Project Report

1. Abstract:

Five primary methods are compared: random forest, decision tree, XGBoost, and logistic regression in the context of breast cancer diagnosis. The random forest approach comprises feature optimization together with bagging techniques for selecting data points. The bagging trees algorithm performs better than the optimal decision tree parameters. Furthermore, accuracy is successfully improved during the training phase by a feature optimization technique.

Apart from the methods mentioned above, neural networks are another powerful tool used in machine learning for breast cancer diagnosis In order to increase accuracy, this provides a comprehensive analysis of several algorithms for the diagnosis of breast cancer achieving accuracy of 96.49%.

2. Introduction:

Millions of women worldwide are afflicted by breast cancer, a common and fatal type of the disease. A successful course of treatment depends on early discovery and precise diagnosis. The purpose of this work is to investigate how well bagging decision trees work for diagnosing breast cancer. The goal is to create a reliable and accurate classification model that will help with early identification and enhance the ability of diagnostics to differentiate between benign and malignant breast tumors. The work emphasizes how early detection can be achieved with little initial investment by using low-cost technology and machine learning algorithms like image processing and neural networks. The work aims to enhance patient outcomes and fight breast cancer by utilizing massive datasets and reliable, accurate models. Neural networks represent yet another potent instrument in machine learning for the diagnosis of breast cancer. This is a thorough examination of multiple algorithms for the diagnosis of breast cancer in an effort to improve accuracy.

3. Related work

In this review and analysis, we will look at the work "Breast Cancer Diagnosis Using Bagging Decision Trees with Improved Feature Selection" by Dudeja, Noonia, Lavanya, Sharma, Kumar, & Ramkumar (2023)[1] . The effectiveness of diagnosing breast cancer by combining enhanced feature selection techniques with the bagging decision tree algorithm is examined in this research. To increase our model's accuracy and interpretability, we will expand on their discoveries and suggest additional enhancements.

4. Method:

The Wisconsin Breast Cancer dataset, also referred to as the "Breast Cancer Wisconsin (Diagnostic) Data Set," was utilized to diagnose breast cancer. There are 569 samples in this collection, and each sample is a breast tissue biopsy. Each digitized image of a biopsy is used to compute 30 features, which include the radius, texture, smoothness, and worst (the mean of the three biggest values) of ten distinct characteristics. The binary target variable indicates if a malignant (1) or benign (0) biopsy diagnosis was made.
Using bagging trees, a technique that creates an ensemble model known as a random forest, decision trees' susceptibility and large variation are addressed. With this method, many decision trees that were separately trained using a bootstrapped resample of the training dataset are combined. Through voting, the predictions made by each individual tree are combined to create a more reliable model with lower variation. Bagging is beneficial for medical applications such as breast cancer diagnosis because it reduces overfitting, increases generalization, and improves accuracy and dependability. The interpretability and robustness of

the technique to noisy data facilitate quick and precise diagnosis, which may enhance patient care and treatment results.

The random forest (RF) technique efficacy was demonstrated by using the dataset to forecast malignant or diseased nodes. This dataset includes a label indicating the malignancy of the tumor in addition to several other features of a breast tissue bulge. Modelling and prediction were performed using the scikit-learn module for Python. The syntax for building and implementing a random forest model in scikit-learn is the same as that of decision trees and logistic regression models; the random-state condition is introduced to ensure that the code is structured for comparison and performs the same split each and every time. The training and testing sets' data points would vary every time without this random state, which would make code testing more challenging.

A New Model Based on Neural Network: An effective computer system is the neural network. To enable transmission amongst the units, NN gathers a sizable collection of units that are connected in some way. These components, which go by the names nodes or neurons as well, are basic parallel processors. Via a connection link, each neuron is linked to every other neuron. Every connection link has a weight connected to it that contains details about the input signal. Because the weight often either excites or suppresses the signal being sent, this is the most helpful information for neurons to solve a specific problem. An activation signal is the intrinsic state that every neuron has. The activation rule and input signals are combined to create output signals, which can then be routed to other units as shown in Figure 1.

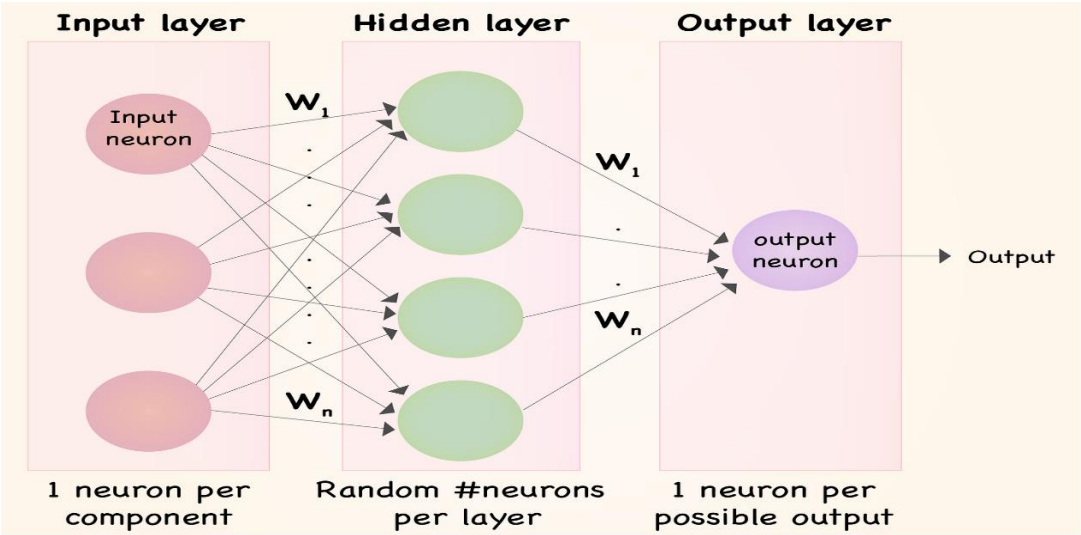

Figure 1 Neural Network Architecture.[1]

5. Experiments:

1. Data Preprocessing:

We used the "Breast Cancer Wisconsin (Diagnostic) Data Set" that we downloaded from Kaggle for this project [3]. To make sure the dataset is consistent and of high quality, we managed missing values, clean up the data, and normalize the variables. In order to determine the most useful features for breast cancer diagnosis, we also investigated the feature selection strategies. Splitting data into 80% for training and 20% for validation as in Figure 2.

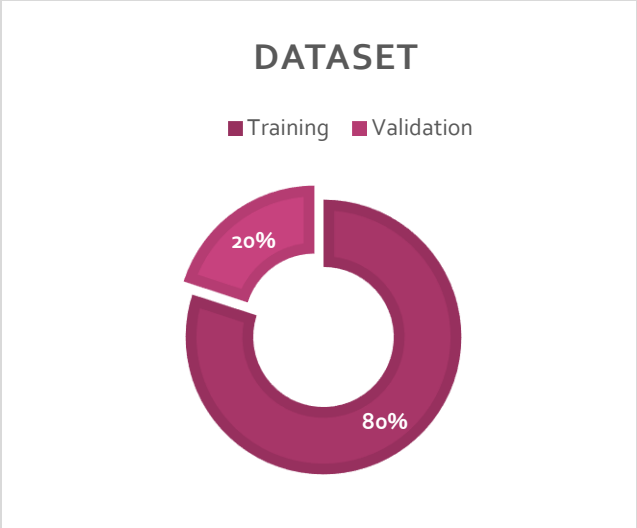

Figure 2 Data splitting.

2. Bagging Decision Tree Model Development:

The bagging decision tree algorithm that the cited research discusses is put into practice. To increase accuracy and decrease overfitting, this ensemble learning method mixes several decision tree models. We used the preprocessed dataset to train the model and assess its effectiveness.

3. Improved Feature Selection:

To determine the features most pertinent to the diagnosis of breast cancer, we employed the refined feature selection methods. The performance and interpretability of the model is improved by this step.

4. Performance Evaluation:

Using relevant metrics like accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), we assessed the performance of our model. To confirm the efficacy of our strategy, we also contrasted the outcomes with the conclusions from the mentioned research.

6. Results and Discussion

the efficacy of those techniques in obtaining precise classification, supporting early diagnosis, and differentiating between benign and malignant breast tumors. The results make a valuable contribution to the field of machine learning and its applications in medicine, especially in the diagnosis of breast cancer. We can see that from achieving accuracy for each algorithm as shown in Table 1 and Figure 3. The results of performance metrics are included in Jupyter Notebook[4].

Table 1 Algorithms Accuracy

	Algorithm	Accuracy
1	Logistic Regression	0.982456
2	Decision Tree	0.947368
3	Random Forest	0.956140
4	XGBoost	0.964912
5	Neural Network	0.9649

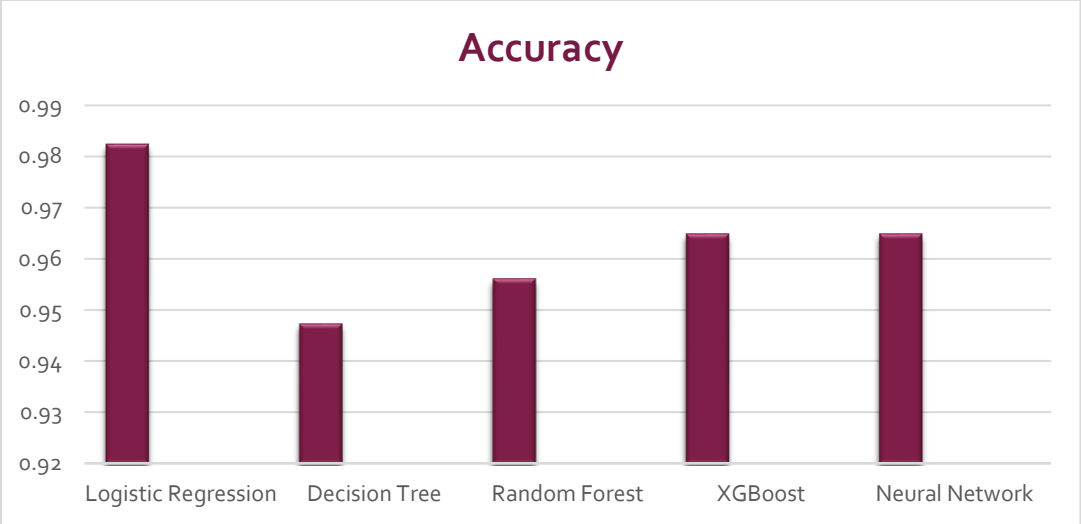

Figure 3 accuracy for five algorithms.

With the Wisconsin Breast Cancer dataset, neural networks have demonstrated impressive outcomes in breast cancer diagnosis. The 569 data points that make up the dataset each correspond to a different feature of a breast tissue mass. Factors like radius, concavity, symmetry, smoothness, compactness, and fractal dimension are a few examples of these properties. A label reflecting the tumor's malignancy or benignity is also provided by the dataset. Neural networks have the capacity to recognize complex patterns and relationships in data, with a total of thirty characteristics at their disposal. The remarkable accuracy rate of 96.49% in the classification of breast tumors has been attained through rigorous training and optimization.

7. Conclusion:

In the context of diagnosing breast cancer, five main approaches are compared: random forest, decision tree, XGBoost, neural network , and logistic regression. The random forest method combines bagging techniques for data point selection with feature optimization. The optimal decision tree parameters are outperformed by the bagging trees algorithm. Furthermore, a feature optimization strategy successfully increases accuracy during the training phase.
Early detection and improve the diagnostic tools' capacity to distinguish between benign and malignant breast tumors. Image processing and neural networks are two examples of low-cost technologies and machine learning techniques that can be used to provide early detection with minimal initial expenditure. By utilizing increasingly effective neural network techniques, such as Convolutional Neural Networks (CNN), we will be able to increase detection accuracy in the future.

References

[1].Dudeja, D., Noonia, A., Lavanya, S., Sharma, V., Kumar, V., Rehan, S., & Ramkumar, R. (2023). Breast Cancer Diagnosis Using Bagging Decision Trees with Improved Feature Selection. Engineering Proceedings, 59, 17. https://doi.org/10.3390/engproc2023059017

[2] https://towardsdatascience.com/understanding-neural-networks-what-how-and-why-18ec703ebd31

[3].https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download

[4] https://colab.research.google.com/drive/1bSnj7st6S30d487n71sXjTIe3LuLljYW?usp=sharing