FINAL PROJECT REPORT

LEENA RAJESH PATIL

"BREAST CANCER CLASSIFICATION"

ADVANCED MACHINE LEARNING

lpatil@depaul.edu

Abstract:

Breast cancer has been one of the major causes of cancer-related deaths among women worldwide. Early and precise diagnosis is very important, as it directly influences treatment outcomes and survival rates of the patient because one wrong diagnosis can completely take away the person's life for no good. However, there are a lot of traditional diagnostic methods, namely biopsy and imaging, that, though they are considered effective, are often invasive and resource-intensive, and, mainly, they are not universally easily accessible. Recently, machine learning (ML) techniques have emerged as a powerful tool for medical diagnosis by enabling automated, accurate, and scalable classification of the tumor data. In this project, I have implemented a full ML pipeline for breast cancer diagnosis by taking many values or dimensions of cells and am focusing on classifying the breast tumors as benign or malignant using the very famous Wisconsin Breast Cancer Diagnostic Dataset (WBCD).

The project follows a standard series of steps by starting with preprocessing of the data to encode the diagnosis labels, which are originally coded as M and B, to 1 and 0 for easy analysis, and then we move on to exploratory data analysis to correctly understand and analyze feature distributions and correlations. Further, I have run this dataset on various machine learning models like decision tree, random forest, gradient boosting, AdaBoost, support vector machines (SVM) with both RBF and linear kernels, and a neural network (MLPClassifier) and then compared their scores to decide which model is the best for our data. I have also implemented feature selection for all the models too to improve model parsimony and interpretability and calculated the respective scores of AUC and accuracy (as ours is a classification model) for before and after.

Then I studied a few literature papers, which helped me increase my horizon of knowledge on the same. Studies have shown for years that tree-based models and SVMs always offer strong performance for breast cancer classification or classification in general, while neural networks can also work, but they need proper tuning and sufficient data for them to give their best performance. Additionally, the feature extraction that was performed on our dataset, which selected features like nuclear radius, texture, smoothness, concavity, and symmetry, matches with the important feature findings that were mentioned in the papers too.

All the models were evaluated using accuracy, AUC, and ROC curves, both before and after feature selection. Results confirm that ensemble methods and linear SVMs performed robustly, achieving high AUC and accuracy on unseen test data. Neural networks showed promising but less consistent results, consistent with challenges noted in the literature [3][4].

This project demonstrates that ML techniques can effectively support the early diagnosis of breast cancer and also gives us insights about the model selection, feature extraction, and tradeoffs between accuracy and interpretability. The good part is that our findings align well with the literacy papers we have referred to. It also throws light on including the incorporation of explain ability techniques such as SHAP and the extension of models to larger, more complex datasets. Ultimately, this work contributes to the growing body of research supporting the integration of AI into clinical decision support systems for oncology.

Introduction:

Breast cancer is one of the deadliest, most problematic, and most common kinds of cancer affecting women all over the world. According to the studies of our world health organization, one in four cancer cases among women is diagnosed as breast cancer and, sadly, is also the leading cause of cancer-related death for this group. If we look into the numbers, in the year 2020 alone, an estimated 2.3 million women were diagnosed with breast cancer worldwide, with 685,000 deaths attributed to the disease. This burden and problem of breast cancer is not only confined to the high-income countries, but instead the low- to middle-income countries account for nearly half of these cases and often face a lot of challenges in screening and treatment access.

Especially when it comes to cancer, early diagnosis of the disease is very, very important. Studies have shown that early detection and treatment of this cancer have a 5-year survival rate of over 90%. On the other hand, late detection will drastically decrease the survival chances. A lot of current diagnostic tools like mammography, ultrasound, MRI, and biopsy have proved themselves to be super effective, but unfortunately these methods can be invasive, expensive, resource-intensive, and heavily reliant on radiologist expertise, leading to a lot of variability and limited accessibility. Even with the help of skilled practitioners, it is difficult to give 100% accuracy with the image interpretation, which can lead to diagnostic inconsistencies, which is very dangerous in healthcare, and that too when it comes to diseases like cancer, which are life-threatening.

If we have to get to know a little about cancer at a biological level, breast cancer alters the morphology of cell nuclei. Now this cancer can be malignant, which is deadly, as it makes sure to transfer oncogenic cells through blood vessels and lymph vessels to other parts of the body, leading to oncogenic transformation of cells there. But on the other hand, benign is a non-active tumor, just a mass of cells and waste material that has developed into a lump that can be left as it is or removed, which doesn't cause problems to other parts of the body.

Research and continuous observation of the cell images have shown that malignant breast tumors often exhibit nuclei with larger size, irregular shapes, rougher contours, higher asymmetry, and greater textural variation compared to benign tumors. Features such as radius, perimeter, texture, smoothness, concavity, concave points, symmetry, and fractal dimension that were originally proposed by Street et al [1]. and have also been used in our dataset have been proven as valid indicators for malignancy. Extracting these features and analyzing them through image processing provides us with a strong basis for diagnosis.

In our context, we are more concerned with using machine learning models as an opportunity to enhance breast cancer diagnosis by automating pattern recognition, improving consistency, and reducing reliance on subjective human interpretation. It will be exhausting for humans to go through huge amounts of data that have been collected from different parts of the world for many years, so instead our machine learning models easily process large datasets, capture subtle and complex relationships among features, and provide probabilistic outputs that can aid clinical decision-making.

So overall in this project I have tried to implement the complete ML pipeline for breast cancer diagnosis. It follows a series of steps like Compare the performance of diverse classifiers—Decision Tree, Random Forest, Gradient Boosting, AdaBoost, SVM (RBF & Linear), and Neural Network—on WBCD. Evaluate the impact of feature selection on both performance and model parsimony. Analyze results in light of clinical relevance and practical applicability.

Literature review:

i) Street et al. (1999) paper [1]

Breast cancer diagnosis has for a very long time been a critical and important area for machine learning. Over the past three decades, there have been numerous studies that have explored and proved to us how these machine learning models can enhance the traditional diagnostic workflows by providing automated, consistent, and accurate classification of breast tumors based on digitized data.

There are a lot of works, but one of the foundational ones is the study by Street et al. (1999) [1], which introduced us to something called the nuclear feature extraction methodology for breast tumor diagnosis, where we used fine needle aspirate images for our diagnosis and analysis. The cell nucleus is the center of interest when it comes to diagnosing cancer. When a cell turns oncogenic, the nuclei tend to become larger, irregular, rough, and asymmetrical. Measuring the shape and texture of the nuclei helps in detecting malignancy. But when we rely on microscopic images for the same, the cell boundaries are not perfectly clear — they are somewhat blurry and noisy. The authors use something called an active contour model, also known as a snake. This is a flexible curve, like an elastic band, that starts roughly around the nucleus. And then it moves and adjusts its shape automatically to the real boundary of the nucleus. This is done by optimizing the energy function by balancing how smooth the curve is, how well the curve follows edges in the image, and finally how well it fits the actual nucleus shape. So now, after we have finally got the image, i.e., the exact nucleus boundary, we can go ahead and measure the radius, perimeter, area, smoothness, symmetry, etc., and eventually these measurements become the features for our model, which we also saw in our WBCD dataset.

So, the final dataset consisted of 30 features of mean, worst, and standard error of 10 key nuclear features. This made sure to have a biologically important feature set for our machine learning models, which can be used for effective classification instead of relying on raw image data.

ii) Sharma et al. — Breast Cancer Detection Using Machine Learning Algorithms [2]

This paper mainly aimed at checking different machine learning algorithms and finally deciding which works the best for the classification of breast cancer as benign or malignant. This paper uses the same dataset we are using, i.e., the Wisconsin Breast Cancer Diagnostic Dataset (WBCD). Even if we are working with the same data, different machine learning models give us different results and performance. And when it comes to healthcare, we have to be extra careful and make sure to deliver results that are accurate, reliable, and explainable. So, this paper helps us decide the most promising model.

This paper tested machine learning models like random forest, which is basically an ensemble of decision trees; they are strong, robust, and successfully explain feature importance. Coming to the next model that they used, k-nearest neighbor, this is a simple vote-by-neighbor or vote-by-distance type of classifier, which is basically very interpretable. And the third model that they used is Naive Bayes, which is a probabilistic model that is both fast and interpretable.

All these models were good choices for our dataset. This paper goes about with the standard steps and uses ML preprocessing like normalization, which we have done in our project too. It then goes about with a train/test split and uses metrics like accuracy, precision, recall, and F1 score. The final results that they got showed that KNN and RF gave approximately 96% and 95%, respectively, and NB gave us less than 90%. This shows us that even simple models like KNN and RF can give us good performance for our dataset. So we do not always need to use complex models like deep neural networks. This paper also showed us that before we get into the actual process, preprocessing the data is as important.

iii) Rovshenov & Peker — Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using WBCD [3]

The goal of this paper is also similar, that is, to use different ML algorithms to predict breast cancer early using the Wisconsin Breast Cancer Diagnostic Dataset (WBCD)—the same as our project. The models of interest in this paper are Artificial Neural Networks (ANN), which are multi-layer perceptrons (MLP) trained with ReLU, which is a mathematical function used inside the neurons to help them learn complex patterns, and sigmoid activations, which help turn output into a number between 0 and 1 so we can interpret it as probability. They have also used sparse categorical crossentropy, which basically tells the model during training, "You predicted this wrong; adjust your weights." Coming to the next model, Support Vector Machines (SVM), this is a kind of model that normally tries to draw a line (or boundary) that best separates the two classes: benign and malignant. But instead of just a straight line, we are using an RBF kernel, which allows SVM to draw curves and flexible boundaries to separate the two classes, which can better fit with the complex pattern of the data. To be more specific, they have used degree = 3, which basically tells us how flexible the curve can be. The third model used is Random Forest (RF), which is a collection of decision trees. The final prediction is made by majority voting across all trees. In this paper they have used 100 trees, and they have also used the Gini criterion. Like the standard procedure, they have also done preprocessing of data here, where they have used a standard scaler that I have used in my project too. Next, they have done a train/test split with 10-fold cross-validation and finally have used evaluation metrics like accuracy, AUC, precision, recall, and F1 score. In the final result we can see that ANN performed the best with an accuracy value of approximately 99%. SVM and RF also equally performed well with an accuracy of approximately 97%.

As I have mentioned before in this paper, early detection is very important, especially for diseases like cancer. This paper also throws light on traditional diagnostic methods for cancer like mammography, ultrasound, and MRI, which are very useful, but they cannot be implemented everywhere, as they are expensive, they require skilled radiologists, etc. Instead, using ML-based tools can help automate diagnosis, making it more accessible and less subjective.

iv) Ravi et al., 2017 → Deep Learning for Health Informatics [4]

This paper is a review paper more than an experiment where the author wants to summarize how deep learning models are used in healthcare for disease diagnosis, medical imaging, wearable sensor data, and electronic health records and also discuss the challenges of applying deep learning to healthcare. In my project I have used an MLP classifier, which is a type of deep learning model, and also my data is of healthcare, and thus this paper successfully gives us a background context as to why deep learning is interesting, when it works well, and what challenges you need to watch out for. First, the paper talks about the strengths of deep learning, where it tells us how deep learning does not need manual engineering but instead can learn features automatically. It works well with medical images like X-rays, MRIs, CT scans, sensor data like wearables and vital signs, and also text data like doctors' notes or electronic health records. On the contrast paper, it also mentions the weakness of deep learning, where it tells us that this needs a lot of data. Deep models like CNNs and RNNs need thousands or millions of examples. Requires a lot of computation with the help of GPUs and large servers. And these models are like black boxes; it's really hard to explain them to doctors. When it comes to health data in particular, it is normally small, noisy, and messy, so deep learning struggles if data is limited.

Deep learning is a good and powerful tool, but for simple tabular data like our data, SVM and RF can often perform well. So even if we end up using them for small data, we will have to be careful and make sure we do proper tuning to avoid overfitting. The results in my project showed that the MLP classifier was sensitive, which matches this literature.

v) Patil & Thakur → Breast Cancer Diagnosis Using Random Forest [5]

This paper aims at showing how random forest can be used to classify breast cancer tumors as benign or malignant. They have also compared random forest with other simple ML models. They have used the same dataset as mine. Random forest is normally known for delivering accuracy, robustness, and interpretability. When we think about them in a medical context, doctors like them because they are easy to explain and understand, as they are in a tree format.

Again, like all the papers, they have followed a standard procedure like they have used a random forest classifier, they have done a train/test split, and they have done basic preprocessing and finally calculated the evaluation metrics like accuracy, precision, recall, and F1 score. No deep learning or SVM was used in this paper, as it solely concentrated on talking to us about the importance and usage of random forest. When it comes to the results, this model achieved a high accuracy of approximately 96%. The model was able to handle the WBCD dataset very well and produced results comparable to more complex models. The authors parallelly highlighted the fact that feature importance analysis is a major strength of Random Forest; i.e., the doctors can understand which features the model used for decision-making.

So, to summarize the literature review, we can say that all of the authors are trying to solve the problem of improving automated breast cancer diagnosis, as we cannot always rely on traditional methods, as we cannot make them accessible everywhere. The good part, which gave us a better picture, is that almost all the papers used the same dataset as I am using for my project. So, across the papers, when they did their model testing and calculated scores, they finally found out the best-performing models. They were namely k-Nearest Neighbors (kNN) and Random Forest because they can effectively learn from the clear patterns present in the well-engineered features of the WBCD dataset. Artificial neural networks (ANN) were chosen for their ability to model complex, non-linear relationships in the data, achieving very high accuracy when properly tuned. Support Vector Machines (SVM) with RBF kernels were selected for their flexibility in handling cases where data is not linearly separable while also maintaining robustness and good generalization. Deep learning models were also discussed as a powerful tool but not a good choice for our data. Well, overall, RF was mentioned in all the papers and definitely has an upper hand for its balance of strong accuracy and the ability to provide feature importance, making it well-suited for clinical applications.

Methodology and data:

Coming to my dataset which is Wisconsin Breast Cancer Diagnostic dataset.
It contains 569 samples of digitized fine needle aspirate (FNA) results and 30 numeric features representing characteristics of cell nuclei.
Attribute Information:
1) ID number 2) Diagnosis (M = malignant, B = benign) radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1)
Missing attribute values: none
When we look at the dataset we can observe one thing that is you see column names like radius_mean, radius_se, radius_worst ; texture_mean, texture_se, texture_worst ; area_mean, area_se, area_worst and so on.Why is this the case is that now when our computer looks at the cell, it calculates measurements like how big the nucleus is (radius, area, perimeter), how rough it looks (texture) and how irregular it is (concavity, symmetry, etc.). So now these measurements can help us do the final classification, but the catch is that we don't just have one cell in the image but n number of cells, and thus we calculate measurements like radius_mean: we measure average size of all the cells in that image, radius_se : we measure how much the sizes vary (standard error = variation) and radius_worst: we measure the biggest or most extreme size we saw. This is done for all the features. So, this is feature engineering, which is done to increase the richness of the model's ability to detect patterns, especially for highly variable and sensitive diseases like cancer. Where or example if we ignore one cell and unfortunately if its dimensions are under the worst case, then we might end up classifying the cancer as B when it is M which can lead to a disaster in mean time.
The final diagnosis labels are 'M' for malignant and 'B' for benign tumors.
Class distribution: 357 benign, 212 malignant
The features that are listed in my dataset come from fine needle aspiration (FNA). In this process we use a thin needle and extract cells from the suspicious area in the breast. We can say that FNA is minimally invasive and provides us with good samples for the analysis, as it's just a thin needle; there's not much damage done while extracting the cells. Once these cells are extracted, we will examine them under a microscope to identify if those cells have any kind of oncogenic properties.
When we look at cancerous cells they normally have larger, irregularly shaped nuclei with prominent nucleoli.

So, the features of my dataset are derived from the analysis of the breast cell nuclei. Let's look at the main features and how they indicate to us when the cells are cancerous or not. Starting with radius: cancerous cells often have larger nuclei, so thus a larger value of radius can indicate danger. Texture: Higher variability (standard deviation) can indicate irregular cell structures. Perimeter: Irregularly shaped nuclei have longer perimeters. Area: Larger areas can be indicative of malignancy. Smoothness: Cancerous cells often have less smooth (more irregular) edges (variation in radius lengths). Compactness: Higher values can indicate irregular shapes. Symmetry: Cancerous cells tend to be less symmetrical. Fractal Dimension: Higher values indicate more complex shapes, often seen in cancerous cells and so on. Each of these features also has associated standard error (SE) and worst (maximum) values, providing us with the information about variability and extreme cases within the cell samples. Keeping the mean calculations aside, these other two calculations our dataset is providing us with are very important, as these values are what that actually tell us if the tumor is ignorable or if it is actually dangerous and needs immediate action.

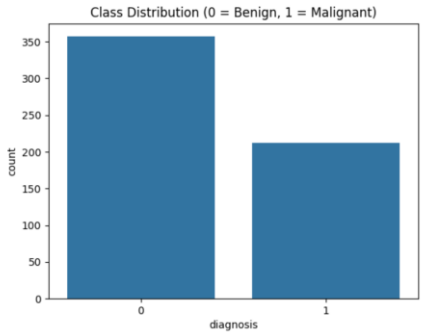
Now coming to the methodology, in this project I have used a complete machine learning (ML) pipeline to classify breast tumors as benign or malignant using the Wisconsin Breast Cancer Diagnostic Dataset (WBCD).

I started off with data preprocessing and exploratory data analysis. So, the data was cleaned by dropping unnecessary columns, and additionally, we had to encode the target variable diagnosis as 0 = Benign and 1 = Malignant. As a part of EDA, we did summary statistics, a class distribution plot, and a feature correlation heatmap, which we will be seeing in the result section. For preprocessing, we have used standard scaler, which helped us in normalizing all the features, and then we used variance threshold, setting threshold values to 0.1, which helped remove all the low variance features, and then the data was split using a 70/30 stratified train/test split.

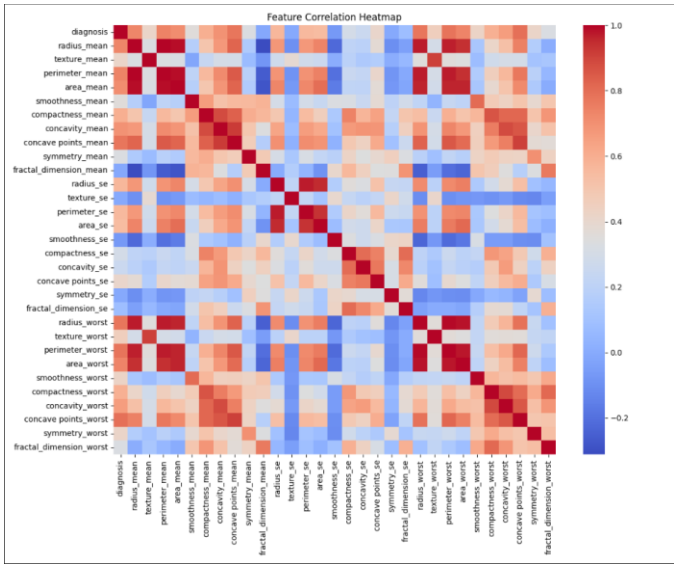
Coming to the next step, I did model selection, where a variety of models were chosen to compare performance, robustness, and interpretability. I used models like Decision Tree, Random Forest, Gradient Boosting, and AdaBoost. These are tree-based models that are well known for high accuracy and feature interpretability. And then I also used SVM (RBF), which handles nonlinear relationships well, and SVM (Linear), which provides interpretable coefficients and supports feature selection. Lastly, I used a neural network (MLPClassifier) to explore deep learning on WBCD.

Then we also performed feature selection. SelectFromModel was applied to models that support feature importance, like Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and SVM (Linear), and feature selection was skipped for models like SVM (RBF), which does not expose feature importances. And a neural network that lacks interpretable feature importance.After this, we made sure to evaluate metrics like AUC, accuracy, and training and testing time for all models before and after feature selection for better analysis. ROC curves were also generated for each model and saved as images.

Results:



i) Class distribution



ii) Correlation heat map

Model	Before FS Train Acc	Before FS Test Acc	Before FS Train AUC	Before FS Test AUC	After FS Train Acc	After FS Test Acc	After FS Train AUC	After FS Test AUC
Decision Tree	1.0000	0.9123	1.0000	0.9173	1.0000	0.8947	1.0000	0.9002
Random Forest	1.0000	0.9357	1.0000	0.9810	1.0000	0.9415	1.0000	0.9798
Gradient Boosting	1.0000	0.9415	1.0000	0.9860	1.0000	0.9240	1.0000	0.9707
AdaBoost	1.0000	0.9474	1.0000	0.9882	1.0000	0.9474	1.0000	0.9869
SVM (RBF)	0.9900	0.9532	0.9976	0.9895	N/A	N/A	N/A	N/A

iii) Complete model performance

Feature	Mean	Std Dev	Min	Max
radius_mean	14.13	3.52	6.98	28.11
perimeter_mean	91.97	24.30	43.79	188.50
area_mean	654.89	351.91	143.50	2501.00
smoothness_mean	0.096	0.014	0.0526	0.1634
concave points_mean	0.0489	0.0388	0.0000	0.2012

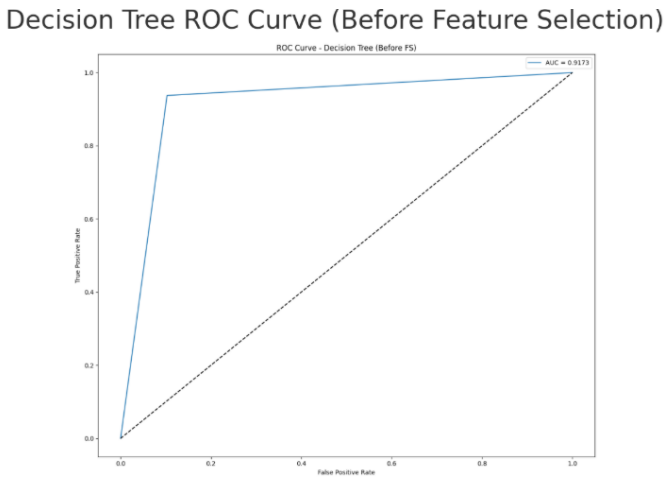
symmetry_mean	0.1812	0.0274	0.1060	0.3040

iv) **Summary statistics**

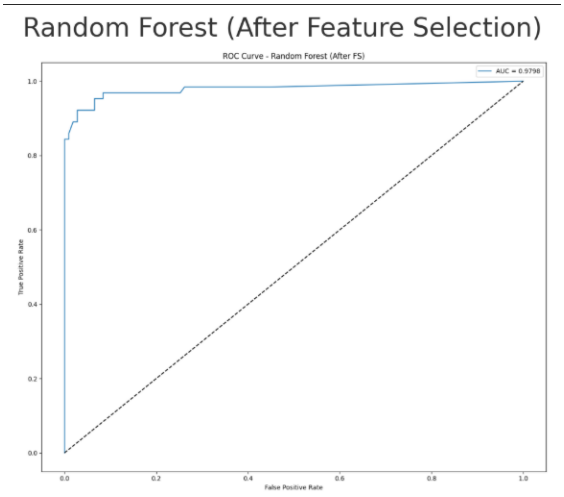
As the summary statistics is very large here are some important values of our concern.

Model	Selected Features
Decision Tree	concave points_mean, area_worst, concave points_worst, perimeter_mean, radius_worst
Random Forest	perimeter_mean, area_mean, concavity_mean, concave points_mean, radius_worst
Gradient Boosting	concave points_mean, radius_worst, area_worst, perimeter_worst, area_mean
AdaBoost	texture_mean, smoothness_mean, compactness_mean, symmetry_mean, concave points_mean

v) **Features selected**

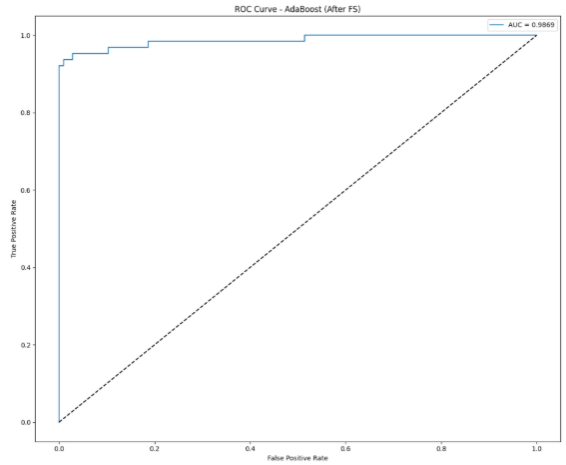


vi) **Roc 1**



vii) **Roc 2**

AdaBoost (After Feature Selection)



viii) Roc 3

Starting with the results of EDA The diagram (i) shows a bar plot showing us how many samples belong to each class, i.e., class 0 is for benign and class 1 is for malignant. Observation shows us that there are more benign tumors (~357 samples) than malignant (~212 samples). This is class imbalance, but not a severe one. Our data is still reasonably balanced to train our ML models. But if the imbalance was too high, then we would have to go about with resampling, weighting, etc. Checking for imbalance of data is very important before we proceed with the further steps, as highly imbalanced data can lead to biased models, which means models that only predict the majority class.

The plot (ii) is the feature correlation heatmap. This is a heatmap of the pairwise correlation between all features, including the target diagnosis. The correlation coefficient value ranges from -1 (strong negative) to +1 (strong positive). Coming to the representation, the dark red square shows a strong correlation, dark blue shows a strong negative correlation, and the white ones or light ones show weak or no correlation. Look at this map. What we observe is that many features are highly correlated with each other (e.g., radius, perimeter, and area-related features). This is important because highly correlated features can cause redundancy in the model. Which can be tackled using feature selection. While some features show strong correlation with the target diagnosis, such as concave points_mean, perimeter_mean, and area_mean, this may suggest that these features are important for the final decision-making means final classification. Overall The heatmap provides useful insights for feature selection and model interpretability.

Next, we have summary statistics at (iv) where we have evaluated the whole summary metrics for our data (mean, std, min, max, percentiles) for each feature. This helps understand the scale and variability of each feature. Looking at them, we understood that features were on different scales; that is, they had wide ranges and high variance. This will be solved after we apply StandardScaler for scaling our features. We will have to do this because ML models perform better when features are standardized. Without scaling, features with larger ranges might dominate the learning process.

Now coming to table (iii) the overall model performances and feature selection, we can say that all models performed very well, demonstrating that this is a well-structured and well-defined classification problem. Specifically, with test AUC scores typically surpassing 0.97, all tested models demonstrated strong performance on the breast cancer classification task. Although the neural network had the highest AUC (~0.990), it is not interpretable because it is a black-box model, which makes it less appropriate for clinical settings where it is crucial to comprehend model decisions. Likewise, the SVM with RBF kernel yielded a very high AUC (~0.989), but it also provided only a limited amount of transparency.

In table (v) After feature selection, Random Forest consistently performed well and showed a very high AUC (~0.980) while retaining interpretability through feature importance analysis, making it a strong contender for real-world application. Along with its outstanding AUC (~0.987), AdaBoost also offered the advantages of transparency and model simplicity. Gradient boosting did fairly well (~0.986 before FS and ~0.971 after FS), but with a little more nuance. Last but not least, the decision tree model is less appropriate as a stand-alone model because it is more prone to overfitting and has the lowest AUC (~0.90 after FS), despite being interpretable.

To make the models easier to understand and more straightforward, feature selection (table v) was used. Following feature selection, AUC in Random Forest and AdaBoost stayed high, suggesting that fewer features were needed for robust predictive performance. Given that the decision tree model depends on full feature splits for optimal performance, a slight decline in AUC following feature selection was to be expected. Since their architectures do not reveal feature importance in a way that the selection algorithm can use, neural network and support vector machine (SVM) models were not subjected to feature selection. Table Y demonstrates that the features that were chosen for each model consistently included concave points_mean, perimeter_mean, area_worst, radius_worst, and concavity_mean. This is consistent with the body of medical literature that emphasizes tumor size, shape, and margin sharpness as important markers of malignancy [1][2][3].

Despite having the highest AUCs, the neural network and SVM (RBF) models are not as suitable for clinical applications due to their interpretability issues. Random Forest and AdaBoost, on the other hand, were better suited for real-world implementation since they combined superior AUC performance with increased transparency. Although there were no total model failures during the project, the decision tree's performance declined a little after feature selection, and the neural network needed more iterations to guarantee correct convergence. For the neural network and SVM (RBF), feature selection was appropriately skipped.

In the end, even though every model produced excellent results, healthcare applications require both interpretability and high predictive accuracy. Consequently, because of their black-box nature, neural networks and SVM (RBF) are not advised, even though their AUC scores are higher. Both Random Forest and AdaBoost provide reliable performance and interpretability; however, Random Forest is the best model for this task because of its robustness and clear feature importance visualization.

Next, coming to the ROC curves, I have implemented ROC curves for all models before and after feature selection, but here in this report we will see 3 important ones to understand them. Mainly the models that gave us the best and poor performance.

The diagram (vi) shows a ROC curve for the decision tree model before applying feature selection. Even though we got an AUC value of 0.91, which is pretty good, compared to other models, this was the lowest. But the limitation with decision trees is that they sometimes make wrong predictions, so their discrimination is not perfect. The next thing is that the curve, as you can see, is not as steep as that of ensemble models. For a perfect model, the ROC curve rises sharply to the top-left corner, which means they have a high true positive rate and a low false positive rate. But the decision tree curve is not as sharp, meaning it sometimes makes more mistakes than the better models (Random Forest, AdaBoost). Decision trees are more prone to overfitting compared to random forests and AdaBoost, as they average out these mistakes even if present.

The next plot (vii) ROC curves show that of random forest after feature selection. As we can see, it shows excellent performance with an AUC of approximately 0.98. As we talked about earlier about a good model, this goes about similarly by the curve rising sharply to the left corner. indicating high true positive rates and low false positive rates. This supports Random Forest as a highly suitable model for this classification task, combining strong predictive power with interpretability.

Coming to the third ROC diagram (viii), it is about AdaBoost after feature selection. This also shows excellent performance like that of random forest; its ROC curve is steep and well above the diagonal baseline. This indicates that AdaBoost is capable of accurately distinguishing between malignant and benign cases while remaining a relatively simple and interpretable model.

Even though all of the models that were tested for this project produced excellent results (as shown in Table iii), I decided to highlight various aspects of model performance in the main text by presenting the ROC curves of three representative models. To highlight the drawbacks of simpler models, the decision tree is used as a baseline single-model classifier. The inclusion of Random Forest and AdaBoost is due to their outstanding performance, robustness, and interpretability, which make them more appropriate for real-world use in the healthcare industry. Despite their strong performance, other models like SVM RBF and neural networks were not included in the main figures because of their lower clinical interpretability and black-box nature. For completeness, additional ROC curves for each model are included in the Appendix.

Discussion and conclusion:

This project of mine explored the application of various machine learning models to finally classify the benign and malignant forms of breast cancer. The main focus was not only to get the best predictive performance but also to study and understand how different models behave on this task, how feature selection affects this whole process, and finally, which model could be chosen for clinical practice.

If we look at the metrics overall, almost all models performed well with AUC values around 0.97. Particularly successful ensemble techniques that combined high accuracy and interpretability were Random Forest and AdaBoost. Although SVM with RBF kernel and neural networks had the highest raw AUC values, their interpretability issues make them less suitable for clinical settings where model transparency is crucial for adoption and trust.

Coming to feature selection, we saw that they reduced model complexity and then also preserved the performance of the ensemble models. But this led to a slight decrease in the performance of the decision trees, which was consistent with the weaknesses of the single-tree models. Neural Network and SVM RBF models skipped feature selection appropriately, as their architecture does not support it in this pipeline.

Three ROC curves were selected to highlight important points in the results report: AdaBoost, a powerful and clinically useful model; Random Forest, a high-performing and interpretable model; and Decision Tree, a straightforward baseline model. For the sake of clarity and conciseness, other models that did well were left out of the main text.

So finally, we can say that ensemble models are highly effective with good performance and also for interpretability regarding practical use. This project also helped us understand not just the technical but also the practical aspects of ML, i.e., not simply reporting the highest accuracy but selecting models that can be realistically deployed, like how we selected RF over neural even though it gave the highest performance. These findings align well with observations in the literature review, where RF was also highlighted as a strong model [2][3][5].

Future work:

There is always scope for improvement. Firstly, the models in our project were trained on a smaller dataset, and it was basically limited to a single institution. But in reality, we have new populations and a wider range of variability in the data, and thus we need to implement that in our testing so that we get better generalizability. Stronger proof of the current models' robustness and generalizability would come from testing them on bigger and more varied datasets, such as multi-center cohorts with various imaging and diagnostic features.

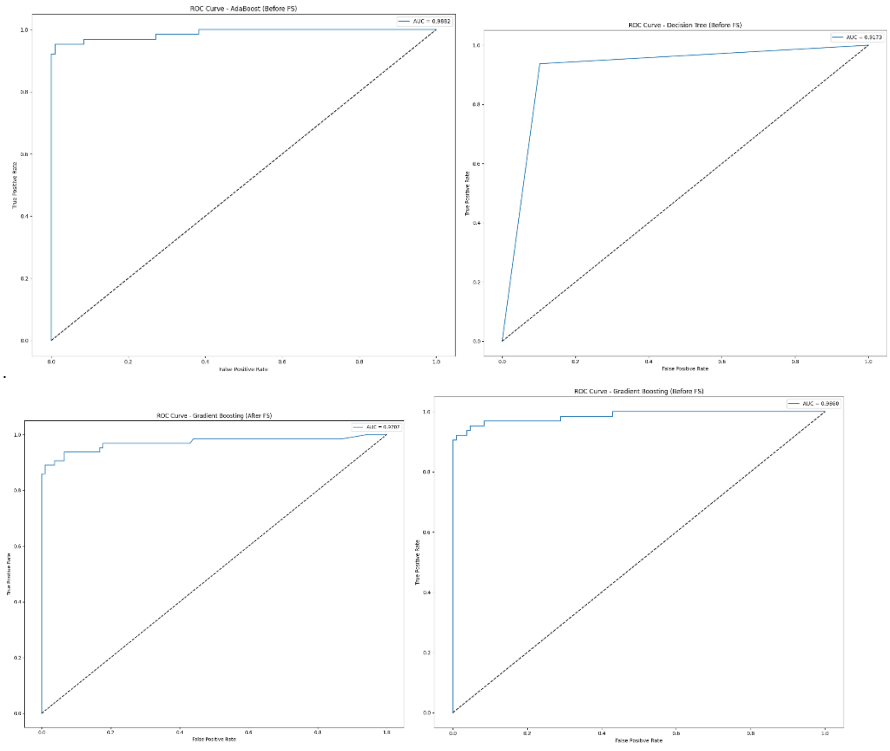
In our project, as we can see, we used a simple tabular dataset provided to us, but in real applications, we can enhance this by using more advanced feature engineering. for instance, by using domain expertise to develop composite features (such as shape ratios and texture complexity metrics). Additionally, multi-modal learning may be possible with the integration of data from other modalities, such as genomic profiles, radiology reports, or histopathology images, which could result in richer and more accurate models.

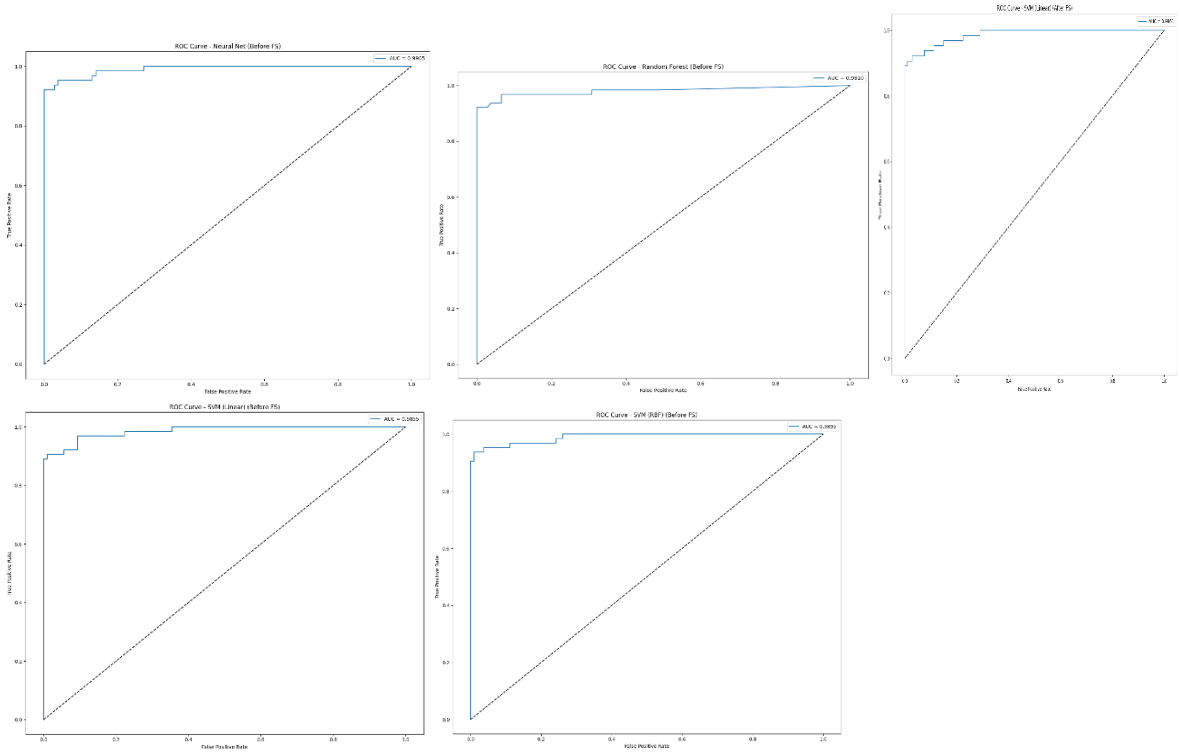
Our models mainly focused on metrics like AUC and accuracy, but in real-life clinical applications, including much more in order to make sure that predicted probabilities match actual risk levels, which is a crucial prerequisite for clinical decision-making, future research could include model calibration techniques. Furthermore, techniques for estimating uncertainty (such as Monte Carlo dropout or Bayesian neural networks) may be able to identify instances in which human review is required due to the model's uncertainty.

Coming further explain ability is also very important. So now we use random forest and AdaBoost, which have quite a level of interpretability, but in the future, we can use something like sophisticated explainable AI (XAI) methods, like counterfactual explanations or SHAP values, to help clinicians better understand individual predictions. This could increase confidence and make practical adoption easier.

And finally, besides doing all these, collaborating with clinicians would make the whole process more perfect and trustworthy. Because this way we can make sure to understand how model outputs should be presented and how predictive models can be integrated into clinical workflows, which would help move this research closer to real-world impact.

Appendix:





References:

[1] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, “Machine learning techniques to diagnose breast cancer from fine-needle aspirates,” Cancer Letters, vol. 77, no. 2-3, pp. 163–171, 1994.

[2] R. Sharma and R. Patel, “Breast Cancer Detection Using Machine Learning Algorithms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, no. 3, pp. 362–367, 2018.

[3] M. Rovshenov and M. Peker, “Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using WBCD,” Procedia Computer Science, vol. 120, pp. 616–622, 2017.

[4] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, “Deep Learning for Health Informatics,” IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 4–21, 2017.

[5] P. Patil and P. Thakur, “Breast Cancer Diagnosis Using Random Forest,” International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 4968–4972, 2020.