Megan Lawson
Miya Livingston
Sanika Rewatkar
DS 3001 Final Report

Introduction

Our project focused on predicting whether a breast tumor was benign or malignant using the Breast Cancer Wisconsin (Diagnostic) Dataset. This is a public dataset on the UC Irvine Machine Learning Repository, originating from a 1993 study on image processing techniques. It consists of 569 entries representing malignant or benign breast masses, with measured characteristics computed from digitized images of fine needle aspirates. The dataset contains 32 columns: 3 statistical features (mean, standard error, worst) for 10 different characteristics of breast masses (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension), resulting in 30 features in total, along with 2 non-feature columns, “id” and “Unnamed: 32” (Wolberg et al., 1993).

The aim of our project was binary classification of tumors as malignant (M, encoded as 1) or benign (B, encoded as 0); as such, our target feature was diagnosis. The metrics we used to evaluate our model’s performance were recall, precision, and F1 score, though we did note accuracy. We chose these measurements because false negatives (missing a positive instance of a malignant tumor) would be incredibly costly to a patient’s health, so recall was the most important metric we measured and the one we aimed to optimize. There is no consensus among the medical community on whether false negatives or false positives are a more serious problem with cancer diagnosis. However, false positives are more a common issue (Taksler et al., 2019), so precision was an important insight. By extension, the F1 score– though typically used for imbalanced data– served in our case as a valuable summary metric for recall and precision, as we aimed to minimize the rate of both false positive and false negative classifications.

Exploratory Data Analysis

When doing exploratory data analysis, we focused on the three different types of measurement we have in our data and the distributions of the data to look for any outliers. First, we made three confusion heatmaps with seaborn to look for any data that is very closely or very loosely correlated to the diagnosis columns. We saw a pattern that the mean and worst measurement of a characteristic had a stronger correlation to the diagnosis as opposed to the standard error for the same characteristic. For ease of visualization, we made a smaller dimensioned version of the three heatmaps to show the pattern (Figure 1). This led us to dropping the standard error columns from the initial runs of our models. Later, these columns would be added back due to PCA revealing that they were more relevant than initially thought.

Secondly, we looked at the distribution of each type of characteristic through box plots. There were the most outliers in the worst characteristics, but upon further research into the differences between malignant and benign tumors, we found that these outlier values for characteristics like texture are typically used as strong indicators for one diagnosis or the other (City of Hope, 2024). Thus, we ended up leaving all of the data points as they are. Looking at the distributions did allow us to see notable differences in the characteristics depending on what the actual diagnosis was (Figure 2). Larger perimeters and areas along with worse texture and more concave points were seen in more malignant tumors compared to benign.

Data Pipeline

In our preprocessing, we first dropped the "id" and "Unnamed: 32" columns. Since the original dataset is clean and has no null or missing values, no imputing was needed. In our earlier plan, we intended to drop all standard error columns, as we mistakenly assumed that they did not contain information that was helpful for classification. However, as PCA later revealed that they were influential to principal components, we ultimately kept these columns in the dataset. Next, we standardized all columns except "diagnosis", on which we then used a Label Encoder to convert M/Malignant to 1 and B/Benign to 0. We then applied an 80-20 train-test split with a random state of 42. As the data was imbalanced (of the 569 entries, 357 are benign and 212 are malignant), we oversampled the minority diagnosis (malignant) using SMOTE. It is important to note that for this project, we used SMOTE on both training and testing data (i.e. we applied SMOTE before creating the train-test split). While we did later attempt to run all our models on synthetic training data with real testing data (i.e. applying SMOTE after creating the train-test split, on only the training data), our final results differed significantly from our original results. As such, the results in this report are from models running on synthetic training and testing data, though in future analyses, we would run our models on synthetic training data and only real testing data.

Model Overview

K-Nearest Neighbors (KNN)

Our first supervised learning model was a K-nearest neighbors (KNN) model, with optimized parameters– n_neighbors (k), weights (weight function), algorithm (for computing nearest neighbors), metric (distance metric), and p (power if using the Minkowski metric)– based on iterative analysis and GridSearchCV. The default parameters for KNN are n_neighbors = 5, weights = "uniform", algorithm = "auto", metric = "minkowski", and p = 2. For the parameters' tuning ranges, we provided the range [3,10] for n_neighbors (avoiding lower and higher k values to prevent overfitting and underfitting), "uniform" and "distance" for weights, "ball_tree", "kd_tree", and "brute" for algorithm, "euclidean", "manhattan", "chebyshev", and "minkowski" for metric, and the range [1,11] for p. While GridSearchCV found that the best parameters when optimized for recall were n_neighbors = 6, algorithm = "ball_tree", metric = "euclidean", and weights = "distance," the optimal parameters found in the iterative analysis differed, with n_neighbors = 7 instead, and all algorithm options leading to the same results. Further comparing the two sets of parameters, while the two k values led to the same recall, k = 7 was able to achieve a higher precision, F1 score, and accuracy. Thus, we proceeded with k = 7 and the default algorithm ("auto") for our analysis.

With the optimized parameters, our model achieved a recall of 0.9865, precision of 0.9733, F1 score of 0.9799, and accuracy of 0.9790 (these metrics were selected for the reasons explained in the introduction). For comparison, we also ran an untuned KNN model with the default parameters, which had a recall of 0.9730, precision of 0.9474, F1 score of 0.9600, and accuracy of 0.9580. This indicates that our tuning was able to improve recall by 1.39%, precision by 2.74%, F1 score by 2.07%, and accuracy by 2.19%. The confusion matrix (Figure 3) shows that the tuned model had 2 false negatives and 1 false positive, whereas the default model had 4 false negatives and 2 false positives (Figure 4).

Random Forest

Our second supervised learning model was a Random Forest model with parameters tuned and optimized using GridSearchCV. As a standard for comparison, we ran a baseline Decision Tree model with a random state of 42 and all other parameters left as default, which returned a recall of 0.9595, precision of 0.9726, an F1 score of 0.9660, and an accuracy of 0.9650. Additionally, we ran a baseline Random Forest model with a random state of 42 and all other parameters left as default (n_estimators = 100, criterion = “gini”, max_depth = None, min_samples_split = 2, min_samples_leaf = 1). This model returned a recall of 0.9730, a precision of 0.9730, an F1 score of 0.9730, and an accuracy of 0.9720. The parameters tuned were n_estimators, criterion, max_depth, min_samples_split, and min_samples_leaf. The tuning ranges for these parameters were 25 to 400 for n_estimators, “gini”, “entropy”, and “log_loss” for criterion, “None” to 40 for max_depth, 2 to 8 for min_samples_split, and 1 to 8 for min_samples_leaf. Five different grids were chosen for grid search, and all models’ recall scores were compared to the baseline from the Decision Tree and the untuned Random Forest model.

The fit that resulted in the highest recall was from the first grid, where n_estimators ranged from 50 to 200, possible criterions were “gini”, “entropy”, and “log_loss”, max_depth ranged from “None” to 20, min_samples_split ranged from 2 to 8, and min_samples_leaf ranged from 1 to 4. The optimal fit had the following parameters: n_estimators = 50, criterion = “entropy”, max_depth = None (default), min_samples_split = 2 (default), and min_samples_leaf = 1 (default). This model returned a recall of 0.9865, a precision of 0.9865, an F1 score of 0.9865, and an accuracy of 0.9860. Notably, as shown in the confusion matrix (Figure 5), the model only had one false positive and one false negative. Furthermore, as shown in the feature importance graph (Figure 6), the most important features were mean radius (11% variation), texture (11% variation), and perimeter (10% variation); this is likely because malignant tumors are larger, rougher, and more irregularly shaped than benign tumors (Rangayyan and Nguyen, 2006). The extremely high scores for this fit could potentially suggest that this is an overfit, and in the future, testing on a larger dataset could verify whether this is an overfit. Additionally, all fits were able to achieve a greater recall than the baseline Decision Tree, though all fits other than the best fit returned recalls equal to that of the baseline random forest (Figure 7).

Principal Component Analysis (PCA)

Our unsupervised learning model was a principal component analysis (PCA) of our data. We used the sklearn PCA function. With 5 components, there were a couple of parameters that we kept as their default values. These included copy = True, tol = 0.0, interated_power = “auto”, n_oversamples = 10, and power_iteration_normalizer = “auto”, which didn’t make sense for us to change because we didn’t meet the requirements to change them (n_oversamples needing svd_solver to be “randomized”) or because they did not benefit the model when tuning. This left us with two primary hyperparameters that we tuned: whiten (False by default) and svd_solver (“auto” by default). When “True”, the whiten parameter uses a mathematical process of multiplying the square root of n_samples with the component vectors and dividing by singular values. This ensures that the outputs are uncorrelated with unit component-wise variances. This was beneficial for us to do since we had strong correlation between certain characteristics like area and radius. We also used “covariance_eigh” for our svd_solver parameter since we had many more samples than features. This solver is very efficient for data sets like this, and its primary drawback of being less numerically stable does not affect our data much since it is all standardized, so there is not a large range of values.

Our first PC explained about 55% of the variance in the data with the second at about 17% (Figure 8). We sorted the features that had the biggest combined impact on the first and second principal component for mean, standard error, and worst values. We saw that standard error, which still contributed less than the other two measurements, contributed significantly more than previously thought. This was true in especially the second PC through the standard error of compactness, concave points, and fractal dimension (Figure 9). This is what led us to including the standard error data for the other two models. The other results aligned with our random forest closely. Concave points, which made up the top two most important features, were in the top three most impactful features on our first and second PC for all three measurement types (Figure 9). The third most important feature was radius worst, which was in the top three most impactful worst features. Mean radius was also in that list for the mean measurements.

Discussion

Comparing our two supervised learning approaches, both the tuned KNN model and tuned Random Forest model achieved the same recall score of 0.9865 (Figure 10). However, while the KNN model had a precision of 0.9733 and F1 score of 0.9799, the Random Forest model had slightly higher metrics, with a precision and F1 score of 0.9865. In their classifications, the KNN model had 2 false negatives and 1 false positive, while the Random Forest model had 1 false negative and 1 false positive. These results indicate similarly strong performance from both models, with the Random Forest model outperforming the KNN model by a small amount, and both models achieving better recalls than many deep learning models.

Meanwhile, our unsupervised approach with PCA was able to effectively identify the 5 components contributing the most variance in the dataset, and further analysis showed that the most influential features in the first 2 primary components were concave points and radius_worst. These insights align with the feature importance results from the Random Forest model, which showed that size (radius, perimeter, area) and texture-related features (texture, smoothness, concavity) had the highest importances.

Ethics

When it comes to ethical concerns related to our models, there are two primary ideas we will consider: integration of the model and the larger concern around medical data. Technology has assisted in diagnoses through tests like X-rays, MRIs, CT scans, blood tests, and more. However, technology that gives recommendations as opposed to just images or data is much newer (Al-Antari, 2023). Machine learning allows programs to analyze large quantities of data at rates much faster than humans. This can give medical professionals more time to work on other tasks and possibly see patterns that are unintuitive for humans. Alternatively, not all algorithms are accurate.

While our accuracy is quite high, it is important to consider the way that mistakes are handled in healthcare when they are made by models as opposed to people. The phenomenon "automation bias" describes human tendency to trust decisions made by automated services with less critical thinking than decisions made by humans. This comes from the belief that machines are more objective than people. In a 2017 study on clinical decision support from an automated prescription assistant, use of the automated tools decreased prescribing errors by 58.8% when the support was correct but increased errors by 86.6% when it was incorrect (Lyell et. al, 2017). Despite being told to check drug information whenever the tool recommended a decision, the clinical students were much more likely to let incorrect information from the tool pass through compared to human-made mistakes.

Applying this to our models, while there are not many cases where the diagnosis of the breast mass will be incorrect, it is important that those inaccuracies are found and that the results are not trusted blindly. Healthcare professionals, who can be far removed from the way that algorithms work, should be educated on the risks of algorithmic decision making if they are using these models.

More generally, many ethical questions can arise from the collection and use health related data. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) protects the medical records of individuals, including any sort of "identifiable health information" (CDC, 2024). For electronic data, they put an emphasis on secure healthcare messaging and data storage. Copies of medical data being sent to different groups for studying increases the number of places where a data breach could happen. Since this data is so personal, privacy is a priority.

Aside from the inherent issues related to collection and storage of healthcare data, there is also a risk of misuse. Marshall Allen explains in ProPublica that healthcare companies are already scrounging up any data they can that indicates "lifestyle" in order to charge rates to customers based on what they feel is "appropriate risk" (Allen, 2018). While the Affordable Care Act prohibits insurers from denying coverage based on pre-existing health conditions, they can still price their plans however they would like. When it is more accessible to train models like ours that make predictive or diagnostic decisions, it becomes easier for healthcare companies to weaponize that power. In the case of our data, things like gender, race, age, and other characteristics are not present, but if they were, they give companies more ways to "predict" who is more likely to have certain health issues and charge them accordingly.

Associate Professor at Oxford Carisa Véliz argues that in the same way that people have the right to not incriminate themselves legally, they should also have the right to not "incriminate" themselves to their healthcare providers (Véliz, 2020). In essence, patients should never have to reveal any more than is medically necessary about their habits, even if they are doing things that are "bad" for them. This argument can be applied to medical data as well. Patients should have the right to not be assessed for their likelihood of certain disease or ailment or to have their data train or be put through a model for such predictions. They should have the right for that analysis to not affect their care. When collecting data for machine learning, these pillars of ethics should be considered.

Conclusion

From our supervised learning model results, it can be seen that Random Forest outperforms KNN for our dataset, having higher values for precision, F1 score, and accuracy (recall values were equal). Furthermore, while KNN models have some advantages in computational efficiency and making no relational assumptions regarding the data, KNN performance degrades with increased dimensionality, and models are sensitive to noise. Random Forest models, on the other hand, can be resource-intensive and likewise sensitive to noise, but they handle increased dimensionality well and reduce the risk of overfitting. Thus, given the superior metric results of our Random Forest model and the advantages of Random Forest when dealing with high dimensionality, we recommend the use of Random Forest with our relatively high dimensional (30 feature) dataset.

Given our binary target, it should be noted that the scope of our analysis is somewhat limited, and our classifications do not take into account or reveal more complex relationships or patterns that may exist in the data beyond just the diagnosis of tumors. With that said, our Random Forest feature importance and PCA results– which showed that concave points and radius measures (among other size and texture-related features) were among the most important features when categorizing our data– align with medical domain knowledge that benign and malignant tumors differ in size, irregularity, and texture. Thus, our predictive results open the door for future exploration of the relationship between the identified high importance features and tumor diagnosis, as well as highlight the potential for more sophisticated predictive models and other diagnostic tools to be created based on these relationships.

References

Al-Antari, M. A. (2023). Artificial Intelligence for Medical Diagnostics—Existing and Future AI Technology! Diagnostics, 13(4), 688. PubMed Central. https://doi.org/10.3390/diagnostics13040688

Allen, M. (2018, July 17). Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates. ProPublica. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates

CDC. (2024). Health insurance portability and accountability act of 1996 (HIPAA). Public Health Law; Centers for Disease Control and Prevention. https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html

City of Hope. (2024, March 18). Benign vs malignant tumors: What's the difference? City of Hope. https://www.cancercenter.com/community/blog/2023/01/whats-the-difference-benign-vs-malignant-tumors

Lyell, D., Magrabi, F., Raban, M.Z. et al. (2017). Automation bias in electronic prescribing. BMC Med Inform Decis Mak 17. https://doi.org/10.1186/s12911-017-0425-5

Taksler, G. B., Keating, N. L., & Rothberg, M. B. (2018). Implications of false-positive results for future cancer screenings. Cancer, 124(11), 2390–2398. https://doi.org/10.1002/cncr.31271

Véliz C. (2020). Not the doctor's business:Privacy, personal responsibility and data rights in medical settings. Bioethics.712–718. https://doi.org/10.1111/bioe.12711

Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

Appendix

Figure 1. Feature Heatmaps.

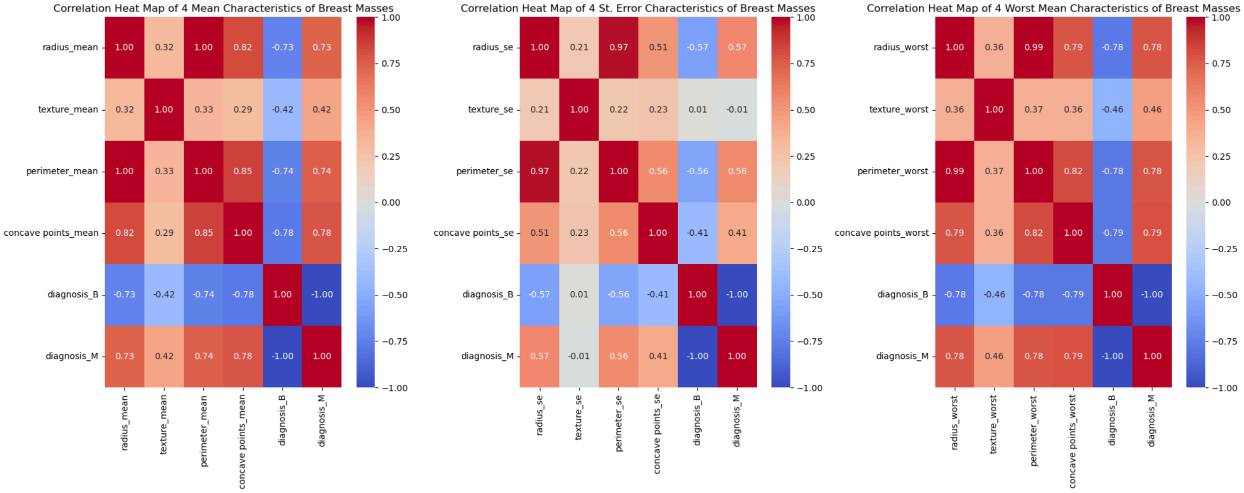

Figure 2. Boxplots for Worst Values by Diagnosis.

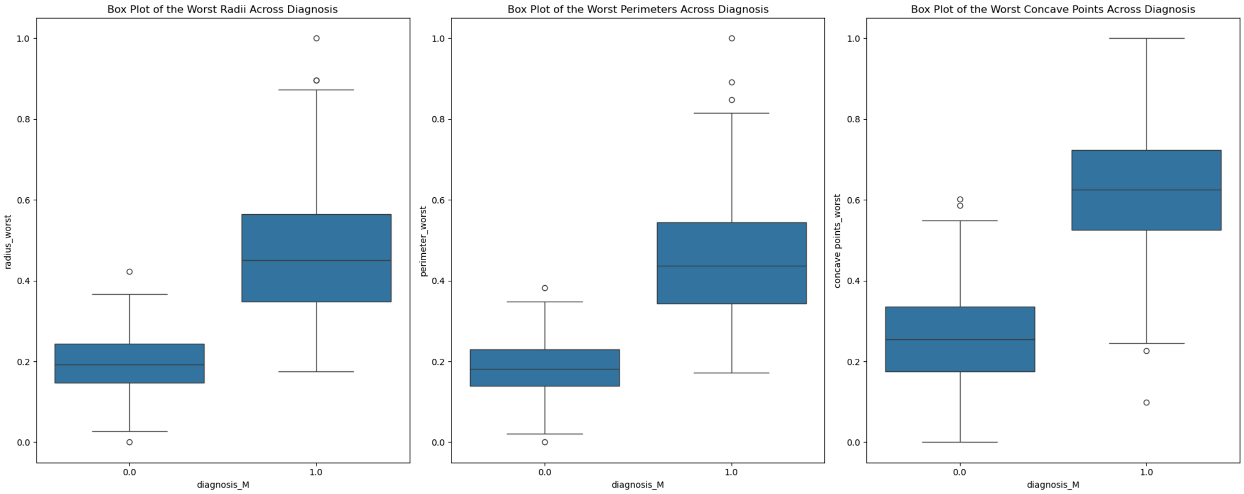

Figure 3. Confusion Matrix for Tuned KNN Model.

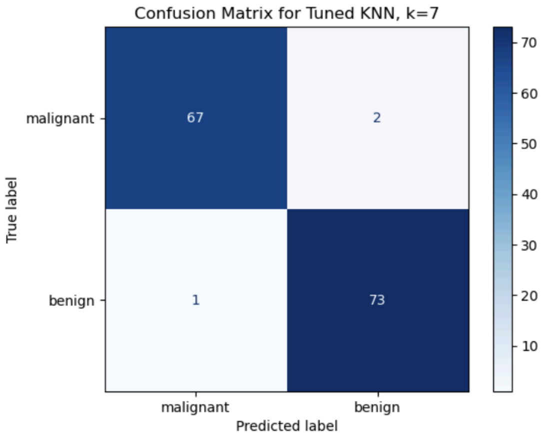

Figure 4. Confusion Matrix for Untuned (Default) KNN Model.

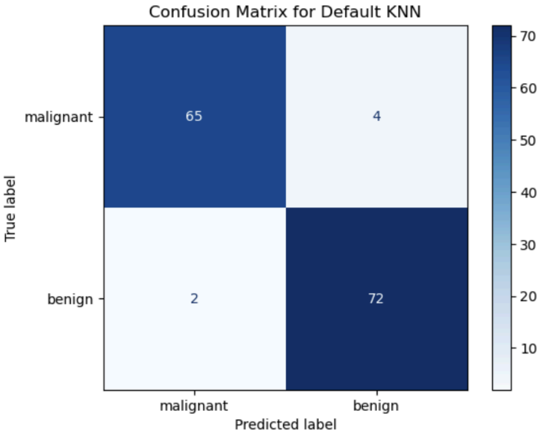

Figure 5. Confusion Matrix for Random Forest Model.

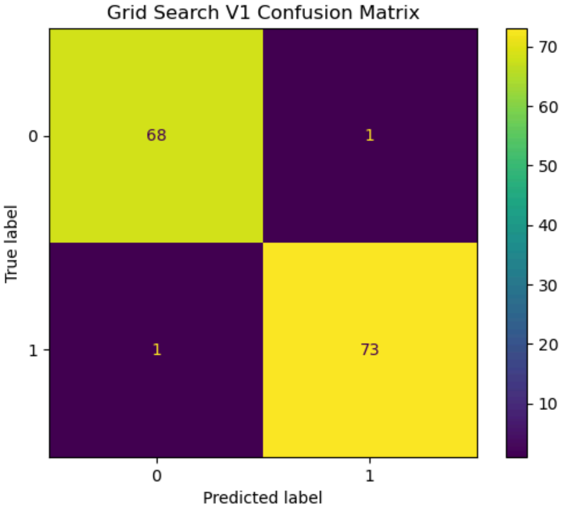

Figure 6. Feature Importance Graph for Random Forest.

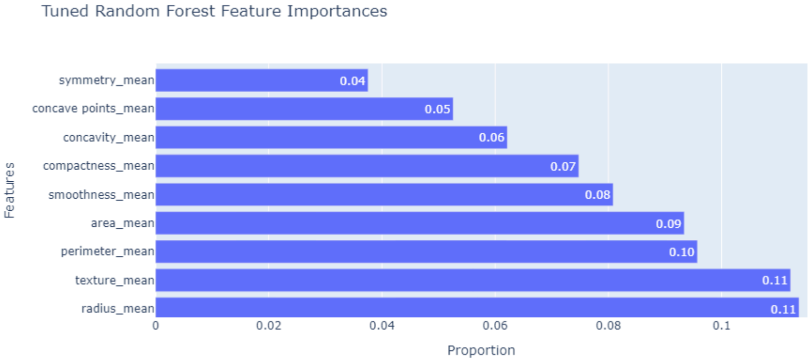

Figure 7. Recalls of Random Forest Fits.

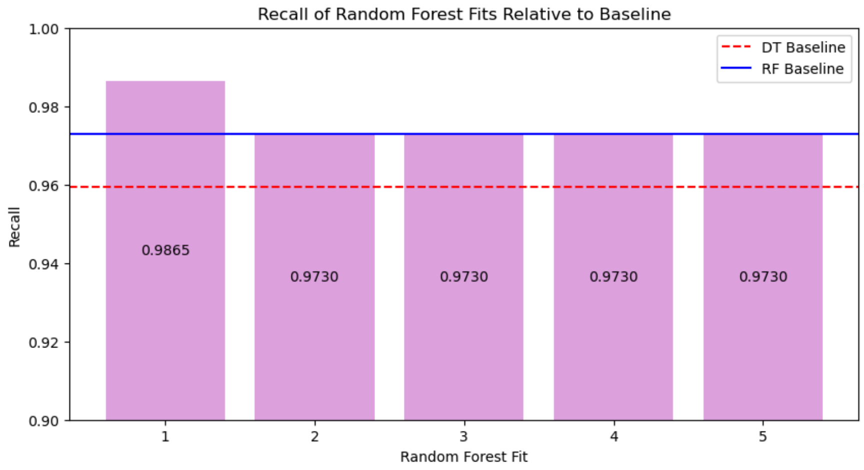

Figure 8. Proportion of Variance Graph.

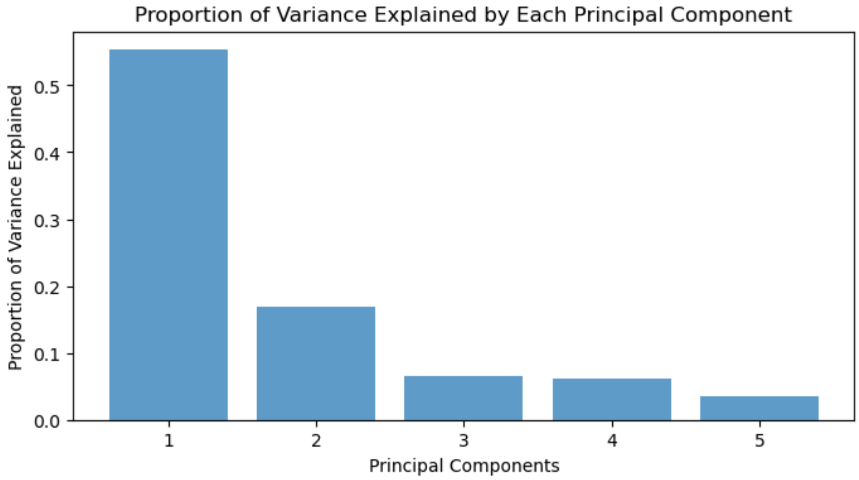

Figure 9. PCA Tables.

Standard Error

	Compactness SE	Concave Points SE	Fractal Dimension SE
PCA 1	0.119489	0.118890	0.039978
PCA 2	0.212479	0.100949	0.159095

Worst

	Radius Worst	Perimeter Worst	Concave Points Worst
PCA 1	0.268637	0.268545	0.365810
PCA 2	-0.241167	-0.206095	0.065860

Mean

	Radius Mean	Concave Points Mean	Fractal Dimension Mean
PCA 1	0.251258	0.327010	0.033251
PCA 2	-0.255467	0.005170	0.399082

Figure 10. Comparing KNN and Random Forest Results

	Tuned KNN	Tuned Random Forest
Recall	0.9865	0.9865
Precision	0.9733	**0.9865**
F1 Score	0.9799	**0.9865**
Accuracy	0.9790	**0.9860**
False Negatives	2	**1**
False Positives	1	1