Jasmine Justin & Katharina Dowlin
Final Report
198:439
May 9 2025

Predicting Cancerous Tumors in Breast using Data from Imaging

Dataset: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

GitHub Repository: https://github.com/JasmineJustin7/CS439FinalProject/tree/main

Overview of Research:

Our goal of this project and research is to conclude if it is possible to accurately predict if a tumor in breast tissue is benign or malignant based on various physical factors of the tumor itself. If this can be done accurately and efficiently, this could prevent any usage of invasive procedures that have to be performed instead. Solving this dilemma would save time, money and prioritize safety when properly diagnosing a patient with cancer without trouble. We approached this binary classification problem using machine learning models including Logistic Regression, Random Forest, Support Vector Machine, and several boosting methods.

Motivation:

Research into both malignant and benign tumors exists to help better understand how breast cancer diagnoses can be caught early and treated properly and sufficiently. Some questions that exist in this field of research include, how benign breast conditions can influence cancer risk as well as the risk of overdiagnosis and overtreatment. Despite being benign, some benign tumors have been linked to an increased risk of developing breast cancer. As for the risk of overdiagnosis and overtreatment, studies show that some breast cancers identified through mammograms never become life-threatening, thus leading to unnecessary treatment or diagnoses. There is also work related to AI's increasing role in diagnostic accuracy which is a topic aligned with the goal of this research problem we want to solve and analyze. According to an article by the *Journal of Clinical Medicine*, AI has shown superior performance compared to the traditional methods of mammography and CT scans when classifying tumors.

Dataset:

The dataset we use for our research comes from Wisconsin University. It is a tabular dataset that includes diagnostic information on tumors both malignant and benign. Some features included within this dataset include radius, texture, perimeter, area, and smoothness of the tumor. We are utilizing a supervised learning approach by which we take features of each tumor and use them to find any important features that are useful in predicting types of tumors, how reliable and influential these features are as well as their usefulness.

Primarily Data Analysis and Data Cleaning:

To begin the data analysis, we removed null values from the data set. We then created a correlation heatmap matrix to see which variables might be highly correlated. This analysis revealed substantial multicollinearity among variables, suggesting the risk of overfitting if all features were used as-is. 13 features had a correlation of .85 or higher with each other. To further look into the multicollinearity of these features we found the VIF of all features. This led to us seeing that radius_mean, for example, had an VIF of 3806 much over the expectable threshold. Many variables had VIFs of over 100, which was a huge cause for concern. To address this issue, we implemented a VIF-based feature reduction process, iteratively removing the variable with the highest VIF until all remaining features had VIF scores below a threshold of 5. After removing redundant features using VIF filtering, we kept 13 key features that each provide unique information about the tumors. These include measurements of size, shape, texture, and symmetry.

Models

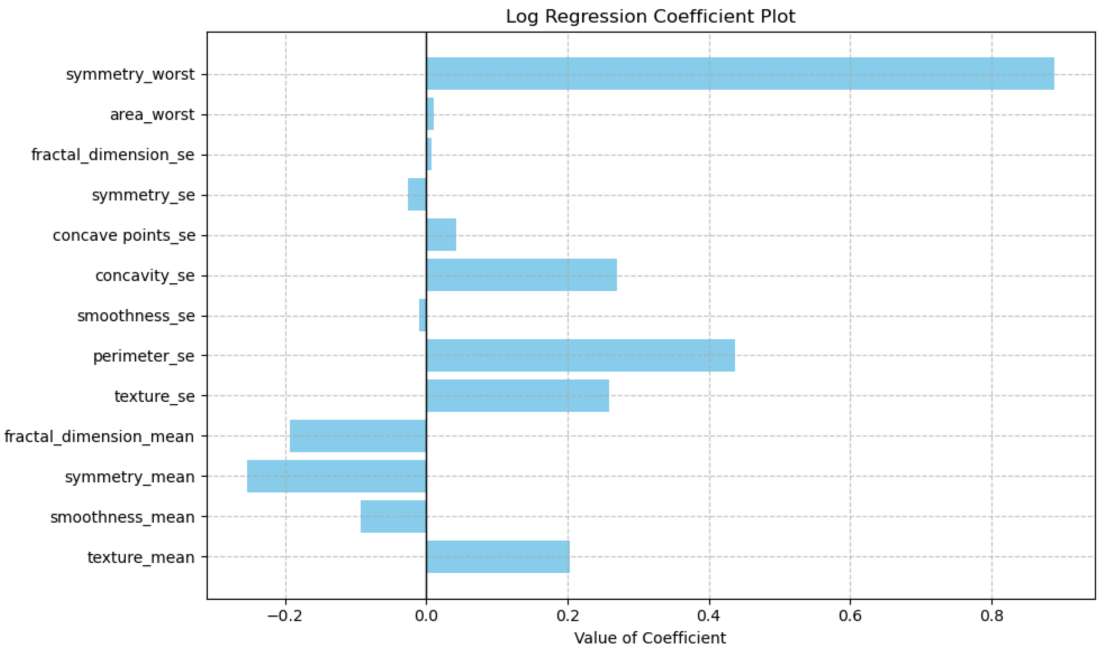
We used several machine learning algorithms. Each model was trained on an 80/20 train-test split and evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

Models - Log Regression

Using log regression on this dataset revealed a lot of significant interpretations of the dataset. First, the classification report revealed good news about the usefulness of the features in the dataset. Recall yielded 0.91, which shows that out of all actual malignant cases, our model correctly identified 91% of them. This is vital because in cancer detection, missing a malignant tumor can be critical in a real-world situation.

Precision yielded a 0.95 which shows that out of all the cases our model predicted as malignant, 95% of them were actually malignant. This shows a high rate of consistency which is beneficial. The less false alarms a model is susceptible to, the less reliable it is.

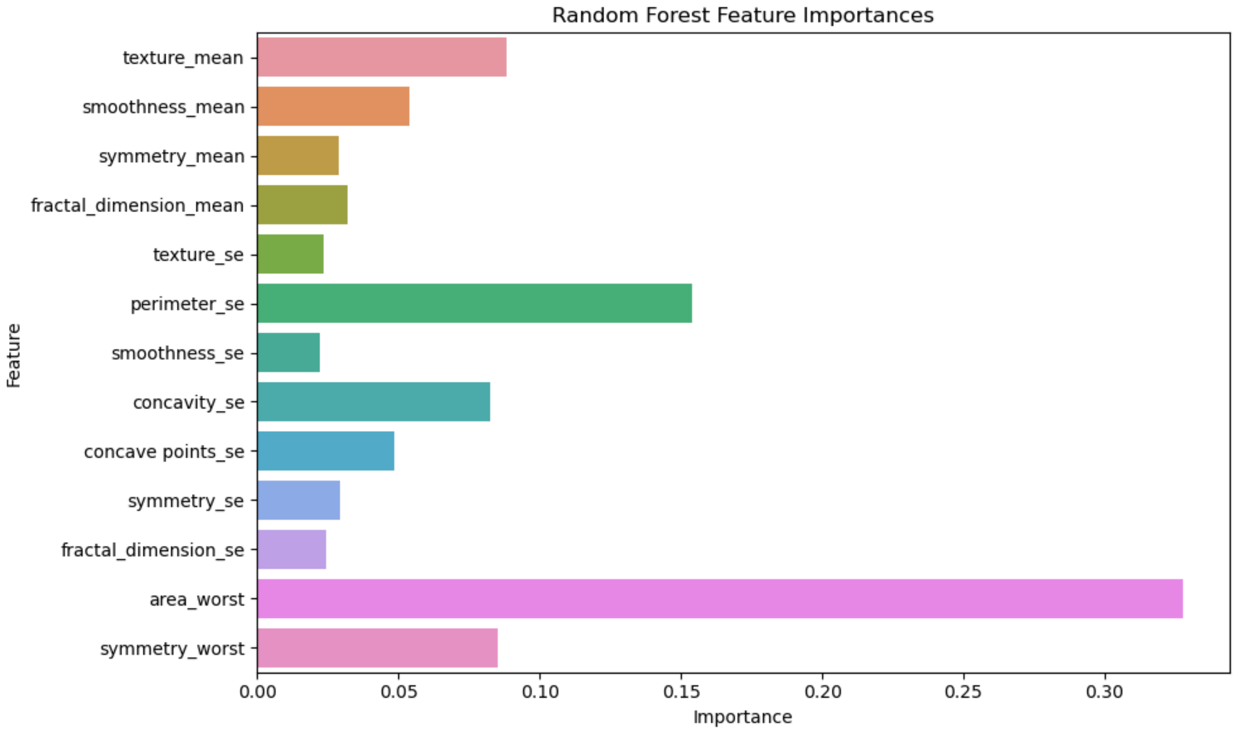
Finally, the F1 score was calculated to be 0.93. The F1 score is a calculation using the precision and recall by using the harmonic mean of the two. An F1 score this high proves great for our goals. The overall accuracy of the model was 94.7% indicating consistent performance across benign and malignant cases. Given these strong baseline results and the simplicity of logistic regression, we determined that additional scaling or preprocessing was not immediately necessary for this model.

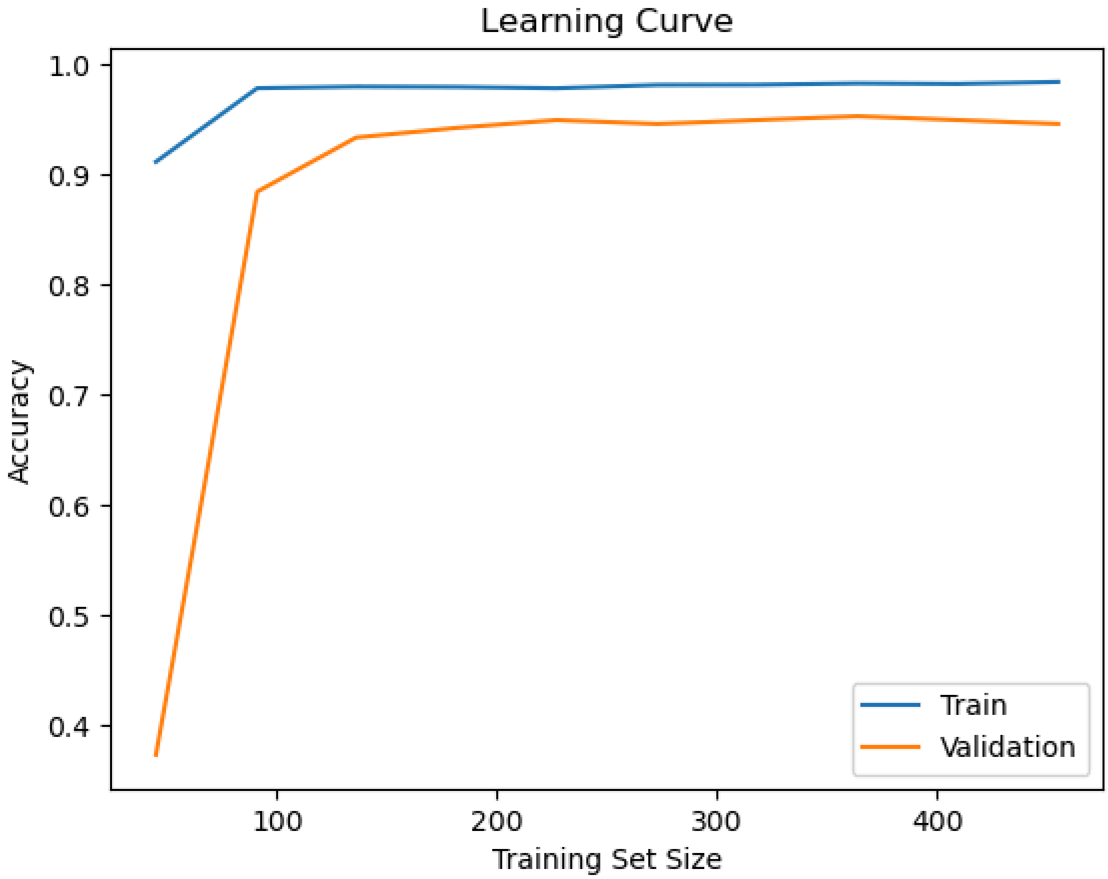


Models - Random Forest

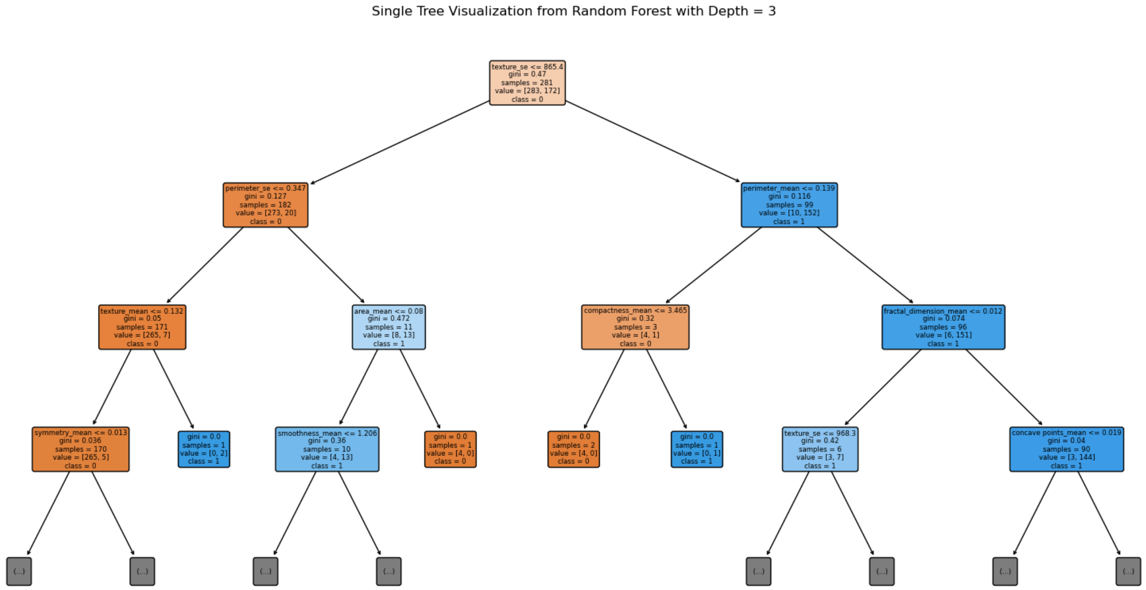
Random Forest was a key model that we used to show what features of the tumors matter most when determining whether or not a tumor is malignant or benign. The top features turned out to be the worst area which had a 35.61% importance, perimeter standard error which had a 17.61%, worst symmetry points having 9.67%, concavity standard error having 8.88%, and mean texture which had a 8.29% importance. Together, these five features make up over 50% of the model's decision power. In conclusion, when predicting whether a tumor is benign or malignant, the model found out that the size and shape of the tumor at its worst point are the most important features it looks for along with the area, radius, concave points and concavity.

Initially, the original model we created had no restrictions, but when the learning curve was graphed it showed a training accuracy of 1.0 across all train set sizes which indicated overfitting of the model. We then changed the model to use 100 decision trees, but each tree is limited to a maximum depth of 5 to prevent it from overfitting the training data. We also set up a minimum of 10 samples to split a node and at least 4 samples in each leaf, helping make the model more stable. This changed the learning curve to the one below which showed the training accuracy no longer being a straight line at 1 meaning it was no longer overfitting the data.


Random Forest Feature Importances
Feature
texture_mean
smoothness_mean
symmetry_mean
fractal_dimension_mean
texture_se
perimeter_se
smoothness_se
concavity_se
concave points_se
symmetry_se
fractal_dimension_se
area_worst
symmetry_worst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Importance


Learning Curve
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
100
200
300
400
Training Set Size
Train
Validation

These findings are in alignment with clinical findings as well. Malignant tumors tend to be larger, more irregular in shape and more complex in structure than benign tumors. Looking back at our expected results prior to conducting our analysis of the dataset, we believed that smoothness of the tumors, area, texture and spread of the tumor would be the most important features to distinguish malignant and benign tumors but our results show that only area was important compared to all other physical features which was a finding we are glad held post analysis.

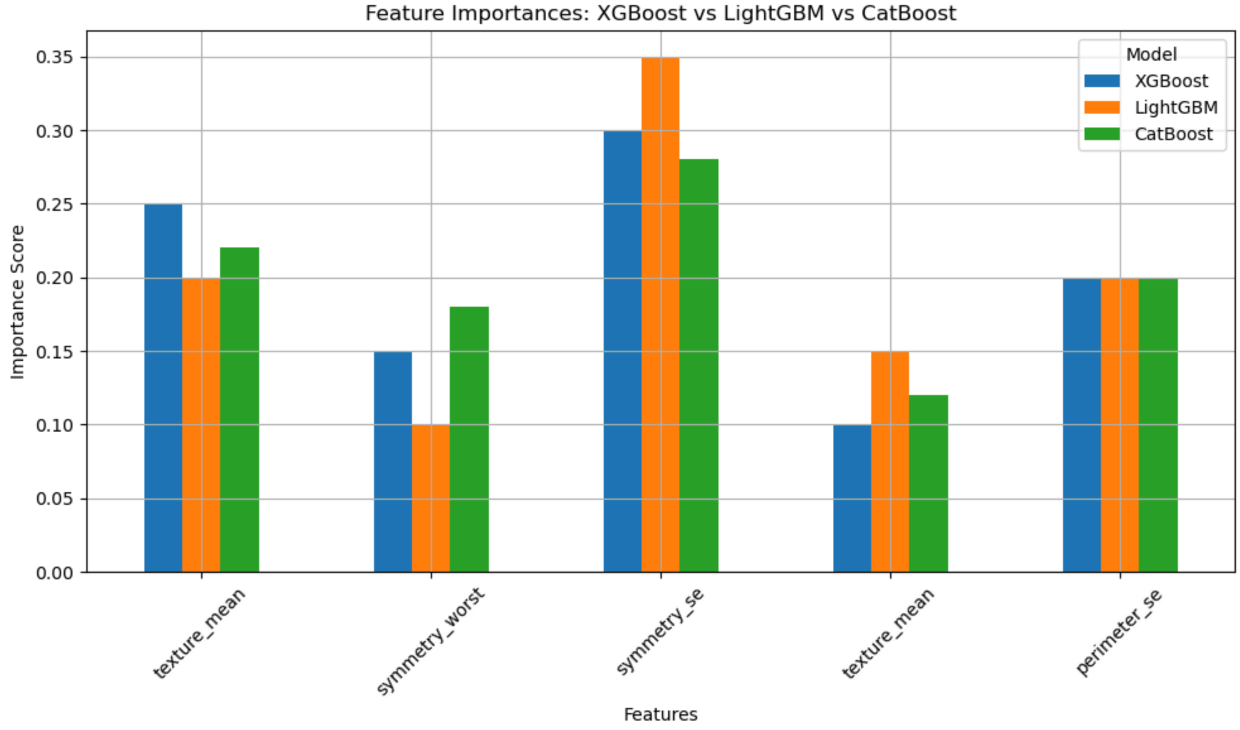

Single Tree Visualization from Random Forest with Depth = 3

Models - SVM

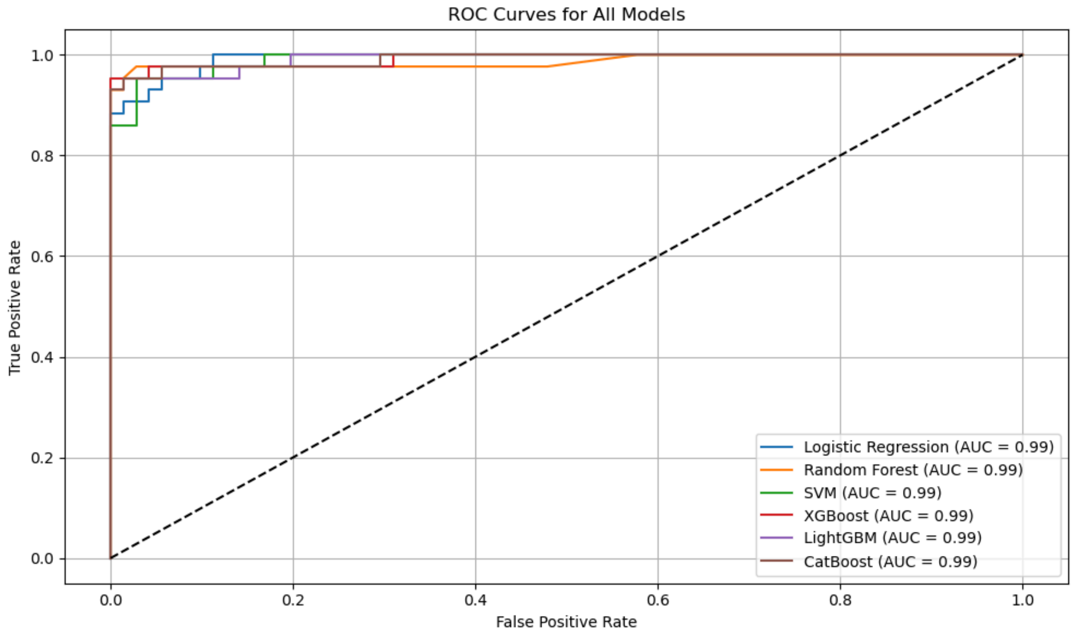
The Support Vector Machine model achieved an accuracy of 95%. The recall for benign tumors was 1 which means it correctly identified all benign tumors. For malignant, the recall was .86 meaning 86% were correctly classified. The model had a precision of 1.00 for malignant tumors, meaning every tumor it predicted as malignant was actually malignant. The F1 score, which balances precision and recall, was 0.92 for malignant tumors and 0.96 for benign, showing a strong overall balance.

Models - Boosting

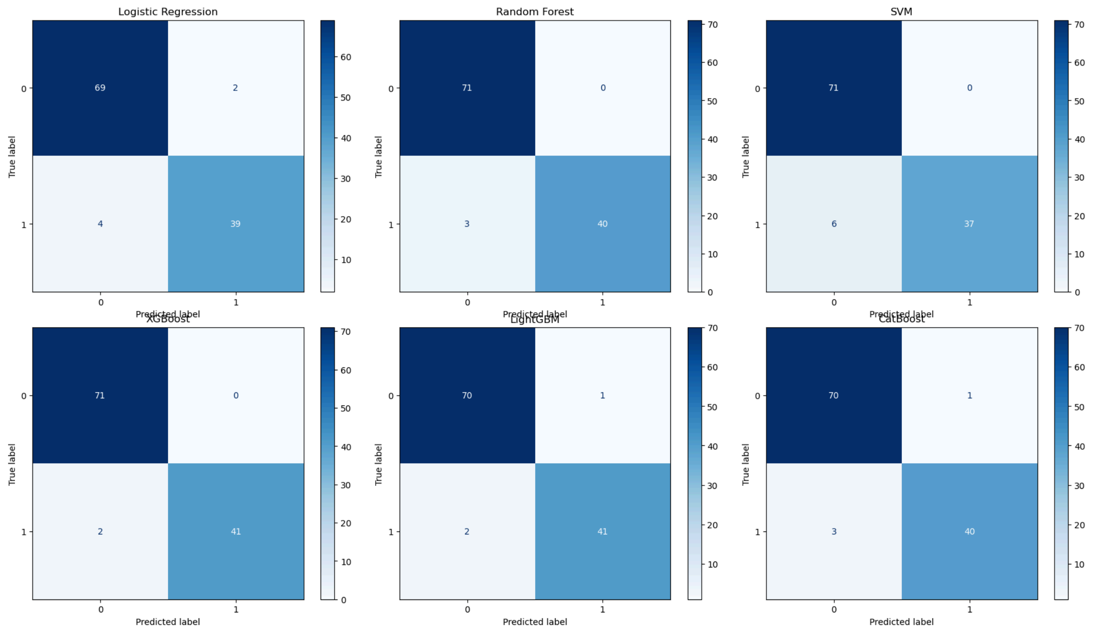
The boosting algorithms (XGBoost, LightGBM, and CatBoost) all performed well in classifying tumors as benign or malignant. Among the three, XGBoost achieved the highest overall performance, with a 98% accuracy and a perfect precision score of 1.00 for malignant cases, indicating that all predicted malignant tumors were indeed malignant. LightGBM followed closely with a 97% accuracy, while CatBoost also performed strongly with a 96% accuracy. Despite slight differences in performance, all three models consistently identified area_worst as the most important feature, confirming that tumor size plays a critical role in predicting malignancy. Other important features shared across models included texture_mean, symmetry_worst, and concavity_se, which describe the tumor's texture, shape, and irregularity which are traits commonly associated with cancerous growths.



Visualization and Analysis:



The ROC curves show that all models (Logistic Regression, Random Forest, SVM, XGBoost, LightGBM, and CatBoost) achieved near-perfect discrimination between malignant and benign tumors, each with an AUC score of 0.99. This indicates that the models are highly effective in distinguishing between the two classes across a range of thresholds.



The confusion matrices across all models show strong predictive performance, with XGBoost and LightGBM performing particularly well by minimizing both false positives and false negatives. XGBoost perfectly identified all benign cases and misclassified only 2 malignant tumors, achieving near-perfect accuracy. LightGBM also misclassified just 3 samples total. Random Forest correctly classified all benign cases and missed only 3 malignant tumors. Logistic Regression and CatBoost showed solid performance with only minor misclassifications, while SVM struggled slightly more, misclassifying 6 malignant tumors meaning it had a lower recall.

Discussion:

In summary, our initial predictions and expectations were that a tumor's features such as smoothness, area, texture, and spread would be the most critical in predicting malignant tumors. While area did prove to be a significant predictor of malignancy, our actual results showed that other factors were more influential across nearly all models. These features – worst area, perimeter, worst symmetry, concavity and mean texture were the top most significant and influential factors. We found these results to be a bit surprising but overall insightful. These results emphasize that the tumor shape and structure are more of an important feature than we had realized, especially at the tumor's worst points. Although there were differences from our expectations, our models performed exceptionally well.

Conclusion:

XGBoost unsurprisingly achieved the highest accuracy of 98%. Additionally, all models we utilized achieved AUC scores of 0.99, confirming strong diagnostic potential. After correcting all signs of overfitting, ultimately, we can conclude with confidence that our results align with existing clinical understanding of physical traits of tumors but we also demonstrated the power of machine learning in striving for more efficient diagnostic accuracy for breast cancer detection. Looking ahead, this research could be extended in many ways. In continuing the trends of using largely numerical models and data, information about the patents age, genetic markers, and family history, could greatly improve the model by allowing us to better understand other factors that can contribute to cancer. We could also work to combine models by seeing where each model might fall short and see if we can minimize error that way. Lastly, testing our models using real world data which can be much more noisy and have more variance, can be a way to ensure that our models are correctly capturing patterns that appear in the real world, so that they could be better applied in the real world.

Works Cited:

Wang, Xin, et al. “A Deep Learning-Based Radiomics Model for Breast Lesion Classification.” Frontiers in Oncology, vol. 11, 2021, doi: 10.3389/fonc.2021.629321.
https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.629321/full

Akram, Muneeb, et al. “A Computer-Aided Diagnosis System for Breast Cancer Classification using CT Images and Deep Learning.” Journal of Clinical Medicine, vol. 12, no. 4, 2023, article 1582, doi:10.3390/jcml12041582.
https://www.mdpi.com/2077-0383/12/4/1582

Sasso, Samantha Lauriello. “Benign Breast Disease can Double your Risk of Breast Cancer, Study Shows.” Health.com, 20 Apr. 2023
Benign Breast Disease and Breast Cancer Risk

Oaklander, Mandy. “Many DCIS Breast Cancer Patients Don’t Need Treatment, Study Finds.” TIME, 27 Aug. 2015
DCIS Breast Cancer: New Study Suggests It's More Serious Than Doctors Thought | TIME

Molteni, Megan. “With Breast Cancer, the Best Treatment may be No Treatment.” Wired, 11 Oct. 2021,
With Breast Cancer, the Best Treatment May Be No Treatment | WIRED