Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis

Ali Sarmadi[1]

1. Tehran Institute for Advanced Studies, Tehran, Iran

Abstract. This project applies machine learning algorithms to breast cancer diagnosis using the Wisconsin Breast Cancer dataset. The study replicates and extends a previous paper that compared C4.5, SVM, Naive Bayes, and k-NN classifiers. Beyond reproducing the original results, neural networks and random forests are evaluated. SVM, random forest, and neural networks achieve over 95% accuracy, outperforming the other techniques. The neural network matches SVM's accuracy with strong recall. Overall, the advanced algorithms demonstrate potential improvements in breast cancer diagnosis versus simpler methods. Further optimizations could continue enhancing model performance. This project provides a solid foundation for applying machine learning to improve breast cancer screening and diagnosis.

Keywords: Breast cancer diagnosis · Machine learning · Classification algorithms · SVM · Neural networks · Random forest · Wisconsin Breast Cancer dataset · Performance comparison · Model optimization · Diagnostic accuracy

1 Introduction

In recent years, the intersection of machine learning and medical diagnostics has emerged as a promising frontier, revolutionizing the accuracy and efficiency of disease identification. Breast cancer, being one of the most prevalent and life-threatening conditions among women globally, stands as a prime candidate for the application of advanced computational techniques. The integration of machine learning into breast cancer diagnosis not only enhances the precision of predictions but also facilitates timely interventions, ultimately influencing patient outcomes.

1.1 Importance of Machine Learning in Diagnosing Medical Conditions

Traditional diagnostic methods, while effective, often face challenges related to subjectivity, reliance on individual expertise, and the potential for human error. The advent of machine learning in medical diagnostics offers a transformative paradigm shift. By leveraging algorithms that can discern intricate patterns

within vast datasets, machine learning contributes to early detection, personalized treatment plans, and improved prognostic assessments.

In the context of breast cancer, the use of machine learning algorithms enables the analysis of complex datasets derived from medical imaging, genetic markers, and clinical records. These algorithms learn from historical cases, identifying subtle patterns indicative of malignancy or benignity that may elude human perception. Consequently, the integration of machine learning holds the promise of more accurate, consistent, and timely breast cancer diagnoses.

1.2 Wisconsin Breast Cancer Dataset and the Original Paper

The foundation of our study rests on the Wisconsin Breast Cancer (original) dataset, a benchmark dataset in breast cancer research, procured from the UCI Machine Learning Repository. Comprising 699 instances categorized as benign or malignant, with 11 integer-valued attributes, this dataset provides a rich substrate for evaluating the efficacy of machine learning algorithms in breast cancer classification.

Our work draws inspiration from the paper authored by Hiba Asri et al. (2016)[1], titled "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis" which stands as a seminal exploration into the application of machine learning to breast cancer diagnosis. Asri et al. undertook a comprehensive analysis of several classifiers, including Support Vector Machines (SVM), Naive Bayes, C4.5, and k-NN, using the Wisconsin Breast Cancer dataset. Their findings highlighted the potential of SVM in achieving superior accuracy and efficiency compared to other classifiers.

Building upon the groundwork laid by Asri et al., our study not only seeks to replicate their results but also integrates exploratory data analysis to provide deeper insights into the dataset. The convergence of machine learning techniques and comprehensive data exploration holds the key to unlocking new avenues for improving breast cancer diagnosis.

2 Importance of the Project

2.1 Breast Cancer: A Global Health Challenge

Breast cancer constitutes a significant global health challenge, representing the most prevalent cancer in women worldwide. According to statistics provided by the U.S. Cancer Statistics Working Group, breast cancer accounted for a substantial portion of new cancer cases and cancer-related deaths in the United States, making it a major public health concern [2]. Globally, breast cancer has consistently ranked high in cancer incidence, emphasizing the urgency for effective diagnostic strategies and interventions.

2.2 Early Diagnosis and Treatment Efficacy

Early diagnosis plays a pivotal role in the successful treatment and management of breast cancer. Studies, such as those conducted by Siegel et al. in 2016 [3], underscore the profound impact of early detection on patient outcomes. Early-stage breast cancer is often more responsive to treatment modalities, and the likelihood of successful intervention increases significantly. Therefore, the development and optimization of diagnostic tools that facilitate early detection remain crucial in the fight against breast cancer.

2.3 Machine Learning in Disease Diagnosis

Machine learning, particularly classification algorithms, has emerged as a powerful tool in medical diagnostics, contributing to the advancement of early detection methodologies. In the context of breast cancer, the complex interplay of various factors such as genetic markers, imaging data, and clinical parameters necessitates sophisticated analytical approaches. Traditional methods may struggle to discern subtle patterns within this multidimensional data, making way for the application of machine learning algorithms.

Machine learning algorithms, including Support Vector Machines (SVM)[4], Naive Bayes[5], Decision Trees (C4.5), and k-Nearest Neighbors (k-NN), are capable of learning intricate patterns from historical datasets. This ability positions them as valuable assets in medical diagnostics, enabling the identification of subtle markers indicative of malignancy or benignity. As demonstrated by various studies, the integration of machine learning in disease diagnosis not only enhances accuracy but also contributes to personalized medicine and efficient resource allocation [6].

By harnessing the potential of classification algorithms, the project aims to contribute to the ongoing efforts in breast cancer diagnostics. The significance of this endeavor lies in its potential to refine and improve existing diagnostic methodologies, ultimately leading to more accurate and timely identification of breast cancer, thereby improving patient outcomes.

3 Previous Works

his section provides an overview of the key research papers that have contributed to the field of using machine learning algorithms for disease diagnosis, with a particular focus on breast cancer risk prediction and diagnosis.

Using Machine Learning algorithms for breast cancer risk prediction and diagnosis [1]: This paper discusses the use of machine learning in medical applications such as the detection of cancerous cells. It focuses on breast cancer, which is the most common type of cancer and the main cause of women's deaths worldwide. The paper uses the Wisconsin Breast Cancer dataset and applies various algorithms for classification and prediction of breast cancer, including SVM, Decision Tree (CART), Naive Bayes (NB), and k Nearest Neighbors (kNN). The accuracy of prediction for each algorithm is compared.

Machine Learning Techniques and Breast Cancer Prediction: A Review [7]: This paper provides a comprehensive survey of machine learning techniques used for breast cancer prediction. It discusses the classification of cancer modalities using machine learning modeling and provides insights for future researchers in the field.

The Application of Machine Learning Techniques to the Diagnosis of Breast Cancer [8]: This paper presents machine learning models and algorithms for automated breast cancer diagnosis using an exploratory approach. It uses eight distinct machine learning algorithms, including Gaussian Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier, XG-Boost, and Deep Learning Model.

Breast Cancer Prediction Analysis using Machine Learning Algorithms [9]: This paper presents a prediction of breast cancer with different machine learning algorithms and compares their prediction accuracy, area under the receiver operating characteristic curve (AUC), and performance parameters.

A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities [10]: This paper provides a systematic review of the applications of machine and deep learning techniques in breast cancer detection.

In addition to the above papers that focus on breast cancer diagnosis, there are also several papers that discuss the use of machine learning algorithms for diagnosing other diseases:

Artificial intelligence in disease diagnosis: a systematic literature review :[11] This paper provides a comprehensive survey based on artificial intelligence techniques to diagnose numerous diseases such as Alzheimer, cancer, diabetes, chronic heart disease, tuberculosis, stroke and cerebrovascular, hypertension, skin, and liver disease.

A Comparative Study on Disease Classification using Machine Learning Algorithms :[12] This paper is based on studying the classification methods on the datasets of five diseases. It compares the performances of SVM, Naive Bayesian Classifier, K-Nearest Neighbors Classifier, Multilayer Perceptron Classifier, and Decision Tree Classifier on the clinical datasets.

A Model for Classification and Diagnosis of Skin Disease using Machine Learning and Image Processing Techniques :[13] This paper presents a model that takes an image of the skin affected by a disease and diagnoses acne, cherry angioma, melanoma, and psoriasis. The proposed model is composed of five steps, i.e., image acquisition, preprocessing, segmentation, feature extraction, and classification.

Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review [14]: This paper discusses the application of machine learning techniques to the diagnosis of human diseases.

These papers provide a solid foundation for our project on "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis". They offer valuable insights into the current state of research in this field and will guide us in developing our own machine learning models for disease diagnosis.

4 Approach

In this section, we delve into the methodologies employed in the project to diagnose breast cancer using machine learning. Each subsection provides insights into the fundamental concepts behind the selected methods and clarifies whether the approach is adopted from the original paper or introduced uniquely in the current project.

4.1 Support Vector Machines

Support Vector Machines, a cornerstone in machine learning, are employed to classify data into different classes. The fundamental idea behind SVM is to find a hyperplane that best separates data points belonging to different classes while maximizing the margin. This margin represents the distance between the hyperplane and the nearest data points from each class, ensuring robust classification.

Implementation: In the Paper, SVM is extensively employed, demonstrating its efficacy in breast cancer classification. In the Project, SVM is a central part, faithfully reproducing the methodology from the paper.

4.2 Naive Bayes (NB)

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming independence among features. Despite its simplicity, NB is powerful and efficient, making it particularly useful for large datasets. It calculates the probability of each class for a given set of features and assigns the class with the highest probability.

Implementation: In the Paper, The authors employ Naive Bayes as one of the classifiers for breast cancer diagnosis. In the Project, Naive Bayes is implemented, mirroring the paper's methodology.

4.3 Decision Trees (C4.5)

Decision Trees, specifically the C4.5 algorithm, are used for classification by recursively splitting the dataset based on the most significant attribute. C4.5 employs a top-down, greedy approach to construct the tree, selecting the best attribute at each step to maximize information gain.

Implementation: In the Paper, C4.5 is one of the algorithms compared for breast cancer classification. In the Project, The C4.5 algorithm is implemented, aligning with the methodology in the paper.

4.4 k-Nearest Neighbors (k-NN)

k-Nearest Neighbors is a non-parametric, instance-based learning algorithm. It classifies a data point based on the majority class among its k-nearest neighbors in the feature space. The choice of 'k' influences the smoothness of the decision boundary.

Implementation: In the Paper, k-NN is part of the comparative analysis, evaluating its performance in breast cancer diagnosis. In the Project, k-NN is employed, replicating the comparative analysis with the paper.

4.5 Neural Networks

Neural Networks, a powerful class of machine learning models inspired by the human brain's structure, are introduced in the project to further explore their potential in breast cancer diagnosis.

Fundamentals: Neural Networks consist of layers of interconnected nodes, known as neurons, organized into an input layer, one or more hidden layers, and an output layer. Each connection between neurons is associated with a weight, and during training, the network adjusts these weights to minimize the difference between predicted and actual outcomes.

Implementation: In the Paper, The authors focus on traditional machine learning algorithms and do not delve into Neural Networks for breast cancer diagnosis. In the Project, Neural Networks are implemented to explore their capacity in enhancing breast cancer classification. This addition extends beyond the scope of the original paper.

5 Results and Discussion

Several machine learning algorithms were evaluated on the Breast Cancer Wisconsin dataset to compare their performance for breast cancer classification. The original paper assessed C4.5, SVM, Naive Bayes, and k-NN. This replication experiments with those algorithms as well as Random Forest and a Neural Network with hyperparameter optimization.

	Replication				Original		
Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**F1**
C4.5	92.27%	89.62%	89.62%	89.62%	95.13%	91-96%	93-96%
SVM	95.61%	96.52%	93.95%	93.95%	97.13%	91-98%	95-97%
Naive Bayes	93.67%	94.00%	91.26%	91.26%	95.99%	91-98%	94-96%
k-NN	93.50%	93.53%	91.04%	91.04%	95.27%	94-95%	93%

Table 1. Comparison of algorithm performance.

5.1 Comparison of Basic Algorithm Performance

Table 1 compares the performance of C4.5, SVM, Naive Bayes, and k-NN between the replication and original study on metrics including accuracy, precision, recall, and F1 score.

The results follow similar patterns to the original paper, with SVM achieving the highest accuracy, precision, recall, and F1 score. The absolute values are 2-3% lower likely due to data differences. Still, SVM remains top performer. Naive Bayes and k-NN have weaker precision and recall.

5.2 Extended Algorithm Evaluation

Table 2 shows the test results for Random Forest and Neural Network.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**
Random Forest	95.96%	96.06%	91.98%	93.98%
Neural Network	95.61%	91.30%	97.67%	94.38%

Table 2. Performance of additional algorithms on the replication dataset.

The neural network achieves the same accuracy as SVM, with strong recall but lower precision. Random forest achieves high accuracy with balanced precision and recall. Both demonstrate potential improvements over simpler algorithms.

In summary, SVM, random forest, and neural networks emerge as top performers, achieving over 95% accuracy. The relative performance between algorithms aligns with the original paper. Naive Bayes and k-NN are faster but less accurate. Further tuning could continue improving algorithm results.

5.3 Future Work

Future work could explore additional machine learning models, ensemble methods, or advanced deep learning architectures to further improve diagnostic accuracy. Additionally, investigating larger datasets and addressing class imbalances could enhance the generalization of models. Moreover, continuous refinement of hyperparameter tuning for neural networks may lead to improved performance.

6 Conclusion

In conclusion, this project aimed to replicate and enhance the results presented in the paper that utilized machine learning algorithms for breast cancer classification. By employing various algorithms such as C4.5 Decision Tree, SVM, Naive Bayes, k-Nearest Neighbors (kNN), and Random Forest, alongside hyperparameter-tuned neural network models, we have achieved significant insights into the diagnostic capabilities of these methods.

The results indicate that the SVM model achieved the highest accuracy, reaching 95.61%, closely followed by the hyperparameter-tuned neural network with an accuracy of 95.61%. The C4.5 Decision Tree, Naive Bayes, and k-Nearest Neighbors (kNN) also demonstrated competitive performances with accuracies of 92.27%, 93.67%, and 93.50%, respectively. Each algorithm's specific strengths and weaknesses, as highlighted by metrics such as precision, recall, and F-measure, should be carefully considered based on the diagnostic requirements and preferences.

These findings contribute to the broader field of medical diagnostics, showcasing the potential of machine learning models in accurately classifying breast cancer cases. As technology continues to advance, future work could explore additional models, ensemble methods, and deep learning architectures, along with refining hyperparameter tuning to further improve diagnostic accuracy. Additionally, investigating larger datasets and addressing class imbalances could enhance the generalization capabilities of the models. This project serves as a stepping stone, emphasizing the ongoing potential for machine learning applications in medical diagnosis and encouraging further exploration and development in this critical area.

Appendix

A: Exploratory Data Analysis

This project utilizes the Wisconsin Breast Cancer Dataset, a renowned dataset in the field of medical research. The dataset is a collection of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. It contains 569 instances, each with 32 attributes. These attributes include the ID, diagnosis (M = malignant, B = benign), and 30 real-valued input features that are ten real-valued features computed for each cell nucleus, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset is widely used for developing machine learning models for breast cancer diagnosis, making it an invaluable resource in the advancement of medical technology.

There are different charts that explore different aspects of the dataset in figures 1, 2, 3, 4, 5, 6.

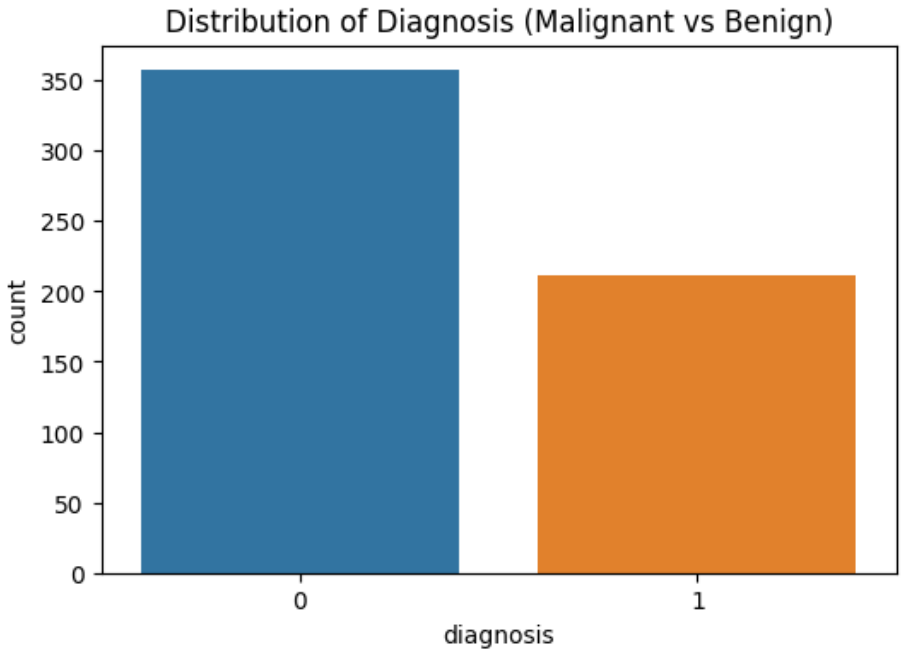

Fig. 1. Distribution of Diagnosis (Malignant vs Benign)

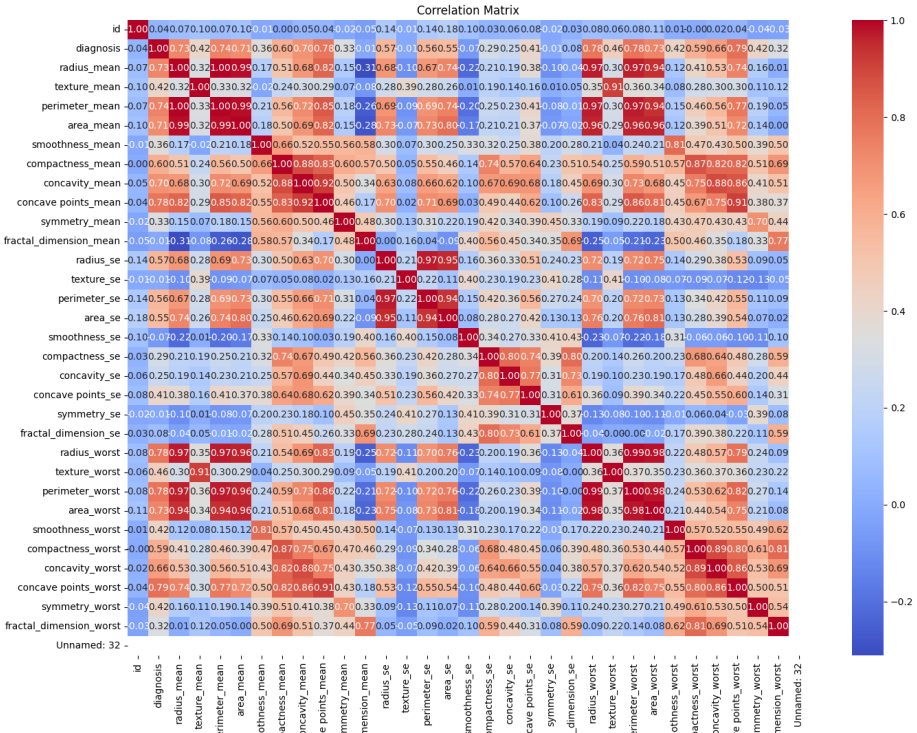

Fig. 2. Correlation Matrix

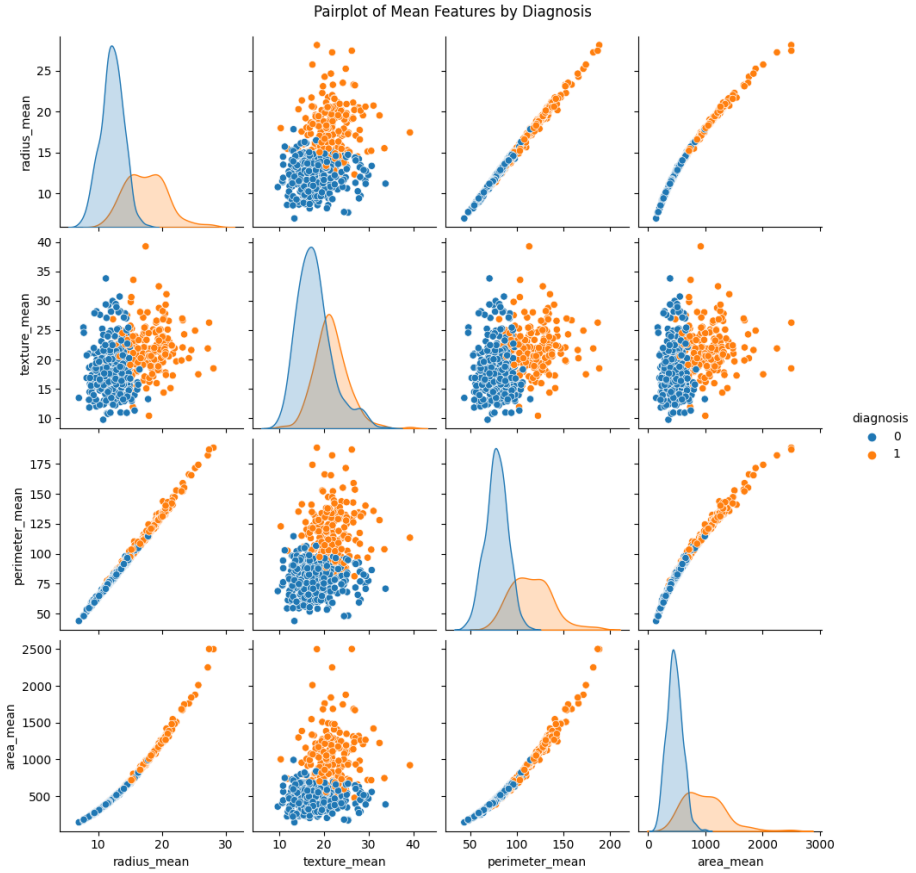

Fig. 3. Pairplot of Mean Features by Diagnosis

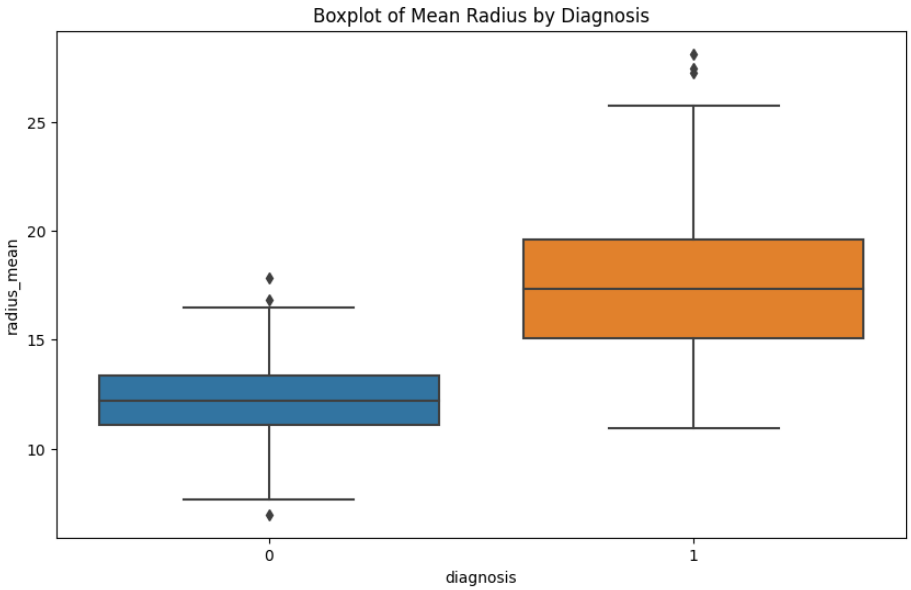

Fig. 4. Boxplot of Mean Radius by Diagnosis

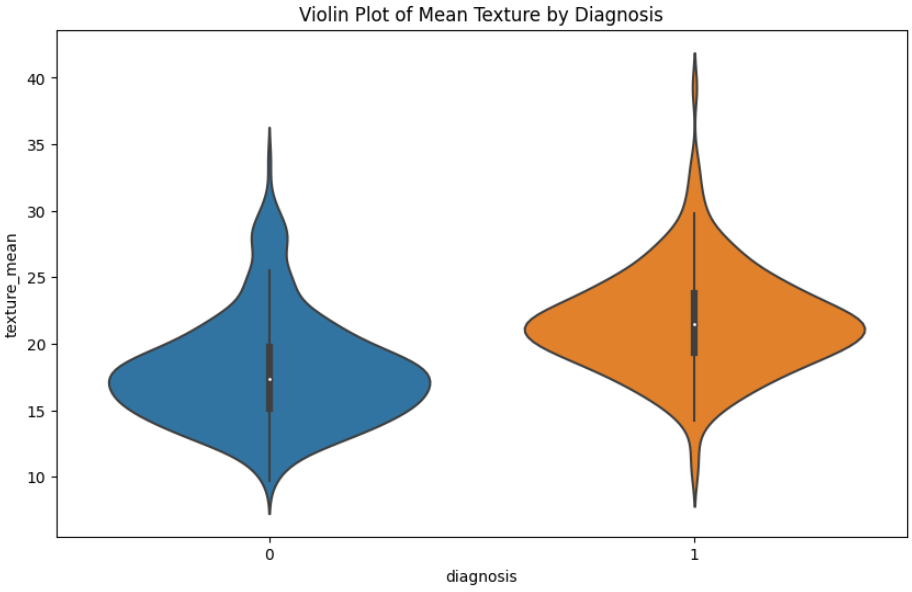

Fig. 5. Violin Plot of Mean Texture by Diagnosis

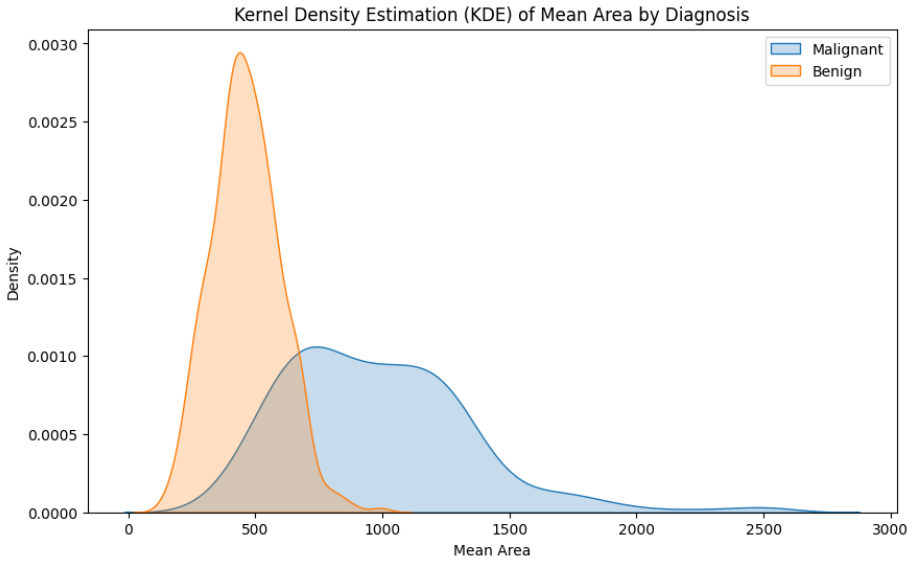

Fig. 6. Kernel Density Estimation (KDE) of Mean Area by Diagnosis

B: Full Results

Classification Methods: The full results of the classification methods can be seen in tables 3, 4, 5, 6, 7, and figures 7, 8, 9, 10.

	Time to Build Model (s)	Accuracy	Precision
C4.5	0.779	0.912	0.865
SVM	0.578	0.914	0.955
NB	0.352	0.937	0.940
kNN	1.952	0.930	0.922
RF	6.153	0.960	0.961

Table 3. time to build the model, accuracy, and precision for each of the five models.

Neural Network: The model was trained using a grid search approach, which is a method for hyperparameter tuning that systematically builds and evaluates a model for each combination of algorithm parameters specified in a grid. This process involved fitting 5 folds for each of the 972 candidates, resulting in a total of 4860 fits.

The model took approximately 838.17 seconds to build. The best hyperparameters found for the model include the following:

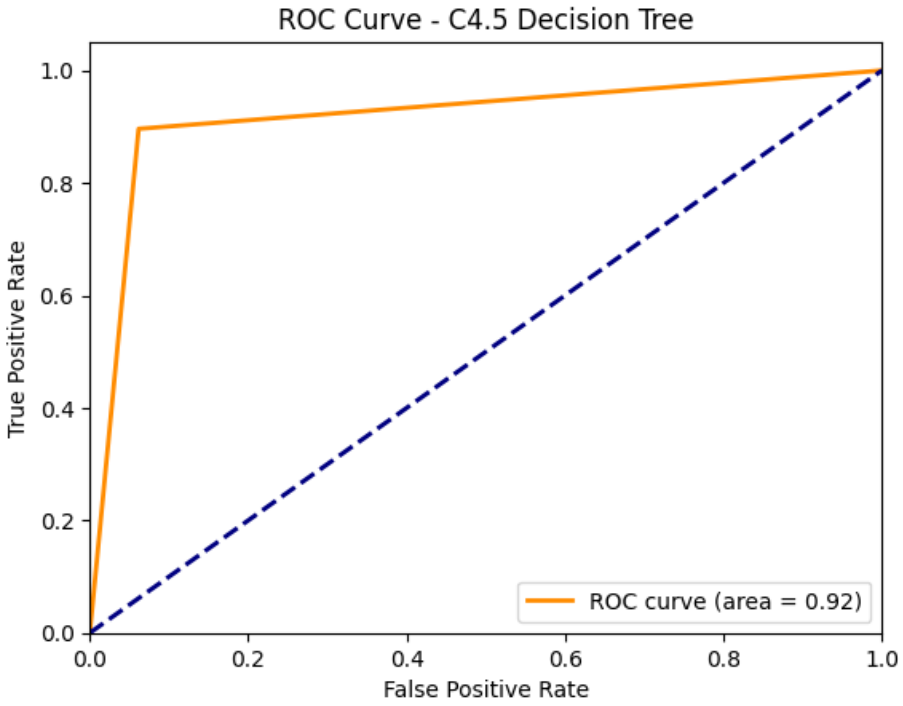

Fig. 7. Receiver Operating characteristic Curve for C4.5 Decision Tree

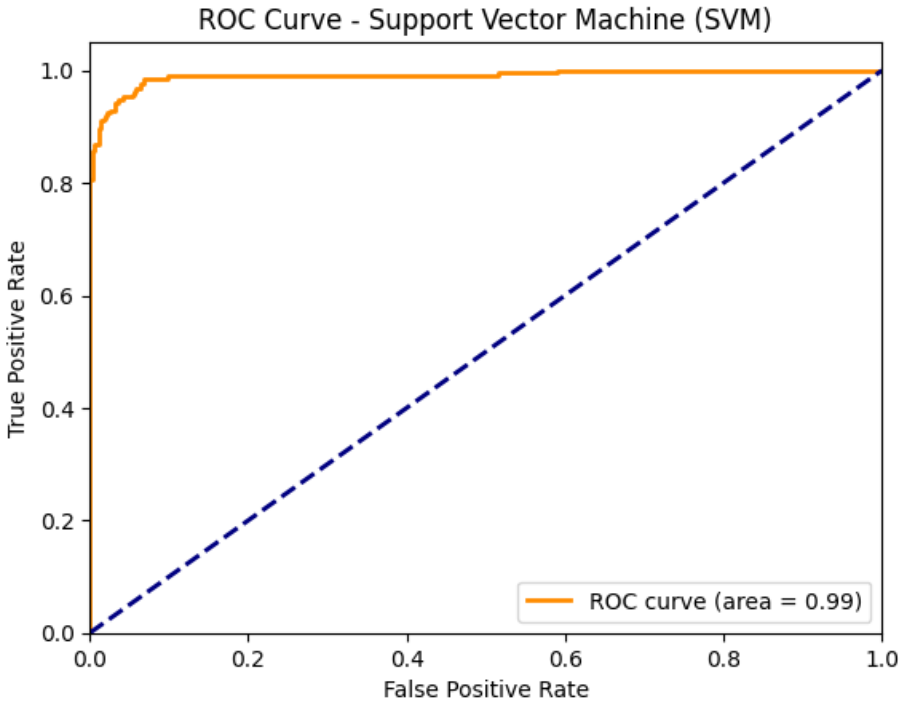

Fig. 8. Receiver Operating characteristic Curve for Support Vector Machine

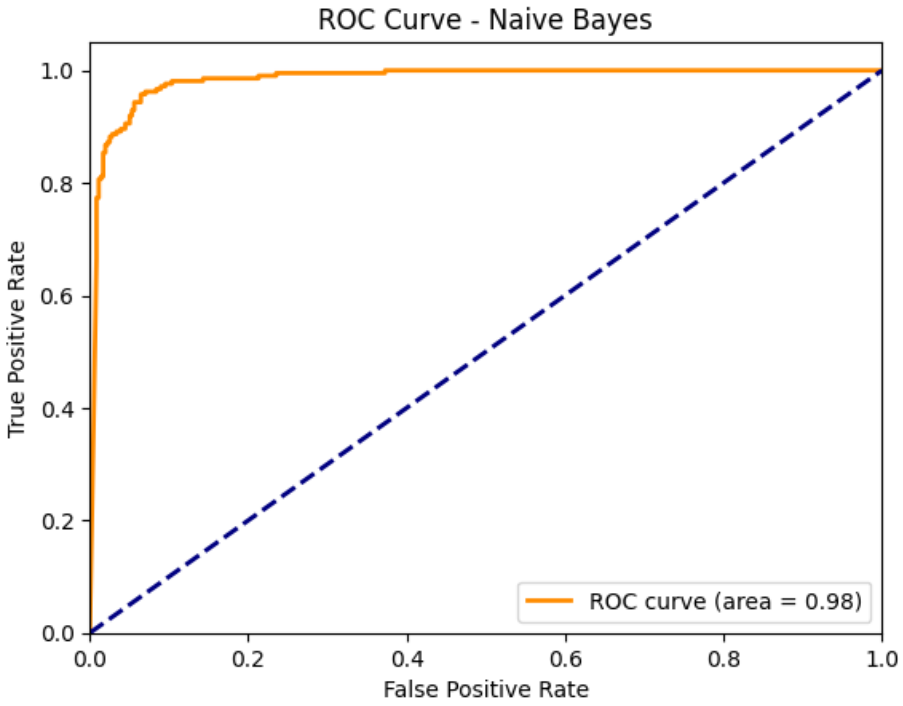

Fig. 9. Receiver Operating characteristic Curve for Naive Bayes

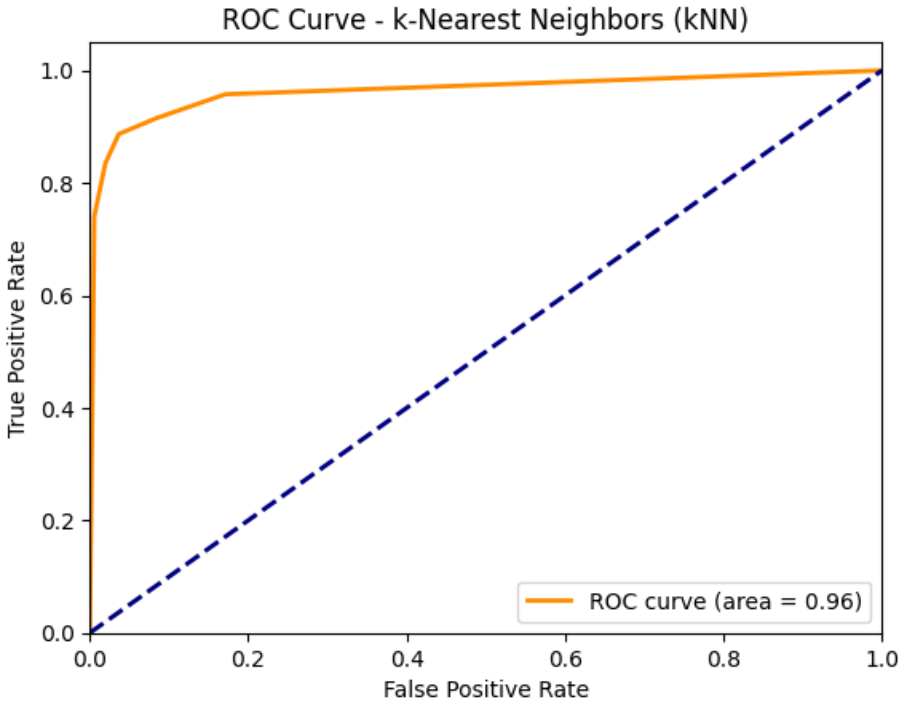

Fig. 10. Receiver Operating characteristic Curve for k-Nearest Neighbors

	Recall	F1-Score	Confusion Matrix
C4.5	0.877	0.871	[[328 29] [26 186]]
SVM	0.807	0.875	[[349 8] [41 171]]
NB	0.887	0.913	[[345 12] [24 188]]
kNN	0.887	0.904	[[341 16] [24 188]]
RF	0.920	0.940	[[349 8] [17 195]]

Table 4. recall, F1-score, and confusion matrix for each model. The confusion matrix is represented as a 2x2 matrix with the format '[[True Positive, False Positive] [False Negative, True Negative]]'.

	Kappa Statistic	MAE	RMSE	RAE	RRSE
C4.5	0.819	0.084	0.290	0.180	0.601
SVM	0.902	0.046	0.214	0.098	0.442
NB	0.863	0.063	0.252	0.135	0.520
kNN	0.848	0.070	0.265	0.150	0.548

Table 5. Kappa Statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) for each of the four models.

	Kappa Statistic	MAE	RMSE	RAE	RRSE
C4.5	0.819	0.084	0.290	0.180	0.601
SVM	0.902	0.046	0.214	0.098	0.442
NB	0.863	0.063	0.252	0.135	0.520
kNN	0.848	0.070	0.265	0.150	0.548

Table 6. Kappa Statistic, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) across the four models.

	Time to Build Model	Accuracy	Confusion Matrix	True Positive Rate
C4.5	0.129	0.923	[[335 22] [22 190]]	0.896
SVM	72.131	0.956	[[350 7] [18 194]]	0.915
NB	0.040	0.937	[[345 12] [24 188]]	0.887
kNN	0.055	0.935	[[344 13] [24 188]]	0.887

Table 7. This table includes the Time to Build Model, Accuracy, Confusion Matrix, and True Positive Rate for each of the four models.

Activation function: 'relu', **Alpha:** 0.01, **Batch size**: 64, **Beta_1**: 0.95, **Beta_2**: 0.999, **Early stopping**: Enabled, **Epsilon**: 1e-08, **Hidden layer sizes**: (128, 32), **Learning rate**: 'constant', **Learning rate initialization**: 0.1, **Momentum**: 0.9, **No change in iteration number**: 10, **Power_t**: 0.5, **Solver**: 'sgd', **Validation fraction**: 0.1.

The model achieved a best accuracy of 0.980 during training and a test accuracy of 0.956. These results suggest that the model is performing well and is effective at making accurate predictions. The high accuracy indicates that

the model correctly identified a high proportion of positive and negative cases. The use of the 'relu' activation function, the 'sgd' solver, and the specific hyperparameters likely contributed to this high level of performance. However, it's important to note that the performance of a model can vary depending on the specific dataset and task at hand. Therefore, these results are specific to the current model and task.

67	4
1	42

Table 8. Confusion Matrix of the Model

References

1. Asri, H., Mousannif, H., Moatassime, H.A., Noel, T.: Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. https://doi.org/10.1016/j.procs.2016.04.224
2. U.S. Cancer Statistics Working Group.: United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report.
3. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer Statistics, 2016. https://doi.org/10.3322/caac.21332
4. Noble, W.S.: What is a support vector machine? Nat Biotechnol. 2006;24(12):1565-1567. https://doi.org/10.1038/nbt1206-1565
5. Rish, I.: An empirical study of the naive Bayes classifier. IJCAI Work Empir methods Artif Intell. 2001;3(November):41-46.
6. Asri, H., Mousannif, H., Al Moatassime, H., Noel, T.: Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud Technol Appl. https://doi.org/10.1109/cloudtech.2015.7337020
7. Kaur, G., Gupta, R., Hooda, N., Gupta, N.R.: Machine Learning Techniques and Breast Cancer Prediction: A Review. https://doi.org/10.1007/s11277-022-09673-3
8. Pratyush, P.,: The Application of Machine Learning Techniques to the Diagnosis of Breast Cancer. https://doi.org/10.1109/AISP57993.2023.10134824
9. Khalid, A.; Mehmood, A.; Alabrah, A.; Alkhamees, B.F.; Amin, F.; AlSalman, H.: Breast Cancer Prediction Analysis using Machine Learning Algorithms. https://doi.org/10.3390/diagnostics13193113
10. Thakur, N., Kumar, P.: A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities. https://doi.org/10.1007/s11042-023-16634-w
11. Caballé-Cervigón, N., Castillo-Sequera, J.L., Gómez-Pulido, J.A., Gómez-Pulido, J.M., Polo-Luque, M.L.: Artificial intelligence in disease diagnosis: a systematic literature review. https://doi.org/10.3390/app10155135
12. Moreno-Ibarra, M.-A.; Villuendas-Rey, Y.; Lytras, M.D.; Yáñez-Márquez, C.; Salgado-Ramírez, J.-C.: Comparative Study on Disease Classification using Machine Learning Algorithms. https://doi.org/10.3390/math9151817
13. Aldera, S.A.: A Model for Classification and Diagnosis of Skin Disease using Machine Learning and Image Processing Techniques. https://doi.org/10.14569/IJACSA.2022.0130531

14. Caballé-Cervigón, N., Castillo-Sequera, J.L., Gómez-Pulido, J.A., Gómez-Pulido, J.M., Polo-Luque, M.L.: Machine Learning Applied to Diagnosis of Human Diseases: A Systematic Review. https://doi.org/10.3390/app10155135