Performance Comparison of Classification Algorithms in Breast Cancer Diagnosis

Mohammed Salman Khan
Lakehead University

Protiva Arafin
Lakehead University

***Abstract*—In this research, there will be comparison of the performance of various supervised Classification Algorithms on a dataset about Breast Cancer Diagnosis. The dataset is taken from the University of Wisconsin Hospitals and has 30 features along with 569 rows and gives a binary target (M or B). The study is going to compare the performance of the Random Forest Classifier, Support Vector Machine (SVM), Gradient Boosting Classifier (XGBoost), k-nearest Neighbors(k-NN), and Naive Bayes Classifier, models for the dataset. The research compare the accuracy, precision, recall, and F1 score for every model. In the end, based on all these metrics, it intends to suggest which model is the best for Breast Cancer Diagnosis .**

INTRODUCTION

Let us first begin by understanding why early breast cancer diagnosis is important. According to the World Health Organization (WHO), it is the most common form of cancer found worldwide with almost 670000 deaths globally[1]. It was the most common form of cancer in almost 157 out of 185 countries globally in 2022. According to the National Breast Cancer Foundation, early diagnosis can ensure a survival rate of around 99%, whereas overtime due to late diagnosis it can drop to as low as 30%[2]. CNNs are the prevalent algorithm used for breast cancer diagnosis, especially for diagnosis direct from images. Datasets like the one this study is using which have numerical data, there is high applicability for SVMs and other algorithms. Thus, this research wants to compare the performance of some of these algorithms. It will be comparing the Random Forest Classifier, Support Vector Machine (SVM), Gradient Boosting Classifier(XGBoost), k-nearest Neighbors(k-NN), and Naive Bayes Classifier models on the metrics of Precision, Accuracy, Recall, and F1 score. About the confusion matrix, it is known that recall is important to prevent false negatives as this study ideally does not want to miss any person with cancer. While accuracy is important, avoiding false positives and hence improving the precision, false positives are less crucial than false negatives in our case. So by F1-Score analysis, it can be said that the model that gives a big recall score will be useful to this research.

Problem Definition

The paper is trying to compare the performance of various algorithms against a breast cancer dataset. It will look to find out based on various performance metrics which algorithm is the best.

The expected outcome for this paper will be that due to the theoretical hypothesis, the Random Forest Algorithm should give the best results due to it using ensemble learning methodology, robustness against overfitting, and ability to process large datasets effectively.

For the comparison of performance, this study will be using precision, recall, and F-1 score. It will be also use 10-cross validation to compare the performance as cross-validation prevents overfitting and provides a more robust estimate of the generalization capacity of the models.

2162-2248 © IEEE

Data Description

- *Dataset Source:–* The dataset this study uses is from the University of Wisconsin Hospitals and is the product of digitization of images of breast cancer images and downloads from Kaggle.

- *Dataset Structure:–* This dataset assigns a numerical value to several image characteristics such as giving a numerical description of the nuclei in the images. It contains 569 rows, where it has around 30 features and a target variable that returns M for malignant and B for benign. The dataset has a class distribution of 357 Benign and 212 Malignant.

- *Preprocessing:–* This study operates some preprocessing steps. It removes the two columns ('id', 'Unnamed: 32') as they are unimportant for the analysis. The target value is in categorical form, and it is converted to binary values (Malignant = 1, Benign =0). Afterwards, it prepares features for PCA. Features are standardized and handle any missing values. This process makes sure that each feature contributes equally. This helps to perform better in many algorithms such as - SVM, KNN, etc. PCA is the process of reducing the number of features and creating new summary features. This study applies PCA to reduce the dataset to 2 principal components, whereas it maintains all the important information.

- *Exploratory Data Analysis:–* This is the process of understanding the data, identifying anomalies, and detecting patterns. Hence, this study analyzes the dataset to understand its characteristics. It implements EDA before applying any algorithms. It displays the first 5 rows, statistics details, data type information, and the number of unique values. Also checks the missing values.

- *Visualization:–*
  After the standardization, figure 1 shows the distribution of the first 12 features. Most of the features show the characteristics of normal distribution. Some of them follow have skewed distributions.

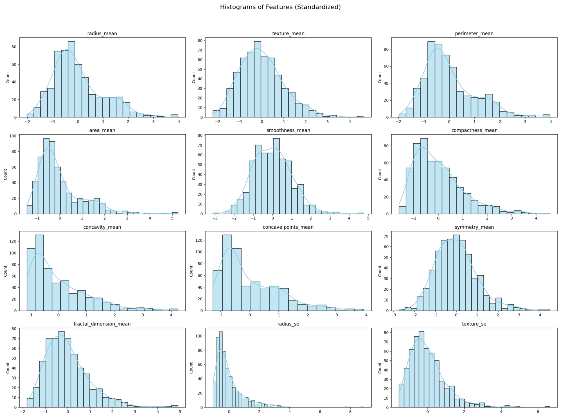


FIGURE 1. Histograms of Standardized Features

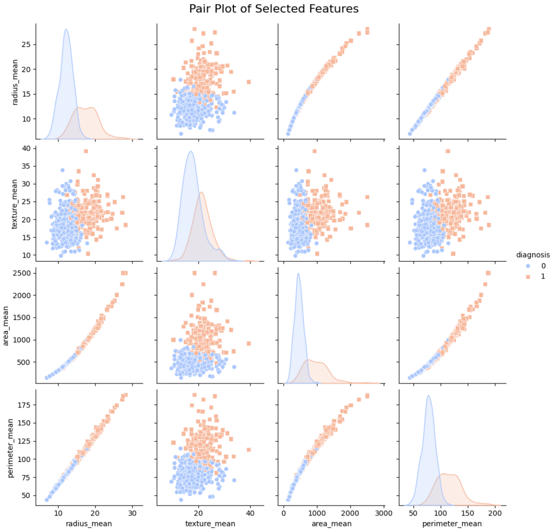


FIGURE 2. Pair Plot of Selected Features

Figure 2 shows the pair plot of selected features. It displays the relationship between selected features and the target variable (diagnosis). In this image, radius_mean and area_mean are strongly correlated, and the radius_mean, parameter_mean states a good separation between malignant and benign classes.

Methodology

This study implements five algorithms and evaluates the performance to compare the results to see which classification model works better on the breast cancer dataset. Here are the descriptions of five algorithms-

- *Random Forest:*– This model uses a collection of decision trees to avoid overfitting and improve classification accuracy. Hyperparameters tuned for this model are – the number of estimators (120) and the maximum leaf nodes (15).

- *Support Vector Machine:*– This model uses the Gaussian distribution to compute the probabilities. Gaussian Naive Bayes works for numerical data. It employs a linear kernel to split the classes with a straight line and c (regularization) is tuned to balance flexibility.

- *XGBoost:*– This model implements gradient boosting on decision trees. Hyperparameters tuned for this model are – the number of estimators (50), the maximum depth (50), and the evaluation matric (logloss).

- *K-Nearest Neighbors:*– This model classifies the data point depending on nearby points. The number of neighbors (5) is tuned for optimization.

- *Naive Bayes:*– This probabilistic model uses the Gaussian distribution to compute the probabilities. Gaussian Naive Bayes works for numerical data.

Model Training

The dataset is split into training and testing as 80-20. PCA is applied to training and testing sets. At the last, this study implements 10-fold cross-validation to ensure the model generalizes well to unseen data and also avoids overfitting.

Evaluation Metrics

In cancer detection, accuracy is important but it is also crucial to prevent false negatives as it will postpone the treatment and can be deadly for a person. However false positives can result in unnecessary treatment, medical costs, and stress. Moreover, the F1-Score balances precision(minimize false positives) and recall(minimize false negatives). As this dataset is moderately imbalanced, this study ensures the performance is not dominated by the majority class through evaluating the average metrics.

Tools and Libraries

This study uses Python, scikit-learn for implementing machine learning algorithms. Also applies numpy, and pandas for data preprocessing, and data manipulation. Moreover, it utilizes seaborn, and malplotlib for visualizing the data distributions through histograms, pair plots, etc.

Results and Discussion

Train-test split strategy

These five algorithms - Random Forest, Support vector machine, XGBoost, K-Nearest Neighbors, and Naive Bayes are evaluated for accuracy.

As per the figure 3, the results are as follows - 99%,99%,97%,97%, and 94%. For overall accuracy, this study states that Random Forest and SVM show the highest accuracy.

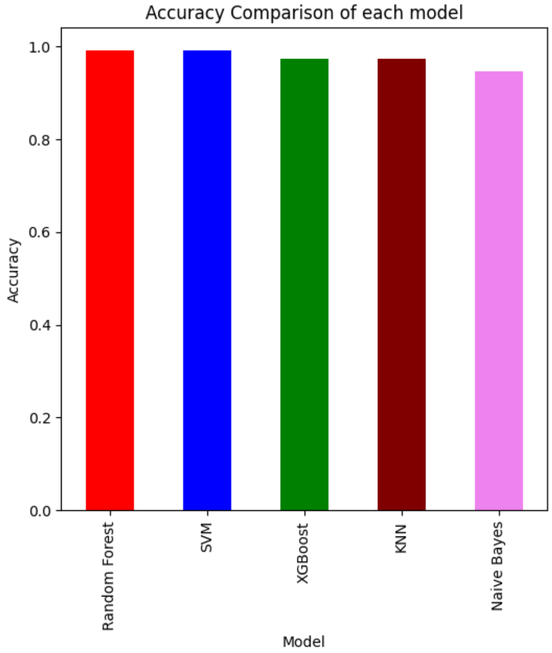
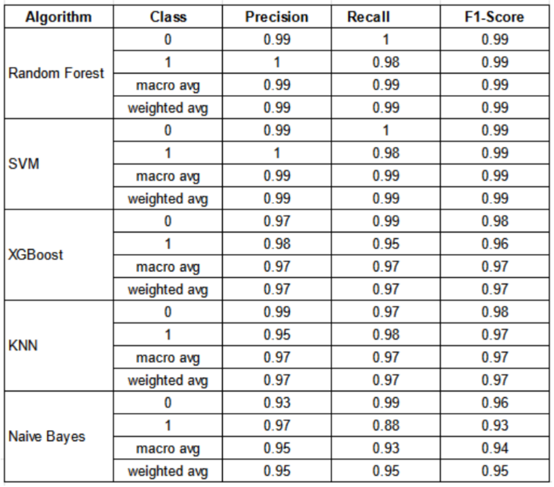


FIGURE 3. Accuracy Comparison of each model

After analysis the all aspects figure 4 states that Random Forest and SVM show the highest performance.

For both classes, these two algorithms achieve the highest precision (0.99 for Class 0 and 1.00 for Class 1), recall(1 for Class 0 and 0.98 for Class 1), and F1-Score value(0.99). These algorithms show consistence performance as they have higher values in macro and weighted averages (both 0.99). Random Forest is easier to interpret and SVM provides a clear decision boundary.



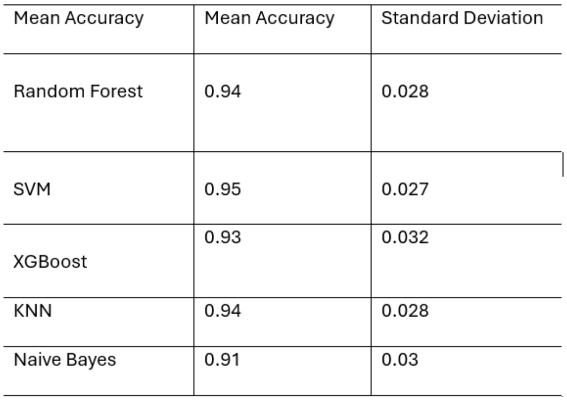
Algorithm	Class	Precision	Recall	F1-Score
Random Forest	0	0.99	1	0.99
	1	1	0.98	0.99
	macro avg	0.99	0.99	0.99
	weighted avg	0.99	0.99	0.99
SVM	0	0.99	1	0.99
	1	1	0.98	0.99
	macro avg	0.99	0.99	0.99
	weighted avg	0.99	0.99	0.99
XGBoost	0	0.97	0.99	0.98
	1	0.98	0.95	0.96
	macro avg	0.97	0.97	0.97
	weighted avg	0.97	0.97	0.97
KNN	0	0.99	0.97	0.98
	1	0.95	0.98	0.97
	macro avg	0.97	0.97	0.97
	weighted avg	0.97	0.97	0.97
Naive Bayes	0	0.93	0.99	0.96
	1	0.97	0.88	0.93
	macro avg	0.95	0.93	0.94
	weighted avg	0.95	0.95	0.95

FIGURE 4. Performance Comparison Table Across Key Metrics

Cross-validation method

K-fold cross-validation (k=10) is implemented because this provides a robust and generalized measure of model performance, reduces variability, and ensures fair comparison.

As per the figure 5, SVM shows higher accuracy than Random Forest with a slightly lower standard deviation. XGBoost and KNN display consistent results, but naive Bayes is weaker than all. Ultimately, SVM's kernel-based approach appears as the most robust and consistent model as this model shows higher mean accuracy and lowest standard deviation.



Mean Accuracy	Mean Accuracy	Standard Deviation
Random Forest	0.94	0.028
SVM	0.95	0.027
XGBoost	0.93	0.032
KNN	0.94	0.028
Naive Bayes	0.91	0.03

FIGURE 5. Cross-Validation Results: Mean Accuracy and Standard Deviation for Each Model.

As per the figure 6, this bar chart displays the mean accuracy of classifiers. SVM achieves the highest accuracy, closely followed by Random Forest and KNN.

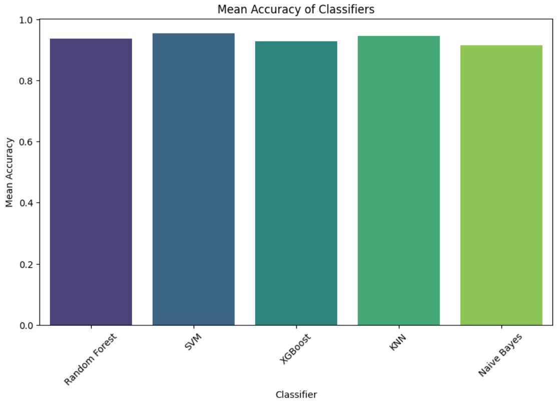


FIGURE 6. Mean Accuracy of classifiers

Conclusion

While the fact that the paper expected Random Forrest to be the most precise model is confirmed after looking at Figure 4. We see that Random Forrest has an F-1 score of about 0.99 which is equaled by the SVM model. Closely followed by XGBoost and KNN model which have a score of 0.98. Naive Bayes had the overall poorest score, which was still relatively high.

The 10-cross validation we used helps to further compare the performance and we can see that in this case, the SVM has the best performance.

So in conclusion we can say looking at both comparison metrics SVM has proved to be the best way overall to look at this classification problem.

Future Work

According to the authors some of the future work that can be done is to first explore how different forms of feature selection can produce a different effect, it would be interesting to see and compare the difference with PCA.

Secondly, there could be a use of techniques like SMOTE (Synthetic Minority Oversampling Technique) to handle class imbalances. There can also be more research done into how deep learning models like neural networks can handle the same issue.

It would be crucial to see if these results would hold up against other datasets. Sometimes such ML models can have a black-box effect and there is a lack of understanding and interpretation of the results. So implementing some model explain-ability techniques like SHAP or LIME could help improve that.

REFERENCES

1. World Health Organization. ”Breast Cancer.” Accessed November 25, 2024. https://www.who.int/newsroom/fact-sheets/detail/breast-cancer.
2. National Breast Cancer Foundation, ”Early Detection of Breast Cancer,” [Online]. Available: https://www.nationalbreastcancer.org/early-detection-of-breast-cancer/. [Accessed: Nov. 25, 2024].