Comparative Analysis of Predictive Models for Tumor Classification Based on Nuclei Morphological Features

Team A2

QST BA 305: Business Decision Making With Data

By Katherine Beaty, Niki Singh, May Al Khalifa, Riley Krisch

Problem

Cancer is the generic term for a large group of diseases that can affect any part of the body and is defined as the rapid and abnormal growth of cells beyond their normal limits. Currently, cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 (WHO, 2025).

Cancer mortality is reduced when cases are detected and treated early. There are two components of early detection: early diagnosis and screening. Early diagnosis allows for the minimal spread of cancer and increases the chance of successful treatment. Screenings are performed with the use of an MRI or CT scan to identify any abnormalities.

Our analysis aims to find the model that performs the best in detection as an early detection tool. This report presents six different models and their performance in detecting malignant tumors from the morphological features of the nuclei.

Data Overview

Our dataset presents tumor classification, benign and malignant, gathered from 569 patients. The dataset includes 33 columns, 30 of which are morphological features of the tumors that describe characteristics of the cell nuclei present in the image. The features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. During FNA, a very thin needle is used to collect a sample of cells, tissue or fluid from an abnormal area or lump. The sample is then examined under a microscope. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. All feature values are recorded with four significant digits.

The outcome variable of our analysis is diagnosis. Diagnosis can either be M or B, representing malignant or benign, respectively. The class distribution is 357 benign, 212 malignant.

Data Cleaning

Column Name	**Type**	**Count**	**Description**
id	Int	569	Id number of tumor
diagnosis	Obj	569	Diagnosis of breast tissue (M = malignant, B = benign)
radius_mean	Float	569	Mean distances from center to points on the perimeter
texture_mean	Float	569	Mean of standard deviations of gray-scale values
permineter_mean	Float	569	Mean size of the core tumor

area_mean	Float	569	Mean area of the tumor
smoothness_mean	Float	569	Mean of local variation in radius length
compactness_mean	Float	569	Mean of perimeter^2/area-1.0
concavity_mean	Float	569	Mean of severity of concave portions of the contour
concave points_mean	Float	569	Mean for number of concave portions of the contour
symmetry_mean	Float	569	Mean for symmetry of tumor
fractal_dimension_mean	Float	569	Mean for “coastline approximation” - 1
radius_se	Float	569	Standard error for the mean of distances from center to points on the perimeter
texture_se	Float	569	Standard error for standard deviation of gray-scale values
perimeter_se	Float	569	Standard error for size of the core tumor
area_se	Float	569	Standard error for size of tumor
smoothness_se	Float	569	Standard error for local variation in radius lengths
compactness_se	Float	569	Standard error for perimeter^2/ area -1.0
concavity_se	Float	569	Standard error for severity of concave portions of the contour
concave points_se	Float	569	Standard error for number of concave portions of the contour
symmetry_se	Float	569	Standard error for symmetry of tumor
fractal_dimension_se	Float	569	Standard error for “coastline approximation” - 1
radius_worst	Float	569	“Worst” or largest mean value for mean of distances from center to points on the perimeter
texture_worst	Float	569	“Worst” or largest mean value for standard deviation of gray-scale values
perimeter_worst	Float	569	“Worst” or largest mean value for size of the core tumor
area_worst	Float	569	“Worst” or largest mean value for tumor area size

smoothness_worst	Float	569	“Worst” or largest mean value for local variation in radius lengths
compactness_worst	Float	569	“Worst” or largest mean value for perimeter^2/area - 1.0
concavity_worst	Float	569	“Worst” or largest mean value for severity of concave portions of the contour
concave points_worst	Float	569	“Worst” or largest mean value for number of concave portions of the contour
symmetry_worst	Float	569	“Worst” or largest mean value for symmetry of tumor
fractal_dimension_worst	Float	569	“Worst” or largest mean value for “coastline approximation” - 1
Unnamed:32	Float	0	unknown

Dropped Columns

Based on the above information, we dropped the “id” column because it is a unique identifier and does not contribute to our classification analysis. We dropped “Unnamed:32” because it was a completely empty column and added no useful information. Additionally, we dropped all standard error columns as, after testing with our models, they did not significantly impact the analysis.

Renamed Columns

To prevent any confusion during analysis, spaces in column names were replaced with an underscore (“_”).

Label Encoded Columns

To make analysis easier for our algorithms, we changed the “diagnosis” column from “M” and “B” to “1” and “0”, respectively.

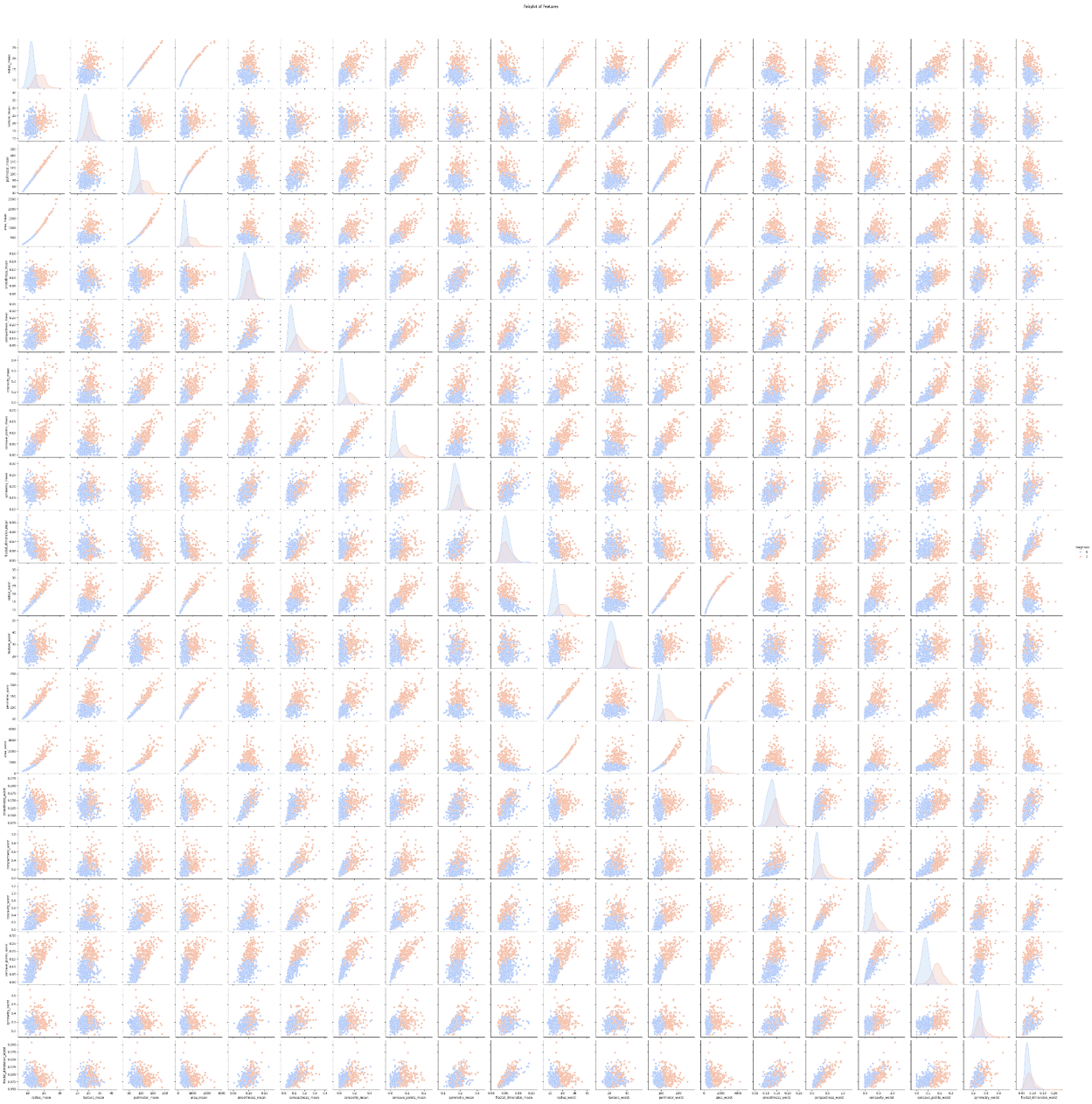
Imbalanced Data Check

The distribution of malignant and benign was 212 malignant (37.3%) and 357 benign (62.7%). This is somewhat statistically unbalanced, so we have used metrics beyond accuracy to measure how well a model performed. We addressed precision and recall to assess how well a model identifies the minority class, an F1-score to find the mean of precision and recall, and a Precision-Recall Curve (PR)/Area under the curve (AUC) to measure how well a model distinguishes between the classes.

Baseline Model

The baseline f1 is 0.387 with a majority class of 0 (Benign) diagnosis.

Data Visualization

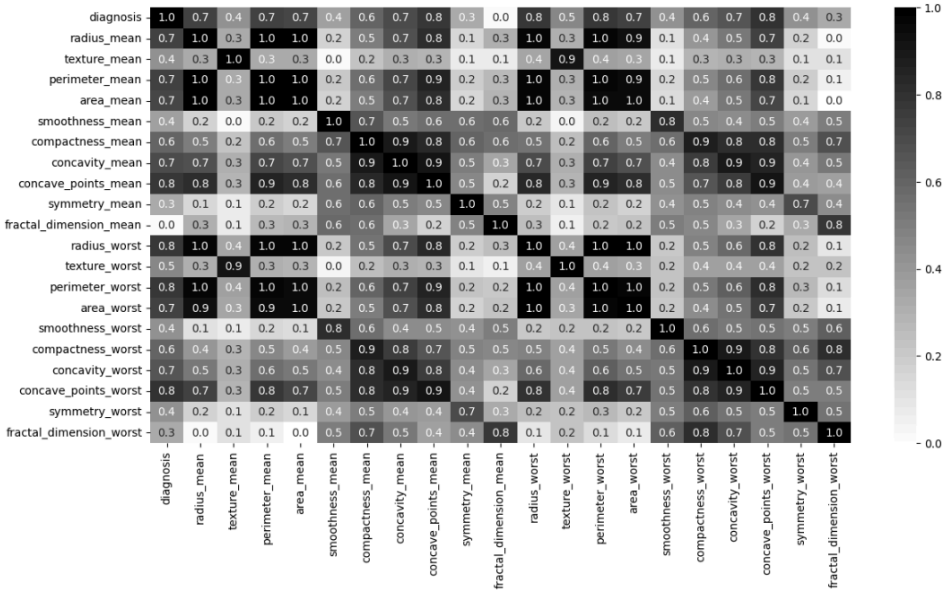


Using pairplots to visualize an overview of the relationship between every pair of features through scatterplots. The diagonal plots represent the distribution of classes in each feature, where many features appear to be right-skewed, specifically, heavy right-skewed distributions include area_worst and area_mean. The off-diagonal scatterplots highlight several insights with some strong linear patterns, some curved non-linear patterns (like perimeter_mean and area_mean), as well as strong distinctions between different class clusters for some feature relationships, meaning these pairs may help more easily/distinctly classify instances.

Principal Component Analysis (PCA)

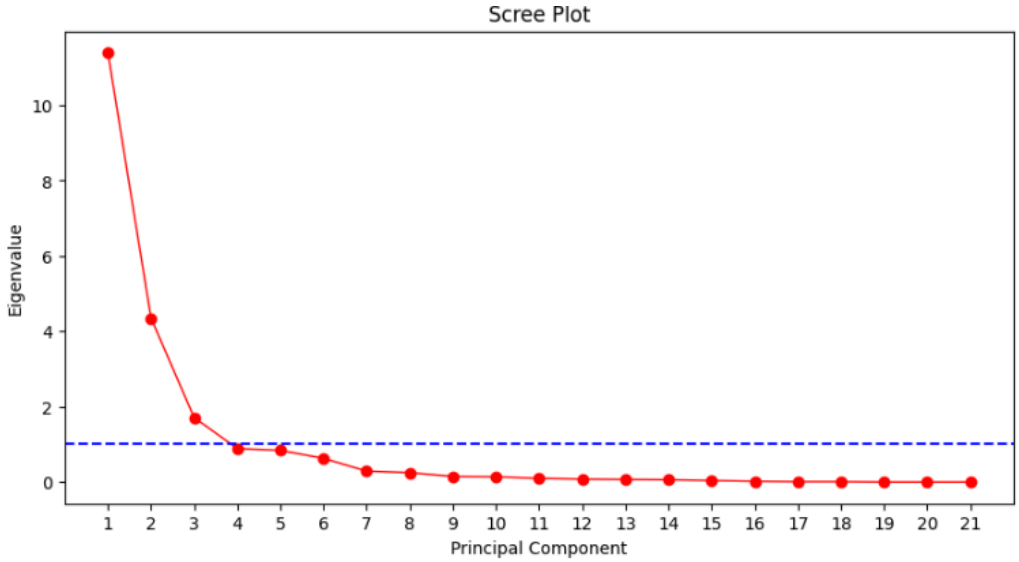
With our dataset, we knew that the features would be highly correlated because each broad feature has three features derived from it: mean, standard error, and worst. To possibly combat the multicollinearity that would arise, we created two versions of our model, one that did not use PCA and one that did use PCA.

Correlation plot before PCA



To decide upon the number of principal components, we performed a scree test in addition to placing the information of variance and cumulative variance into a table.

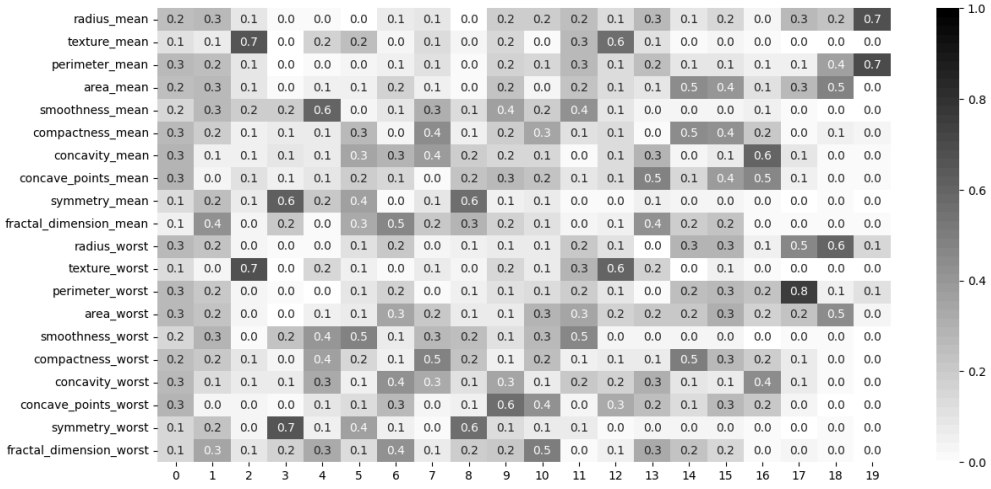
Scree Test



	% of variance explained	Cumulative % explained
0	0.532	0.532
1	0.219	0.751
2	0.086	0.837
3	0.045	0.882
4	0.040	0.922
5	0.030	0.952
6	0.013	0.966
7	0.008	0.973
8	0.007	0.980
9	0.005	0.985
10	0.004	0.989
11	0.004	0.993
12	0.003	0.995
13	0.002	0.997
14	0.001	0.999
15	0.001	0.999
16	0.001	1.000
17	0.000	1.000
18	0.000	1.000
19	0.000	1.000

Although the scree test recommended 3 principal components, we decided on 4 PCA components to ensure that close to 90% of the original data information is retained.

Correlation plot of principal features to original features



\We found that the first few principal components (especially the first 4) were relatively interpretable, showing strong correlations with meaningful original features such as size and texture. Later components, however, were less interpretable due to weaker correlations.

Modeling

Decision Tree

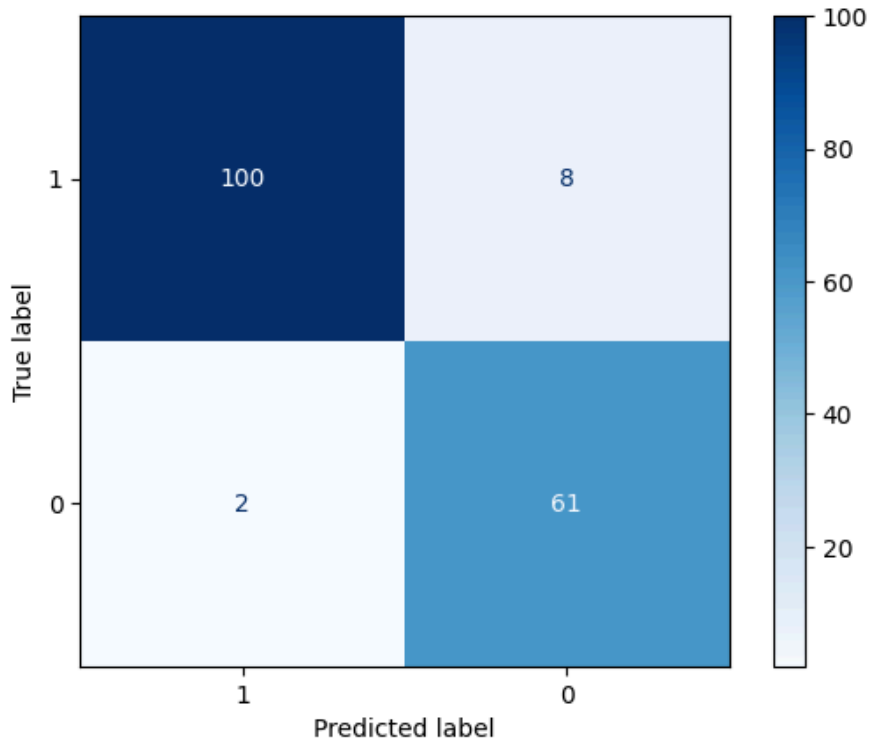
Full Tree (No PCA)

Parameters

Number of Features	Number of Nodes	Number of Leaves	Maximum Depth
20	41	21	8

Results

Confusion Matrix



Precision	Recall	F1 Score
0.884058	0.968254	0.924242

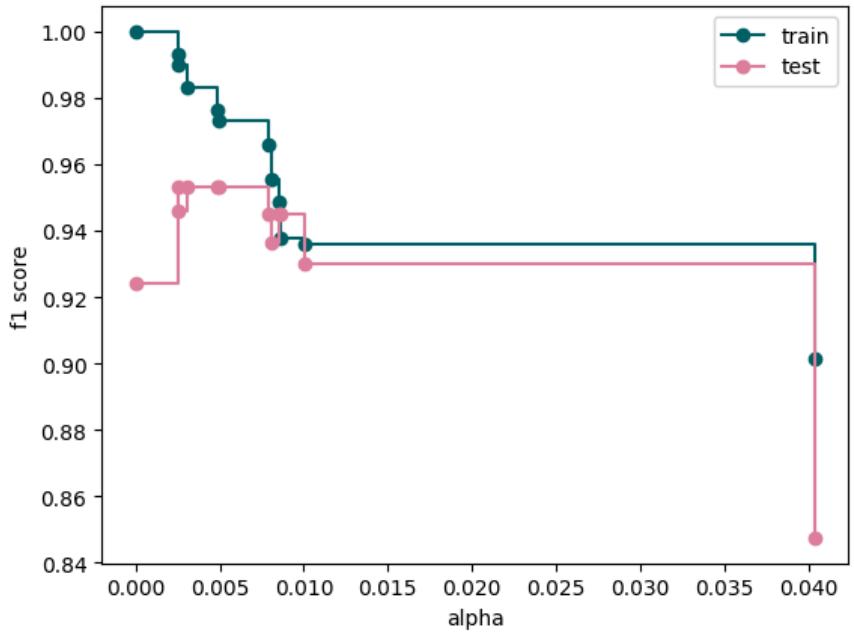
Overall, the model performed well, especially in recall. Our precision score is not great, but not bad by any means. This means some benign tumors are being misclassified as malignant. The tree isn't excessively deep, which helps reduce overfitting and keeps the model relatively interpretable. Although the depth is moderate, 20 features in a tree of 41 nodes could start to memorize noise, especially if some features are highly correlated, which we know are from the pre-PCA correlation plot. Additionally, the model still has the opportunity to overfit with only the stopping criteria of reaching a pure leaf node or cycling through all of the columns.

Pruned Tree (No PCA)

Parameters

Number of Features	Number of Nodes	Number of Leaves	Maximum Depth
20	9	5	3

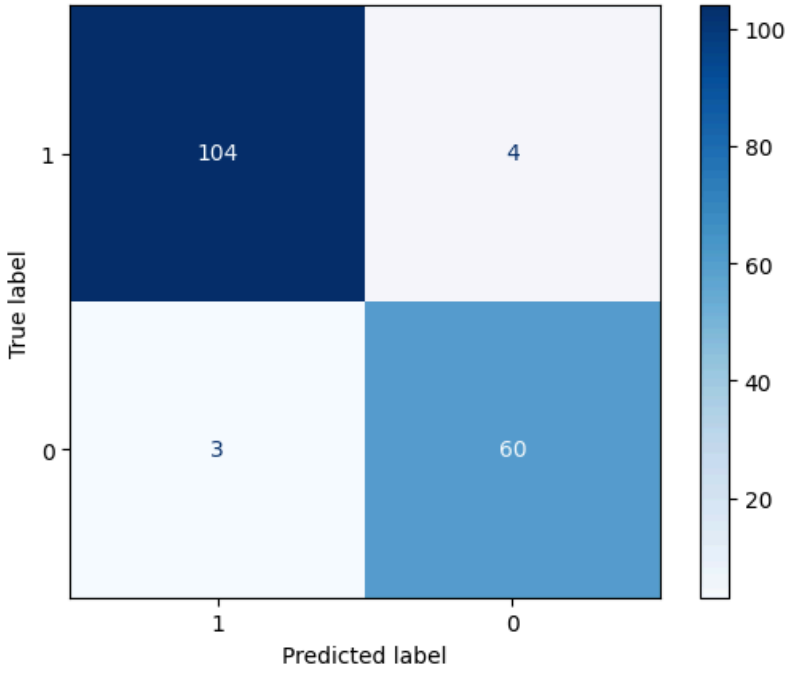
Plotting Tree predictive recall as a function of alpha



Note: We removed one outlier alpha from the graph as it drastically reduced the recall of the model and made the graph hard to interpret.

Results

Confusion Matrix



Precision	Recall	F1 Score
0.937500	0.952381	0.944882

To combat the possibility of overfitting, we performed a pruned tree using cost-complexity pruning with F1 score. This approach ensured that our final model maintained high predictive performance on the testing set while avoiding unnecessary complexity, making it more interpretable and robust for clinical application. However, we noted that the recall fell slightly while precision greatly improved compared to the full tree.

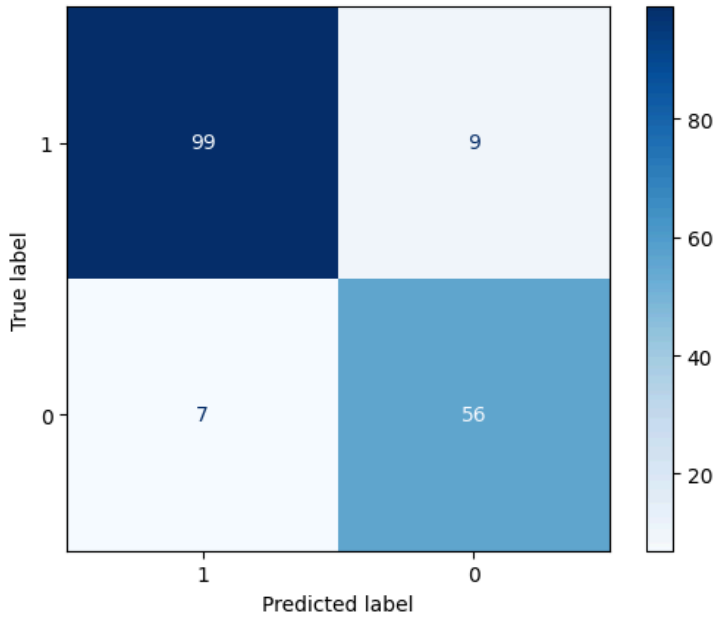
Full Tree (PCA)

Parameters

Number of Features	Number of Nodes	Number of Leaves	Maximum Depth
4	43	22	7

Results

Confusion Matrix



Precision	Recall	F1 Score
0.861538	0.888889	0.875000

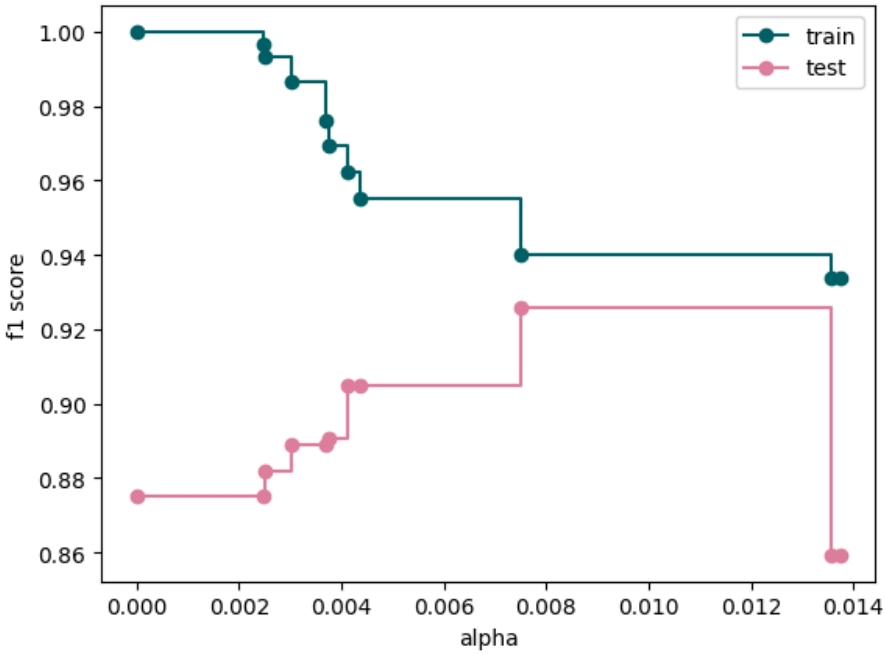
We applied PCA to reduce the dimensionality from 30 to 4, retaining approximately 87% of the total variance. This preprocessing step was performed after train-test splitting and included data scaling. The decision tree, compared to the non-PCA full tree, did not perform as well. Additionally, the use of PCA reduced model interpretability, as the tree's decisions are now based on abstract components rather than understandable tumor nuclei features.

Pruned Tree (PCA)

Parameters

Number of Features	Number of Nodes	Number of Leaves	Maximum Depth
4	7	4	3

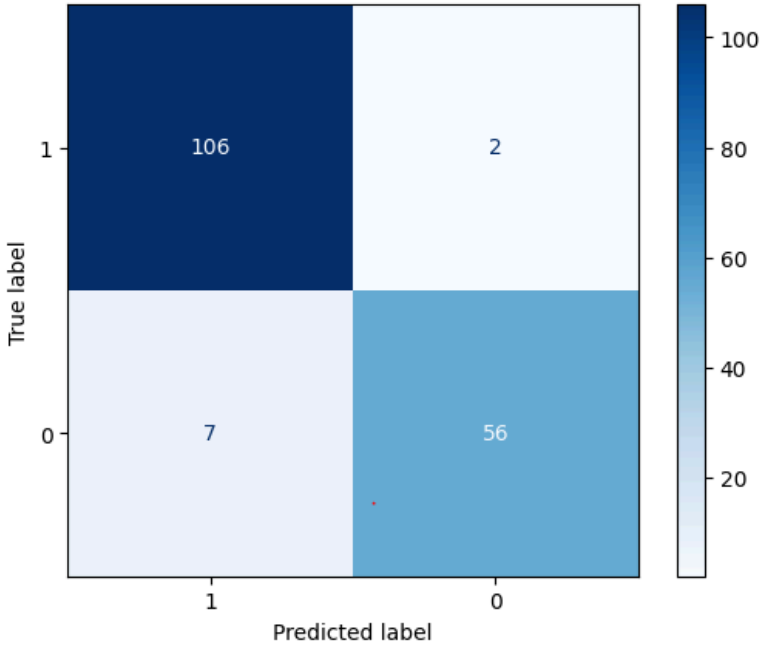
Plotting Tree predictive recall as a function of alpha



Note: We removed one outlier alpha from the graph as it drastically reduced the recall of the model and made the graph hard to interpret.

Results

Confusion Matrix



Precision	Recall	F1 Score
0.965517	0.888889	0.925620

To, once again, combat overfitting, we created a pruned tree with PCA as well. It had pretty good scoring, performing better in precision and therefore F1 in turn. Similar to the full tree with PCA, we reduce the interpretability of the tree in hopes of getting a better F1 score. While pruning improved generalizability and minimized overfitting, the overall performance slightly improved compared to the unpruned model.

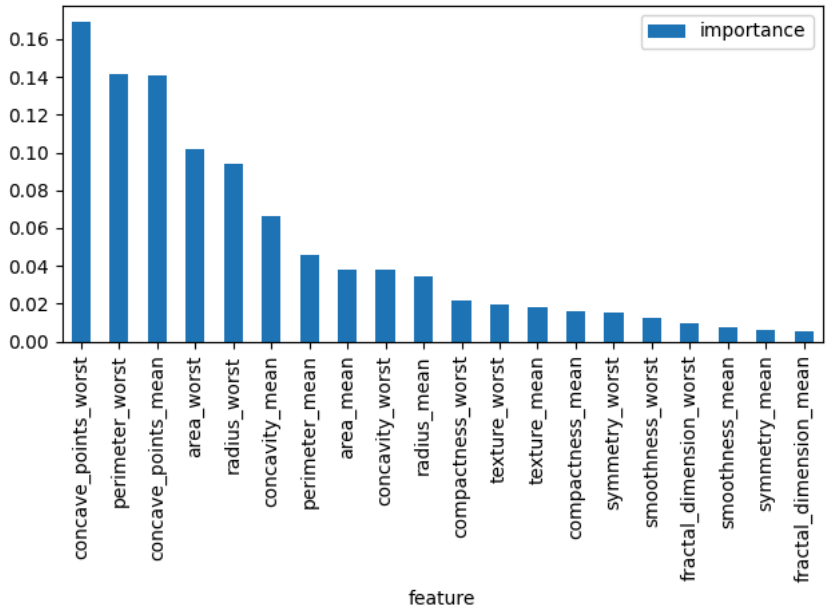
Random Forest

No PCA

Parameters

Number of Trees	Criterion
10,000	Gini

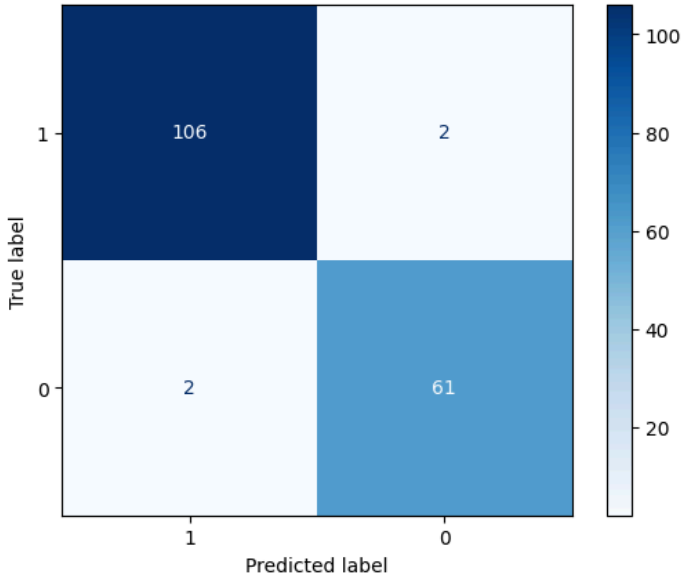
Feature Importance



Based on the above graph, we can conclude that concave_points_worst had the largest impact on the model's predictions. Summing the top 5 most important features accounts for about 65% of the entire importance.

Results

Confusion Matrix



Precision	Recall	F1 Score
0.968254	0.968254	0.968254

The random forest model performed well across the board, and the high precision and recall mean the model is both accurate and reliable across classes. Since we are using the original features, the feature importances are interpretable, which is valuable for clinical settings. Unfortunately, using this method with more data or more features can increase the computational time and possibly make it infeasible to use on a large scale. Additionally, while random forests are robust, having highly correlated or noisy features can still slightly affect generalization in edge cases.

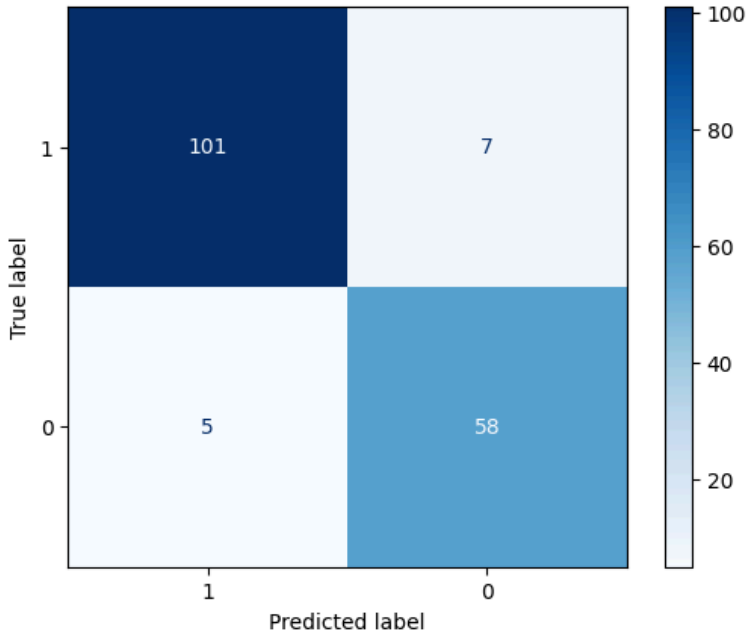
PCA

Parameters

Number of Trees	Criterion
10,000	Gini

Results

Confusion Matrix



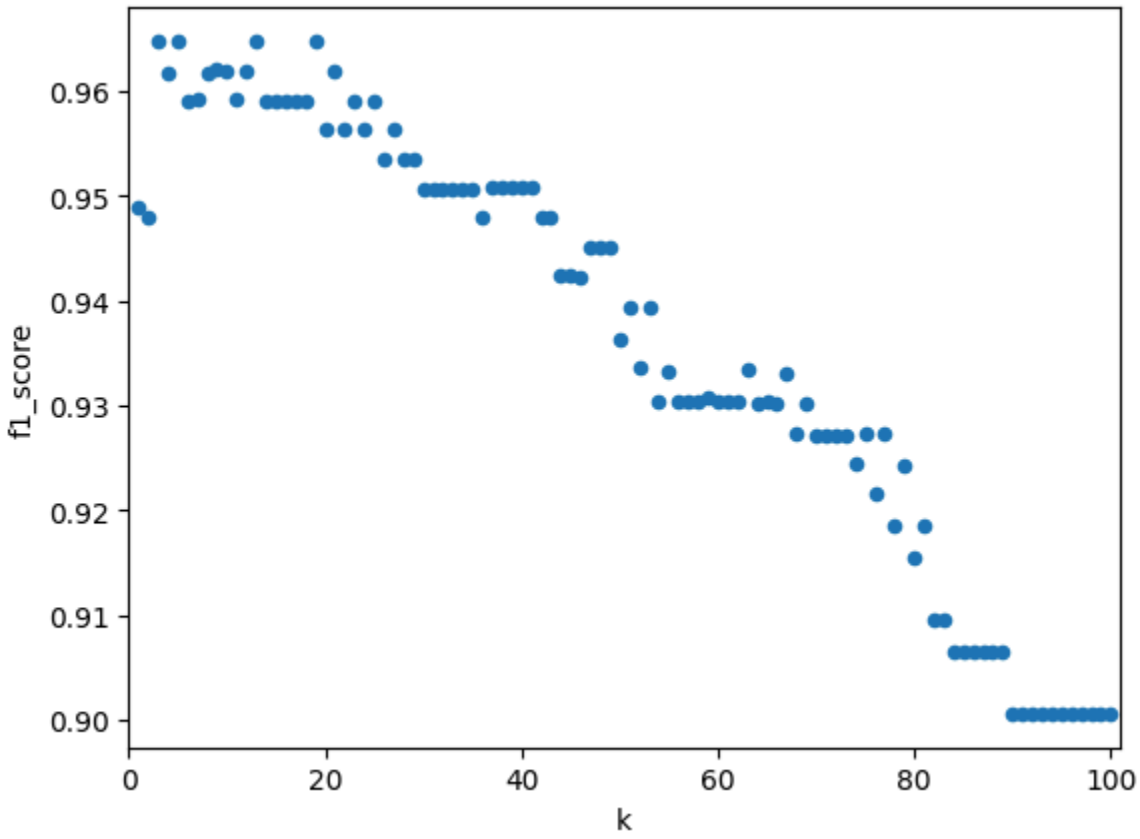
Precision	Recall	F1 Score
0.892308	0.920635	0.906250

The PCA random forest did slightly worse in precision but better in recall compared to its non-PCA counterpart. Further, since we are only using 4 principal components, it leads to faster training and inference, which is especially helpful with large datasets or limited resources. However, the use of PCA reduced model interpretability, as the classification relied on transformed components rather than directly observable tumor nuclei features. The model also sacrificed a small portion of data variance (13%) to achieve dimensionality reduction and efficiency.

K Nearest Neighbor

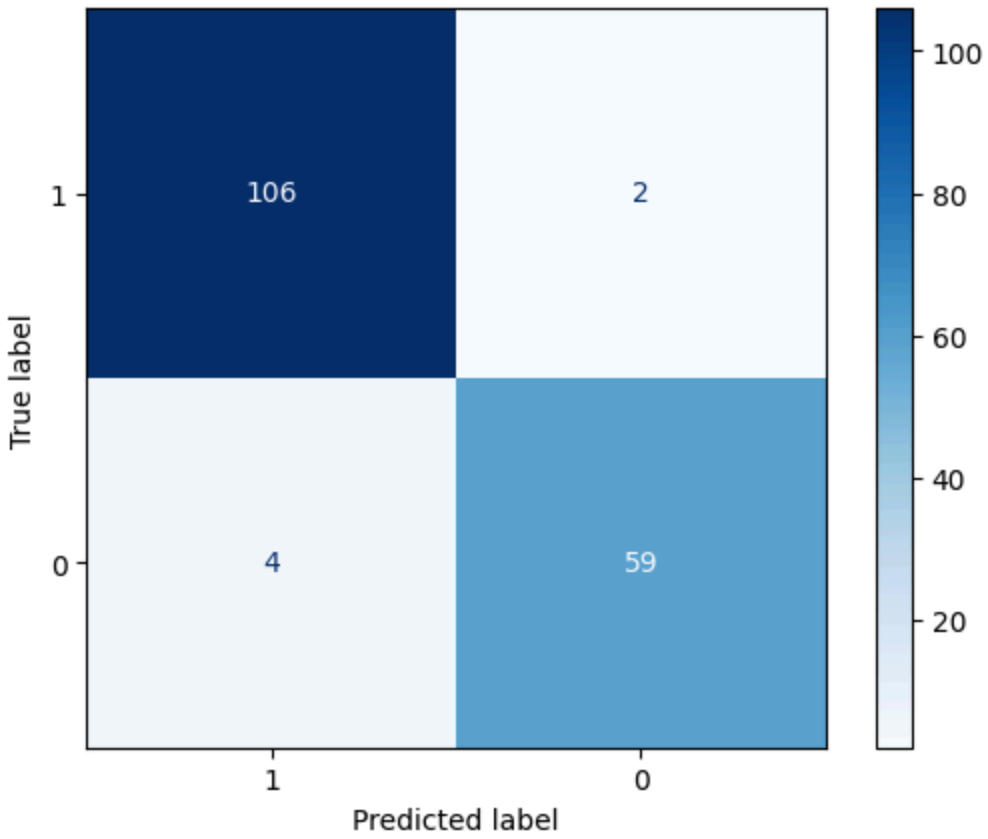
A KNN classification model was implemented, and k values ranging from 1 to 100 were tested to find the best k that produced a model with the highest F1 score.

No PCA



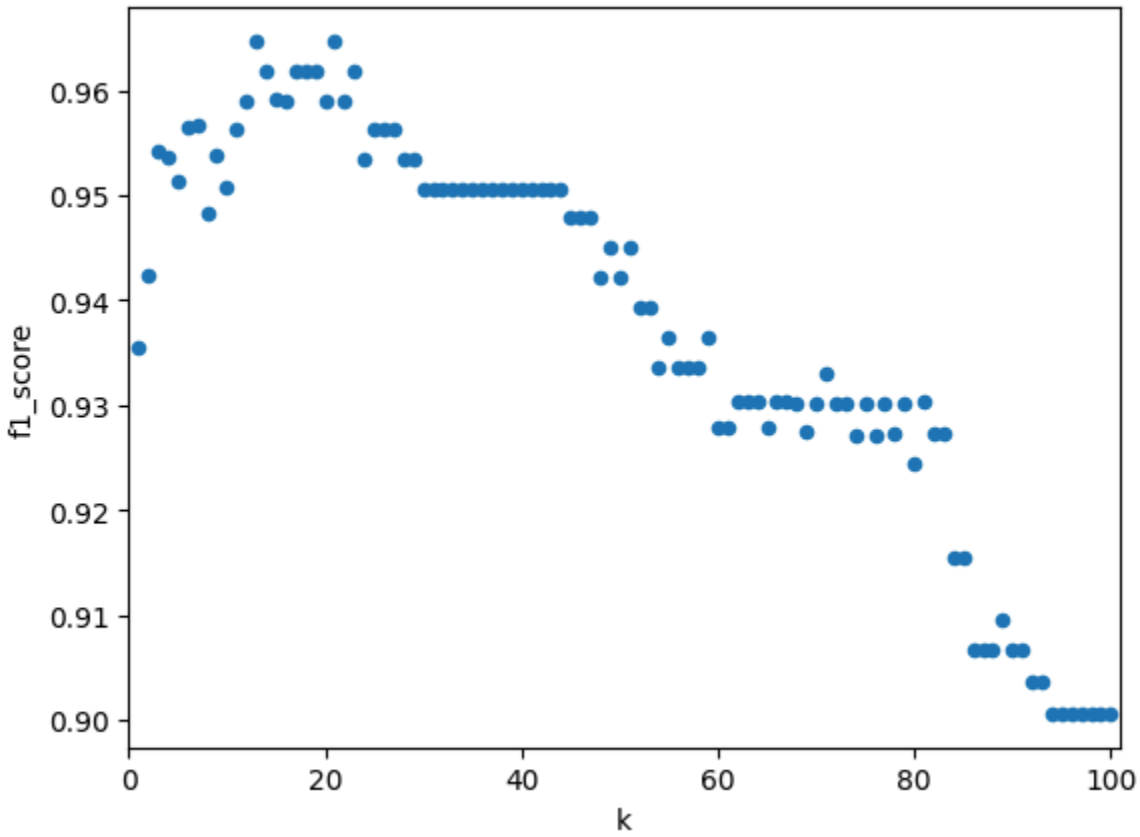
The best k found that had the highest resulting F1 was k = 13. Following this discovery, a new KNN model was trained with the best k = 13 to ensure we obtain the best F1. The resulting model acquired statistics:

Precision	Recall	F1 Score
0.967213	0.936508	0.951613



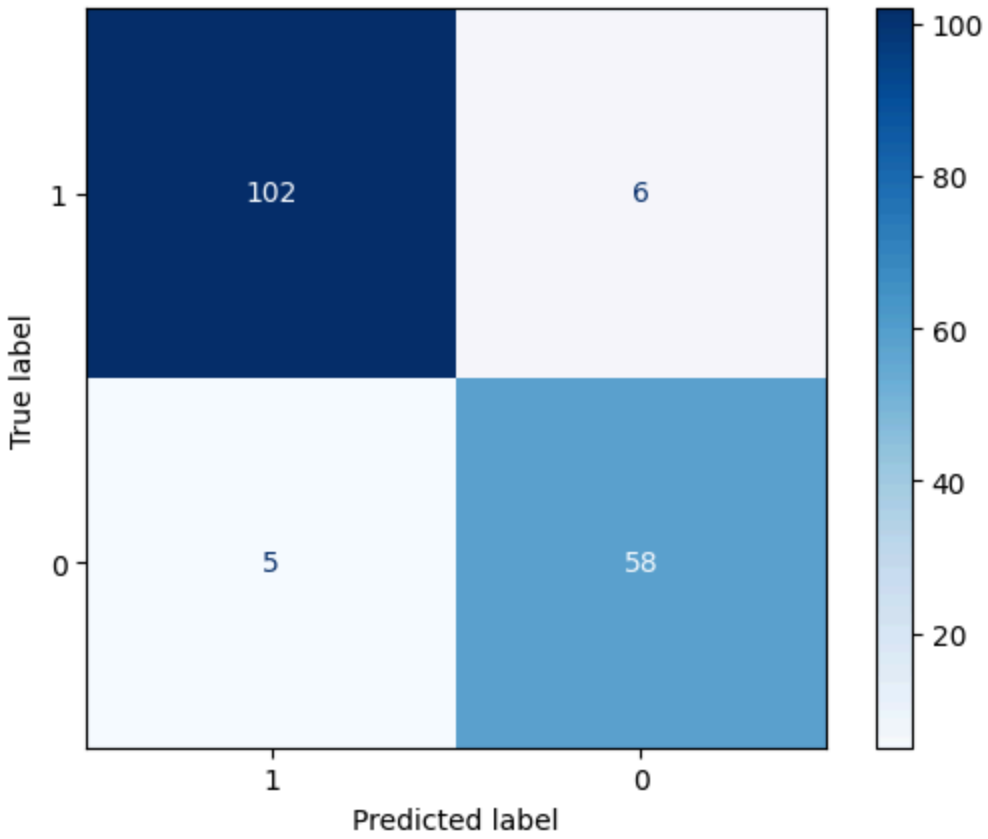
KNN is one of the middle models in terms of F1 score at 0.91, with random forest, logistic regression, and neural network all performing better overall.

PCA



For the PCA KNN model, the best k found that had the highest resulting accuracy was k = 13. Following this discovery, a new KNN model was trained with the best k = 13 to ensure we obtain the best MSE. The resulting model acquired statistics:

Precision	Recall	F1 Score
0.906250	0.920635	0.913386



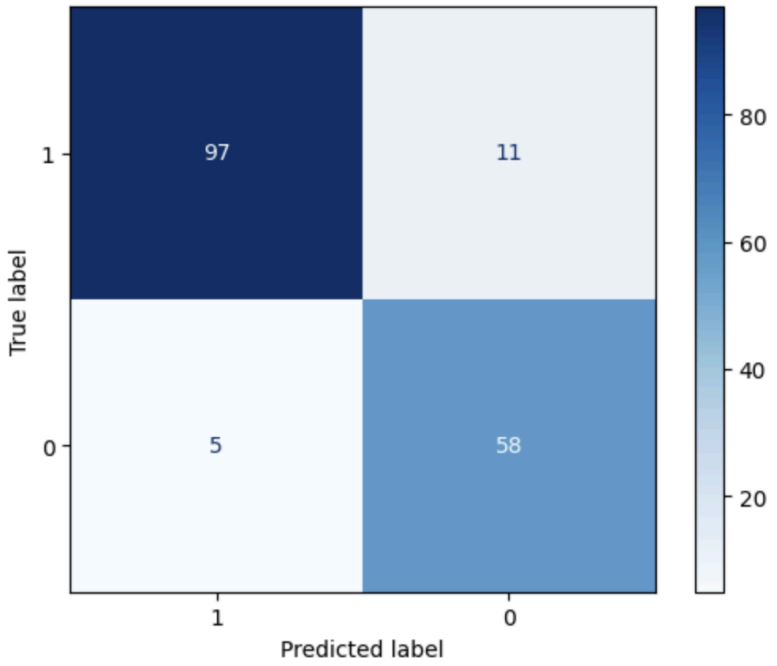
After applying PCA to the KNN model, KNN seems to be one of the middle models in terms of F1, precision, and recall, and PCA has not improved the model in any of those, instead slightly lowering performance.

Naive Bayes

Due to our dataset possessing features that were continuously numerical, it made the most sense to utilize Gaussian Naive Bayes instead of Multinomial Naive Bayes.

No PCA

Results

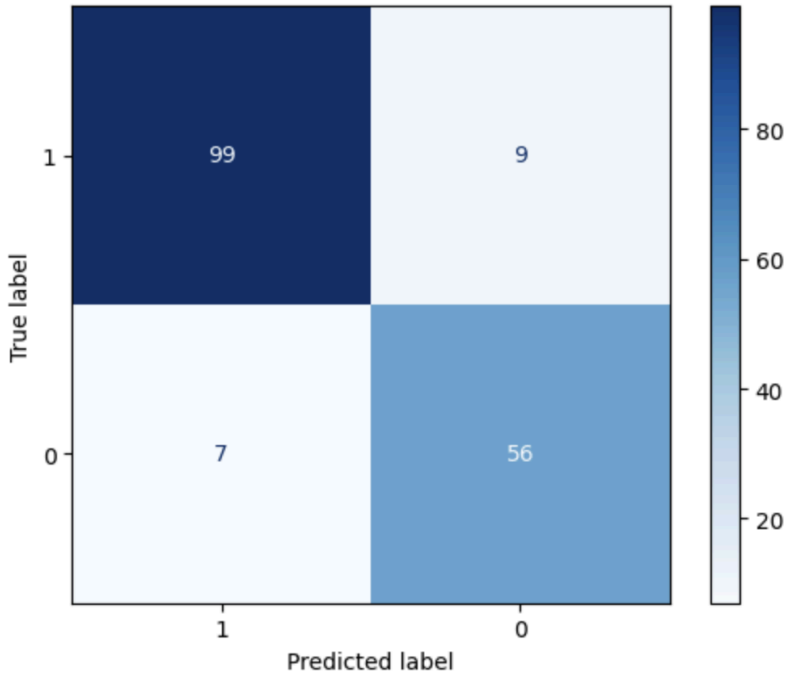


Precision	Recall	F1 Score
0.840580	0.920635	0.878788

Out of all the models without PCA, this model had the worst precision, recall, and F1 score. While it did perform reasonably well, particularly in identifying malignant tumors, indicating a high recall value, its relatively low precision shows that this model is prone to false positives, predicting some benign tumors as malignant. It also did not balance precision and recall as well compared to other models.

PCA

Results



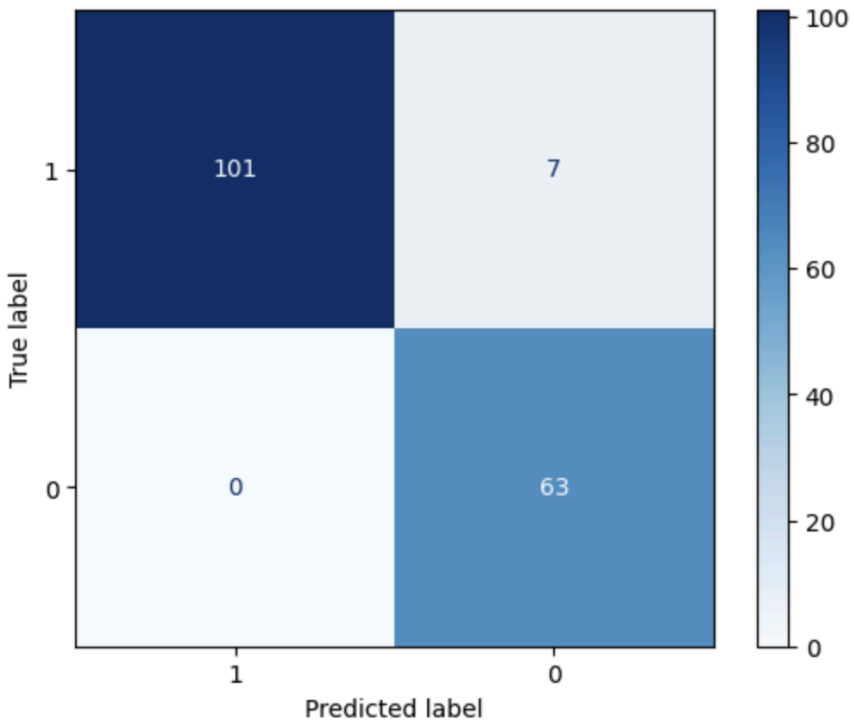
Precision	Recall	F1 Score
0.861538	0.888889	0.875000

When PCA was applied to the Naive Bayes model, there was an overall drop in the performance. There was a slight decrease in F1 scores from 87.8% to 87.5% and an even bigger decrease in the recall value from 92.0% to 88.9%. This overall decline in performance could be attributed to the fact that PCA transforms the original features into correlated components, which violates the Naive Bayes' assumption of feature independence. Additionally, PCA might have removed data that Naive Bayes depended on in the non-PCA model.

Logistic Regression

No PCA

Results

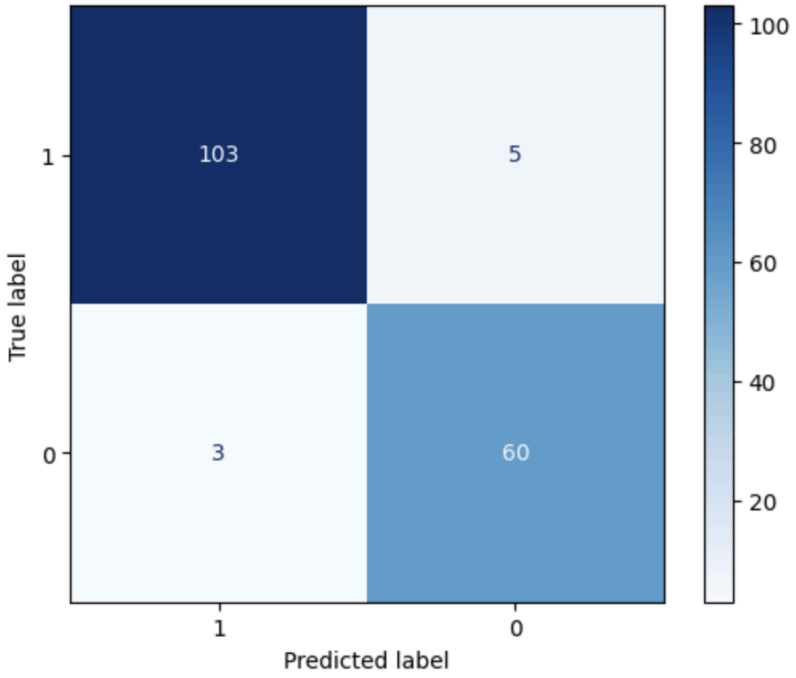


Precision	Recall	F1 Score
0.950000	0.967593	0.956938

The Logistic Regression model produced a strong performance overall. With a precision of 95.0%, a recall of 96.8%, and an F1 score of 95.7%, the Logistic Regression is among one of the top-performing models, only slightly behind Random Forests and Neural Networks. Its strong performance demonstrates that even without the complexity of deep learning methods, it can accurately classify tumors when provided with well-structured features.

PCA

Results



Precision	Recall	F1 Score
0.947388	0.953042	0.950058

When PCA was applied to the Logistic regression, we see only the slightest drop in the F1 Score. However, among the models with PCA applied, it is still ranked in the top three of models. With the small decline, this could suggest the PCA removed some useful features that were being utilized in the non-PCA version of the model.

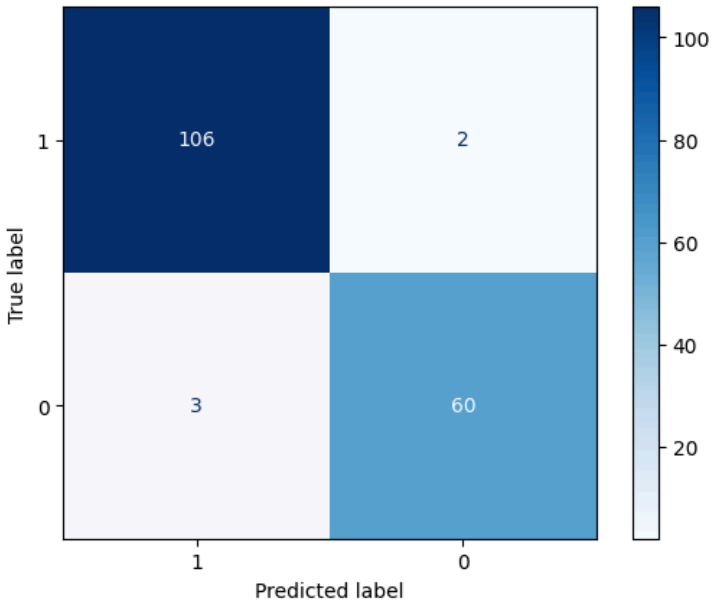
Neural Network

The table below describes the parameters used to train, test, and validate our model.

Parameter	Value	Notes
Framework	Keras (TensorFlow backend)	High-level neural network API
Input Features	30 (No PCA) / ~10 (PCA)	PCA reduced dimensionality based on variance threshold (~95%)
Hidden Layers	1	Simple feedforward network
Neurons in Hidden Layer	16	Tuned for moderate complexity
Activation (Hidden)	ReLU	Enables non-linear decision boundaries
Output Layer	1 neuron	Binary classification (Malignant/Benign)
Activation (Output)	Sigmoid	Outputs probability between 0 and 1
Loss Function	Binary Crossentropy	Suitable for binary classification
Optimizer	Adam	Adaptive learning rate, commonly used
Epochs	50	Number of full training passes
Batch Size	32	Number of samples per gradient update
Validation Split	0.2 (20%)	Portion of training data used for validation
Early Stopping	Not explicitly used (recommended)	Helps prevent overfitting
Preprocessing	StandardScaler (both), PCA (only in PCA version)	Standardizes features; PCA reduces input dimensions

No PCA

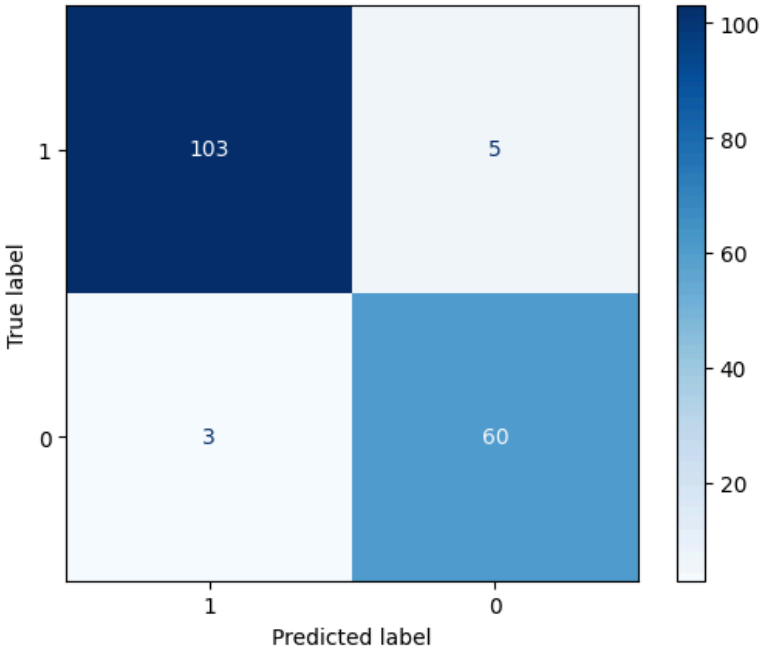
Results



Precision	Recall	F1 Score
0.967742	0.952381	0.960000

PCA

Results



Precision	Recall	F1 Score
0.907692	0.936508	0.921875

The neural network performed strongly in both the non-PCA and PCA versions, with the non-PCA model achieving slightly higher precision and F1 score, while the PCA model showed a small gain in recall. This suggests that while dimensionality reduction helped simplify the model and may reduce overfitting, it slightly compromised precision, likely due to the loss of some feature-specific detail. Overall, both versions performed well, and the choice between them depends on the balance between interpretability, computational efficiency, and the clinical need to minimize false positives or false negatives.

Cost-Benefit Analysis

The cost-benefit matrix sets benign to cost 1 if wrongly classified and malignant to cost 5 if wrongly classified. This matrix focuses on costs with no benefits, as it is the medical professional's duty to give the correct classification, so no benefit and costs are incurred if they violate this duty. True negatives, where a benign tumor is correctly classified, yield no downstream costs, while a false positive, where a benign tumor is classified incorrectly as malignant, will cause patient stress, distrust, and the need for further testing, adding various costs, so we settle for cost 1. This is even worse and more serious for a false positive where a malignant tumor is missed and deemed benign when it is not, as patients in need will be delayed in treatment, leading to more concerning consequences, so it must have the highest cost at 5. Lastly, a true positive does its duty and correctly catches a malignant tumor, allowing early treatment with no cost penalty.

For non-PCA KNN, Naive Bayes, and Logistic regression, each test threshold for the cost-benefit analysis yielded the same net cost of 108, so it was left out as it did not add to our analysis. For our Random Forest under the cost-benefit matrix, the most optimal threshold of 0.35 yielded the lowest penalty with a cost of 7, and Logistic Regression with PCA matched it at minimum cost 7 with threshold 0.36. So, to minimize costs, these selected models should be more sensitive with a lower cutoff to reduce the higher cost of misclassifying a malignant tumor. This tradeoff successfully minimizes overall cost. A close second was a Neural network with PCA that had a threshold of 0.4 and a minimum cost of 8.

Model	Full decision tree	Pruned Decision Tree	Random Forest	Neural Network
Threshold	0.01	0.05	0.35	0.01
Min cost	18	19	7	12

Models with PCA	Full decision tree	Pruned Decision Tree	Random Forest	KNN	Naive Bayes	Logistic Regression	Neural Network
Threshold	0.01	0.43	0.23	0.24	0.12	0.36	0.4
Min cost	44	37	18	12	19	7	8

Challenges

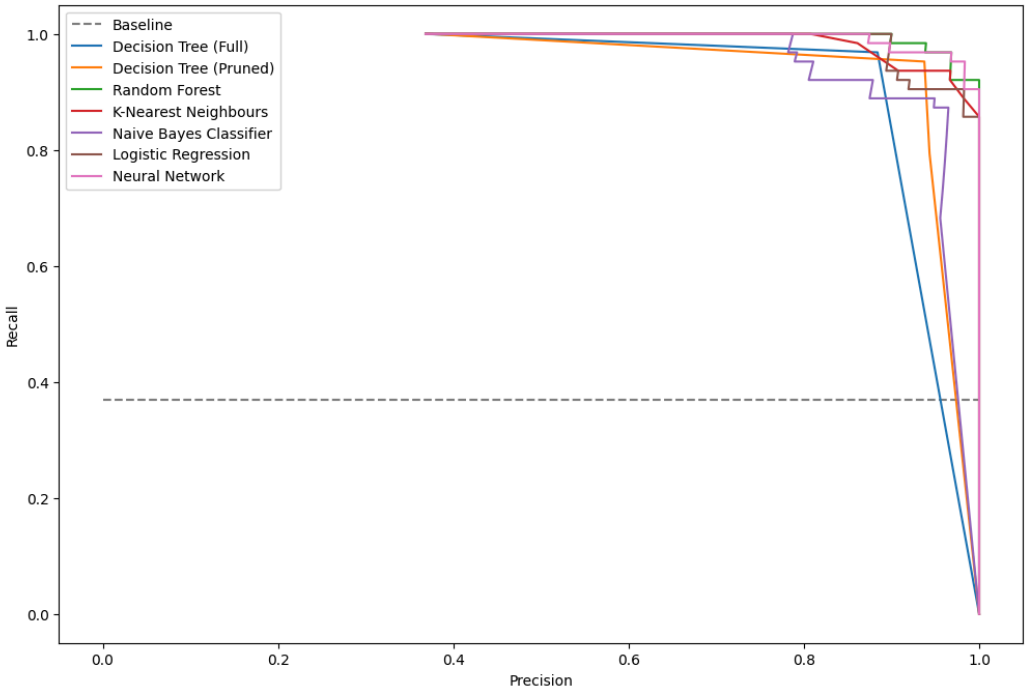
Since we initially weren't sure which metric we were going to use for our analysis, we started with accuracy. However, when we went to change over to the F1 score, due to the imbalance in the dataset, and false negatives and false positives being costly because of the clinical setting, we overlooked important places where accuracy needed to be changed out, like choosing k for k-nearest neighbors.

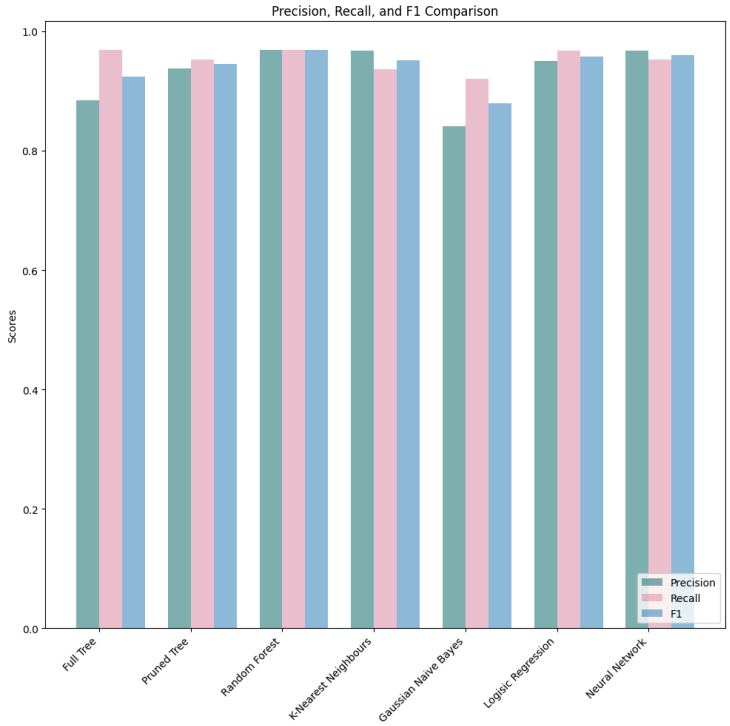
Due to an oversight, we initially trained the entire dataset using PCA rather than just x, train, and test, so our values were overfitting as we were leaking information from the test set into the training set. This made it so our models appeared to be doing better than they actually were. We adjusted this once we realized, and PCA is only applied to the x sets now.

Finally, choosing between PCA and non-PCA was difficult as we needed to address several possible issues. While PCA helped reduce dimensionality and potentially improved performance for some models by eliminating noise and multicollinearity, it also introduced interpretability issues. In a clinical context, models that operate on transformed features may be harder for healthcare professionals to understand and trust, since PCA components are abstract combinations of original medical features. This creates a barrier for model adoption in real-world settings, where transparency and explainability are crucial, especially when patient outcomes are at stake.

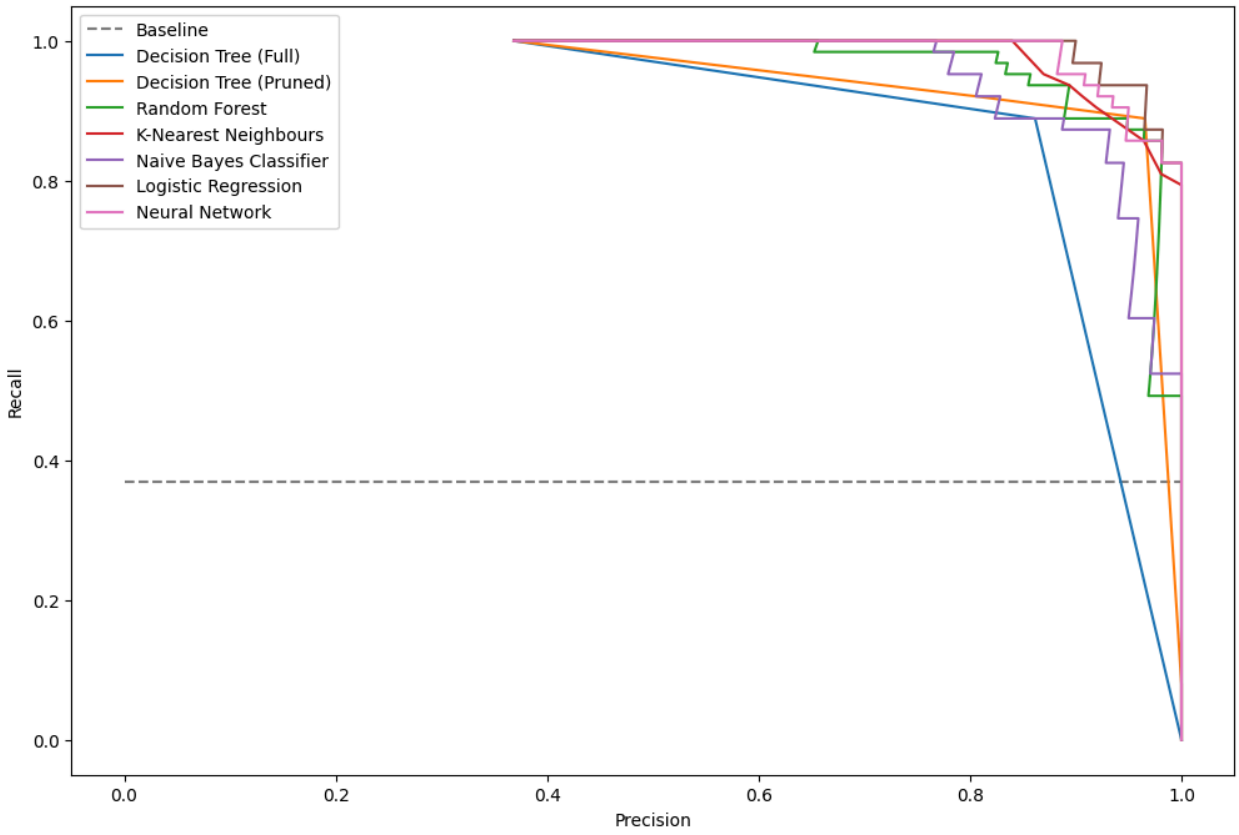
Solution

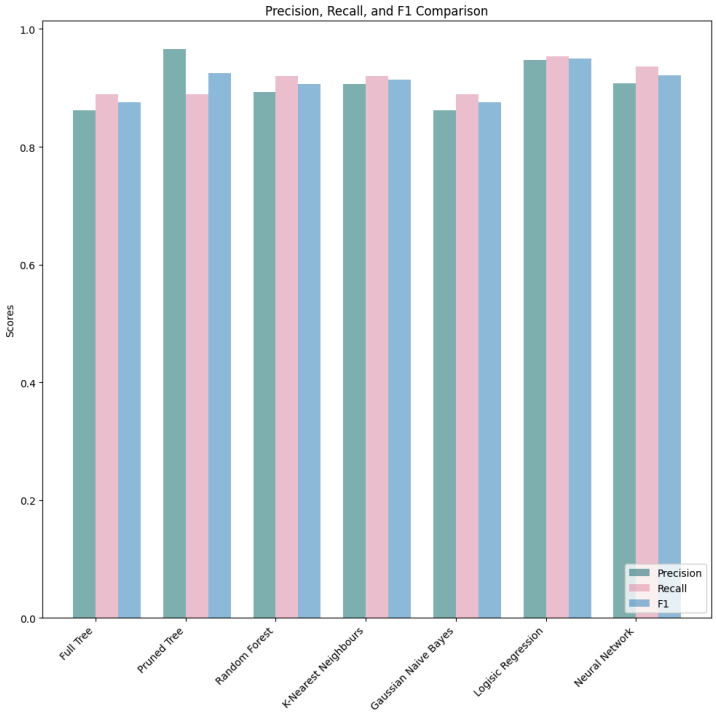
Results of all models: No PCA





Results of all models: PCA


Baseline
Decision Tree (Full)
Decision Tree (Pruned)
Random Forest
K-Nearest Neighbours
Naive Bayes Classifier
Logistic Regression
Neural Network
Recall
Precision
0.0
0.2
0.4
0.6
0.8
1.0


Precision, Recall, and F1 Comparison
Scores
Precision
Recall
F1
Full Tree
Pruned Tree
Random Forest
K-Nearest Neighbours
Gaussian Naive Bayes
Logisic Regression
Neural Network

After looking at the top-performing models, we decided to choose the Non-PCA Random Forest as the ideal model for classifying Benign or Malignant Tumors for Breast Cancer. This model produced incredibly strong results, with the F1 score being 96.8%. Although Neural Networks also produced strong results, it is important to understand the setting to which the model will be applied. In this case, the models will be applied in a clinical setting where many clinicians might not have a strong computer science background. In a study that analyzed how clinicians felt about the implementation of predictive models, many voiced concerns on how it is often hard to interpret and understand how these “black box models” work (Parikh et. al, 2022). Although Random Forests are more complex than a single decision tree, individual trees can still be examined, and feature importance can be assessed, making the model more transparent than a neural network. Given its strong performance and relative interpretability, the Non-PCA Random Forest is the most appropriate choice for this clinical application as it addresses clinicians' concerns and performs well.

Bibliography

Parikh, R. B., Manz, C. R., Nelson, M. N., Evans, C. N., Regli, S. H., O'Connor, N., Schuchter, L. M., Shulman, L. N., Patel, M. S., Paladino, J., & Shea, J. A. (2022). Clinician perspectives on machine learning prognostic algorithms in the routine care of patients with cancer: a qualitative study. Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer, 30(5), 4363–4372. https://doi.org/10.1007/s00520-021-06774-w

World Health Organization. (2025, February 3). *Cancer*. https://www.who.int/news-room/fact-sheets/detail/cancer