Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.15, No.1, 2024

Enhancing Tumor Classification Through Machine Learning Algorithms for Breast Cancer Diagnosis

Lawrence Agbota[1], Edmund Agyemang[1,2,3*], Priscilla Kissi-Appiah[1], Lateef Moshood[1], Akua Osei-Nkwantabisa[1], Vincent Agbenyeavu[1], Abraham Nsiah[4], Augustina Adjei[5]

1. School of Mathematical and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA.
2. University of Ghana, Department of Statistics and Actuarial Science, Ghana
3. Department of Computer Science, Ashesi University, No.1 University Avenue, Berekuso, Ghana.
4. Department of Statistics, Ball State University, Muncie-USA
5. Ghana Health Service, Koforidua-Ghana.

* E-mail of the corresponding author: edmundfosu6@gmail.com

Abstract

In cancer diagnosis, machine learning helps improve cancer detection by providing doctors with a second perspective and allowing for faster and more accurate determination and decisions. Numerous studies have used both classic machine learning approaches and deep learning to address cancer classification. In this study, we examine the efficacy of five commonly used machine learning algorithms; both traditional and deep learning models namely, Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), Decision Tree and Deep Neural Networks (DNN). We analyze their ability to properly classify tumors as Benign or Malignant using the Wisconsin breast cancer dataset (WBCD). Random Forest classifier was employed to reduce model complexity, successfully narrowing down the number of features to 17 through cross-validation and achieving a validation score of 96.84%. Subsequently, a grid search was used to determine the maximum tree depth, resulting in five. The Synthetic Minority Oversampling Technique (SMOTE) was employed as a resampling tool to balance the Benign and Malignant categories adequately solving the class imbalance problem encountered in classification problems. After evaluating the overall performance for the unbalanced data, Random Forest emerged as the best classification model with an accuracy of 98.20%, followed by Logistic Regression with an accuracy of 97.40%. However, after applying SMOTE, both Random Forest and Logistic Regression emerged as the best models both with an accuracy of 94.70%. Both Random Forest and Logistic Regression models had an outstanding performance with an area under the curve (AUC) value of 0.997 and 0.994 respectively.

Keywords: Breast Cancer, Random Forest, Logistic Regression, Support Vector Machines, Deep Neural Networks, Synthetic Minority Oversampling Technique.
DOI: 10.7176/CEIS/15-1-08
Publication date: June 30th 2024

1 Introduction

Breast cancer remains one of the most common cancers among women worldwide, significantly impacting public health [1]. It is the leading cause of cancer-related deaths among women, affecting millions each year. Despite advances in treatment and early detection, the diagnosis of breast cancer at later stages can significantly diminish survival rates and increase treatment complexities. Early and accurate diagnosis is crucial for effective treatment and better patient outcomes. Given its high prevalence and the severe implications of delayed or incorrect diagnosis, there is a pressing need for innovative and more reliable diagnostic methods. This backdrop sets the stage for exploring enhanced machine learning techniques that can potentially transform the landscape of breast cancer diagnosis, promising more accurate, timely interventions. Breast cancer if left unchecked, the tumors can spread throughout the body and become fatal. Breast cancer cells begin inside the milk ducts and/or the milk-producing lobules of the breast. The earliest form (in situ) is not life-threatening and can be detected in early stages. Cancer cells can spread into nearby breast tissue (invasion). This creates tumors that cause lumps or thickening [2]. Invasive cancers can spread to nearby lymph nodes or other organs (metastasize). Metastasis can be life-threatening and fatal. Treatment is based on the person, type of cancer, and its spread. Treatment combines surgery, radiation therapy, and medications

[3]. In 2022, there were 2.3 million women diagnosed with breast cancer and 670,000 deaths globally. Female gender is the strongest breast cancer risk factor. Approximately 99% of breast cancers occur in women and 0.5–1% of breast cancers occur in men [2]. Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life [4]. The treatment of breast cancer in men follows the same principles of management as for women.

A tumor is an abnormal mass or growth of tissue that serves no specific purpose. It can develop when cells grow and divide too quickly [5]. Tumors can be located anywhere in the body. They grow and behave differently depending on whether they are benign (non-cancerous) or malignant (cancerous) [6]. A benign tumor is composed of cells that do not threaten to invade other tissues. The tumor cells are contained within the tumor and are not significantly different from the surrounding cells. Malignant tumors are composed of cancer cells that can develop uncontrollably and infiltrate surrounding tissues. The cancer cells in a malignant tumor tend to be abnormal and very different from the normal surrounding tissue [7]. A biopsy process allows a healthcare worker to obtain a sample of cells to determine whether a tumor is benign or malignant [8]. The cells will next be tested by a pathologist, a specialist who specializes in tissue examination. This includes examining the sample under a microscope. This is the most definitive way to determine tumor status, and the answer is usually clear-cut. But sometimes, the diagnosis is uncertain. It is also possible that cancer could be present, but the biopsy missed the area with malignant cells [2]. In lieu of this, machine learning (ML) which is rapidly transforming the field of healthcare, offering new ways to enhance diagnostic accuracy and patient treatment outcomes become useful. ML capacity to analyze large datasets and uncover patterns undetectable to the human eye makes it especially valuable in medical diagnostics, where early detection can be lifesaving. By leveraging ML, healthcare professionals can make more informed decisions, streamline workflows, and reduce the burden of manual tasks. This technological shift not only promises to improve clinical outcomes but also to revolutionize patient care by making it more data-driven and efficient.

2 Related Works

Supervised classification is one of the most common tasks undertaken by Intelligent Systems. A significant variety of techniques have been created based on Artificial Intelligence (Logic-based techniques, Perceptron- based techniques) and Statistics (Bayesian Networks, Instance-based techniques) [9]. [10] used an ensemble classification mechanism to diagnose breast cancer tumors and compared performance with the hard voting (majority-based voting) mechanism to the state-of-the-art algorithm. Results show that the hard voting mechanism shows better performance with an accuracy of 99.42% as compared to the state-of-the-art algorithm.
[11] used six different classification methods: Multilayer Perceptron, Decision Tree, Random Forest, Support Vector Machine, and Deep Neural Network Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis. Results indicate that the DNN classifier had the greatest performance in accuracy level (92%), indicating better results in relation to traditional models. [12] proposed data exploratory techniques (DET) and developed four different predictive models to improve breast cancer diagnostic accuracy. Prior to modelling, four-layered essential DET, e.g., feature distribution, correlation, elimination, and hyperparameter optimization, were deep- dived to identify the robust feature classification into malignant and benign classes. Results showed that SVM with polynomial kernel gained 99.3%, LR with 98.06%, KNN acquired 97.35%, and EC achieved 97.61% accuracy with the Wisconsin breast cancer dataset (WBCD). [13] predicted breast cancer using twelve classification algorithms: AdaBoost, J-Rip, LR, lazy learner, decision table, IBK, J48, lazy K-star, multiclass classifier, multilayer perceptron, random forest, Na¨ıve Bayes, and random tree. The results showed that other than Na¨ıve Bayes classification, all the algorithms performed very well with accuracy greater than 94% and that lazy and tree classifications outperformed other classification algorithms, with 99% accuracy.

[14] asserted that although ensemble learning enables the improvement of performance of a base learner, it decreases the bias or variance. [15] proposed a new ensemble classification algorithm, CWV-BANNSVM, by combining Boosting Artificial Neural Network (BANN) along with two SVMs to improve the performance for WBCD. In contrast to traditional ensemble learning, [16] proposed a novel dynamic ensemble learning algorithm to automatically determine the number of neural networks and their architecture. Different training sets were used for each neural network hence guaranteeing better learning from the whole training data samples. The proposed DEL was trained several times to find the correct values of learning rate parameter

and the correlation strength parameter by using an incremental training approach. [17] stated that improvement is possible when using the ensemble boosting method. The method was integrated with a Radial Basis Function neural network algorithm and performance was increased to an accuracy of 98.4% for the WBCD dataset. Most of the literature makes use of both the traditional and the deep learning approach. This was reviewed that the accuracy is higher when the occurrence of true positives (TPs) and true negatives (TNs) is high compared to the false positives (FPs) and false negatives (FNs). Aside from accuracy, precision and recall are critical for performance reporting. However, for medical diagnostics, the performance of artificial intelligence systems should prioritize false negatives over false positives, as missing the diagnosis of a disease might have major consequences for patients.

In the quest to enhance breast cancer diagnosis, several machine learning algorithms are being employed, each offering unique strengths. Logistic Regression is widely used for its simplicity and effectiveness in binary classification problems. Support Vector Machines (SVM) are favored for their ability to handle high-dimensional data, making them suitable for complex diagnostic imaging tasks. Random Forest (RF) and Decision Trees provide robustness and ease of interpretation, crucial for medical applications where understanding the decision process is important. Deep Neural Networks (DNN) excel in pattern recognition, learning directly from pixel- level data in imaging studies, which is pivotal for identifying subtle anomalies indicative of early-stage tumors. These algorithms form the backbone of modern computational approaches to medical diagnostics, driving forward the capabilities of automated tumor classification systems. The main objective of this study is to examine the efficacy of five machine learning algorithms namely, Logistic Regression, Support Vector Machines (SVM), Random Forest, DNN and Decision tree classifier in properly classifying tumors as benign or malignant utilizing the Wisconsin breast cancer dataset (WBCD). This comparative analysis aims to discover the best performing algorithm for breast cancer diagnosis, which could provide useful insights to the medical field.

2.1 Breast Cancer Diagnostic Workflow

A schematic representation of the diagnostic workflow for breast cancer using machine learning and deep learning is depicted in Figure 1. The process begins with the feature extraction phase, where significant characteristics of the dataset are identified. These features are then fed into a traditional machine learning classifier or a deep convolutional neural network, which is part of the deep learning approach. Finally, the model generates a prediction indicating whether the observed patterns suggest a benign or malignant tumor.

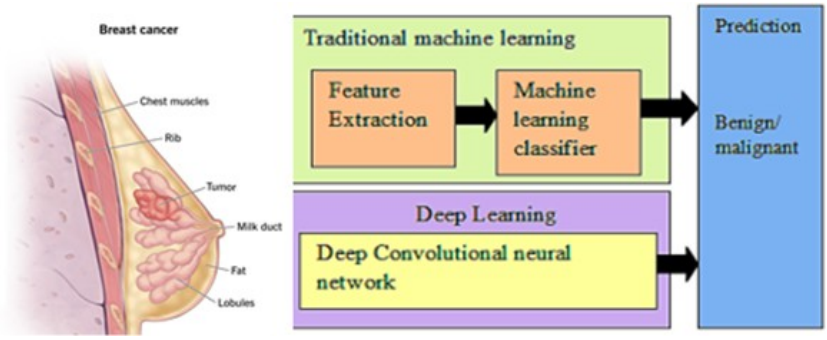


Figure 1: Diagram illustrating the process of diagnosing breast cancer using traditional machine learning and deep learning methods.

2.2 Types of Tumors

Figure 2 shows the fundamental differences between benign and malignant tumors. Benign tumors are non-cancerous and do not spread to other tissues, whereas malignant tumors are cancerous and have the potential to spread to different parts of the body, a process known as metastasis.

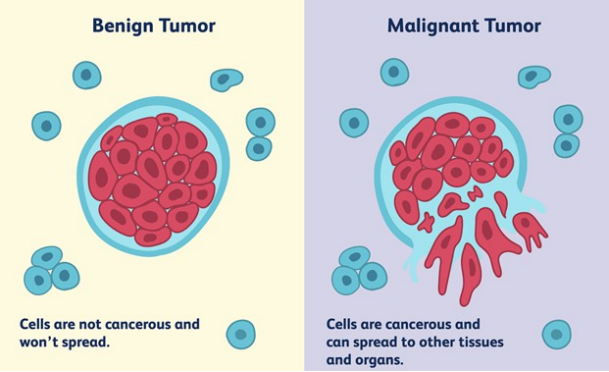


Figure 2: Diagram illustrating the types of tumors.

3 Data and Methods

The dataset was provided in CSV file format and retrieved from Kaggle at https://www.kaggle.com/datasets/ uciml/breast-cancer-wisconsin-data. Diagnosis is the target and is classified as malignant (M) and benign (B). The number of sample class distributions for each benign and malignant is 357 and 212 respectively.

3.1 Data Preprocessing

The data was split into 80% training and the remaining 20% for testing. The training data consisted of approximately 455 samples and the testing data was comprised of 114 samples. Before training the data, we scale the features using MinMaxScaler normalization technique to make all features contribute equally to the result of our predictions. We also conducted extensive hyper-parameter tuning for each of the five classification models. Label encoder was employed to transform the response variable into 0 and 1, where 1 corresponds to a malignant tumor and 0 corresponds to a benign tumor as represented in (1) by:

$$Diagnosis = \begin{cases} 1, & if\ Malignant\ Tumor \\ 0, & if\ Benign\ Tumor \end{cases} \quad (1)$$

Synthetic Minority Over-sampling Technique (SMOTE) was employed to address the class imbalance problem in the breast cancer data. SMOTE works by creating synthetic samples from the minority class instead of simply duplicating existing samples [18]. This is achieved by randomly selecting a point from the minority class and computing the difference between this point and its nearest neighbors. The method then creates new points along the line segments joining the selected points in the feature space. By doing this, SMOTE adds variety to the training data, which helped in achieving a more balanced dataset and thereby improving the performance of a classifier. SMOTE helps in overcoming the overfitting problem which tends to occur when duplicating minority class samples. It also ensures that the decision boundary for the minority class is not too tight, allowing the model to generalize better on unseen data. Hence, SMOTE is particularly useful in scenarios like breast cancer diagnosis, where it is critical to detect the less frequent, but more dangerous cases effectively.

3.2 Classification Models

1. Logistic Regression (LR) is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. LR is represented in (2) by:

$$p(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}} \quad (2)$$

where **x** represents the feature vector, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the model, and $p(y = 1|x)$ is the probability that the target y is 1 given $\boldsymbol{x}$. LR was employed in the study as a classification tool.

2. Support Vector Machines (SVM) are a set of supervised learning methods used for classification, regression, and outlier detection. The objective of the SVM algorithm is to find a hyperplane in an N-

dimensional space that distinctly classifies the data points quantified in (3) as:

$$\boldsymbol{w}.\boldsymbol{x} + b = 0 \tag{3}$$

where **w** is the normal vector to the hyperplane and b is the bias. The optimal hyperplane maximizes the margin between the two classes. SVM was employed as a classification tool in this study with three different kernels namely linear, polynomial and radial basis function (RBF).

3. Random Forest (RF) is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class; that is the majority vote (mode) of the classes of the individual trees represented in (4) by:

$$\text{Random Forest Classification} = \text{mode}\,(\text{Tree}_1(\mathbf{x}),\ \text{Tree}_2(\mathbf{x}),\ \ldots,\ \text{Tree}_n(\mathbf{x})) \tag{4}$$

where **x** is the input feature vector and each Tree_i is an individual decision tree classifier.

4. Classification Tree is a type of decision tree that is used for classifying instances. It makes decisions by splitting data based on feature values. The tree structure consists of nodes and leaves, where nodes represent feature choices and leaves represent decisions or classifications. The prediction for an instance x using a classification tree can be represented in (5) as:

$$f(x) = \bigcup_{i=1}^{n} c_i.I(x \in R_i) \tag{5}$$

where n is the number of leaves in the tree, c_i is the class predicted by leaf i, I is the indicator function, and R_i is the region of the feature space associated with leaf i.

5. Deep Neural Network (DNN) a type of artificial neural network with multiple layers between the input and output layers which can model complex non-linear relationships [19]. The output of a DNN with L layers for an input vector x is given by the composition of multiple non-linear functions in (6) by:

$$f(x) = f^{(L)}\left(f^{(L-1)}\left(\ldots f^{(2)}\left(f^{(1)}(x)\right)\ldots\right)\right) \tag{6}$$

where $f^{(l)}$ denotes the function of the $l-th$ layer of the network. Each layer typically computes the following transformation given in (7) by:

$$f^{(l)}(x) = \sigma\left(W^{(l)}(x) + b^{(l)}\right) \tag{7}$$

where $W^{(l)}$ and $b^{(l)}$ represent the weights and biases at layer l, respectively, and σ is a non-linear activation function like such as tanh, ReLU or sigmoid.

3.3 Definitions of Classification Rates and Evaluation Performance

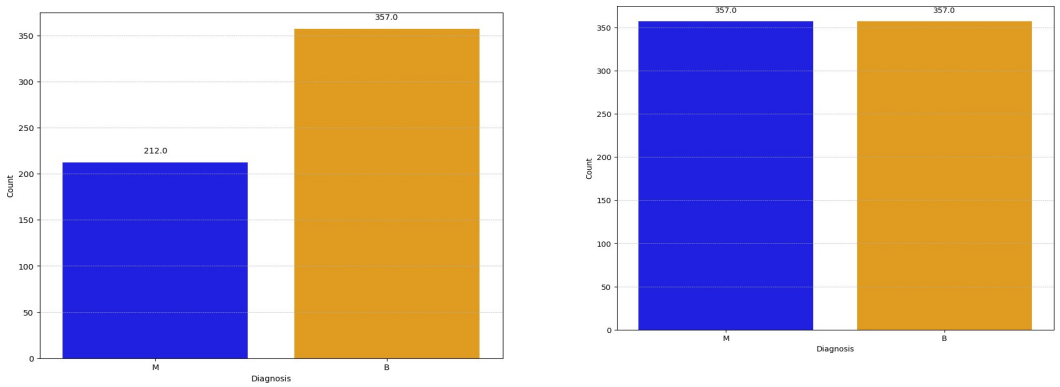
Table 1: Classification Outcomes and Performance Statistics

Metric	Definition
True Positive (TP)	Correct positive prediction
True Negative (TN)	Correct negative prediction
False Positive (FP)	Incorrect positive prediction
False Negative (FN)	Incorrect negative prediction
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Sensitivity/Recall	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
F1-Score	$\frac{2(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$

4 Results & Discussion

In this section, we present the results and findings of the study.

4.1 Diagnostic Distribution with Unbalanced and Balanced Class



(a) Diagnosis count without SMOTE (b) Diagnosis count with SMOTE

Figure 3: Diagnosis count with and without SMOTE

Figure 3(a) indicate that there are 212 malignant cases, which is a significant concern as these represent cases where cancer has been confirmed. Also, there are 357 benign cases, which, while not indicative of cancer, still require attention to ensure they do not progress. In Figure 3(b), SMOTE has been applied to balance the cancer diagnosis counts. After balancing, both classes have the same counts of 357 solving the data imbalance problem commonly encountered in classification problems as evident in Figure 3.

4.2 Analysis of Feature Interactions

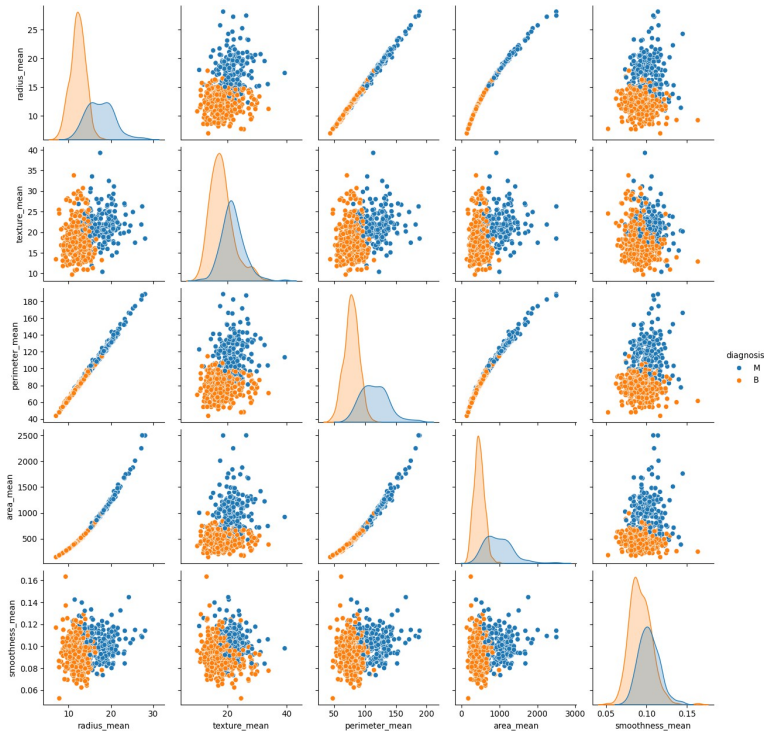


Figure 4: Bivariate relationships and univariate distributions.

From Figure 4, we observe that certain features exhibit a positive linear relationship, suggesting that as one feature increases, the other tends to increase as well. This is particularly evident in the scatter plot correlating the mean area of cells to their mean perimeter. Such a relationship is expected due to the geometric nature of these measurements. Also, certain scatter plots do not exhibit a clear relationship, indicating that the features may provide independent information valuable for classification purposes. The distributions along the diagonal offer insight into the variability of each feature across benign and malignant cases, where significant overlap may indicate a less discriminative feature in isolation. On particular note of the separation of malignant and benign cases in some of the feature combinations; a pronounced separation suggests that these feature pairings could be potent predictors for classification models, as they provide clear boundaries between the two diagnoses. It is these patterns that we seek to exploit in developing machine learning models that can accurately classify tumors based on their feature profiles, ultimately aiding in early and effective diagnosis.

4.3 Exploratory Analysis of Radius and Texture Mean by Diagnosis

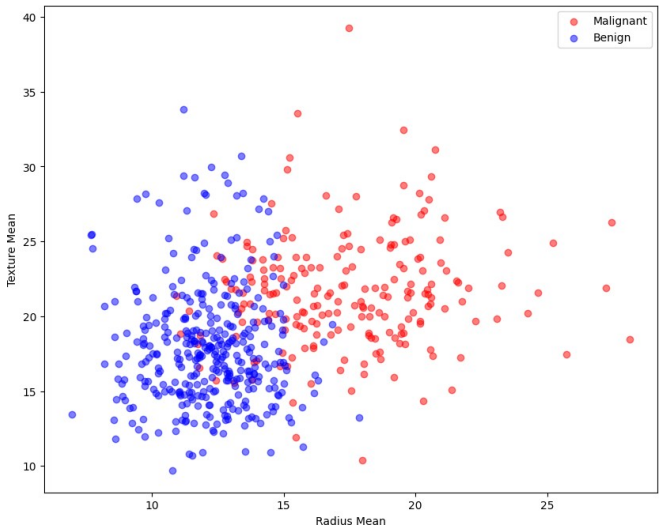
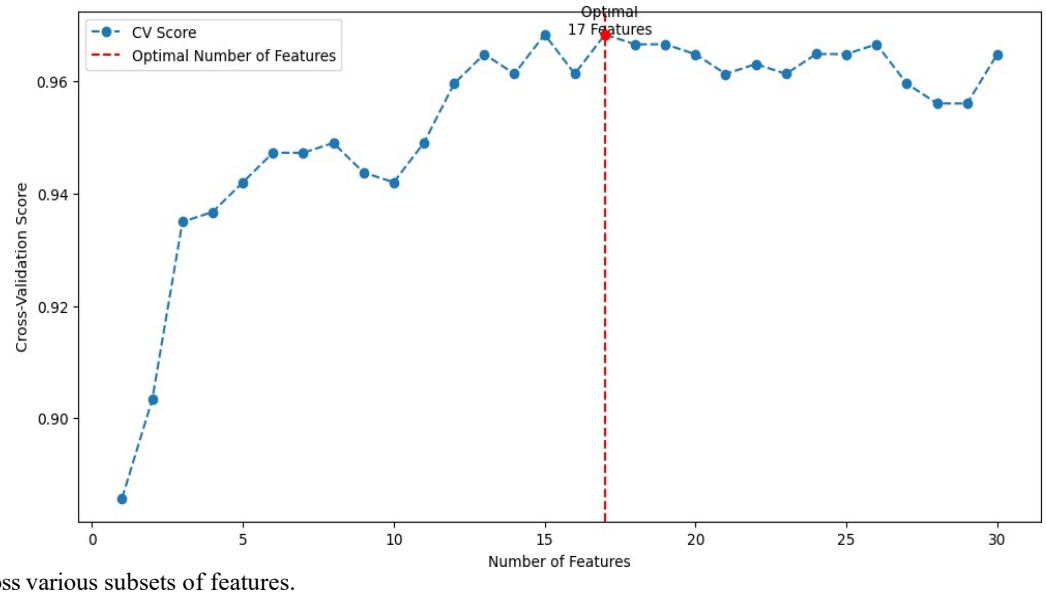


Figure 5: Scatter plot of radius mean versus texture mean by diagnosis

From Figure 5, there is a discernible pattern where points representing malignant tumors tend to exhibit higher values in both radius and texture mean, as opposed to the benign category. This observation aligns with clinical expectations that malignant tumors generally present more irregularities and larger sizes compared to their benign counterparts. The clustering of blue points corresponding to benign tumors is indicative of a relatively tighter grouping, suggesting less variability within these measurements. In contrast, the red points, signifying malignant cases, are more dispersed, reflecting a greater heterogeneity that could be attributed to aggressive tumor growth patterns.

4.4 Optimizing Feature Selection

Within the scope of enhancing our model's performance, we find the optimal set of features by importing the_cross validation score function from the sklearn.model selection library, setting the cross-validation fold count to 5. This systematic approach enabled us to evaluate our Random Forest classifier's performance



across various subsets of features.

Figure 6: Model's cross-validation score with optimal features.

Figure 6 illustrates the model's increasing cross-validation score as more features are incorporated, identifying a plateau that indicates the optimal feature count. The analysis revealed that the highest cross-validation score of 0.9684 resulted in an Optimum number of features of 17. The selected features, which are deemed to be most significant for our classification task as depicted in Figure 7 are Perimeter Worst, Radius Worst, Concave Points Worst, Concave Points Mean, Area Worst, Concavity Mean, Perimeter Mean, Area Mean, Area SE, Concavity Worst, Radius Mean, Texture Worst, Radius SE, Perimeter SE, Texture Mean, Smoothness Worst, and Compactness Worst. These features collectively form the core attributes for effectively distinguishing between malignant and benign tumors, thereby enhancing the predictive precision of our model.

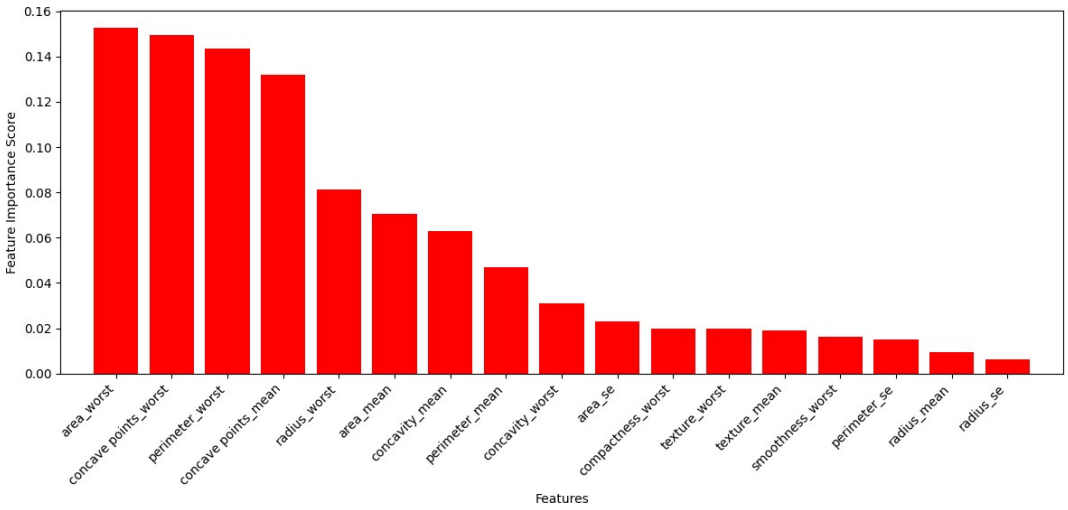


Figure 7: Feature Importance of the 17 optimal features.

4.5 Model Optimization via Hyperparameter Tuning

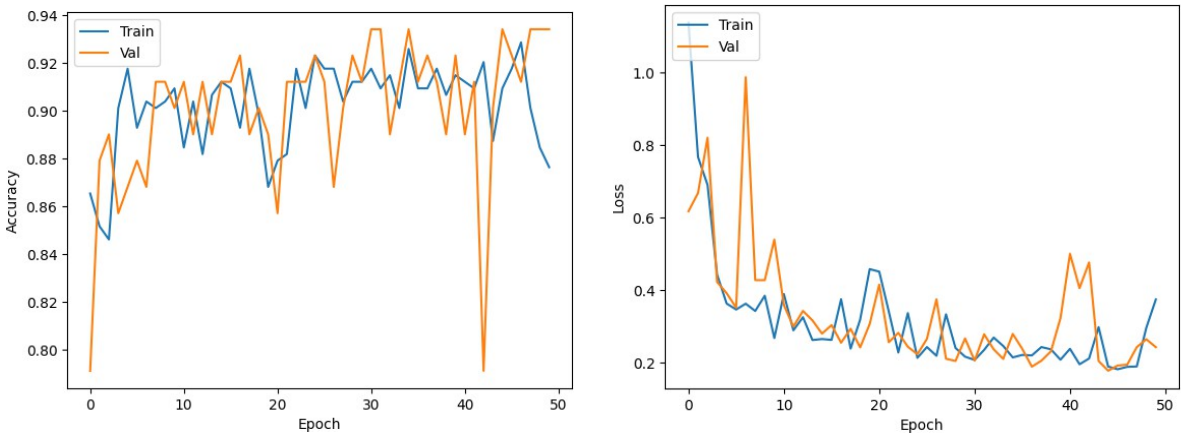
Before arriving at the optimal models, a thorough optimization process was conducted using grid search cross- validation to fine-tune the hyperparameters. For random forest, the grid search tested various combinations of maximum tree depth, minimum samples required to split a node, and minimum samples required at a leaf node. This comprehensive search across the hyperparameter space included depth values ranging from 3 to 20, with minimum sample splits and leaves tested at several critical thresholds. The grid search determined the best-performing model to have a 'max depth' of 5, along with 'min samples leaf' and 'min samples split' both 2. This model configuration achieved a cross-validation score of 0.96, signifying a high level of predictive accuracy while maintaining a balance between model complexity and generalization capability.

Table 2: Gridsearch with 5-fold cross-validation

Model	Max. Depth	Min. Sample Leaf	Min. Sample Split	Accuracy
Random Classifier	5	2	2	0.96
Decision Classifier	10	5	2	0.93

We chose the Random Classifier model from Table 2 because it demonstrated the highest accuracy of 0.96 in our tests, showing superior performance over the Decision Classifier.
Additionally, using grid search for support vector classifier (SVC), we set up the parameters and the distributions to sample to tune hyperparameters for an SVM. These are the cost or regularization parameter and was set as [0.1, 1, 10, 100] and also the kernel coefficients as [1, 0.1, 0.01, 0.001]. The best kernel coefficient, gamma after the search resulted in 0.001 with a regularization parameter of 1. We then proceed to fit the SVM model for the 3 kernel functions: Linear, RBF and Polynomial. The deep neural network (DNN) models were built using TensorFlow and Keras, leveraging the Sequential and Dense classes for model architecture. Label encoding was employed to transform non-numerical labels into a numerical form, while one-hot encoding was applied to convert class vectors into a binary class matrix. The DNN model was structured to use a range of neurons per layer and varying learning rates for the Adam optimizer. The DNN model employed ReLU activation for hidden layers and softmax for the output layer, optimizing for categorical crossentropy loss and monitoring accuracy. Hyperparameter tuning was executed using RandomSearch with a focus on maximizing validation accuracy.

The model was then trained using 50 epochs, with the best-performing model selected based on its validation accuracy. The best validation accuracy achieved during hyperparameter tuning was approximately 93.41%, indicating the model's proficiency in classifying the validation data. The tuning was completed in a time-efficient manner, taking only about 1 minute and 3 seconds. Over the span of 50 training epochs, the model's performance consistently improved. It started with a high initial loss, yet decent accuracy, and progressed through typical mid-training adjustments. By the end of 50 training epochs, the loss had substantially decreased, and accuracy had correspondingly increased as evident in Figure 8, evidencing effective learning from the training data. Upon evaluation with test data, the model's accuracy was about 92.98%, which is remarkably close to the validation accuracy, showcasing its capability to generalize well. The model's test loss stood at 0.3619, underscoring its efficacy.



(a) Model Accuracy Over Epochs (b) Model Loss Over Epochs

Figure 8: **Model Accuracy and Loss Over Epochs**

Figure 9 shows the optimized decision tree after hyperparameter tuning. The tree's maximum depth of five was carefully chosen to prevent overfitting, ensuring that the model remains generalizable to new data. This depth allows the tree to capture sufficient complexity in the data patterns without becoming overly specialized to the training data. Here, each node represents a decision based on the features of the breast cancer dataset, and the terminal nodes indicate the resulting classification as either benign or malignant.

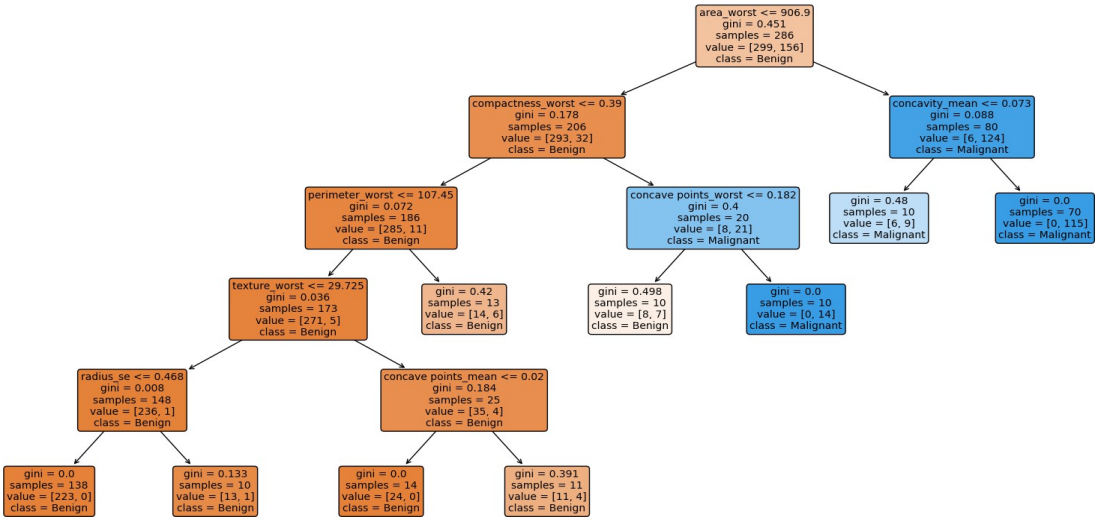


Figure 9: Optimal Decision tree with maximum depth of 5.

4.6 Confusion Matrices for Classification Models With and Without SMOTE

The confusion matrices of the various models under consideration for the classification of breast cancer tumors are summarized in Table 3 for unbalanced data (without SMOTE) and balanced data (with SMOTE).

Table 3: With and Without SMOTE Confusion Matrices for Classification Models

Models	**Without SMOTE**		**With SMOTE**	
Random Forest	Predicted Benign	Predicted Malignant	Predicted Benign	Predicted Malignant
Actual Benign	66	1	62	5
Actual Malignant	1	46	1	46
Logistic Classifier				
Actual Benign	66	1	62	5
Actual Malignant	2	45	1	46
Decision Tree				
Actual Benign	63	4	59	8
Actual Malignant	2	45	3	44
SVM (Linear)				
Actual Benign	66	1	63	4
Actual Malignant	7	40	4	43
SVM (Polynomial)				
Actual Benign	64	3	62	5
Actual Malignant	2	45	2	45
SVM (RBF)				
Actual Benign	61	6	59	8
Actual Malignant	3	44	3	44
DNN				
Actual Benign	59	8	59	8
Actual Malignant	2	45	3	44

From Table 3, the without SMOTE confusion matrix indicates that Random Forest model is highly effective in classifying cases of breast cancer. It correctly identified 66 out of 67 actual benign cases and 46 out of 47 actual malignant cases, demonstrating high accuracy and precision. Misclassifications are minimal, with only 1 benign case incorrectly identified as malignant and 1 malignant case incorrectly identified as benign, suggesting the model is well-calibrated for both sensitivity and specificity in this context. However, after applying SMOTE, Random Forest model correctly identified 62 out of 67 actual benign cases and 46 out of 47 actual malignant cases. The number of correct predictions of the actual benign cases dropped by 4 while the number of incorrect predictions increased by 4. The without SMOTE confusion matrix for Decision Tree classifier successfully predicted 63 cases as benign when they were indeed benign (True Negative), and it correctly identified 45 cases as malignant that were actually malignant (True Positive). However, there were 2 cases where malignant tumors were incorrectly predicted as benign (False Negative), and 4 cases where benign tumors were mistakenly classified as malignant (False Positive). These results suggest that while the decision tree classifier is quite accurate, attention should be paid to the False Negatives due to the critical nature of early and correct diagnosis in breast cancer treatment. Likewise, after the application of SMOTE, decision tree classifier model correctly identified 59 out of 67 actual benign cases and 44 out of 47 actual malignant cases. The number of correct predictions of the actual benign cases dropped by 4 while the number of incorrect predictions of actual benign cases increased by 4. Similar trends were observed by the confusion matrix of the logistic classifier and SVM (linear) but with additional 1 and 3 correct predictions for the malignant class respectively. Both SVM (polynomial) and SVM (RBF) also share similar characteristics with actual benign cases decreasing by 2 while the number of incorrect predictions of actual benign cases increasing by 2. The number of TP and FN remains the same for both SMOTE and without SMOTE. For the DNN classifier, the number of TN and FP remains the same for both SMOTE and without SMOTE. While the number of FN increased by 1, the number of TP decreased by 1.

4.7 Machine Learning Models Performance Evaluation

Table 4: Model Comparison Results For With and Without SMOTE

Models	Accuracy	Precision	Recall	Specificity	F1-Score
Without SMOTE (Testing)					
Random Forest	**0.982**	**0.979**	**0.979**	**0.985**	**0.979**
Logistic Classifier	**0.974**	**0.978**	**0.957**	**0.985**	**0.967**
Decision Tree	0.947	0.918	0.957	0.940	0.938
SVM (Linear)	0.930	0.976	0.851	0.985	0.909
SVM (Polynomial)	0.956	0.938	0.957	0.955	0.947
SVM (RBF)	0.921	0.880	0.936	0.910	0.907
DNN	0.912	0.849	0.957	0.881	0.900
With SMOTE (Testing)					
Random Forest	**0.947**	**0.979**	**0.979**	**0.925**	**0.979**
Logistic Classifier	**0.947**	**0.979**	**0.979**	**0.925**	**0.979**
Decision Tree	0.904	0.846	0.936	0.881	0.889
SVM (Linear)	0.930	0.915	0.915	0.940	0.915
SVM (Polynomial)	0.939	0.900	0.957	0.925	0.928
SVM (RBF)	0.904	0.846	0.936	0.881	0.889
DNN	0.904	0.846	0.936	0.881	0.889

The Random Forest classifier performed very well on the original unbalanced class with good model evaluation metrics but after the class was balanced using SMOTE, both the Random Forest classifier and the Logistic regression models performed above the other models with equal accuracies of 94.70%. We therefore concluded that both Random Forest and Logistic classifiers were the most appropriate models for breast cancer data classification. This decision is supported by the model's superior performance across metrics, including precision, recall, specificity, F1-score and AUC. The high model evaluation metrics of both the Random Forest model and the Logistic before and after the application of SMOTE indicate that these models can correctly classify the majority of the cases in the given dataset. The precision and recall values show the model's strength in minimizing false positives and false negatives, which is crucial in medical diagnostic applications. Moreover, the high F1-score suggests a balanced classification performance for both classes, (benign and malignant) as it is to be noted that accuracy is not a good measure for unbalanced data classification.

5 Models Comparison with ROC Curves

Here, we compare the AUC values of the classification models.

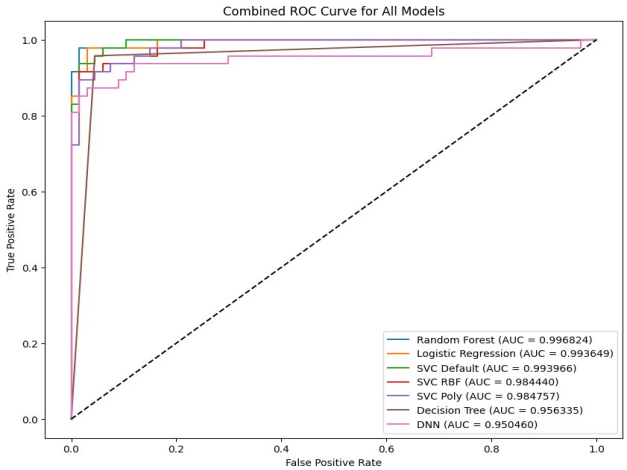


Figure 10: Comparison of AUC for All Models

The Receiver Operating Characteristic (ROC) curve, a graphical plot illustrating the diagnostic ability of the breast cancer data is presented in Figure 10. It shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. The area under the curve (AUC) provides a measure of the model's ability to distinguish between the malignant and benign classes. In Figure 10, we observe the following AUC scores: Random Forest has an AUC of approximately 0.9968, Logistic classifier has 0.9936, SVC with linear kernel has 0.9939, SVC with RBF kernel is at 0.9844, SVC with Polynomial kernel has 0.9847, the Decision Tree model has 0.9563, and the Deep Neural Network (DNN) is at 0.9504. All models significantly outperform the random chance classifier, which is represented by the diagonal black dashed line with an AUC of 0.5. The ROC curves are closer to the top-left corner of the plot, which indicates a high true positive rate and a low false positive rate, suggesting that the models have a strong discriminatory power for the positive class. The proximity of these curves and their AUC scores close to 1.0 reflect the excellent predictive power of the models on the given breast cancer data. The Random Forest and SVC with default parameters marginally outperform the other models, indicating their superiority in this specific classification task.

6 Conclusion

The application of machine learning for breast cancer classification over the past few years has yielded significant insights into the efficacy of various algorithms in distinguishing between benign and malignant tumors. By employing five prevalent machine learning models such as Logistic regression, Support Vector Machines (SVM), Random Forest, DNN and Decision Tree classifier on the breast cancer data, this study not only reinforced the capability of these models in clinical settings but also highlighted the superior performance of the Random Forest classifier. The Random Forest model's good model evaluation metrics for the unbalanced class underscore its robustness, driven by its ability to manage the complexity through the reduction of features to seventeen optimal predictors. This feature reduction, achieved via cross-validation, facilitated a more streamlined model that outperformed its counterparts, Logistic Regression and SVM (polynomial). The Logistic model followed closely demonstrating its effectiveness, particularly in terms of sensitivity and specificity. Meanwhile, the rest of the models, though slightly lagging in some evaluation statistics still proved to be a viable option for cancer classification. However, after applying SMOTE to balance the data, both Random Forest and Logistic Regression emerged as the best models with equal accuracies of 94.70%. Random Forest and Logistic Regression models performed outstandingly with an area under the curve (AUC) value of 0.997 and 0.994 respectively. The grid search methodology was used to fine-tune the Random Forest model further by determining an optimal maximum tree depth of five, enhancing the model's predictive accuracy and making it the most suited model for clinical application. The success of these models, particularly the Random Forest and Logistic Regression, not only confirms their potential in supporting diagnostic processes but also suggests a pathway for future research where these models could be integrated into real-world clinical workflows to augment the diagnostic capabilities of medical practitioners. In conclusion, this study has shown how machine learning presents itself as a promising avenue for enhancing the accuracy, efficiency, and reliability of cancer diagnostics. Future studies may focus on integrating these models with other diagnostic tools and technologies to provide a holistic diagnostic framework, potentially increasing the survival rates and improving patient outcomes through early and accurate detection. Further research should also investigate the model's applicability in a clinical environment and explore the potential of combining the classifier with other modalities, such as imaging data, to enhance diagnostic accuracy. Additionally, researchers should aim to develop methods that improve the interpretability of the Random Forest classifier without compromising its performance.

Declaration of competing interest.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement
The authors acknowledge the invaluable contributions of the anonymous reviewers and editors, whose insightful comments greatly enriched this work. Also, the authors acknowledge the enormous support of the University of Texas Rio Grande Valley (UTRGV) Presidential Research Fellowship and Deans's Graduate Research Assistantship fund.

Data Availability

The data used to support the findings of this study are available on Kaggle and can be assessed at https: //www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

References

[1] Louise Wilkinson and Toral Gathani (2022). Understanding breast cancer as a global health concern. *The British journal of radiology*, 95(1130):20211033.

[2] Redhwan Ahmed Al-Naggar (2014). Principles and practice of cancer prevention and control. *OMICS International: USA*.

[3] Lance A Liotta (1992). Cancer cell invasion and metastasis. *Scientific American*, 266(2):54–63.

[4] Stella Winters, Charmaine Martin, Daniel Murphy, and Navkiran K Shokar (2017). Breast cancer epidemiology, prevention, and screening. *Progress in molecular biology and translational science*, 151:1–32.

[5] Judah Folkman (1976). The vascularization of tumors. *Scientific American*, 234(5):58–73.

[6] Anupam Saini, Manish Kumar, Shailendra Bhatt, Vipin Saini, and Anuj Malik (2020). Cancer causes and treatments. *Int. J. Pharm. Sci. Res*, 11:3121–3134.

[7] William G Stetler-Stevenson, Sadie Aznavoorian, and Lance A Liotta (1993). Tumor cell interactions with the extracellular matrix during invasion and metastasis. *Annual review of cell biology*, 9(1):541–573.

[8] Meredith V Brown, Jonathan E McDunn, Philip R Gunst, Elizabeth M Smith, Michael V Milburn, Dean A Troyer, and Kay A Lawton (2012). Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. *Genome medicine*, 4:1–12.

[9] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190.

[10] Adel S Assiri, Saima Nazir, and Sergio A Velastin (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6):39.

[11] Fabiano Teixeira, João Luis Zeni Montenegro, Cristiano Andre Da Costa, and Rodrigo da Rosa Righi (2019). An analysis of machine learning classifiers in breast cancer diagnosis. In *2019 XLV Latin American computing conference (CLEI)*, pages 1–10. IEEE.

[12] Abdur Rasool, Chayut Bunterngchit, Luo Tiejian, Md Ruhul Islam, Qiang Qu, and Qingshan Jiang (2022). Improved machine learning-based predictive models for breast cancer diagnosis. *International journal of environmental research and public health*, 19(6):3211.

[13] Junho Lee, Wu Wang, Fouzi Harrou, and Ying Sun (2020). Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management*, 208:112582.

[14] Moloud Abdar and Vladimir Makarenkov. Cwv-bann-svm ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement*, 146:557–570, 2019.

[15] Kazi Md Rokibul Alam, Nazmul Siddique, and Hojjat Adeli (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, 32(12):8675–8690.

[16] Ahmed Hamza Osman and Hani Moetque Abdullah Aljahdali (2020). An effective of ensemble boosting learning method for breast cancer virtual screening using neural network model. *IEEE Access*, 8:39165–39174.

[17] Joanna Didkowska, Klaudia Barańska, Marta Julia Miklewska, and Urszula Wojciechowska (2024). Cancer incidence and mortality in poland in 2023. *Nowotwory. Journal of Oncology*.

[18] Dina Elreedy and Amir F Atiya (2019). A comprehensive analysis of synthetic minority

oversampling technique (smote) for handling class imbalance. *Information Sciences*, 505:32–64.

[19] Vahid Asghari, Yat Fai Leung, and Shu-Chien Hsu (2020). Deep neural network-based framework for complex correlations in engineering metrics. *Advanced Engineering Informatics*, 44:101058.