JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 08 Issue 01
2024

Research Article
https://doi.org/10.56979/801/2024

A Comparative Study of Advanced Machine Learning Ensemble Techniques for Classification of Breast Cancer

Tayyaba Yasmeen[1], Muazzam Ali[1*], M U Hashmi[1], M Adnan Hashmi[2], and Zeeshan Mehmood[3]

[1]Department of Basic Sciences, Superior University, Lahore, 54000, Pakistan.
[2]Higher Colleges of Technology, Abu Dhabi.
[3]Faculty of Sciences, Superior University, Faisalabad, 37000, Pakistan.
[*]Corresponding Author: Muazzam Ali. Email: muazzamali@superior.edu.pk

Received: May 21, 2024 **Accepted:** September 21, 2024

Abstract: The most prevalent and the life-threatening diseases include Breast cancer worldwide. An early detection and diagnosis which are accurate is essential for improving in survival rates. In the real-time applications, models of machine learning provide powerful tools for aiding in medical professionals in diagnosing the breast cancer very accurately and efficiently. This research focuses on application of advanced ensemble techniques, such as the Random Forest, the Support Vector Machines (SVM), the K-Nearest Neighbors (KNN), the Stacking, the Boosting, and the Blending, to classify cancer of breast as either benign or malignant. After a comprehensive analysis, the Boosting emerged as the highest-performing model with an accurate precision of 0.8493 and a ROC AUC of 0.9051. These findings are align with the Sustainable Development Goal (SDG) 3**,** which advocates for health care and well-being, that highlighting the importance of an accessible, data-driven education in healthcare and decision of the support systems. Hence, our work suggests that these models can significantly enhance precision in diagnostic, reducing the burden on systems of healthcare, especially in underserved areas.

Keywords: Breast Cancer Classification; Machine Learning; Ensemble Techniques; Random Forest; Boosting Algorithm; Medical Diagnostics AI.

1. Introduction

Breast cancer is a kind of cancer that begins in the cells of the breast. It may manifest as a lump, a change in shape, or fluid, among other symptoms. BRCA1/2 mutations [4] are among the genetic risk factors. Other lifestyle variables include alcohol and tobacco usage. However, changing one's lifestyle could only prevent 25% of cases. Breast cancer became the most common type of cancer in 2020 when more cases were diagnosed than lung cancer [1]. Due to the high expense of consultations and the shortage of qualified specialists, many women are unable to receive the treatment they require. Developing countries such as India also have a significant shortage of medical professionals. Therefore, the implementation of automated clinical decision support systems might potentially mitigate this issue. But one of the main barriers to the adoption of such systems is the general lack of understanding regarding the role that machine learning models play in medicine. This essay seeks to address this problem by offering findings that bolster the argument in favor of integrating these models into the therapeutic process. Early detection improves prognosis, however undiagnosed breast cancer frequently results in death [2]. Consequently, an automated a technique for diagnosing breast cancer might significantly reduce the amount of deaths. Additionally, such a technology might support medical professionals in the field by validating their diagnosis. The current process of manually examining mammography findings is not scalable in countries with huge populations like India.

The Random Forest method regularly produces better models for the detection of different illnesses, as seen in [5]–[9]. In [5] diagnose coronary heart disease with an accuracy of 83.85%. Additionally, they demonstrate how well Random Forests detect breast cancer. Random Forests perform better than current machine learning techniques for categorizing neuroimaging data related to Alzheimer's disease, according [6]. A review of [10]–[13] provides evidence in favor of using machine learning to diagnose breast cancer. Evidence that the Support Vector Machine and Random Forest algorithms detect malignant tumors with

above 96% accuracy [12]. A Deep Neural Network outperforms common supervised models like K-Nearest Neighbors and Decision Trees. [11], with an accuracy of 97.21%. Neural networks are not recommended for the selected dataset, though, because they tend to over fit on smaller data sets. Several additional supervised machine learning methods also shown efficacy in identifying malignant breast cancers, as demonstrated by [14]–[17] and [24]. In [14] use several Decision Tree classifiers to diagnose breast cancer with accuracies of 95%. Using the Wisconsin breast cancer dataset, Desai and Shah [15] implement a Multilayer Perceptron (MLP) classifier and find an accuracy of 91.9%. A Support Vector Machine version, known as the Least Square Support Vector Machine (LS-SVM), is used by Polat and Güneş [16]. They report a final accuracy of 98.53%, confirming that using SVMs for tumor identification is feasible. Finally, K-Nearest Neighbors (KNN) is used [17]. In comparison to current models, they assert to have improved classification outcomes by 1.17%.

This research uses supervised machine learning approaches to develop such a system. Given the increased likelihood of mortality associated with malignant tumors, a classification model has to be developed. The only goals of treatment for benign tumors are to get rid of the malignant cells and keep them from coming back. Since malignant tumors can spread to other parts of the body, it's important to diagnose people with them at an early stage of the illness.

2. Materials and Methods

2.1. Data Gathering

For this research, the dataset of Breast Cancer was utilized and taken from Kaggle. This dataset contains the detailed information about the breast cancer diagnosis, in which it includes various patient characteristics and results in diagnostic. The all data was collected from the clinical environments where the patients were screened for cancer of breast using medical imaging, followed by the laboratory analyses of the biopsy results.

This dataset includes the variables that describe:

- Characteristics of patients such as age, gender, and the history of family of breast cancer.
- Features of diagnostic from imaging techniques such as mammography, including the tumor size, its shape, its texture, and the margins.
- Findings in pathological such as the appearance of malignant cells, the hormone receptor status, and the tumor grading.
- Outcome data, which includes whether that cancer is benign or malignant, as well as rates of survival and the recurrence statistics.

This dataset was processed by the removing or any missing or unauthentic data points to ensure accuracy during the phase of analysis. The Data was first cleaned, then standardized, and then prepared for the modeling. This dataset comprises a number of the rows, in which each representing an individual case of patient. The given below is a brief display of sample rows some from the dataset.

Table 1. Data Set of the Breast Cancer

ID	Age	Tumor_Size	Diagnosis	Texture	Smoothness	Concavity	Outcome
1	50	23	Malignant	12.5	0.25	0.35	Survived
2	45	18	Benign	8.6	0.18	0.22	Recurred
3	60	25	Malignant	13.2	0.28	0.37	Deceased

In this, each row shows the key data points for individual cases of patient, including their age, the tumor characteristics, the diagnosis type, and outcome status. This table is representing a snapshot of how the dataset is organized for the statistical analysis and the machine learning modeling in the process of research.

2.2. Data Preprocessing

In order to prepare the dataset for training and testing in our machine learning model the following preprocessing was done. These steps were essential for preparing the data and ensuring the model could interpret it accurately:

Data Upload: The dataset was then loaded and imported in the research environment by following the instructions given by a custom function which uses the files.upload() method. The data was then loaded into a Pandas Data Frame which is a commonly used data structure to work with tables and data frames. This process helped in the first stage of data cleaning of the given dataset for future analysis.

Label Encoding: For the same reason, as the Random Forest Classifier or other models used in the present research process only work with numerical variable, all the categorical variables were encoded using the

Label Encoder from the learn. Preprocessing. Since we have some categorical variables like the diagnostic labels for each tree measured, the Label Encoder was used to convert these categories into integer for machine learning interpretation during training of the model.

Feature and Target Separation: The variables were assigned and classified into features set, commonly known as independent variables and the target set, also widely known as the dependent variable. To validate this assumption, the last column of the dataset was considered as the target variable (cancer type –benign or malignant), and all the other columns as features including characteristics of a patient and cancer diagnostic results. This separation was important for the next step of training and testing the model.

Train-Test Split: To assess the effectiveness of the chosen model, it was decided to divide the dataset into a training and testing part using the train_test_split tool from the sklearn.model_selection library. The data was split in the ratio of the training set at 80% while the test set comprised 20%. It was chosen that a random seed of 42 will be initiated in order that the result set will always be the same, making it possible to compare them at different steps and with various models.

These preprocessing steps provided the groundwork for correct model training by properly organizing the data and standardizing data formats that can be well interpreted by a machine learning model.

2.3. Machine Learning Models and Ensemble Techniques.

For this research, we utilized the following classifiers and ensemble techniques: The machine learning based models includes Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Stacking model, Boosting, and Blending models. Every of the methods has its strengths which makes it ideal for using with our breast cancer dataset.

2.3.1. Random Forest

Random Forest is one of the ensemble learning techniques that makes k number of decision trees while training phase and uses the class with the maximum occurrence for classification. When it comes to feature either numerical or categorical Random Forest may be preferred for our dataset because of the properties in how it deals with high dimensions and average over many trees to minimize over fitting. Random Forest algorithm has a very high tolerance to noise and naturally knows what to do when there are missing values which is an important factor when analyzing medical data since it often contains noise. It is also relevant to mention that with a large number of records Random Forests can reach serious computational problems [17]. In addition, they can only provide a skewed result for the given dataset when the number of benign and malignant cases differ greatly. A Random Forest Classifier is composed of a collection of classification trees h;

$$\{\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{T}, \boldsymbol{\varphi_k}), \quad \boldsymbol{k} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \boldsymbol{K}\}$$

Where $\boldsymbol{\varphi_k}$ represents identically and independently distributed random vectors, and each tree casts a unit vote for the most likely class at input $\boldsymbol{x}$

2.3.2. Support Vector Machine (SVM)

SVM is a very useful classification algorithm that is based upon identifying a hyper plane that will provide the greatest separation between classes. SVMs are ideally designed for use in high dimensional spaces and come in handy when there is ample separation between the classes [18]. There are probably non-linear relations between features in the included dataset, which is breast cancer dataset. SVMs, particularly ones using non-linear kernels such as the Radial Basis Function (RBF), are capable of identifying such relationships and generalizing well. SVMs are highly expensive and time consuming for large dataset and several iterations may be needed to optimize the kernel, and/ or the level of regularization. It is also less preferred when there are such situations as overlapping classes or much noise in the data.

2.3.3. K-Nearest Neighbors (KNN)

KNN is easy to understand and interprets, which tends to classify an object based on the outcomes of the nearest points. This is a non-parametric technique of testing which does not involve assuming any of the all or any of the values based on any definite pattern. KNN can be applied in medical data base where patients needing a similar treatment or those that have similar symptoms, would be treated similarly [19]. It is simple and efficient to use, which means it can be beneficial when starting with data analysis by using big data. Therefore, large data sets make the prediction time lengthy, due to the fact that KNN computes the distance of each data point during the prediction. Besides this, improving the value of k and choosing the most appropriate distance measure is an important criterion in KNN. It also does not possess well balanced data and can contain noise. Mathematically it is represented by

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2.3.4. Stacking

Stacking is a meta-model, learned by combining multiple predictions from base models like Random Forest, SVM, KNN for instance. This approach builds on the proposition and usage of capabilities from many models. Stacking can enhance the predictiveness of models if these models contain different kinds of adequacy. For instance, SVMs are more accurate in linear structure and Random Forest is better in handling interactions between features. At the same time, the combination of such models can result in the improvement of the total result. Stacking also needs CHAMPs and meta-models to be appropriately chosen and adjusted in their parameters. It can also be numerically costly and can lead to the issue of overfitting where the relationship between the training and testing dataset is overly learn if not well regularized or the base models are too related [20].

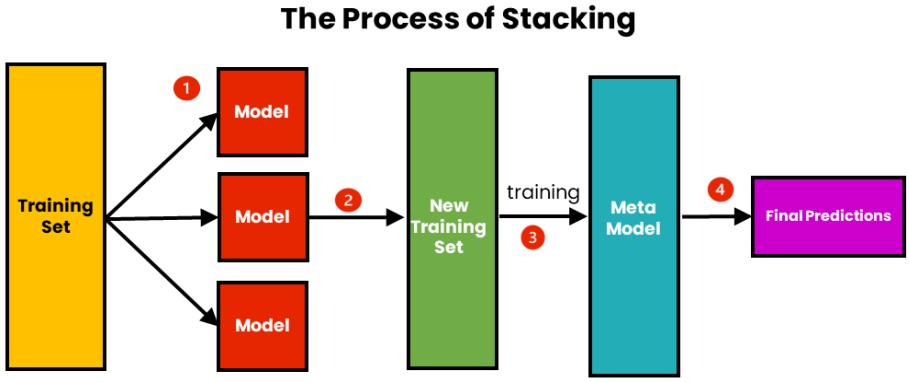

Figure 1. The Process of Stacking

2.3.5 Boosting

Boosting is an ensemble technique in which separate models are created one after the other, and each new model tries to rectify the mistakes made by its predecessor model. Some of the common operation boosting algorithms are Ada-boost and Gradient Boosting. Boosting is named after the intention of decreasing bias and increasing the accuracy of weak learners. Boosting, in general, is useful in that it helps alleviate both bias and variance, and therefore is useful on small dataset like breast cancer data where minor misclassification, for instance mistaking a malignant tumor for a benign one, can have fatal consequences. Boosting model sporting XGBoost Compared with Boosting, the errors made when diagnosing new cases are significant if the information used to train the model contains outliers or outliers are included in the set. However, it could be easily prone to overfitting if not regularized enough and training could take a long time if the dataset is large enough [21].

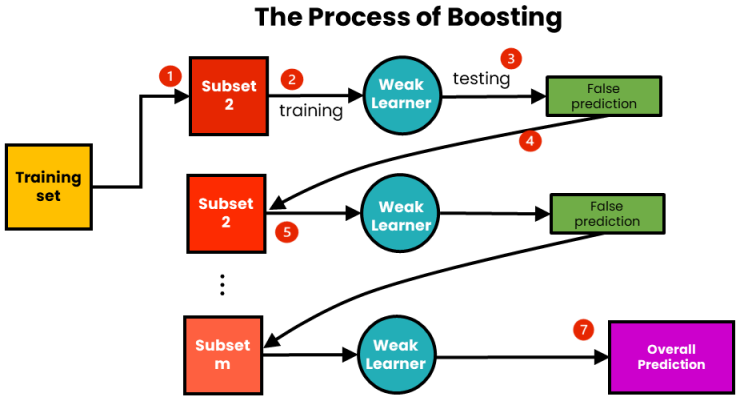

Figure 2. The Process of Boosting

2.3.6. Blending

Well, blending is very similar to stacking but the main difference is how the two splits the dataset. In blending, probabilistic forecasts from the base models are computed using a validation dataset and these probabilistic outcomes are subsequently fed to the meta-model. Compared to stacking, blending is easier

to apply and still yields better performance by averaging over different models' results. In cases involving datasets such as this one, blending empowers us to merge the forces of more than one models without overburdening the formula [22]. Blending usually is inferior to stacking since it utilizes less training data for the base models (as a result of the holdout set). Moreover, if the holdout set is not a random subset from the entire dataset then blending might introduced errors.

2.3.7. Models Fitness for the Dataset

These models were used as they all tackle issues relevant to medical data: the dependencies between features, the presence of noise and missing values. If we use a number of individual classifiers such as Random Forest, SVM, KNN, boosting, stacking, blending, then the improved results of the models can be obtained. I mean this variety also helps us determine the best model combinations that are likely to give the most accurate results time and again for breast cancer cases.

2.3.8. Limitations

Despite the strengths of these models, there are several limitations:

- **Imbalanced Data:** For instance, KNN and Random Forest are among those that perform poorly on imbalanced data, in which one class, say, benign tumors, contains the majority of the data samples.
- **Computational Cost:** Some of the methods are computationally expensive, so it may be extremely time consuming to work with large data sets like Boosting, Stacking and SVMs with non-linear kernels.
- **Overfitting:** Problems like Boosting and Stacking are also prone to overfitting training data if not well regulated and this is made possible by the complexity of most models or due to lack of regularization.
- **Interpretability:** While such models can give high levels of accuracy such as the Random Forest or Boosting, they are considered black box models due to their lack of interpretability which is very important in medical practice since practitioners need to understand why the model took a certain decision.

2.4. Evaluation Metrics

While applying machine learning models for breast cancer datasets, it is highly essential to measure different values in order to estimate and compare the performance of such methods. Every metric provides information on different aspects of model behavior, especially if we are working with health information, where false positives or false negatives can have catastrophic outcomes.

2.4.1. Accuracy

That is why, in medical datasets, accuracy can be very misleading, especially when we have a significantly more number of instances of some class than instances belonging to other class (like, xxx cases of benign tumors and only yyy cases of malignant tumors). Even if we achieve a high accuracy level, the number of screened and missed malignant cases will remain a problem. Hence, the accuracy, as the parameter of correctness, has to be used combined with other values.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

2.4.2. Precision

Accuracy is the fraction of positive predictions that were actually positive, in other words, true positives to total positives.

$$recall = \frac{TP}{TP + FN}$$

Accuracy defines the fitness of the model to correctly predict malignant tumor. Accuracy is very important in medical operations because people who do not have cancer should not be told they have the disease.

2.4.3. Sensitivity or True Positive Rate or Recall

Precision is the measure of the true positive rates to the current positive predictions of the model (true positive and false positive).

$$precision = \frac{TP}{TP + FP}$$

Recall in medical diagnostics is important because it quantifies the specific model's capability of making correct diagnoses about malignant tumors. A low recall necessarily means that the model is omitting malignant cases in the population, which could potentially worsen the patients' condition. In breast cancer detection the patients' recall is high in the servicing centers since omission of the breast cancer is an irreversible mistake.

2.4.4. F1 Score

The F1 score in other words is the reciprocal of the average of the static precision and the actual recall. F-measure is again a merging of precision and recall, which is particularly important in a medical situation where false negative results, or false positive ones are acceptable. The Coefficient of error is especially useful when dealing with unbalanced datasets because it offers one measure of the error that takes into account both kinds of mistakes.

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

2.4.5. Cohen's Kappa

Kappa coefficients of Cohen is the coefficients that estimate the pattern of probabilities when the set of predicted classifications and actual classifications matching with each other. Cohen's Kappa offers an improved estimation of the system's performance compared to accuracy, particularly when confronted with imbalanced data. This is preferred to the earlier method because it gives a clear picture of how much better the model is a performing compared to hope and chance.

2.4.6. The Matthews correlation coefficient (MCC) is defined as follows:

MCC is a correlation coefficient for two variables whose result take all of TP, TN, FP, and FN into consideration. This is considered as a balanced measure and is acceptable even if data is skewed towards one class. Specifically, MCC raises its capabilities when applied to breast cancer datasets which could include many more benign cases than malignant ones. It also offers a fair assessment of the model predictions in case one class outweighs all other classes within the dataset.

2.4.7. Realization of ROC Curve and AUC (Area Under the Curve)

ROC (Receiver Operating Characteristic) which depicts true positive rate (recall) against false positive rate (1 – specificity). The AUC – Area Under the Curve is a measure of the performance of the model across all classification thresholds. An area under the ROC curve from 0 onwards, where a closer to 1 suggests better model accuracy. ROC curve and AUC gives a graphical representation with respect to the ability of the model to completely segregate classes of objects (benign and malignant). A high AUC means that the model is effective in segregating between the two classes at different threshold, which is desirable in medical diagnosis, where it is often important to deferentially adjust between sensitivity and specificity.

2.4.8. PR Curve and AUC

This curve defines precision against recall. The AUC (Area Under the Curve) is calculated as the space under this curve, which shows the interaction of recall with exactness. This curve comes handy especially when dealing with imbalanced classes, as it is the case with this dataset where malignant cases are far much fewer than benign ones. This curve shows how the model is doing in the positive cases (malignant tumors) and the PR AUC gives a single measure of how good this is.

3. Results and Discussion

Considering evaluation metric results derived from Random Forest, SVM, KNN, Stacking, Boosting, and Blending models on breast cancer dataset the insights into the performance metric of the models is as follows: It is found out that each classifier shows varying characteristics because of the metrics calculated during classification. The following is a♎ comparison made based on the methodology of the two studies.

Table 2. Performance Metrics of the Selected Models

Model	Accuracy	Precision	Recall	F1 Score	Cohen's Kappa	MCC	ROC AUC	PR AUC
Random Forest	0.913	0.8289	0.525	0.6429	0.5962	0.6164	0.8836	0.727
SVM	0.8994	0.8197	0.4167	0.5525	0.5025	0.5391	0.9007	0.7063
KNN	0.8845	0.7455	0.3417	0.4686	0.4136	0.4535	0.8011	0.5478
Stacking	0.8845	0.7455	0.3417	0.4686	0.4136	0.4535	0.8011	0.5478
Boosting	0.9143	0.8493	0.5167	0.6425	0.597	0.6209	0.9051	0.7402
Blending	0.913	0.8472	0.5083	0.6354	0.5895	0.6144	0.9061	0.7379

3.1. Accuracy

The accuracy results are moderate overall although random forest, boosting and blending models leads the lot with nearly 91% accuracy. The high AC means that these models are able to correctly classify if the tumor is Benign or Malignant. As seen in the methodology section, accuracy when dealing with imbalanced data can be misleading since the model may score high accuracy due to over predicting the majority class (benign tumors). This means that to evaluate the models, it is necessary to consider precision, recall, and other related indicators.

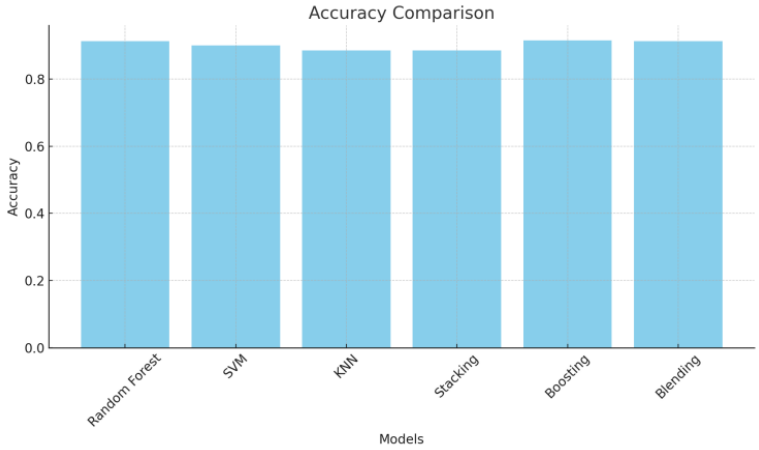

Figure 3. Comparison Graph of Accuracy

3.2. Precision

Accuracy determines the capacity of the model in correctly identifying the malignant tumors amongst the cases that have been correctly identified to be positive. Hence, models with the highest precision include Boosting and Blending; these models are the most careful and produce the least number of false positives. Random Forest and SVM exhibit equally good performance, while KNN and Stacking Register a lower precision suggesting that, a percent higher numbers of benign tumors were classified as malignant.

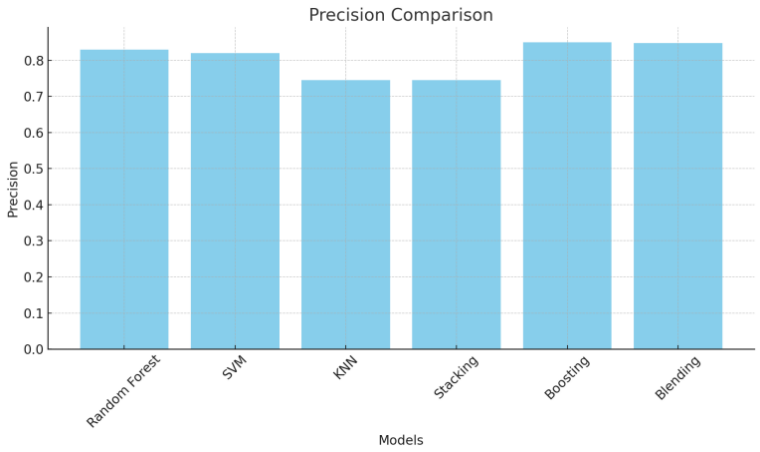

Figure 4. Comparison Graph of Precision

3.3. Recall

A good performance of a model in predicting actual malignant cases of the disease is what is being referred to as recall. Among the nine models Random Forest, Boosting, and Blending yield the highest recall. High recall is more important in medical diagnostics since false negatives, cases claiming to be negative while they actually contain malignant cases, are less desirable. Specifically, recall results of SVM and KNN are much lower than those of the other models, which indicates that two models missed many true positive cases of malignant tumor detection. This could be so because the corruption sensitivity of SVM has been well documented as well as the robustness of KNN to imbalanced datasets as explained in the methodology.

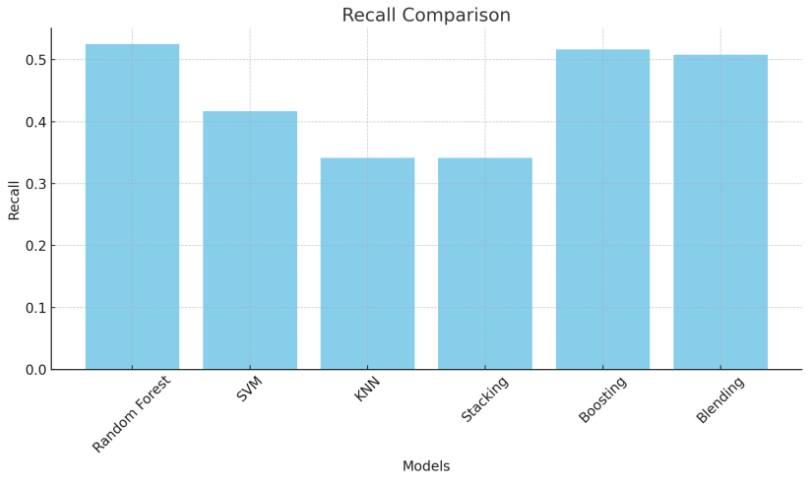

Figure 5. Comparison Graph of Recall

3.4. F1 Score

The F1 score is the accuracy that is calculated as a harmonic mean of precision and recall values. Boosting and blending have higher F1 scores meaning that these models have the best trade-off between recall, which is the ability to identify malignant tumors, and precision, which is the ability to do so with as few false positives as possible. The recall measure for the SVM and KNN model is slightly low, and has impacted the F1 score indicating the model tendency to miss quite a few number of malignant cases.

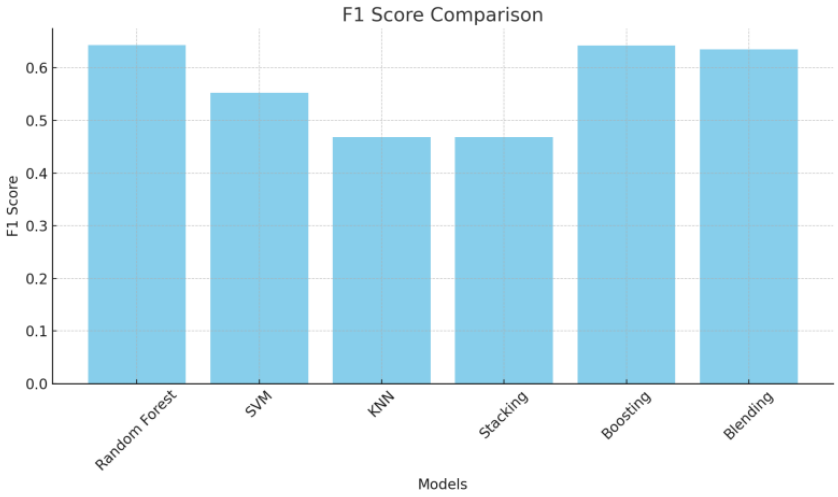

Figure 6. Comparison Graph of F1 Score

3.5. Cohen's Kappa

For classifying the results, Cohen's Kappa is used which determines how much the two sets of classifications, one predicted and the other actual, have coincided taking in account of chance coincidence also. From the analysis of the results in Table 3, high Kappa values indicate that Boosting and Random Forest are most accurate than the other models followed by Logistic Regression, whereas, SVM and KNN are less accurate having lowest Kappa values. When compared to the other models, the Kappa scores for both KNN and Stacking show a poor model reliability in classifying for malignant mainly due to the small number of attributes in the data set. These lower Kappa scores suggest that these models are closer to a coin toss for some of these classifications, particularly for malignant tumors.

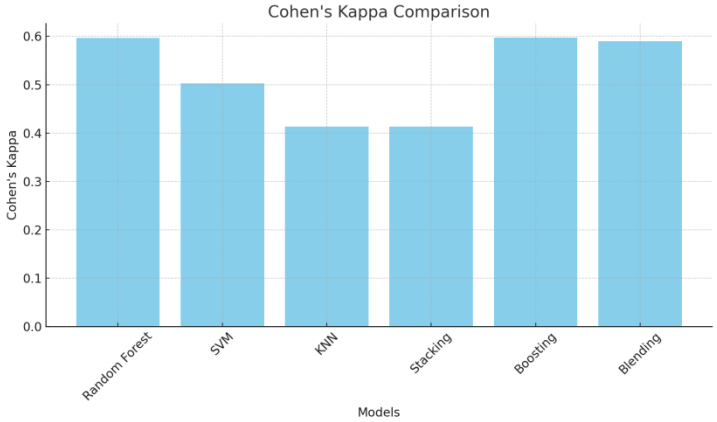

Figure 7. Comparison Graph of Cohens Kappa

3.6. Matthews Correlation Coefficient (MCC)

MCC is a centered measure that separates True Positives, True Negatives, False Positives and False Negatives. As with all other measurements, there is a parity where both Boosting measurements and Random Forest measurements yield the highest results, meanwhile, Blending also gives rather mediocre but acceptable results when compared to the best performers in the list. Out of all the models, Stacking and KNN have the lowest MCC which means these models are not doing well in terms of both false positive and false negative. The relatively the higher MCC for Random Forest, Boosting, and Blending indicate that these models are capable of classifying both the benign and malignant case scenarios with better accuracy and both types of errors are minimum.

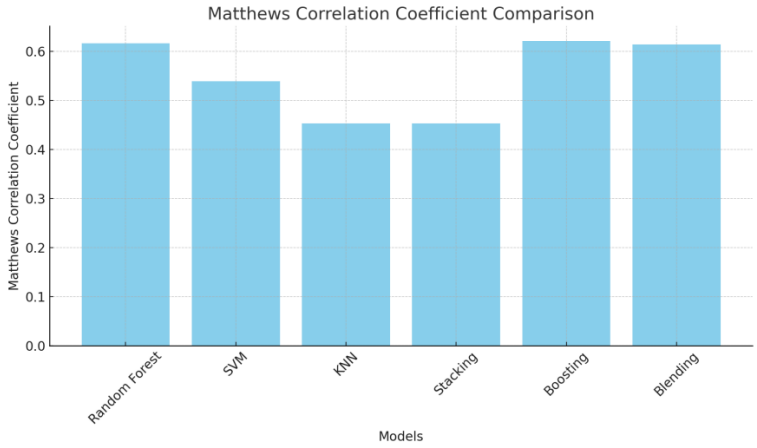

Figure 8. Comparison Graph of MCC

3.7. ROC AUC and Precision-Recall AUC

ROC AUC means the power of the model to classify instances as benign or malignant using all the classification margins. Ranking this in descending order of ROC AUC, we see that SVM, followed by Boosting and Blending, are the most accurate in distinguishing between the two classes of analysis. Random Forest also gives a pretty good score, but nothing like what SVM and Blending offered. KNN and Stacking have the lowest AUC meaning they perform worst in terms of discrimination while predicting new cases as benign or malignant over different thresholds.

The PRC-AUC is critical when the datasets are imbalanced since precision and recall dominate over accuracy. The four criteria we calculated demonstrate that Boosting, Blending, and Random Forest models have fairly high PR AUC and thus accurately contribute to predicting malignant cases with a relatively low number of false positives. KNN and Stacking have the lowest scores of Precision-Recall AUC, which again indicates that these methods are not suitable for this type of problem due to poor handling of the unbalanced dataset and, therefore, poor result accuracy for malignant tumors.

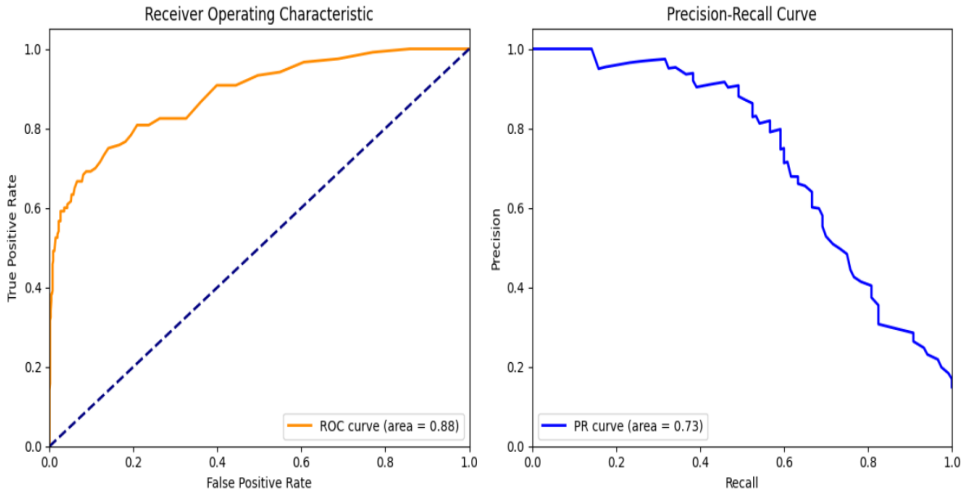

Figure 9. ROC-Curve and PR-Curve for Random Forest Tree

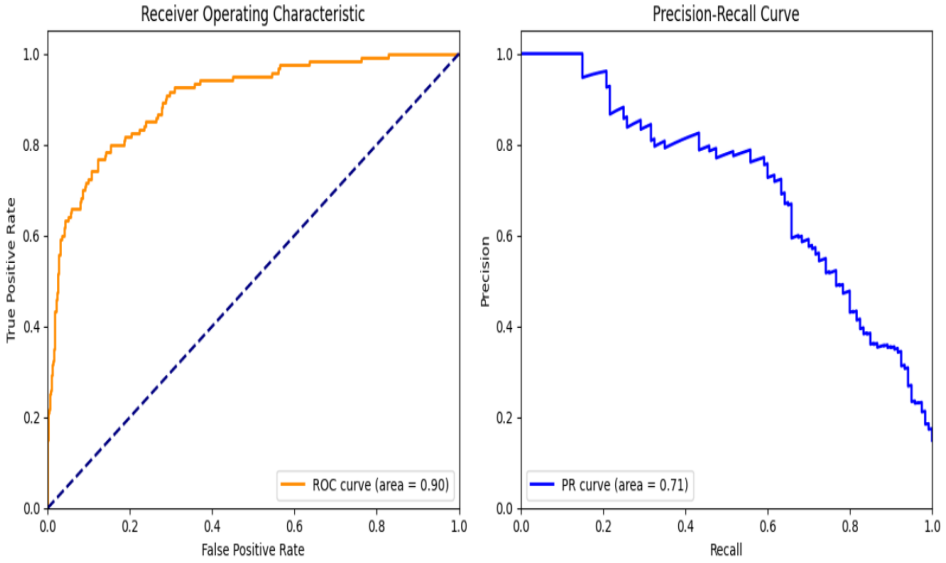

Figure 10. ROC-Curve and PR-Curve for SVM

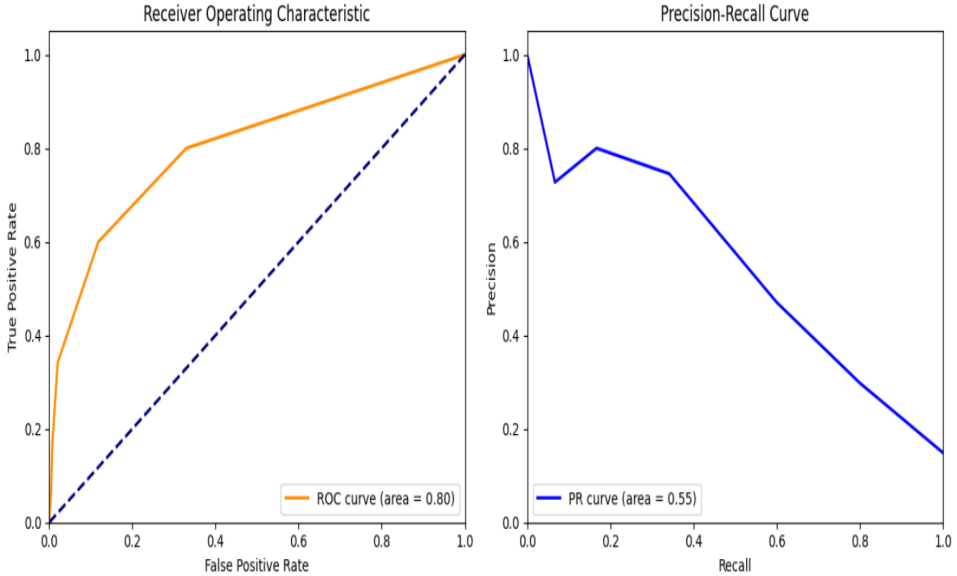

Figure 11. ROC-Curve and PR-Curve for KNN

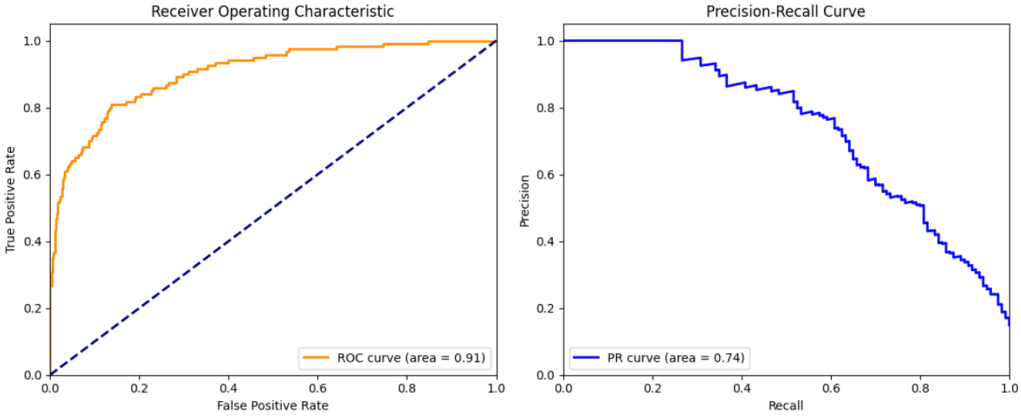

Figure 12. ROC-Curve and PR-Curve for Stacking

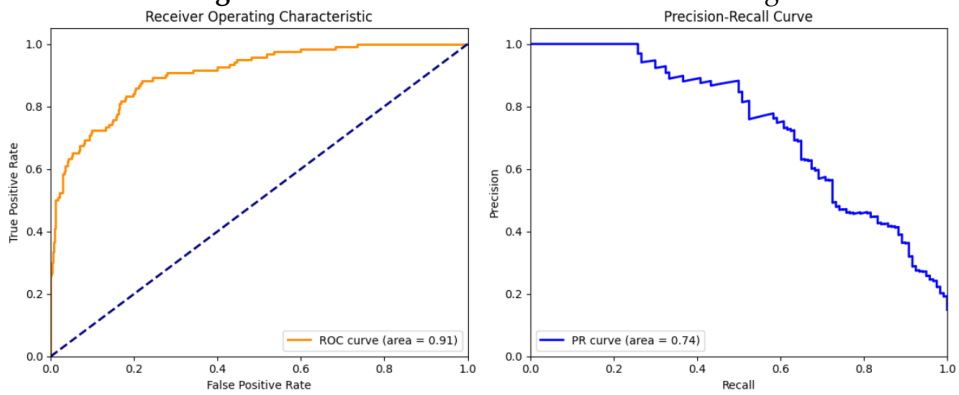

Figure 13. ROC-Curve and PR-Curve for Boosting

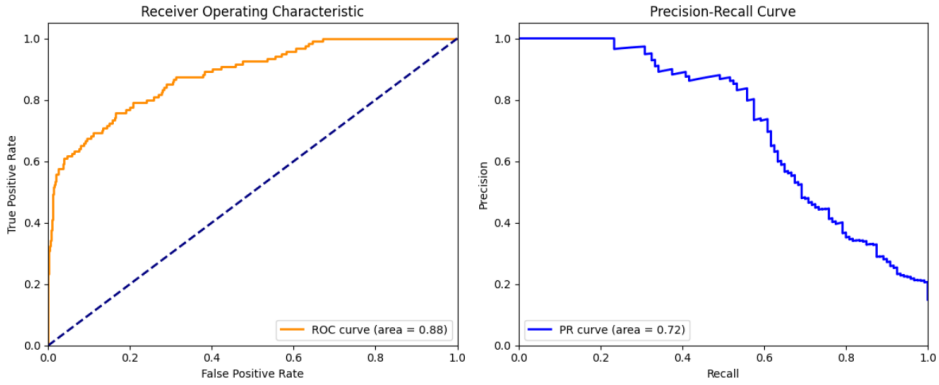

Figure 14. ROC-Curve and PR-Curve for Blending

3.8. Overall Insights

Random Forest has again and again shown better performance across all the metrics without any exception for Boosting and blending also show good results almost in all the values. These models provide a good F-score, which means that we lose less information when reconstructing medical datasets, and do not include too much noise in the reconstruction. Hence, Boosting appears to provide the most comprehensive improvement the lowest amount of bias and variance in most cases. Although SVM provides the maximum ROC AUC value, this classifier does not possess a high recall and, hence, is less appropriate for detecting malignant cases. It is good in detection of class as is evident from the size of squares around squares, but it misses many actual positive cases which is very important especially in medical diagnosis. Decision Tree, Random Forest, and Gradient Boosting and stacking, these models are the worst performing models in virtually each of the metrics considered here. This is evident in their precision scores where they seem to have poor capability in recognizing malignant cases, therefore, they are less suitable for this given dataset. They also fail to manage imbalanced data well enough, as is seen with the lower MCC and the Kappa value.

The models that are suitable to solve this breast cancer dataset include Random Forest, Boosting, and Blending models because such models excel when solving medical datasets with imbalanced classifications between benign and malignant cases. The evaluation metrics reveal that both precision and recall have to

be kept in check on such datasets, and since Boosting comes under the paradigm of Bagging, it works amazingly due to the sequential learning process and the errors that are likely to occur are rectified during the further phases of learning. Similarly though SVM excels in class differentiation it has low recall, clearly in medical diagnosis missing positive examples are not option. Similarly, KNN and Stacking are unfit in this dataset since they performed poorly based on fundamental measures like recall and the F1 score and MCC.

4. Conclusions

In this study, for breast cancer classification we evaluated various models of machine learning, focusing on the ensemble techniques. The best performance is Boosting and Random forest across more metrics, excelling that particularly in precision and recall, for real-time applications of diagnostics making them suitable. Our findings that demonstrate the potential of models of machine learning to aid early in a detection and diagnosis, thus improving the patient outcomes. However these models, need the further validation in clinical settings of real-world to ensure the generalization across in diverse populations. Research in future should explore the integration of that additional features, such as data of genomic, and consider the interpretability of model to facilitate better understanding among practitioners in medical.

Data Availability Statement: The dataset that is used in this study, publicly available on the Kaggle and the data in processed can be shared upon the reasonable request ahead to the corresponding author.
Acknowledgement: Thanks to the institutions by authors that provided the infrastructure for carrying out to this research. And Special thanks to our technical team for the supporting that data processing tasks and also to the professionals of healthcare who provided the valuable insights into the applications of real-time of these models.
Conflict of Interest: No conflict of interest declare by authors regarding the publication of this research paper.

References

1. Cao, W., Qin, K., Li, F., & Chen, W. (2024). Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (GLOBOCAN). Journal of the National Cancer Center.
2. Wang L. Early Diagnosis of Breast Cancer. Sensors (Basel). 2017 Jul 5;17(7):E1572.
3. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
4. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet. 1998 Mar;62(3):676–89.
5. Chunduru, A., Kishore, A.R., Sasapu, B.K. and Seepana, K., 2024. Multi chronic disease prediction system using CNN and random forest. SN Computer Science, 5(1), p.157..
6. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience [Internet]. 2017 [cited 2022 Apr 14];9. Available from: https://www.frontiersin.org/article/10.3389/fnagi.2017.00329
7. Nabavi Monfared, S.S. and Yousefi, M.R., 2024. Improving the Accuracy of Early Detection of Parkinson's Disease Using Brain Signals Based on Feature Selection in Machine Learning. Computational Intelligence in Electrical Engineering..
8. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making. 2011 Jul 29;11(1):51.
9. Al-Jamimi, H.A., 2024. Synergistic Feature Engineering and Ensemble Learning for Early Chronic Disease Prediction. IEEE Access..
10. Anji Reddy V, Soni B. Breast Cancer Identification and Diagnosis Techniques. In: Rout JK, Rout M, Das H, editors. Machine Learning for Intelligent Decision Science [Internet]. Singapore: Springer; 2020 [cited 2022 Apr 14]. p. 49–70. (Algorithms for Intelligent Systems). Available from:
11. Vaka AR, Soni B, K. SR. Breast cancer detection by leveraging Machine Learning. ICT Express. 2020 Dec 1;6(4):320–4
12. Naji MA, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. Procedia Computer Science. 2021 Jan 1;191:487–92.
13. Priyanka KS. A Review Paper on Breast Cancer Detection Using Deep Learning. IOP Conf Ser: Mater Sci Eng. 2021 Jan;1022(1):012071.
14. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput & Applic. 2013 Dec 1;23(7):2387– 403.
15. Khan, M. F., Iftikhar, A., Anwar, H., & Ramay, S. A. (2024). Brain Tumor Segmentation and Classification using Optimized Deep Learning. Journal of Computing & Biomedical Informatics, 7(01), 632-640.
16. Desai M, Shah M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). Clinical eHealth. 2021 Jan 1;4:1–11.
17. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694–701
18. Zhang, C., Liu, Y. and Tie, N., 2023. Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron. Forests, 14(2), p.254.
19. Ke, T., Ge, X., Yin, F., Zhang, L., Zheng, Y., Zhang, C., Li, J., Wang, B. and Wang, W., 2024. A general maximal margin hyper-sphere SVM for multi-class classification. Expert Systems with Applications, 237, p.121647.
20. Prasad, B.S., Gupta, S., Borah, N., Dineshkumar, R., Lautre, H.K. and Mouleswararao, B., 2023. Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach. Preventive Medicine, 174, p.107619.
21. Suha, S.A. and Islam, M.N., 2023. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. Heliyon, 9(3).
22. Bukhari, O., Agarwal, P., Koundal, D. and Zafar, S., 2023. Anomaly detection using ensemble techniques for boosting the security of intrusion detection system. Procedia Computer Science, 218, pp.1003-1013.
23. Hasan, M., Abedin, M.Z., Hajek, P., Coussement, K., Sultan, M.N. and Lucey, B., 2024. A blending ensemble learning model for crude oil price forecasting. Annals of Operations Research, pp.1-31.