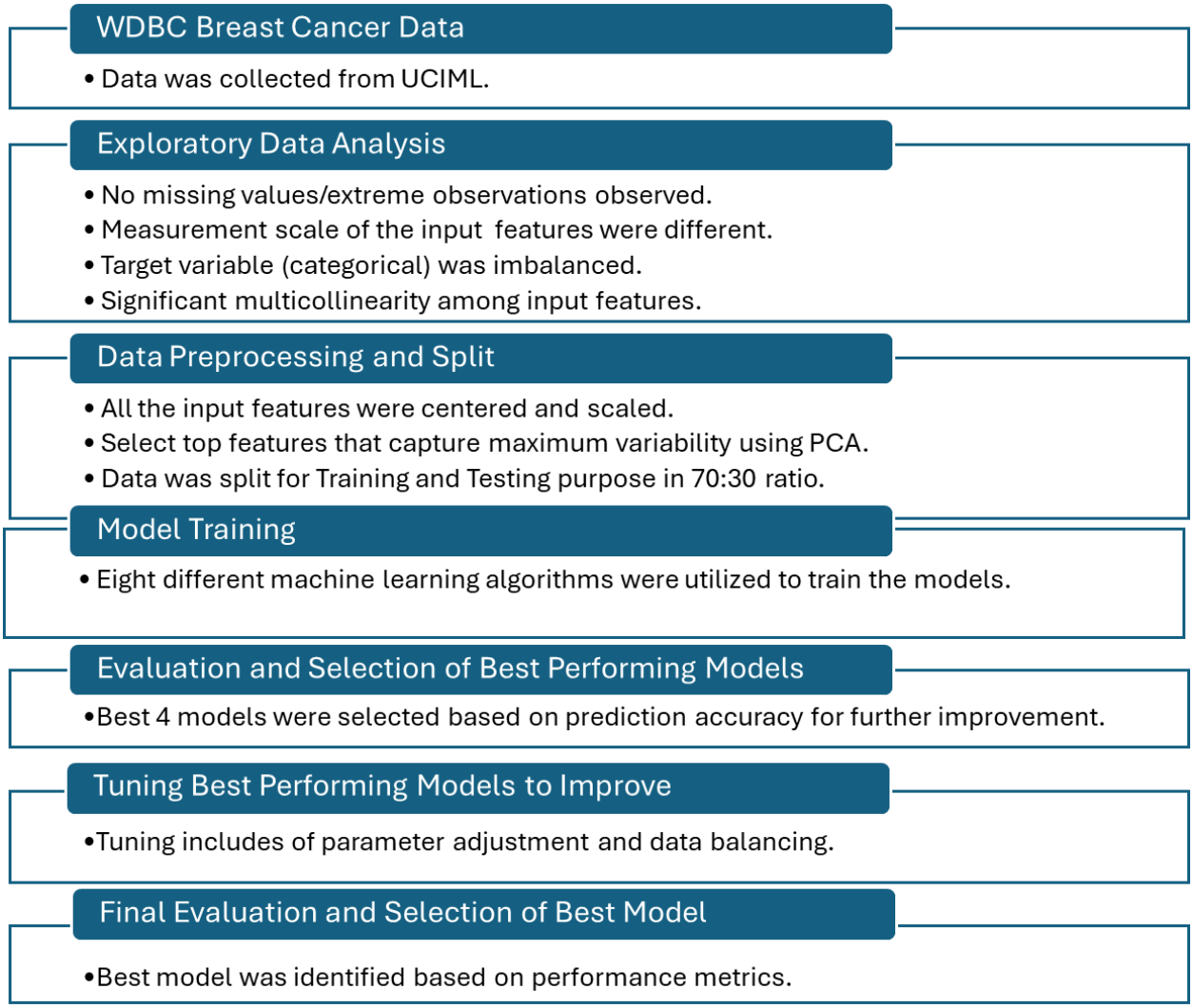

Diagram 1: *Workflow of the Project*

Third step of the workflow is Data Preprocessing and splitting the dataset into Training and Validation set. At this step, we standardized all the input features to ensure that no single feature would disproportionately influence the model due to its scale. Also, we applied using principal component analysis to pick the top features which would capture the maximum variability. For feature selection, we used top 2 PCA that captures 99.9% variability. Mean of absolute factor loading of the top 3 components were used to select the best 10, top 15, top 20 and top 25 features.

For model fitting and evaluation, we split the dataset into training and validation set. The training dataset contained 398 cases while validation dataset contained 171 cases. Fig. 2 presents the distribution of training and validation dataset after splitting.

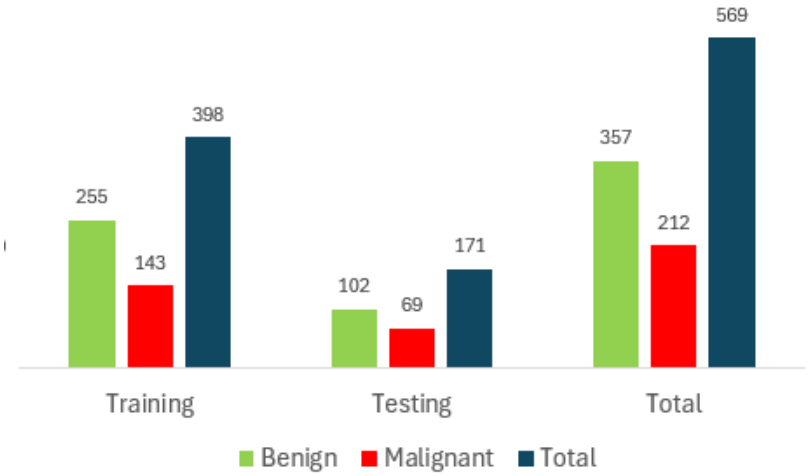

Fig. 2 *Distribution of Training and Validation Dataset after split*

We balanced training dataset using 3 different techniques: under-sampling, k-medoid clustering and over-sampling. Under-sampling randomly reduces the size of the majority class to match the minority class, while K-medoid clustering divides the majority class into k clusters (where k equals the number of instances in the minority class) and selects the center or medoids of each cluster. Synthetic Minority over-sampling technique (SMOTE) was used for over-sampling the dataset. SMOTE identifies the k-nearest neighbors in the feature space for each minority instance and generates synthetic instances by interpolating between the minority instance and its neighbors. Specifically, it randomly selects one of the k-nearest neighbors for each minority instance and creates a synthetic instance along the line segment connecting the two points. This process iterates until the desired balance between minority and majority classes is achieved. Due to its random nature, generating a SMOTE sample requires careful consideration to achieve the desired balance between minority and majority classes. Fig. 3 captures the Distribution of Diagnose cases in Training Dataset after data balancing. We kept validation dataset intact throughout the process to maintain the integrity of the evaluation process.

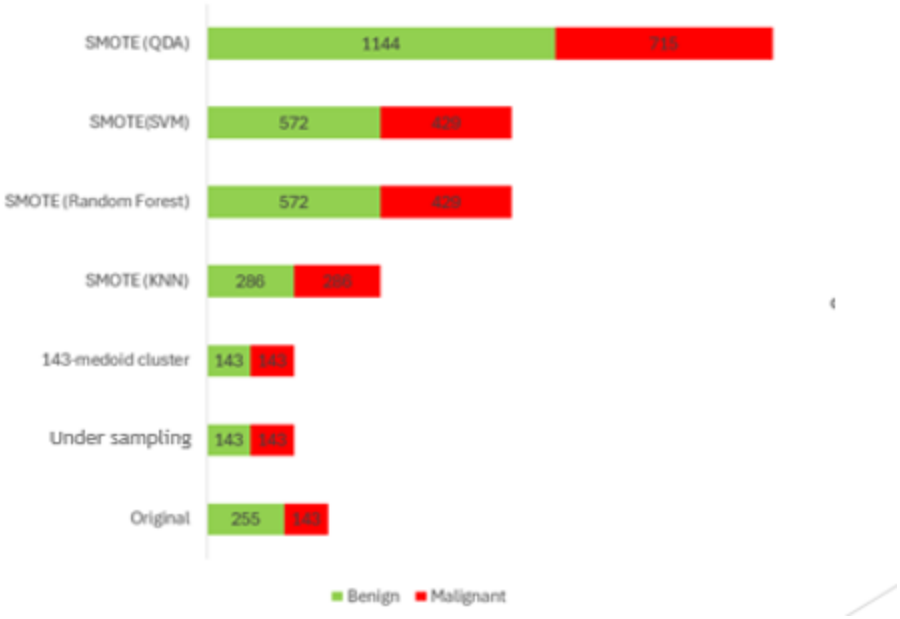

Fig. 3 *Distribution of Diagnose cases in Training Dataset after data balancing*

To train the model using 8 different machine learning models, such as Support vector Machine, K-Nearest Neighbor (KNN), Random Forest, Gradient Boosting, Decision Tree, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Naïve Bayes. Initially, the top 4 models were selected based on accuracy of the models on the original dataset. To improve the performance and the prediction accuracy of the models, we fine-tuned the parameters and utilized balanced data using three different techniques mentioned above. Finally, the best-performing model was identified based on performance metrics analysis.

Results and Discussion

Descriptive Analysis

Features	count	mean	median	sd	min	max	q25	q75	IQR
radius_mean	569	14.13	13.37	3.52	6.98	28.11	11.70	15.78	4.08
texture_mean	569	19.29	18.84	4.30	9.71	39.28	16.17	21.80	5.63
perimeter_mean	569	91.97	86.24	24.30	43.79	188.50	75.17	104.10	28.93
area_mean	569	654.89	551.10	351.91	143.50	2501.00	420.30	782.70	362.40
smoothness_mean	569	0.10	0.10	0.01	0.05	0.16	0.09	0.11	0.02
compactness_mean	569	0.10	0.09	0.05	0.02	0.35	0.06	0.13	0.07
concavity_mean	569	0.09	0.06	0.08	0.00	0.43	0.03	0.13	0.10
concave_points_mean	569	0.05	0.03	0.04	0.00	0.20	0.02	0.07	0.05
symmetry_mean	569	0.18	0.18	0.03	0.11	0.30	0.16	0.20	0.03
fractal_dimension_mean	569	0.06	0.06	0.01	0.05	0.10	0.06	0.07	0.01
radius_sd	569	0.41	0.32	0.28	0.11	2.87	0.23	0.48	0.25
texture_sd	569	1.22	1.11	0.55	0.36	4.89	0.83	1.47	0.64
perimeter_sd	569	2.87	2.29	2.02	0.76	21.98	1.61	3.36	1.75
area_sd	569	40.34	24.53	45.49	6.80	542.20	17.85	45.19	27.34
smoothness_sd	569	0.01	0.01	0.00	0.00	0.03	0.01	0.01	0.00
compactness_sd	569	0.03	0.02	0.02	0.00	0.14	0.01	0.03	0.02
concavity_sd	569	0.03	0.03	0.03	0.00	0.40	0.02	0.04	0.03
concave_points_sd	569	0.01	0.01	0.01	0.00	0.05	0.01	0.01	0.01
symmetry_sd	569	0.02	0.02	0.01	0.01	0.08	0.02	0.02	0.01
fractal_dimension_sd	569	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00
radius_worst	569	16.27	14.97	4.83	7.93	36.04	13.01	18.79	5.78
texture_worst	569	25.68	25.41	6.15	12.02	49.54	21.08	29.72	8.64
perimeter_worst	569	107.26	97.66	33.60	50.41	251.20	84.11	125.40	41.29
area_worst	569	880.58	686.50	569.36	185.20	4254.00	515.30	1084.00	568.70
smoothness_worst	569	0.13	0.13	0.02	0.07	0.22	0.12	0.15	0.03
compactness_worst	569	0.25	0.21	0.16	0.03	1.06	0.15	0.34	0.19
concavity_worst	569	0.27	0.23	0.21	0.00	1.25	0.11	0.38	0.27
concave_points_worst	569	0.11	0.10	0.07	0.00	0.29	0.06	0.16	0.10
symmetry_worst	569	0.29	0.28	0.06	0.16	0.66	0.25	0.32	0.07
fractal_dimension_worst	569	0.08	0.08	0.02	0.06	0.21	0.07	0.09	0.02

Table. 1 *Descriptive Summary of 30 real valued features.*

The summary of cytological features from the WDBC dataset are mentioned in Table 1 showing the mean, standard deviation and worst(largest) of the cell nuclei. The average radius of the largest cells is around 16.27 units even though the maximum value lies at 36.04 units and minimum at 7.93.

This shows that among the ten randomly selected biggest cell nuclei throughout the image analysis process, there is a significant diversity in their radius. The significant ranges displayed by the "area_mean" and "area_worst" (from 143.5 to 2501.0 and 185.2 to 4254.0, respectively) suggest that there are wide variations in tumor sizes within the population. The comparatively narrow ranges and standard deviations of features like "smoothness_mean" and "compactness_mean" imply that these attributes vary less amongst tumors. Minimum values of 0 for variables such as "concavity_mean" and "concave_points_mean" may suggest measurement abnormalities or the absence of these traits in certain malignancies. The comparatively lower standard deviations and IQRs of features like "symmetry_mean" and "fractal_dimension_mean" suggest that certain tumor traits may be more consistent among patients. This may suggest that tumor texture and complexity vary less than tumor size or shape.

Certain tumors have very compact and concave characteristics, as evidenced by the high maximums of the "compactness" and "concavity" metrics, particularly at their worst values, in comparison to their means. These excessive numbers may indicate tumor forms that are more aggressive.

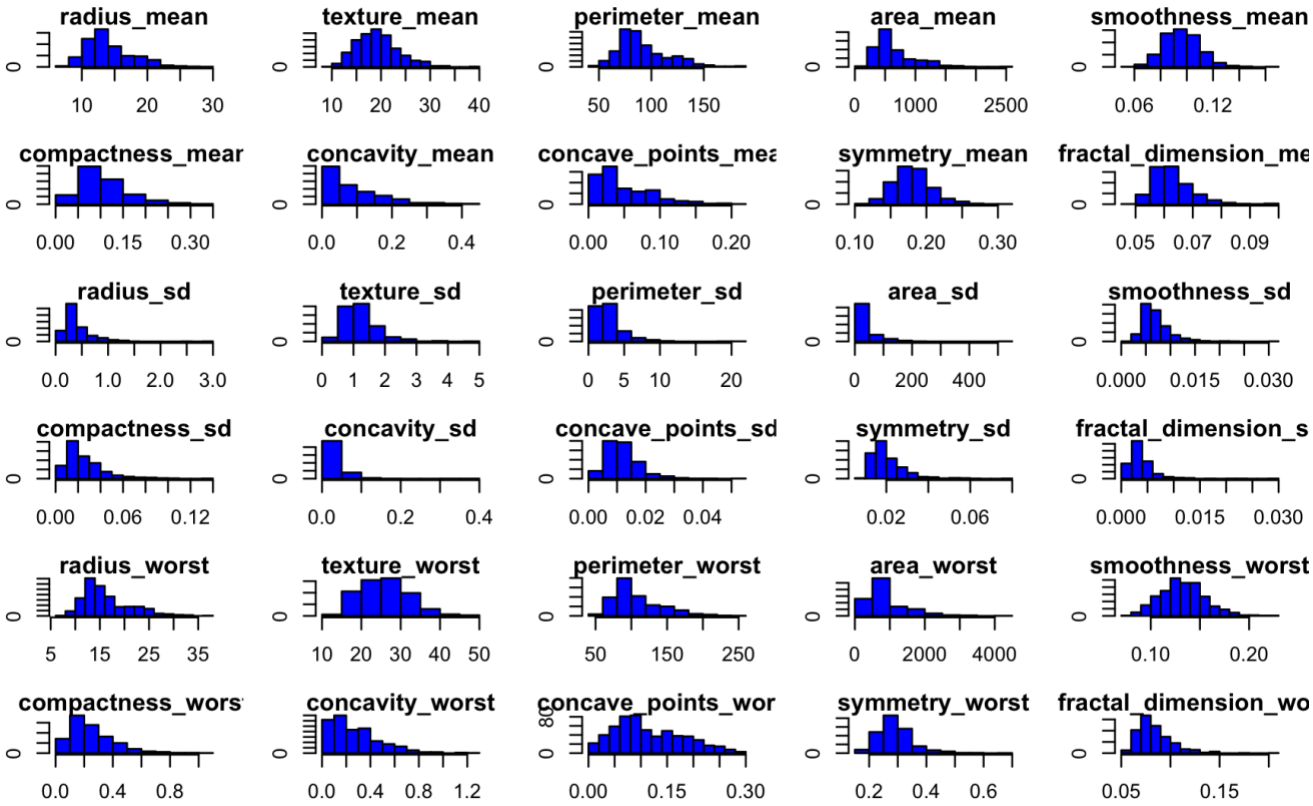

Fig. 4 *Dataset distribution of all the variables present in the Cancer Dataset.*

The histograms from Fig.4 provide a detailed visual summary of the distribution of the features. Most tumors have a smaller average radius, but there is a long tail towards larger sizes. Features such as radius_mean, perimeter_mean, area_mean, concavity_mean, concave_points_mean, area_worst, concavity_worst, concave_points_mean are right skewed indicating there are outliers for these features and will be able to capture the aggressive tumor cases in predictive modelling when used in the training dataset. The features texture_mean, smoothness_mean, symmetry_mean and fractal_dimension_mean are left skewed.

Before we jump into the modeling, we wanted to gain insight into the data through visualization. As it is very challenging to plot 30 feature in a single diagram, we opted to visualize the data using first two principal components (Fig. 5). We observe that while benign and malignant cases are not entirely linearly separable, a significant portion exhibits separability.

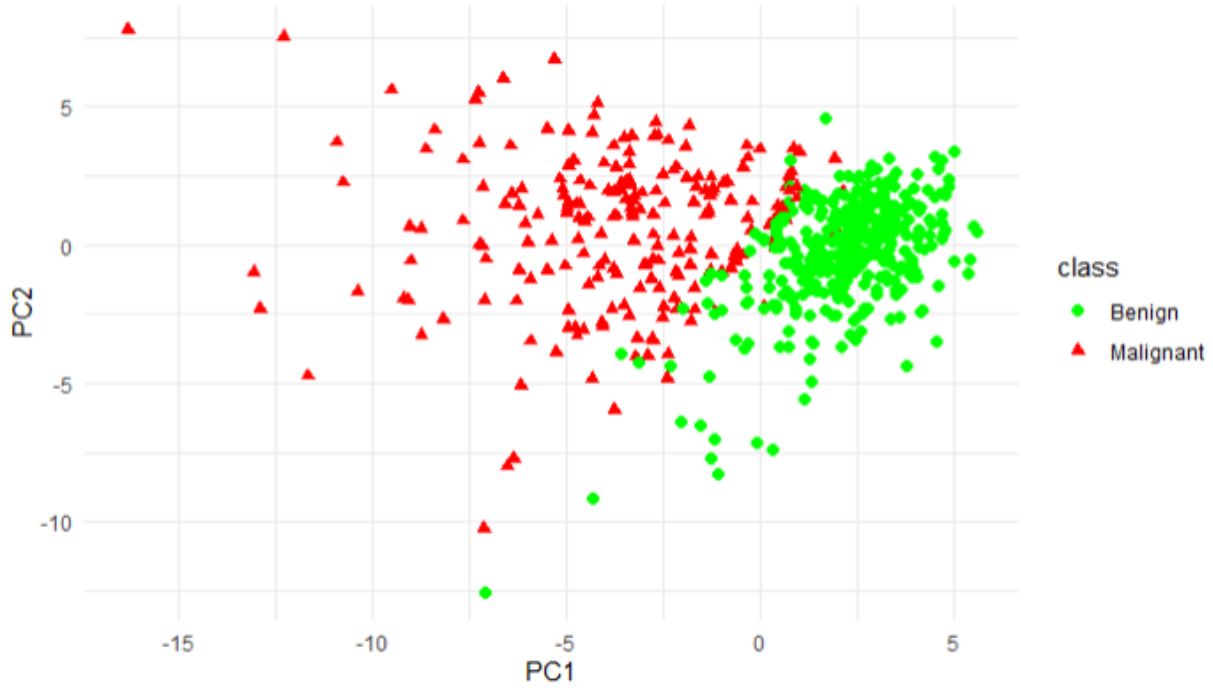

Fig. 5 *Dataset distribution of all the variables present in the Cancer Dataset.*

Model Evaluation

With the original training and test dataset, we modeled and evaluated all the eight machine learning techniques. Table 2 summarizes the model performance based on accuracy. We used all the features, and also utilized top 10, top 15, top 20 and top 25 features to fit the models (table 2). We identified SVM as the best performing model with top 20 features (accuracy:98.25%). It is followed by QDA (accuracy: 97.66%), KNN (accuracy: 96.46%), Random Forest (accuracy: 96.49%) and GBM (accuracy: 96.49%). So, we selected SVM, QDA, KNN and Random Forest for further improvement of accuracy with fine-tuning the parameters and applying balanced data. Though Random Forest and GBM had same accuracy, we choose Random Forest as this model is less prone to overfitting compared to GBM.

Model	Prediction Accuracy (%) with Imbalanced Data				
	All features	Top 10	Top 15	Top 20	Top 25
SVM	97.66	93.57	95.32	98.25	97.08
QDA	97.66	94.74	95.32	95.91	95.32
KNN	96.49	89.47	92.98	92.4	95.32
Random Forest	96.46	92.4	94.74	96.49	96.49
GBM	95.32	93.57	93.57	96.49	96.49
Naïve Bayes	92.98	90.06	92.98	94.74	93.57
LDA	92.98	92.98	94.74	94.74	94.74
Decision Tree	91.81	90.06	87.72	91.81	91.81

Table 2: *Accuracy of all 8 models used.*

Table 3 presents model performance accuracy of the selected models after fine-tuning and utilizing balanced datasets. We observe that SMOTE balancing techniques in all cases provide best accuracy. Using SMOTE-balanced training dataset, KNN could predict the case in validation dataset with 100% accuracy. It is followed by SVM (accuracy: 98.25%), Random Forest (98.25%) and QDA (97.66%).

Model	Prediction Accuracy (%)			
	Original Data	Under-sampling	K-medoid clustering	SMOTE
KNN (k=5)	96.49	98.83	97.66	100
SVM (Kernel=radial)	97.66	97.08	98.25	98.25
Random Forest	96.46	95.91	95.91	98.25
QDA	97.66	94.15	97.08	97.66

Table 3: *Accuracy of top 4 selected*

We also observed sensitivity, specificity, F-1 score and AUC (area under the curve) of the models. Precision represents the proportion of true positive predictions among all positive predictions made by the model. Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions among all actual positive instances in the data. F1 score is the harmonic mean of precision and recall. The F1 score ranges from 0 to 1, where a higher value indicates better model performance. For each model, obviously precision and recall was very high, as was F-1 score.

It's evident that KNN emerges as the top-performing model across various performance metrics except for AUC, where it exhibited a notably lower score of only 0.5031. The AUC metric, which essentially measures the performance of a binary classifier averaged across all possible decision thresholds, can be seen as less informative from a practical standpoint. When considering a classifier deployed with a specific threshold, what it does in a theoretical and abstract scenario, when averaged across all thresholds, might hold little relevance for practitioners. Given our study's rigorous procedures for model fitting and validation, we cast serious doubt on the utility of AUC in evaluating machine learning techniques. However, the AUC of other models was pretty high, nearly 1.

After KNN, the second-best performing model was Random Forest. The F-1 score of the model was 0.9854, and AUC was 0.9987 which means the predictive power of the model was 99.87%.

Model	Prediction Accuracy				
	Accuracy(%)	Precision (%)	Recall (%)	F1-score(0-1)	AUC
KNN (k=5)	100	100	100	1	0.5031
SVM (Kernel=radial)	98.25	99.01	98.04	0.9852	0.9987
QDA	97.66	98.04	98.04	0.9804	0.9953
Random Forest	98.25	98.06	99.02	0.9854	0.9987

Table 4: *Accuracy, Sensitivity, Specificity and AUC of top 4 selected models.*

Conclusion

In conclusion, this study delved into the classification of benign and malignant cases using machine learning techniques applied to the Breast Cancer Wisconsin Diagnostic Dataset (WDBC). Through thorough Exploratory Data Analysis (EDA), we gained insights into the dataset's structure, feature distributions, correlations, and potential challenges. Our findings underscore the importance of robust preprocessing steps to ensure accurate model performance.

We identified KNN and Random Forest as the top-performing models, with KNN notably achieving 100% prediction accuracy. However, the discrepancy between its high accuracy and low AUC raises questions about the practical applicability of AUC in evaluating machine learning algorithms. Additionally, the consistent improvement in prediction accuracy across all models with the SMOTE balancing technique highlights its efficacy in handling imbalanced datasets.

Moving forward, our study suggests the need for further research to explore alternative evaluation metrics that better capture model performance, especially in scenarios where class imbalances exist. Moreover, investigating the underlying reasons for the observed discrepancy between accuracy and AUC in KNN models could provide valuable insights into refining model evaluation methodologies.

References

[1] Sadad, Tariq, et al. “Fuzzy C-means and region growing based classification of tumor from mammograms using hybrid texture feature.” Journal of computational science 29 (2018): 34-45.

[2] Kumar, Abhinav, et al. “Deep feature learning for histopathological image classification of canine mammary tumors and human breast cancer.” Information Sciences 508 (2020): 405-421.

[3] Olvi, Nick, et al. “Breast cancer diagnosis and prognosis via linear programming.” Operations Research, Jul. - Aug., 1995, Vol. 43, No. 4 (Jul. - Aug., 1995), pp.
[4] 570-577

[5] Sadad, Tariq, et al. “Fuzzy C-means and region growing based classification of tumor from mammograms using hybrid texture feature.” Journal of computational science 29 (2018): 34-45.

[6] Vijayarajeswari, R., et al. “Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform.” Measurement 146 (2019): 800-805.

[7] Roth, Holger R., et al. “Improving computer-aided detection using convolutional neural networks and random view aggregation.” IEEE Transactions on medical imaging 35.5 (2015): 1170-1181.

[8] Szegedy, Christian, et al. “Going deeper with convolutions.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[9] Saba, Tanzila, et al. “Brain tumor detection using a fusion of handcrafted and deep learning features.” Cognitive Systems Research 59 (2020): 221-230.

[10] Ejaz, Khurram, et al. “An unsupervised learning with feature approach for brain tumor segmentation using magnetic resonance imaging.” Proceedings of the 2019 9th International Conference on Bioscience, Biochemistry, and Bioinformatics. 2019.

[11] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[12] Bennett, K. P. “Decision Tree Construction Via Linear Programming.” Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992.