

Comparative Analysis and Hypertuning of Machine Learning Models for Breast Tumor Classification Evaluation; Single Models vs. Stacked Ensembles

Roberto Ruiz Felix¹, Simon Tran², Haiyan Wang²

Mathematical Methods in Data Science²

1. College of Health Solutions, Arizona State University, Arizona, 85004, USA
2. School of Mathematical and Natural Sciences, Arizona State University, 85281, USA

Email: rruizfel@asu.edu

Index Terms—Hypertuning, Stacking Classifiers, Random Forest, Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Area under the Curve (AUC), Confusion Matrix, Accuracy

Abstract

Objectives:

1. To evaluate the predictive performance of machine learning models in classifying breast tumors as benign or malignant.
2. To identify the most important features contributing to tumor classification through model-specific and aggregated feature importance methods.
3. To optimize model performance using hyperparameter tuning and assess improvements for Random Forest, Support Vector Machine, K-Nearest Neighbors, and XGBoost models.
4. To compare baseline and optimized versions of each model to determine their relative effectiveness.
5. To implement stacking techniques with the best-performing models to enhance overall classification metrics, including AUC, accuracy, and confusion matrix results.

This approach aims to develop a robust, data-driven methodology for improving breast tumor classification using machine learning techniques.

Introduction: Machine learning (ML) has revolutionized medical diagnostics, particularly in breast cancer detection, where accurate classification of malignant and benign tumors is vital. This study systematically compares individual ML models—Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (kNN), and Extreme Gradient Boosting (XGB)—and explores the potential of ensemble methods to enhance classification performance.

Methodology: Using a dataset of 569 breast tumor samples (357 benign, 212 malignant), we developed and optimized ML models through hyperparameter tuning. Feature importance was evaluated across models, but no explicit feature selection was performed. Performance metrics, including accuracy, precision, recall, F1-score, and AUC, were used to compare models before and after optimization. Stacked ensembles of the best-performing models were constructed to assess improvements in classification accuracy and robustness.

Results: Individual models achieved strong performance, with accuracies around 96%. kNN had the highest AUC (0.99705), followed by RF (0.99607). Stacked ensembles significantly improved metrics, with the XGB & SVM pair achieving the highest accuracy (98%) and an AUC of 0.9938, and the XGB & kNN pair attaining the highest AUC (0.9980) with 97% accuracy. These results demonstrate the effectiveness of ensemble learning in improving tumor classification.

Conclusion: Ensemble learning enhances ML model performance for tumor classification, surpassing standalone models in accuracy and AUC. This study underscores the promise of stacked models in medical diagnostics and highlights future directions, including feature engineering and ensemble optimization, to further refine classification systems. More specifically, we created a hypertuned stack model that outperforms single models, hypertuned and not, as well as those in related literature; XGBoost & SVM pair with a logistic regression meta-model.

Introduction

Breast Tumors: Breast cancer is a condition characterized by the uncontrolled growth of abnormal breast cells, leading to the formation of tumors [1]. Both men and women have breasts, but women have more breast tissue than men which is why breast tumors are common in women. The breast is composed of 15 to 20 sections called lobes, which are organized in a pattern similar to the petals of a daisy. Each lobe contains smaller structures known as lobules, ending in tiny-milk producing bulbs. Thin tubes called ducts, connect the lobes, lobules, and bulbs, leading to the nipple [2]. The spaces between the lobules and ducts are filled with fat. Although there are no muscles within the breast itself, underlying muscles cover the ribs beneath each breast. The breast is also supplied with blood vessels and lymphatic vessels, which drain into small lymph nodes. These nodes are clustered in areas such as under the arms, above the collarbone, and within the chest, as well as throughout other regions of the body [2][3]. Understanding the structure of the breasts, it is important to understand how breast cancer begins. Healthy cells grow and divide in a controlled manner to replace old cells or heal damage. When cells become abnormal and grow uncontrollably, they may form a mass or better known as a tumor. Tumors can either be benign, the noncancerous type, or malignant, the cancerous type. Benign tumors typically grow slowly without invading surrounding tissues, whereas malignant tumors are capable of spreading and affecting other areas of the body [1][4].

Diagnosis Statistics: In 2022, there were 2,296,840 new cases of breast cancer among women, as it is not reported for men [5]. Among the 10 countries with the highest rates of breast cancer, the United States was the second highest in the world, accounting for 274,375 cases and 42,900 deaths [5]. In 2024, breast cancer remained the most commonly diagnosed cancer among women in the United States, accounting for approximately 30% of all new cancer cases in women [6]. This year, an estimated 310,720 women will be diagnosed with invasive breast cancer and 56,500 women will be diagnosed with a non-invasive form of the disease known as Ductal carcinoma in Situ (DCiS), with a projected 42,500 women to die from the disease [6][7]. This means, over the span of two years, there will be a 13.25% increase in breast cancer patients within the United States, with a 0.94% decrease in deaths. Taking a more micro approach, more specifically in the state of Wisconsin, 5380 women are estimated to be diagnosed with invasive breast cancer, 1,175 additional women are diagnosed with a DCiS as of 2022, and 720 women died; 6555 women [8]. Although DCiS is just one of the many types of the non-invasive forms of the disease, this form is compared with the national average. Assuming a similar pattern in Wisconsin and the national average, 7423 women are expected to have breast cancer in 2024 with 713 expected to die. This is assuming a trend similar to that of the national statistics without the consideration of geographical, demographic, and other factors; so there are limitations to these numbers. While the increase in diagnosed cases reflects advancements in awareness and screening, it also highlights the importance of continued research into the causes, risk factors, and potential therapies for the disease.

Machine Learning in Medicine: The integration of machine learning (ML) into medicine is revolutionizing the healthcare landscape, offering innovative solutions to improve diagnosis, treatment, and patient outcomes. As medical data become increasingly complex and abundant, traditional approaches to analysis and decision-making are often insufficient. Machine learning, with its ability to analyze large columns of data and identify patterns that are not readily apparent to the human eye, presents an opportunity to enhance clinical decision-making, predict diseases, and personalize treatment plans. In particular, ML algorithms are being harnessed to address some of the most challenging and time-sensitive aspects of medicine, including early diagnosis, image analysis [9], genomics [10], and drug discovery [11]. Among the diverse applications of ML, one of the most crucial is in diagnostic imaging and the prediction of disease outcomes. This has a profound impact on conditions like breast cancer, where accurate tumor classification is vital for improving patient prognosis [12]. This is where binary classification becomes especially relevant, exploring only two levels; in this case, benign or malignant. By leveraging tumor characteristics, such as size, shape, and other clinical features, machine learning models can be developed to predict the likelihood of malignancy. This study employs the models of Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost).

Decision Boundary Models: Among the ML community, SVM has gained popularity due to its ability to perform well in situations where the relationship between features and outcome is

non-linear, which is the case in our data [13]. The core principle behind SVM is that a decision boundary divides the data points of one class from another. In its simplest form, a hyperplane is a linear classifier, meaning it aims to create a straight line that separates the two levels correctly [14]. The ideal hyperplane is the one that maximizes the margin between the two classes, ensuring that the distance from the hyperplane to the nearest data point on either side is as large as possible. While SVM excels in scenarios where the relationship between features and outcomes is non-linear, as in the case of our dataset, other machine learning models provide different perspectives on the classification task. One such model is the kNN algorithm, which focuses on finding an optimal decision boundary, as opposed to a single line. The fundamental idea behind this is that a new data point is classified based on the majority class of its k-nearest neighbors in the training dataset [15]. This makes kNN a non-parametric method, meaning it does not assume any underlying distribution for the data, which is particularly useful when dealing with complex or unknown distributions [16]. In binary classification, kNN is especially effective because it is straightforward to interpret. By simply examining the class distribution of the k-nearest neighbors, one can easily understand the decision-making process of the algorithm.

Tree-based Ensemble Models: As an ensemble method, Random Forests build multiple decision trees, each trained on a random subset of the data through a technique called bootstrapping [17]. By averaging the results of these trees, RF significantly reduces overfitting, a common problem in decision trees, which ensures better generalization to unseen data or the testing data in this case. One key advantage of RF in binary classification is their ability to manage large and complex datasets with both quantitative and qualitative variables [18]. Since they operate by splitting data into smaller subsets based on the most important features, they mitigate the risk of overfitting, which is critical when working with small, noisy datasets often seen in medical data. In the case of our imbalanced data, RF ensures a balanced evaluation of the tumor classification task, where both false negatives and false positives are minimized. A similar tree-based model is Extreme Gradient Boosting (XGB), which offers additional advantages, particularly in terms of performance and handling large datasets; a limitation with our sample size. Since our dataset has many features, XGB excels in learning complex patterns within the data, which is essential for our dataset since subtle variations in cell structure and tissue characteristics need to be identified [19]. Another significant benefit of using XGB is in handling class imbalance, since it mitigates this by using techniques such as weighted data points and boosting, which prioritize misclassified instances and enhance the model's sensitivity to the minority class [20]. This ensures that even the rarer malignant cases are correctly identified, which is crucial for timely and accurate diagnosis.

Objective: Tumor classification plays a critical role in cancer diagnosis and treatment planning, with the accurate differentiation between benign and malignant tumors being essential for patient care. Recent advancements in machine learning (ML) have shown promising potential for automating and improving classification tasks in medical diagnostics. This study evaluates four commonly used machine learning models—Random Forest, XGBoost, Support Vector Machine

(SVM), and K-Nearest Neighbors (kNN)—for their effectiveness in classifying tumors as benign or malignant. Furthermore, we investigate the use of stacking models, which combine multiple classifiers to leverage their strengths and improve predictive performance. Each model's performance is measured using several key metrics, including accuracy, precision, recall, F1-score, and AUC (Area Under the Curve). The results demonstrate that while individual models provide competitive performance, stacking classifiers offer a significant improvement in classification accuracy. This study highlights the potential of ensemble learning techniques, particularly stacking, in enhancing tumor classification systems and providing more reliable tools for medical diagnostics. The findings also suggest avenues for future research in further optimizing these models and applying them to other medical classification tasks.

Related Work

Feature Engineering: Feature engineering plays a vital role in machine learning models. Effective feature extraction can enhance the model's ability to distinguish between malignant and benign tumors. In this scenario, features such as texture, shape, and margin characteristics of tumors are often used to make predictions. However, many studies fail to fully explore the importance of selecting the informative features and handling high-dimensional data, often leading to overfitting or loss of predictive power. Narrowing down related work to the same or similar breast cancer dataset, several studies have applied different methods to identify the most significant features for breast cancer classification. Spillai used p-values derived from ANOVA testing to select features with statistically significant differences between benign and malignant tumors [23]. This approach focused on highlighting features with notable variations, such as “concave points”, “perimeter”, and “radius”, which are biologically significant in understanding cell characteristics. It was found that eliminating features with low predictive value, such as “standard error” features, significantly improved model accuracy. By reducing the number of features to the most relevant ones, accuracy improved. Similarly, Yura Ueno employed hypothesis testing to test the null hypothesis that malignant and benign tumors do not differ in their feature values. Using a Welch's t-test, the study found that malignant tumors exhibited significant differences in features like “mean radius”, “mean texture”, and “radius error”, confirming the importance of these features in differentiating between tumor types [24]. This demonstrates that features with higher correlations to the target variable are more likely to show significant differences in the t-test, supporting the notion that careful feature selection based on statistical significance can lead to better model performance. In contrast, Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al focused on eliminating highly correlated features through correlation-based methods and used Recursive Feature Elimination (RFE) to rank features based on their importance. The study found that reducing the number of features through these methods improved classification accuracy. The RFE process identified 11 key features that were deemed essential, including “area mean”, “concavity mean”, and “smoothness worst” [26]. These studies demonstrated that feature selection can boost performance, but they also have limitations. For instance, ANOVA-based methods may overlook interactions between

features, while hypothesis testing might not capture non-linear relationships, and correlation-based methods can be limited by the assumptions of linearity and may not consider feature interactions effectively.

Machine Learning: It is clear that the choice of algorithm can significantly affect the accuracy and reliability of predictions. It is clear in Table 1 that Support Vector Machines (SVM) have been a popular choice, as well as k-Nearest Neighbors (kNN) and Naive Bayes (NB), but these models may not perform as well as SVM on datasets with high-dimensional data. A significant example of this is the study done by Tarannum R, Wood J, and Ensari T, which compared the performance of SVM, kNN, and Naive Bayes. They found that SVM outperformed the other models with an accuracy of 96% [21]. However, this study did not optimize the other models or consider other evaluation metrics, which would provide a more holistic performance assessment. Similarly, Yura Ueno's study also focused on logistic regression and achieved a high classification accuracy, 93.36% [24]. Following this trend, other studies have found that decision boundary models have proved to be their best model, whether they were compared to others in their study, or backed by other literature [21][22][23][24][25][27]. However, this focus on simpler models like SVM and Logistic Regression, while useful for initial classification, may not capture the full potential of more complex machine learning techniques. On the other hand, more complex models like Random Forests (RF), Artificial Neural Networks (ANN), and Gradient Boosting (GBoost), were employed. In the work of Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al, multiple models were used, including, RF, ANN, GBoost, SVM, and MLP, with the MLP model achieving the highest accuracy of 99.12% [26]. Although this result is impressive, the study relied solely on accuracy as the evaluation metric which limits a comprehensive evaluation of the model's performance, particularly, in situations where false positives and false negatives are crucial. In real-world applications, especially in medical diagnostics, focusing solely on accuracy can be misleading. For example, a model may achieve high accuracy by simply predicting the majority class, but fail to accurately detect malignant cases, leading to dangerous misdiagnosis. Evaluating performance based on other measures is important in this context to ensure that the model is accurately identifying malignant tumors, as well as minimizing false positives and false negatives. Additionally, Recursive Feature Elimination was used to select the most important features, which enhances the performance of the machine learning models by eliminating irrelevant or redundant features. However, despite the improved accuracy, many studies do not employ cross-validation [21][22][23][25], which is essential to avoid overfitting and ensure the generalizability of the model. This is a common limitation across the studies specific to this dataset, which often rely on static evaluation without accounting for variability across different subsets of the data. Table 1 demonstrates the machine learning models discussed in each study as well as associated limitations.

Metrics: In medical applications, metrics play a crucial role in assessing the performance of machine learning models, misclassification can have severe consequences. Accuracy is

commonly used but often fails to provide a complete picture of model performance, especially when dealing with imbalanced datasets. For instance, a model that predicts only the majority class could still achieve high accuracy without being useful in practice. Spillai's study reported an accuracy of 96%, but did not provide other evaluation metrics such as precision, recall, or a confusion matrix which would have given a better understanding of how well the model distinguished between benign and malignant tumors [23]. The confusion matrix, which includes true positives, false positives, true negatives, and false negatives, can help assess how well a model performs in real-world scenarios, highlighting the trade-off between sensitivity (recall) and specificity (precision). Ultimately, machine learning models cannot be used in a clinical setting without proving their application in the real world, for this reason, confusion matrix is the ultimate metric for measuring a model's performance in classifying breast tumors. In other studies, such as the one by Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al, a more detailed evaluation was conducted with metrics like True Positives, False Positives, Recall, Precision, and F-Measure. The results for SVM were particularly strong, with recall values of 0.97 for benign and 0.96 for malignant tumors [26], indicating that the model was able to correctly identify most of the malignant tumors. The F-Measure, which combines precision and recall into a single metric, was 0.96 for benign and 0.95 for malignant tumors, demonstrating a good balance between precision and recall. However, the lack of cross-validation in these studies is a limitation, as it is essential for ensuring that the model generalizes well to unseen data. Cross-validation helps mitigate the risk of overfitting, which can occur in this case due to the limited sample size. Furthermore, studies that do not incorporate area under the curve (AUC) [21][22][23][24][25][26][27] or Receiver Operating Characteristic (ROC) [21][22][23][24][25][26] analysis may overlook important information about how well a model distinguishes between classes across various threshold settings. These additional metrics provide deeper insights into the model's true performance and can guide the choice of the most suitable model for deployment in real-world applications. Thus, the machine learning models used in the studies show promising results, there is still a need for comprehensive evaluation using a range of metrics. More detailed analyses, including cross-validation, precision, recall, and AUC, are necessary to better understand model behavior and ensure their reliability in clinical settings. Table 1 demonstrates the metrics discussed in each study as well as associated limitations.

Reference	Models	Best (Model) Metrics	Limitations/Notes
[21]	SVM, kNN, NB	Best Model: SVM Accuracy: 96%	Models were not optimized, and only used accuracy to evaluate model performance.
[22]	RF, Logistic Regression	Best Model: Logistic Regression Accuracy: 99.6%	Accuracy was the only metric used to evaluate model performance.
[23]	SVM	Best Model: SVM Accuracy: 62% (Pre feature elimination)	No decision-tree based model was used, and accuracy was the only metric to evaluate the model.

Reference	Models	Best (Model) Metrics	Limitations/Notes
		Accuracy: 95% (Post feature elimination)	
[24]	Logistic Regression	Best Model: Logistic Regression Accuracy: 93.96% FN Rate: 0.09333 Confusion Matrix: 3FP, 4FN	No decision-tree based model was used. Furthermore, this study uses a 5-fold approach.
[25]	Linear Classifier	Best Model: NA No metrics were shown	There is a clear lack of validations, only three features were selected and it is discussed that removing features discards informative ones.
[26]	RF, ANN, GBoost, SVM, MLP	Best Model: Multi-Layer Perceptron Accuracy: 99.12%	This study relied on accuracy as its only metric. However, the literature review shows evidence that 5-fold is best.
[27]	SVM, NB, kNN	Best Model: Support Vector Machine TP: 0.97(B), 0.96(M) FP: 0.03 (B), 0.02 (M) Recall: 0.97 (B), 0.96 (M) F-Measure: 0.96 (B), 0.95 (M) Precision: 0.98 (B), 0.95 (M)	Individual classifiers were not optimized and remained standard. Furthermore, results, not shown, were based on static evaluation metrics and did not incorporate cross-validation.

Table 1. Related Work Comparison/Limitations

Methodology

Feature Importance: Feature importance is determined using machine learning models that can assess the contribution of each feature to the overall prediction. This approach allows for a more comprehensive understanding of the dataset, as different models may highlight different aspects of the data. In particular, the use of several models allows for cross-validation of feature importance, ensuring that the most influential features are consistently recognized across different algorithms. For the tree based-models, RF and XGboost, feature importance is computed by evaluating how much each feature contributes to reducing the error in the model's predictions. These models perform feature selection naturally during the training process by considering how splits in the data improve the accuracy of predictions. The more a feature contributes to accurate predictions, the higher its importance score. These scores can be directly extracted from the model and are used to rank features based on their contribution. Below, in

figure 2a and 2b, we can see the most important features in order. Examining both graphs we can assess the top five important features along with their important score in Table 2. For linear

Figure 1a. RF Feature Importance

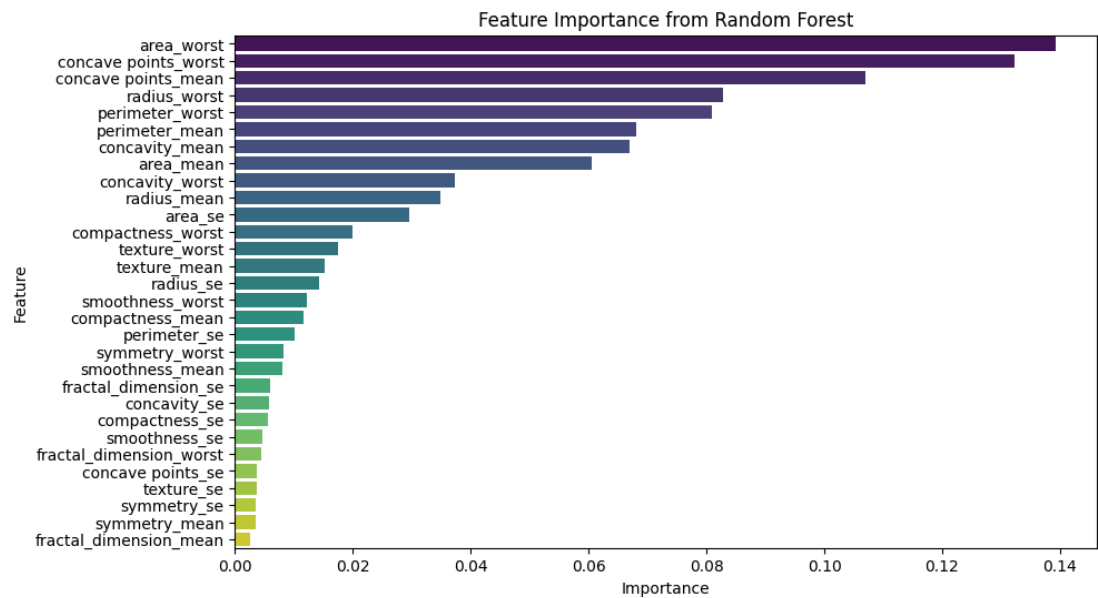
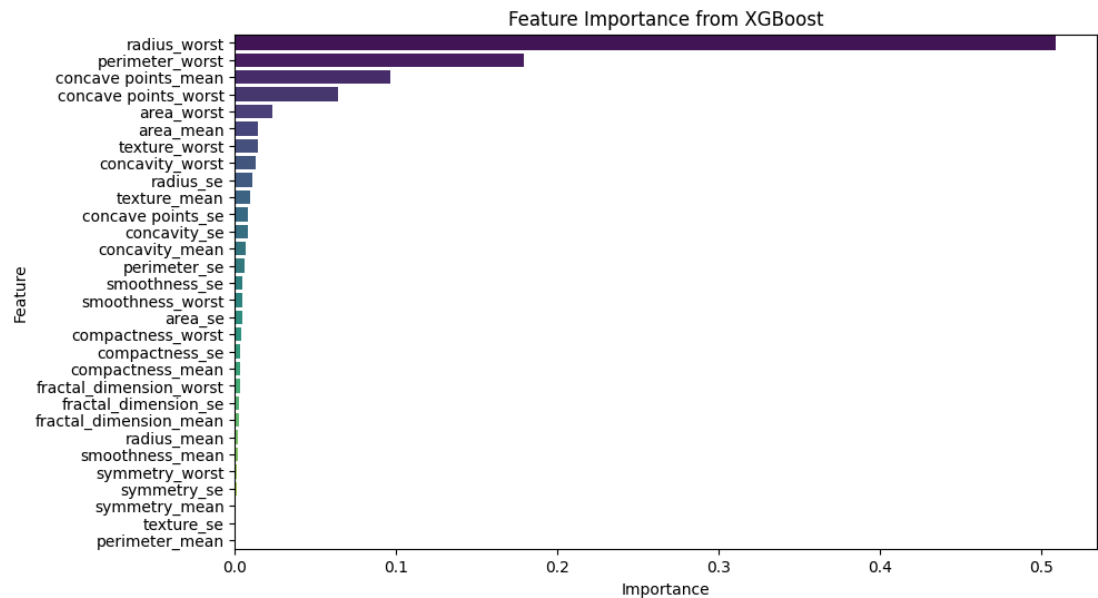


Figure 1b. XGBoost Feature Importance



models, Logistic Regression and Support Vector Machines (SVM), feature importance is determined by examining the model’s coefficients. In a linear model, each feature is assigned a weight/coefficient that reflects its contribution to the decision boundary. Larger absolute values of these coefficients indicate that the corresponding feature has a stronger influence on the model’s performance. To ensure that the features with larger scales do not dominate the model,

the weights are often normalized, and the absolute values are used to assess their relative importance. The feature importance graphs for the linear models are displayed in figures 2c and 2d, as seen below. Once the feature importances are computed for each model, these values are

Figure 1c. SVM Feature Importance

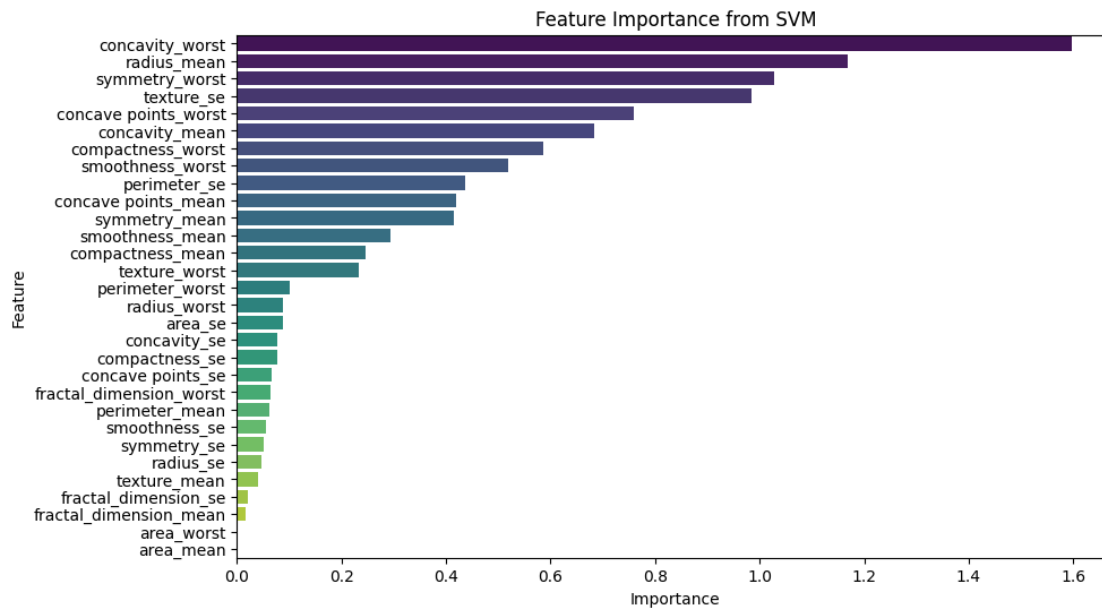
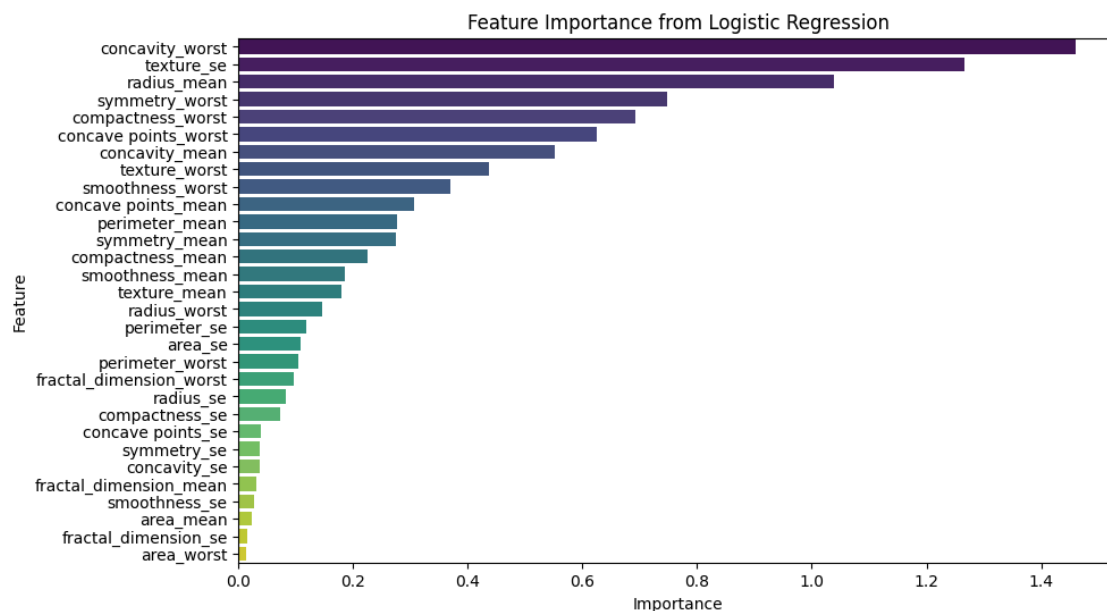


Figure 1d. Logistic Regression Feature Importance



aggregated to form a consolidated ranking. This aggregation allows for a more robust understanding of which features are consistently important across different modeling techniques. Features that are ranked highly across multiple models are considered to be more reliable

indicators of tumor classification, as they have proven to be influential in various contexts. Examining Table 2a, at a glance, it may seem that “area_worst”, “concave points_worst”, “concave points_mean”, “radius_worst”, “radius_mean”, and “concavity_worst” are the features that appear the most often. However, it is important to remember that two models are decision-boundary based whereas the others are decision-tree based. If we break this down, “concave points_worst” is the only feature that appears across all models implying that it has proven its importance in various contexts. Aggregating feature importances also helps to mitigate

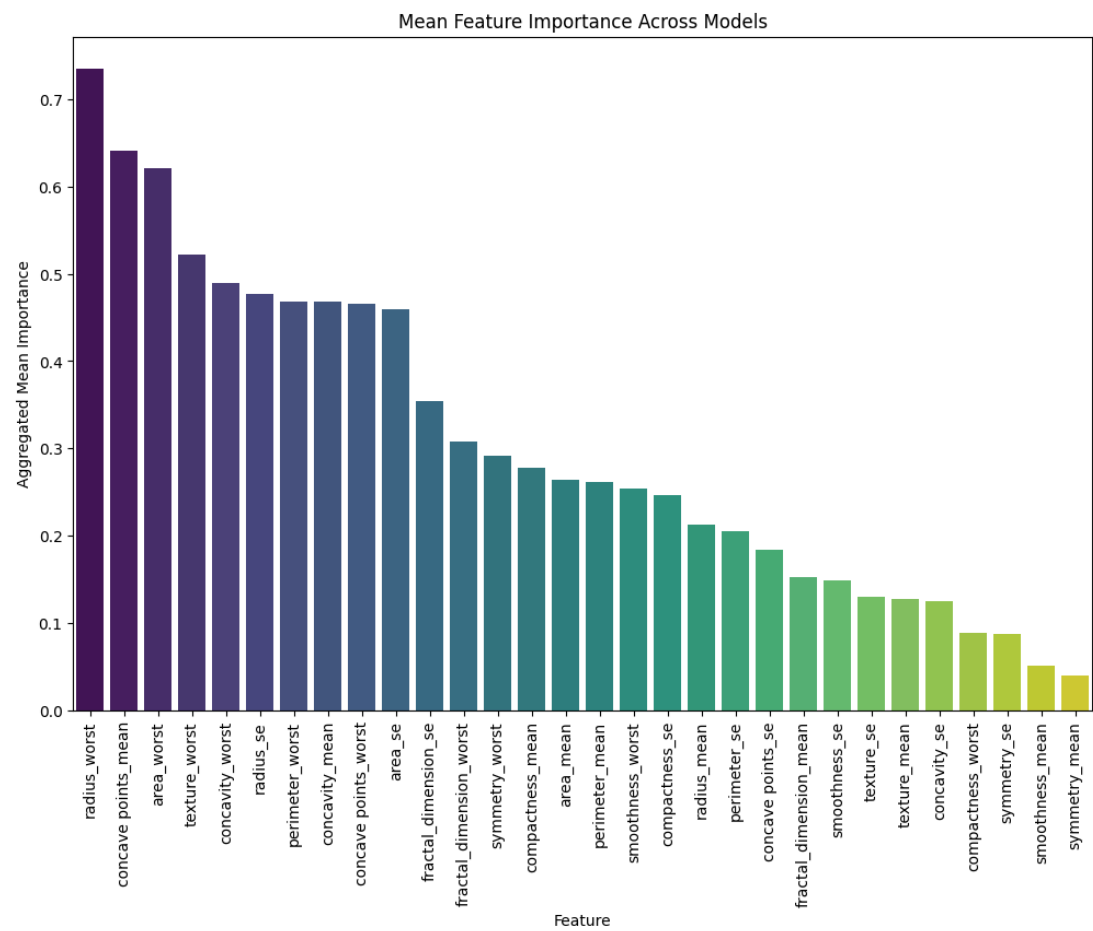
Figure 2a. Feature Importance per model

	RF	XGBoost	SVM	Logistic Regression
Rank	Importance Score			
1	area_worst: 0.139357	radius_worst: 0.509064	concavity_worst: 1.598302	concavity_worst: 1.461328
2	concave points_worst: 0.132225	perimeter_worst: 0.178712	radius_mean: 1.169144	texture_se: 1.26565
3	concave points_mean: 0.107046	concave points_mean: 0.096523	symmetry_worst: 1.027734	radius_mean: 1.038392
4	radius_worst: 0.082848	concave points_worst: 0.063827	texture_se: 0.985240	symmetry_worst: 0.746968
5	perimeter_worst: 0.080850	area_worst: 0.023247	concave points_worst: 0.758662	compactness_worst: 0.692349

the potential biases that may arise from any single model’s interpretation, ensuring that the final ranking reflects a holistic view of the data. After normalization, the feature importances can be compared on a consistent scale. Normalizing the scores ensures that features from different models, which may use different scoring systems, can be compared directly. The normalized importance scores are then averaged across all models to obtain a mean importance value for each feature. This aggregated mean importance provides an overall ranking of the features, identifying those that are most influential in the classification task. In the context of medical research, this is crucial for several reasons. First, this allows us to identify key features that help in distinguishing benign from malignant tumors, which can inform medical decisions such as diagnosis and patient monitoring, regardless of model employed. Secondly, identifying key features helps in refining machine learning models, ensuring that the most relevant information is included while eliminating irrelevant or redundancy features. Although not stated, an implication here is that this makes machine learning more feasible and cost-effective since feature extracting means a model requires less computational power, meaning it can be more accessible for implementation in a clinical setting. Lastly, feature importance enhances the interpretability and trustworthiness of machine learning models, which is particularly important in healthcare settings where clinical experts must understand and validate the model’s predictions. Visualizing

the aggregated, Figure 1, feature importance, we gain an intuitive understanding of which features play the most significant role in predicting tumor type, the top three being “radius_worst”, “concave points_mean”, and “area_worst”. These three features have the greatest impact on the model’s ability to differentiate between benign and malignant tumors, and therefore, they may be of particular interest for further medical investigation. If we examine

Figure 1. Aggregated Feature Importance



these further, the delta between the first and second ranked, 0.094269, and the delta between second and third ranked, 0.019459, we see nearly a 5 time difference between the two. This indicates that “radius_worst” carries much greater weight. The precise aggregated importance score of the top ten features are best represented in Table 2.

Table 2. Aggregated Feature Importance Score

Rank	Feature	Aggregated Importance Score
1	radius_worst	0.735269

2	concave points_mean	0.641
3	area_worst	0.621541
4	texture_worst	0.522288
5	concavity_worst	0.489791
6	radius_se	0.477435
7	perimeter_worst	0.468732
8	concavity_mean	0.468013
9	concave points_worst	0.466250
10	area_se	0.459462

Model Comparison: The approach to developing and evaluating machine learning models in this study is structured and comprehensive, aiming to identify the most effective model for classifying tumor data. The process integrates model evaluation, hyperparameter tuning, and metric-based comparison to ensure reliable and accurate predictions. The first step involves testing standard machine learning models, including RF, XGBoost, SVM, and kNN. These models are chosen for their diverse approaches to learning from data and their proven performance in classification tasks. To evaluate these models, we aimed at expanding the metrics used in previous studies and examining three metrics: confusion matrix, accuracy, and Area Under the Curve (AUC). The confusion matrix is central to model evaluation as it provides a breakdown of the actual model performance in a real-world application. This detailed insight allows for an understanding of the model's strengths and weaknesses. Furthermore, this allows for models to be further tuned for different scenarios. For instance, if false positives are more prone, this can be used as a screening tool for diagnosis which takes precautions one step further. Accuracy is the next metric used and it is straightforward why we want something to compare our models with that of previous research. Furthermore, accuracy is considered alongside other metrics to provide a holistic view of model performance. Lastly, AUC is used to measure the model's ability to distinguish between classes. It is particularly useful because it evaluates performances across all classification thresholds, giving a more comprehensive assessment compared to accuracy alone. We used these parameters to evaluate the performance of the baseline models. These models were initially tested in their default configurations, without any tuning, to establish a foundational understanding of their capabilities and to provide a benchmark for comparison. Examining our dataset, we have a moderately sized, high-dimensional classification problem. Thus, we consider these characteristics to inform the setup of our hyperparameter tuning for model tuning:

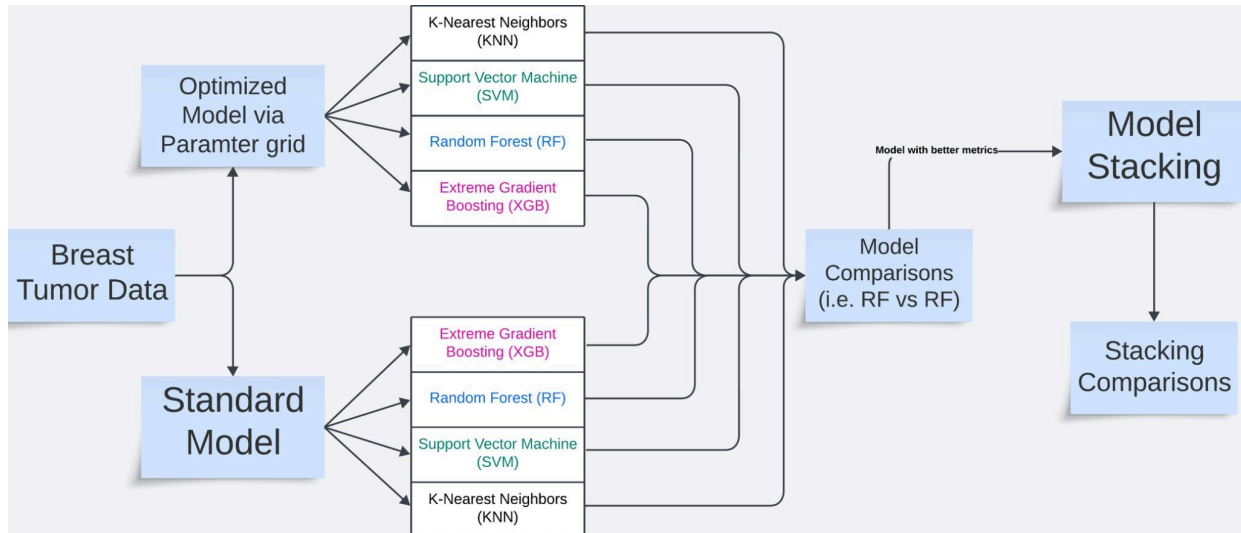
- **Random Forest:** number of trees, maximum depth, minimum sample requirements
- **XGBoost:** Learning rate, number of boosting rounds, maximum tree depth, and subsampling ratios.

- **Support Vector Machine:** The penalty parameter, kernel type, and kernel-specific parameters.
- **k-Nearest Neighbors:** Number of neighbors, weighting strategy, and distance metrics.

The parameter grids are carefully chosen to match the dataset's size and complexity, ensuring that the tuning process is computationally feasible while exploring meaningful variations in model configurations. Grid search with cross-validation is employed to tune each model. Cross-validation ensures that the evaluation is not biased by a particular train-test split and provides a robust estimate of model performance. The tuning process focuses on maximizing metrics to prioritize the ability to discriminate between classes. After tuning, the models are re-evaluated using the test set, and their performance metrics are compared. Here, the confusion matrix, accuracy, and AUC are key criteria for comparison.

Model Stacking: The final step involves selecting the best-performing model based on the evaluation metrics. The chosen model balances high accuracy, a well-calibrated confusion matrix, and a superior AUC score, ensuring it is both accurate and reliable. This methodology ensures that the selection process is rigorous, data-driven, and aligned with the dataset's characteristics and the classification task's critical requirements. The flow diagram of this process is shown below in figure 2.

Figure 2. Flow Diagram



Once the models are selected, combinations of two base models are tested for their energy when stacked. Each pair is integrated into a stacking ensemble, where the individual model predictions serve as inputs to a meta-model. For this research, a Logistic Regression model was used since it is best for binary classification when interpretability is needed. Furthermore, this works well since the base models already handle complex relationships, and since this is not a complex non-linear model, it solves the issue of overfitting on the training set of the base model.

Experiment setups and results

Setups: This experiment was divided into two phases: initial model evaluation and stacking-based evaluation. The following subsections detail the hardware, software environment, dataset preparation, and experimental design.

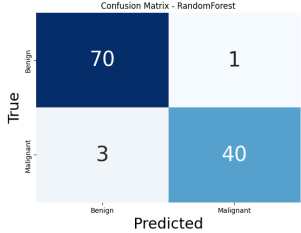
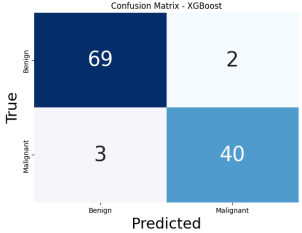
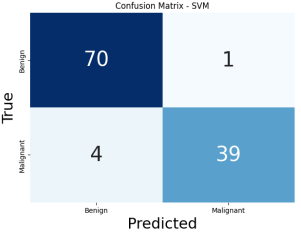
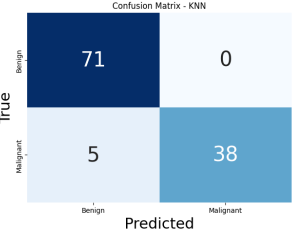
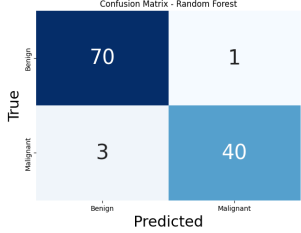
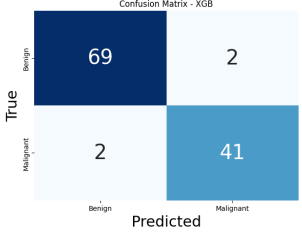
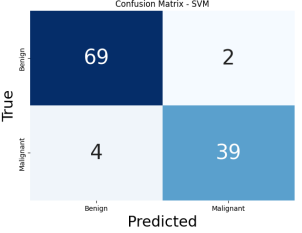
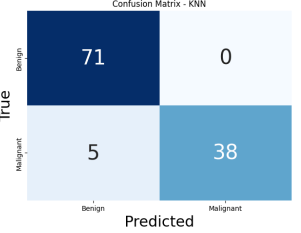
Hardware and Software Environment: All experiments were conducted on a MacBook Pro equipped with the Apple M1 Pro chip. The computational environment utilized the following specifications, Processor: Apple M1 Pro (8-core CPU, 10-core GPU), Ram: 16 GB. The software stack consisted of, Python Version: Python 3.12.1, Integrated Development Environment: Visual Studio Code.

Dataset Preparation: The dataset used in this experiment 569 breast tumor samples (357 benign, 212 malignant) with 33 columns. Continuous features were normalized to ensure uniformity across dimensions. Furthermore, the data was divided 80% training and 20% testing using a stratified sampling strategy to preserve class distributions. Regarding missing values, there were no missing values found in this dataset.

Single Model Results: Evaluating the results from the four standard machine learning models, all models had an accuracy of around 96%. The RF model had a precision of 0.96 for benign tumors and 0.98 for malignant ones. Similarly, it had a recall value of 0.99 for benign and 0.93 for malignant indicating the model's excellent ability to distinguish between classes. The XGBoost model had a precision value of 0.96 for benign and 0.95 for malignant. Recall for benign was 0.97, and 0.93 for malignant. This suggests that although lower than RF, it still offers robust performance regarding its classification ability. The SVM model had a precision of 0.95 for benign and 0.97 for malignant. Recall was 0.99 for benign and 0.91 for malignant, this indicates better performance than the XGBoosty model but just under RF. Lastly, kNN had a precision of 0.93 for benign and 1.00 for malignant. Recall was 1.00 for benign and 0.91 for malignant, indicating the highest performance amongst all models. The table for these values can be found in supplemental materials, as well as the associated f-score and supporting actual data values. However, the main three metrics used are found below in table 3. Evaluating the results from the four hypertuned machine learning models, it is clear that all models performed well, achieving high accuracy rates of around 96%, similar to that of the standard models. The RF model achieved an accuracy of 96.49%. It exhibited a precision of 0.96 for benign tumors and 0.98 for malignant tumors, reflecting its high accuracy in correctly identifying both classes. The recall for benign tumors was 0.99, which indicates that the model was very effective at identifying the true benign cases. However, the recall for malignant tumors was slightly lower at 0.93, suggesting that while the model performed well, it occasionally missed some malignant cases. The overall F1-scores were also strong, further emphasizing the model's balance between precision and recall. Similarly, the XGBoost model also achieved an accuracy of 96.49%, demonstrating robust classification performance. The precision for benign tumors was 0.96, and for malignant tumors, it was 0.95, slightly lower than that of Random Forest. The recall for benign tumors was 0.97,

indicating that the model did a very good job identifying benign cases. However, its recall for malignant tumors (0.93) was the same as Random Forest, showing that while it was effective at identifying malignant cases, it had room for improvement. The F1-scores were well-balanced, with a minor dip in precision for malignant tumors compared to the Random Forest model. The Support Vector Machine (SVM) model, with an accuracy of 94.74%, performed slightly worse than both Random Forest and XGBoost but still achieved commendable results. Its precision was 0.96 for benign tumors and 0.97 for malignant tumors, indicating that it was slightly less precise in predicting malignant tumors compared to the other models. The recall for benign tumors was 0.99, but for malignant tumors, it dropped to 0.91, the lowest among all the models tested. This suggests that while SVM was very good at identifying benign cases, it was less reliable in detecting malignant tumors compared to Random Forest and XGBoost. The F1-scores, though still strong, reflected this disparity, particularly for malignant tumors. Finally, the k-Nearest Neighbors (kNN) model achieved an accuracy of 95.61%. This model demonstrated perfect precision for malignant tumors (1.00), meaning it correctly identified all malignant cases. However, its precision for benign tumors was 0.93, which was lower than that of the other models. The recall for benign tumors was perfect (1.00), indicating that the model correctly identified all benign cases. However, recall for malignant tumors was slightly lower at 0.91, similar to SVM. Despite the slightly lower precision for benign tumors, kNN demonstrated the best overall performance in terms of malignant tumor detection due to its perfect precision for malignant cases. In conclusion, Random Forest and XGBoost emerged as the top performers, each achieving an accuracy of 96.49% and displaying strong precision and recall values. Random Forest had a slight edge with its higher recall for malignant tumors, while XGBoost showed slightly lower precision for malignant cases but still performed robustly. SVM was competitive but slightly less effective at identifying malignant tumors, while kNN showed great strength in identifying malignant cases with perfect precision, although it had lower precision for benign tumors. These results demonstrate the strengths and weaknesses of each model, with Random Forest and XGBoost being the most well-rounded and reliable choices for classifying tumors. The main metrics, accuracy, confusion matrix, and area under the curve, are displayed in Table 3 down below. Here, it is easy to compare the performance of the standard models against that of the hypertuned ones. Comparing each model individually, we can examine and decide which model, hypertuned or standard, would be best to use in our stacked approach. Examining the Random Forest Model, they have the exact same confusion matrix and accuracy scores so these metrics are negligible for now. The precision, recall, and F1-scores were identical across both models, indicating that hyperparameter tuning did not significantly impact the core classification performance. The only notable difference between the models was the slight improvement in AUC, where the **hypertuned model performed marginally better** at 0.99607 compared to the original model's 0.99525. This improvement suggests that the hypertuning process made a small but meaningful difference in the model's ability to distinguish between the benign and malignant classes. When comparing the XGBoost model, the hypertuned version

Table 3. Hypertuned vs. Standard Model Metrics

	Random Forest	XGBoost	SVM	kNN
	Standard Models			
Accuracy	0.96491	0.95614	0.95614	0.95614
AUC	0.99525	0.99083	0.995774	0.99591
Confusion Matrix				
	Hypertuned Models			
Accuracy	0.96491	0.96491	0.94737	0.95614
AUC	0.99607	0.99181	0.98854	0.99705
Confusion Matrix				

correctly identified one more malignant tumor than the standard model, indicating that it is more sensitive. This leads to a 0.00877 increase in accuracy and a 0.00097 increase in AUC score indicating a better ability to distinguish between the classes. Thus, the **hypertuned XGBoost** model will be used to stack. Contrary to this, our standard SVM model identifies one benign tumor correctly more than the hypertuned model. Paired with a greater accuracy of 0.00877 and a greater AUC of 0.007234, it becomes clear the standard model is better. However, if we examine Table 3a and 3b, found in the supplemental materials, the standard SVM model achieved a precision of 0.95 for benign tumors and 0.97 for malignant tumors, demonstrating strong accuracy in correctly identifying malignant cases. After hyperparameter tuning, the precision improved slightly to 0.96 for benign tumors and 0.98 for malignant tumors. This improvement indicates that the Hypertuned SVM model was better at avoiding false positives, particularly for malignant cases, highlighting its enhanced reliability. Similarly, the standard SVM model had a recall of 0.99 for benign tumors and 0.91 for malignant tumors. This indicates that while it was highly sensitive to benign cases, it had a slightly lower ability to identify all malignant tumors. The Hypertuned SVM model retained the same recall of 0.99 for benign tumors, ensuring continued sensitivity to benign cases, while improving the recall for malignant tumors to 0.93. This increase in recall demonstrates the hypertuned model's improved capacity to

identify malignant cases, reducing the likelihood of false negatives. Lastly, the standard SVM model achieved F1-scores of 0.97 for benign tumors and 0.94 for malignant tumors, reflecting strong performance with a slight imbalance in handling malignant cases. After hyperparameter tuning, the F1-score for benign tumors remained the same at 0.97, but the F1-score for malignant tumors improved to 0.95. Overall, both models demonstrated strong classification performance, with the Hypertuned SVM showing notable improvements in precision, recall, and F1-scores for malignant tumors. This improvement suggests that the hyperparameter tuning process optimized the model's sensitivity and precision, reducing errors in classifying malignant cases. The trade-offs between metrics were minimal, as the hypertuned model retained the same level of performance for benign tumors while improving for malignant tumors. This makes the **Hypertuned SVM** a more balanced and reliable option, particularly in scenarios where correctly identifying malignant cases is critical. Lastly, comparing our kNN models, there is no difference in the accuracy nor the accuracy metrics so we must analyze deeper. Examining the AUC, the hypertuned model achieved 0.00114 higher AUC underscoring that this increases the model's ability to distinguish between classes while retaining the overall accuracy. Taking a look at the F1-Score, the kNN, the F1-scores were 0.97 for benign tumors and 0.94 for malignant tumors. In the Hypertuned kNN, the F1-scores remained consistent for benign tumors at 0.97 and improved to 0.95 for malignant tumors. These results highlight the hypertuned model's enhanced balance in correctly identifying and predicting both classes. In a similar fashion, the baseline kNN model had a recall of 1.00 for benign tumors, demonstrating perfect sensitivity, and 0.88 for malignant tumors, showing a tendency to miss some malignant cases. In the Hypertuned kNN, recall for benign tumors slightly decreased to 0.99 but improved significantly to 0.93 for malignant tumors. This shift indicates that the hypertuned model was better at capturing malignant cases while maintaining high sensitivity to benign cases, balancing performance across both classes. In conclusion, the **Hypertuned kNN** model demonstrated superior performance compared to the baseline kNN by enhancing recall and F1-scores for malignant tumors while maintaining strong precision and accuracy. These improvements make the hypertuned model more suitable for applications where correctly identifying malignant cases is critical, as it minimizes the likelihood of false negatives while maintaining high overall performance. Overall, the hypertuned models performed better than the standard models, so they will all be used in the stacking portion.

Hybrid Model Results: The stacking approach used Logistic Regression as the meta-model to combine the predictions of various hypertuned base models yielded diverse and competitive results. Each hybrid model combination was evaluated with the exact same metrics as the individual models. The XGBoost & SVM combination emerged as the best overall performer, achieving the highest accuracy and near-perfect F1-scores. Similarly, the XGBoost & KNN pairing demonstrated the highest AUC, emphasizing its ability to separate classes effectively. Hybrid models such as RandomForest & XGBoost and RandomForest & KNN also provided

Table 4. Stacked Model Metrics

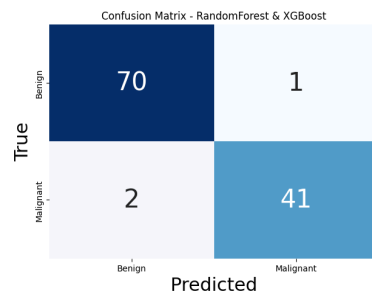
RF & XGBoost

XGBoost & kNN

Accuracy 0.97368

AUC 0.9967

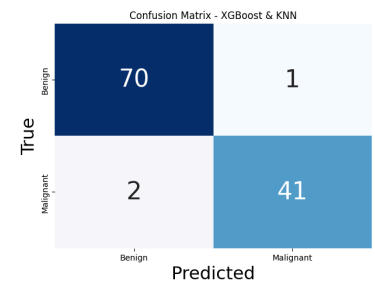
Confusion Matrix



Accuracy 0.97368

AUC 0.9980

Confusion Matrix

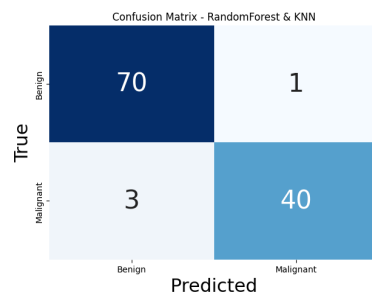


RF & kNN

Accuracy 0.96491

AUC 0.9971

Confusion Matrix

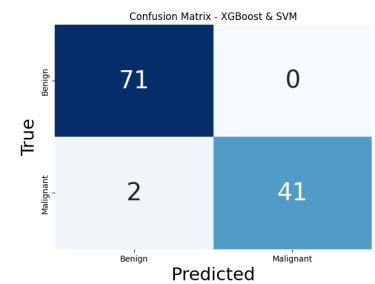


XGBoost & SVM

Accuracy 0.98246

AUC 0.9938

Confusion Matrix

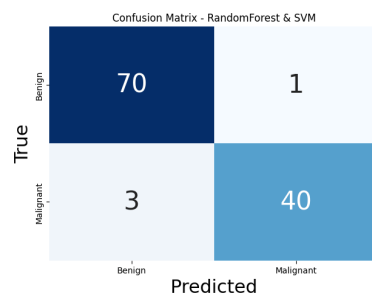


RF & SVM

Accuracy 0.96491

AUC 0.9938

Confusion Matrix

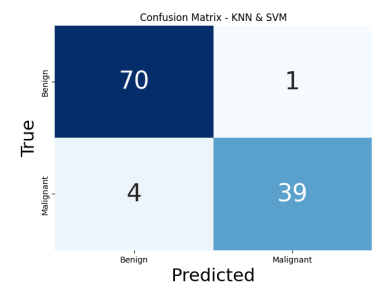


kNN & SVM

Accuracy 0.95614

AUC 0.9934

Confusion Matrix



strong performance, showcasing the benefits of combining ensemble methods with more specialized classifiers. Deeper metrics such as precision, recall, and F1-score are seen in Table 4a, found in the supplemental materials.

Random Forest & XGBoost: This pairing demonstrated an excellent balance between precision and recall, with high values for both classes. The precision of 0.97 for benign cases and 0.98 for malignant cases signifies the model's ability to minimize false positives, while its recall of 0.99 for benign cases and 0.95 for malignant cases highlights its robustness in identifying true cases. These metrics resulted in F1-scores of 0.98 and 0.96 for benign and malignant cases, respectively, reflecting the model's strong overall performance. The high accuracy of 97.37% and an AUC of 0.9967 underline its reliability, particularly in maintaining a low rate of misclassification.

Random Forest & kNN: This hybrid model also performed well, with a precision of 0.96 for benign cases and 0.98 for malignant cases. While its recall for benign cases remained strong at 0.99, the recall for malignant cases was slightly lower at 0.93. These values produced F1-scores of 0.97 for benign cases and 0.95 for malignant cases. Although the model maintained a high accuracy of 96.49% and an AUC of 0.9971, the slight drop in recall for malignant cases suggests a minor trade-off, with the potential for missing a small number of true malignant cases.

Random Forest & SVM: The RandomForest & SVM hybrid closely mirrored the performance of RandomForest & KNN, achieving precision values of 0.96 for benign cases and 0.98 for malignant cases. Recall values were also consistent at 0.99 for benign cases and 0.93 for malignant cases, with F1-scores of 0.97 and 0.95, respectively. Despite an identical accuracy of 96.49%, the slightly lower AUC of 0.9938 hints at a marginal reduction in the model's ability to differentiate between classes compared to the RandomForest & KNN hybrid.

XGBoost & kNN: This hybrid model stood out for its well-rounded performance, achieving precision scores of 0.97 for benign cases and 0.98 for malignant cases, along with recall values of 0.99 and 0.95, respectively. These metrics led to high F1-scores of 0.98 for benign cases and 0.96 for malignant cases. With an accuracy of 97.37% and the highest AUC of 0.9980 among all models, this pairing excelled in balancing sensitivity and specificity, ensuring a strong classification performance with minimal false predictions.

XGBoost & SVM: The XGBoost & SVM hybrid emerged as the top performer, achieving nearly perfect classification metrics. Its precision for malignant cases reached 1.00, indicating an ability to completely avoid false positives, while recall for benign cases was similarly perfect at 1.00. These values resulted in exceptional F1-scores of 0.99 for benign cases and 0.98 for malignant cases. With the highest accuracy of 98.25% among all hybrids and a competitive AUC of 0.9938, this model showcased an unparalleled ability to balance the trade-offs between precision and recall.

kNN & SVM: The KNN & SVM hybrid demonstrated strong precision at 0.95 for benign cases and 0.97 for malignant cases. Its recall values, however, showed a slight decline for malignant cases, dropping to 0.91, resulting in an F1-score of 0.94 for this class. This indicates a minor tendency to miss some true malignant cases. The model maintained an accuracy of 95.61% and

an AUC of 0.9934, but the slight imbalance between precision and recall suggests room for improvement in its classification performance.

Discussion

The present study aimed to evaluate the efficacy of hybrid stacking models in classifying cases as benign or malignant using a relatively small dataset. By combining machine learning algorithms through a meta-model approach, the study demonstrated significant improvements in precision, recall, F1-scores, and accuracy across various hybrid combinations. Notably, combinations involving XGBoost, particularly the XGBoost & SVM pairing, consistently outperformed others in terms of both balanced classification metrics and low misclassification rates. These findings highlight the potential of ensemble techniques for enhancing predictive performance in challenging classification tasks. However, despite the promising results, several limitations warrant discussion. One key limitation is the relatively small size of the dataset used.

Limitations: A limited dataset restricts the capacity for comprehensive testing and generalization of the models to unseen data. While the models achieved high performance on the available dataset, the results may not translate effectively to larger or more diverse datasets, potentially limiting their practical applicability. Furthermore, the small dataset size may lead to an overrepresentation of certain patterns or biases, which could distort the evaluation of model performance. Another limitation relates to the absence of explicit feature selection in the study design. Although feature importance was evaluated using aggregated metrics derived from machine learning models, no formal feature selection process was applied to refine the input variables. Feature selection can play a critical role in improving model performance by reducing dimensionality, enhancing interpretability, and minimizing overfitting. The lack of such an approach in this study may have limited the models' ability to focus on the most discriminative features, potentially impacting their overall effectiveness.

Future Direction: Future research should address these limitations to build upon the findings of this study. Expanding the dataset is essential for enhancing the robustness and generalizability of the models. (1) A larger and more diverse dataset would allow for more rigorous testing, providing a clearer picture of the models' true capabilities and their potential for real-world application. In addition, a larger dataset would support the use of advanced validation techniques, such as k-fold cross-validation with a greater number of folds, to obtain more reliable estimates of model performance. (2) Another promising avenue for future work is the incorporation of a formal feature selection process. Identifying and extracting the most informative features could lead to more efficient and interpretable models, potentially improving classification performance. Methods such as recursive feature elimination, principal component analysis (PCA), or correlation-based feature selection could be explored to determine the optimal subset of features. By emphasizing the most relevant features, future studies may enhance the predictive power of hybrid stacking models while reducing computational complexity. (3) Lastly,

future research could explore the impact of different hyperparameter tuning strategies and meta-model algorithms to further optimize hybrid stacking frameworks. While this study employed logistic regression as the meta-model, other options such as gradient boosting or neural networks might provide additional performance gains. Incorporating explainable AI techniques to interpret the models' decisions could also provide valuable insights for domain experts, particularly in sensitive applications such as healthcare.

Observations and Implications: The comparison reveals that stacked models generally provide meaningful improvements over hypertuned single models, particularly in terms of accuracy and F1-scores. These enhancements can be attributed to the complementary strengths of the combined classifiers, which help mitigate individual model weaknesses. Among the stacked models, XGBoost & SVM consistently demonstrated the best performance across all metrics, making it a strong candidate for applications requiring high precision and recall. However, the performance improvements were not uniform across all stacked models. Certain combinations, such as RF & SVM and RF & kNN, performed similarly to their single-model counterparts, suggesting that the effectiveness of stacking depends heavily on the synergy between the paired classifiers. These findings emphasize the importance of careful model selection when designing ensemble methods.

Comparison

Single vs. Stacked Model: The accuracy of the hypertuned single models was robust, with Random Forest, XGBoost, and kNN achieving 96.49%, while SVM fell slightly short at 94.74%. AUC values for the single models were equally strong, with kNN leading at 0.99705, closely followed by Random Forest at 0.99607. These metrics reflect the high reliability of individual models in distinguishing between benign and malignant cases. However, the stacked models generally outperformed these single models, demonstrating the advantages of combining classifiers. For instance, the XGBoost & SVM combination achieved the highest accuracy of 98.25% and an AUC of 0.9938. Similarly, other hybrid combinations, such as XGBoost & kNN and RF & XGBoost, achieved accuracy rates of 97.37%, showing consistent improvements over their individual components. Precision values for both single and stacked models were notably high, particularly for malignant cases. Hypertuned single models like Random Forest, XGBoost, and SVM achieved precision scores of 0.98 for the malignant class, demonstrating their ability to minimize false positives effectively. The stacked models, however, enhanced these metrics further. The XGBoost & SVM pair achieved a perfect precision score of 1.00 for malignant cases, illustrating its superior ability to classify positive instances without error. Recall, which measures the models' ability to correctly identify malignant cases, showed a slight improvement in the stacked models compared to the single ones. Hypertuned single models typically achieved a recall of 0.93 for malignant cases, leaving room for improvement in minimizing false negatives. In contrast, the XGBoost & SVM stacked model achieved a recall of 0.95, reducing false negatives and demonstrating greater sensitivity in detecting malignant cases. This

improvement is crucial in applications where the cost of missing malignant cases is high. The F1-score, which balances precision and recall, further highlights the advantages of the stacked models. Hypertuned single models consistently achieved macro-average F1-scores of 0.96, demonstrating their balanced performance across both classes. However, the XGBoost & SVM stacked model raised this macro-average to 0.98, representing a more robust and balanced classification. Other stacked combinations, such as RF & XGBoost and XGBoost & kNN, also showed slight improvements, with macro-average F1-scores of 0.97, emphasizing their ability to classify benign and malignant cases effectively.

Our approach vs. Related Work

Evaluation Metrics and Optimization: Many of the related works focused primarily on accuracy as the sole metric for evaluating model performance. For instance, studies by [21], [22], and [26] relied solely on accuracy, which provides a limited understanding of model effectiveness, particularly in datasets where class imbalance may exist. While [27] extended evaluation to precision, recall, and F1-scores, their models were not optimized and relied on standard configurations, potentially underestimating model capabilities. In contrast, our study evaluates performance using a broader set of metrics, including precision, recall, F1-score, accuracy, and AUC, ensuring a comprehensive analysis of each model's strengths and weaknesses. Furthermore, we employed hyperparameter optimization for all single classifiers, enabling a more accurate representation of their potential. This contrasts with most studies, which used default configurations, as seen in [21], [23], and [27]. By optimizing parameters, we achieved consistent improvements across all evaluated models, ensuring a fair comparison with the stacked models.

Stacked Models and Feature Importance: Unlike previous studies, our approach explored hybrid models by stacking classifiers, combining their strengths to enhance classification performance. For instance, while [21] and [23] relied on single models like SVM and kNN, our study demonstrated that combining classifiers, such as XGBoost & SVM, outperformed these standalone models in metrics like recall and F1-score. This innovative approach highlights the advantages of ensemble methods in achieving higher robustness and accuracy. Feature selection is another critical area where our approach diverges from previous studies. Although we did not explicitly perform feature elimination or extraction in this study, we assessed feature importance using aggregated feature importance derived from multiple classifiers. In contrast, [23] demonstrated the significant impact of feature elimination on performance but did not explore how decision-tree-based models could contribute to this process. Our findings suggest that incorporating explicit feature selection techniques in future work could further enhance model performance, as informative features could be prioritized while redundant ones are discarded.

Cross-Validation and Generalization: Many of the related works lacked robust validation methodologies. For instance, studies like [24] and [26] used 5-fold cross-validation, while others such as [22] and [27] did not incorporate cross-validation at all. Cross-validation is critical for

assessing a model's ability to generalize across unseen data. Our study utilized a hold-out method but emphasized the potential benefits of adopting more robust techniques, such as k-fold or stratified cross-validation, in future research to ensure generalizability.

Decision-Tree Model Inclusion: Decision-tree-based models, such as Random Forest and XGBoost, have been underrepresented in previous studies. For instance, [23] and [24] did not explore decision-tree-based methods, despite their demonstrated effectiveness in handling feature interactions and non-linear data patterns. Our study addressed this gap by including decision-tree models in both single and stacked configurations, providing a more comprehensive exploration of classifier performance. The inclusion of these models revealed their competitive edge, particularly when combined with other algorithms, as seen in the RF & XGBoost and XGBoost & SVM pairs.

Limitations in Related Work: Previous studies suffered from several limitations, including small sample sizes, lack of feature selection, and reliance on static evaluation metrics. For instance, [25] only utilized three features and acknowledged that eliminating features discarded informative ones, which may have impacted model performance. Additionally, studies like [21], [22], and [26] did not optimize models, potentially leading to suboptimal results. Our study sought to address these gaps through model optimization and comprehensive evaluation but also recognizes that the dataset size in our research remains a limitation.

Conclusion

This study evaluated the performance of individual machine learning models—Random Forest, XGBoost, Support Vector Machines, and K-Nearest Neighbors—and their stacked combinations for tumor classification. While individual models exhibited high accuracy and strong metrics across precision, recall, and F1-score, the stacked ensembles demonstrated a marked improvement in classification performance, with the **XGB & SVM pair achieving the highest accuracy** of 98% and the XGB & kNN pair achieving the highest AUC of 0.9980. These findings underscore the efficacy of ensemble methods in leveraging the strengths of multiple models for enhanced robustness and predictive accuracy. Despite the promising results, the study faced limitations, including a relatively small dataset size, which constrained comprehensive testing and generalization. Additionally, feature selection was not explicitly performed, which could further refine model performance in future studies. Addressing these limitations and incorporating advanced techniques such as feature engineering and validation on larger datasets will be crucial to improving model reliability and applicability in clinical contexts. This research contributes to advancing automated tumor classification, offering insights into optimizing machine learning frameworks for medical diagnostics.

Data Availability

The dataset used in this study, the Breast Cancer Wisconsin (Diagnostic) dataset, is publicly available from the University of California Irvine Machine Learning Repository. It can be accessed through <https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>.

Author Information

Corresponding Author

Roberto Ruiz Felix - *Department of Biomedical Informatics, College of Health Solutions, Arizona State University, Arizona, 85004, USA.*

Email: rruizfel@asu.edu

Authors

Haiyan Wang - *School of Mathematical and Natural Sciences, Arizona State University, 85281, USA.*

Simon Tran - *School of Mathematical and Natural Sciences, Arizona State University, 85281, USA.*

Authors Contributions

The research and analysis in this paper were conducted by Roberto Ruiz Felix as part of the requirements for MAT 422: Mathematical Methods in Data Science under the supervision of Haiyan Wang and Simon Tran. Roberto Ruiz Felix was responsible for data preprocessing, implementing machine learning models, performing evaluations, and writing the manuscript. Haiyan Wang and Simon Tran provided guidance on research design, and methodology.

This format gives proper credit to everyone involved and aligns with academic norms for acknowledging contributions.

Acknowledgements

We would like to thank University of California Irvine for providing open access to sample and clinical information used in this study, and the University of Wisconsin Hospitals, Madison for collecting the data.

Supplemental Material

Table 3a. Standard Model Metrics

	Precision	Recall	F1-Score	Support
		Random Forest		
Benign (0)	0.96	0.99	0.97	71

	Precision	Recall	F1-Score	Support
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114
XGBoost				
Benign (0)	0.96	0.97	0.97	71
Malignant (1)	0.95	0.93	0.94	43
Macro Avg.	0.96	0.95	0.95	114
Weighted Avg.	0.96	0.96	0.96	114
SVM				
Benign (0)	0.95	0.99	0.97	71
Malignant (1)	0.97	0.91	0.94	43
Macro Avg.	0.96	0.95	0.95	114
Weighted Avg.	0.96	0.96	0.96	114
kNN				
Benign (0)	0.93	1.00	0.97	71
Malignant (1)	1.00	0.88	0.94	43
Macro Avg.	0.97	0.94	0.95	114
Weighted Avg.	0.96	0.96	0.96	114

Table 3b. Hypertuned Model Metrics

	Precision	Recall	F1-Score	Support
Random Forest				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114

	Precision	Recall	F1-Score	Support
XGBoost				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114
SVM				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114
kNN				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.96	0.96	0.96	114

Table 4a. Stacked Model Metrics

	Precision	Recall	F1-Score	Support
RF & XGBoost				
Benign (0)	0.97	0.99	0.98	71
Malignant (1)	0.98	0.95	0.96	43
Macro Avg.	0.97	0.97	0.97	114
Weighted Avg.	0.97	0.97	0.97	114
RF & kNN				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43

	Precision	Recall	F1-Score	Support
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114
RF & SVM				
Benign (0)	0.96	0.99	0.97	71
Malignant (1)	0.98	0.93	0.95	43
Macro Avg.	0.97	0.96	0.96	114
Weighted Avg.	0.97	0.96	0.96	114
XGBoost & kNN				
Benign (0)	0.97	0.99	0.98	71
Malignant (1)	0.98	0.95	0.96	43
Macro Avg.	0.97	0.97	0.97	114
Weighted Avg.	0.97	0.97	0.97	114
XGBoost & SVM				
Benign (0)	0.97	1.00	0.99	71
Malignant (1)	1.00	0.95	0.98	43
Macro Avg.	0.99	0.98	0.98	114
Weighted Avg.	0.98	0.98	0.98	114
kNN & SVM				
Benign (0)	0.95	0.99	0.97	71
Malignant (1)	0.97	0.91	0.94	43
Macro Avg.	0.96	0.95	0.95	114
Weighted Avg.	0.96	0.96	0.96	114

Index

Table 1. Related Work Comparison/Limitations: Compares related work similar to this dataset and identifies strengths/weaknesses to use and improve on.

Table 2. Aggregated Feature Importance Score: Displays the top ten most important features based on the aggregated and standardized scores given by the applied machine learning models.

Table 3a. Standard Model Metrics: Gives the comprehensive metrics used beyond the top three to evaluate the standard models used.

Table 3b. Hypertuned Model Metrics: Gives the comprehensive metrics used beyond the top three to evaluate the hyperuned models used.

Table 4. Stacked Model Metrics: Returns the three metrics mainly used to assess the performance of the stacked models; accuracy, AUC, and confusion matrix.

Table 4a. More Stacked Model Metrics: Gives a classification report and deeper metrics used beyond the top three to evaluate the stacked models used.

References

1. Breast cancer [Internet]. World Health Organization; 2024 [cited 2024 Nov 28]. Available from: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer#:~:text=Breast%20cancer%20is%20a%20disease,producing%20lobules%20of%20the%20breast.>
2. Anatomy of the breasts [Internet]. John Hopkins University ; 2024 [cited 2024 Nov 28]. Available from: <https://www.hopkinsmedicine.org/health/wellness-and-prevention/anatomy-of-the-breasts>
3. Schockney LD. Breast anatomy [Internet]. National Breast Cancer Foundation ; 2024 [cited 2024 Nov 28]. Available from: <https://www.nationalbreastcancer.org/breast-anatomy/>
4. Breast anatomy and how breast cancer starts [Internet]. National Breast Cancer Foundation; 2023 [cited 2024 Nov 28]. Available from: <https://nbcf.org.au/about-breast-cancer/diagnosis/breast-cancer-anatomy/>
5. Breast cancer statistics [Internet]. World Cancer Research Fund; 2024 [cited 2024 Nov 28]. Available from: <https://www.wcrf.org/preventing-cancer/cancer-statistics/breast-cancer-statistics/>
6. Breastcancer.org. Breast cancer facts and statistics [Internet]. Breastcancer.org; 2024 [cited 2024 Nov 28]. Available from: <https://www.breastcancer.org/facts-statistics>
7. Breast cancer statistics: How common is breast cancer? [Internet]. American Cancer Society ; 2024 [cited 2024 Nov 28]. Available from: <https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html>
8. Breast cancer facts [Internet]. Wisconsin Breast Cancer Coalition; 2022 [cited 2024 Nov 28]. Available from: <https://www.wibreastcancer.org/resources/breast-cancer-facts/>

9. Pinto-Coelho L. How artificial intelligence is Shaping Medical Imaging Technology: A survey of innovations and applications. Won Lee S, editor. *Bioengineering*. 2023 Dec 18;10(12):1435. doi:10.3390/bioengineering10121435
10. Monaco A, Pantaleo E, Amoroso N, Lacalamita A, Lo Giudice C, Fonzino A, et al. A Primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal*. 2021 Jul 31;19:4345–59. doi:10.1016/j.csbj.2021.07.021
11. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*. 2019 Apr 11;18(6):463–77. doi:10.1038/s41573-019-0024-5
12. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *New England Journal of Medicine*. 2019 Apr 4;380(14):1347–58. doi:10.1056/nejmra1814259
13. Sidey-Gibbons JA, Sidey-Gibbons CJ. Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*. 2019 Mar 19;19(1). doi:10.1186/s12874-019-0681-4
14. Zeng W, Jia J, Zheng Z, Xie C, Guo L. A comparison study: Support vector machines for binary classification in Machine Learning. 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI). 2011 Oct 17;1621–5. doi:10.1109/bmei.2011.6098517
15. Zhang S. Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*. 2022 Oct 1;34(10):4663–75. doi:10.1109/tkde.2021.3049250
16. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in classification. *Lecture Notes in Computer Science*. 2003;986–96. doi:10.1007/978-3-540-39964-3_62
17. Kirasich K, Smith T, Sadler B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. 2018 Aug 25;1(3).
18. Kumari R, Kr. S. Machine learning: A review on Binary Classification. *International Journal of Computer Applications*. 2017 Feb 15;160(7):11–5. doi:10.5120/ijca2017913083
19. Bansal A, Kaur S. Extreme gradient boosting based tuning for classification in Intrusion Detection Systems. *Communications in Computer and Information Science*. 2018 Oct 31;905:372–80. doi:10.1007/978-981-13-1810-8_37
20. Shi H, Wang H, Huang Y, Zhao L, Qin C, Liu C. A hierarchical method based on weighted extreme gradient boosting in ECG Heartbeat Classification. *Computer Methods and Programs in Biomedicine*. 2019 Apr;171:1–10. doi:10.1016/j.cmpb.2019.02.005
21. Tarannum R, Wood J, Ensari T. Breast Cancer Classification with Machine Learning [Internet]. Arkansas Tech University ; 2024 [cited 2024 Nov 29]. Available from: https://orc.library.atu.edu/cgi/viewcontent.cgi?article=1097&context=atu_rs
22. Breast cancer prediction: How machine learning can support the process [Internet]. Supper & Supper GmbH; 2023 [cited 2024 Nov 29]. Available from:

- <https://www.linkedin.com/pulse/breast-cancer-prediction-how-machine-learning-can-support-mipge/>
23. Spillai. Wisconsin Breast Cancer Diagnosis: SVM Analysis [Internet]. University of Massachusetts Amherst; 2023 [cited 2024 Nov 29]. Available from: <https://spillai.sites.umassd.edu/2023/12/16/wisconsin-breast-cancer-diagnosis-svm-analysis/>
 24. Ueno Y. Breast Cancer Classification with Logistic Regression [Internet]. 2024 [cited 2024 Nov 29]. Available from: <https://deepnote.com/app/yura-ueno/Breast-Cancer-Classification-a1f36dc3-3558-4c42-af70-94dfa8d2fd79>
 25. Mangasarian OL, Wolberg WH. Machine Learning for Cancer Diagnosis and Prognosis [Internet]. University of Wisconsin-Madison; [cited 2024 Nov 29]. Available from: <https://pages.cs.wisc.edu/~olvi/uwmp/cancer.html>
 26. Aamir S, Rahim A, Aamir Z, Abbasi SF, Khan MS, Alhaisoni M, et al. Predicting breast cancer leveraging supervised machine learning techniques. *Computational and Mathematical Methods in Medicine*. 2022 Aug 16;2022:1–13. doi:10.1155/2022/5869529
 27. Asri H, Mousannif H, Moatassime HA, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*. 2016;83:1064–9. doi:10.1016/j.procs.2016.04.224