BioMedInformatics

MDPI

Article

Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study

Emilija Strelcenia * and Simant Prakoonwit

Department of Creative Technology, Bournemouth University, Fern Barrow, Poole BH12 5BB, UK; sprakoonwit@bournemouth.ac.uk

* Correspondence: estrelcenia@gmail.com

Abstract: Breast cancer is among the most common cancers found in women, causing cancer-related deaths and making it a severe public health issue. Early prediction of breast cancer can increase the chances of survival and promote early medical treatment. Moreover, the accurate classification of benign cases can prevent cancer patients from undergoing unnecessary treatments. Therefore, the accurate and early diagnosis of breast cancer and the classification into benign or malignant classes are much-needed research topics. This paper presents an effective feature engineering method to extract and modify features from data and the effects on different classifiers using the Wisconsin Breast Cancer Diagnosis Dataset. We then use the feature to compare six popular machine-learning models for classification. The models compared were Logistic Regression, Random Forest, Decision Tree, K-Neighbors, Multi-Layer Perception (MLP), and XGBoost. The results showed that the Decision Tree model, when applied to the proposed feature engineering, was the best performing, achieving an average accuracy of 98.64%.

Keywords: breast cancer; classification; machine learning; class imbalance issue; neural networks; breast cancer dataset

Citation: Strelcenia, E.; Prakoonwit, S. Effective Feature Engineering and Classification of Breast Cancer Diagnosis: A Comparative Study. *BioMedInformatics* **2023**, *3*, 616–631. https://doi.org/10.3390/biomedinformatics3030042

Academic Editors: Hans Binder and Alexandre G. De Brevern

Received: 18 May 2023
Revised: 6 June 2023
Accepted: 25 July 2023
Published: 2 August 2023

1. Introduction

Breast cancer is a type of cancer among women caused by ecological risk factors and genetic interplay. This type of cancer is caused by irregular patterns of cells in breast tissue, which creates tumours. Tumours can be both malignant and benign, where benign are not cancerous while malignant are cancerous [1,2]. A statistical report published by the International Agency for Research on Cancer (IARC) in 2020 reported that breast cancer has now surpassed lung cancer as the most commonly diagnosed cancer [3]. Similarly, the World Health Organization (WHO) in its report stated that there were 685,000 deaths related to breast cancer and 2.3 million women were diagnosed with breast cancer in 2020 alone [4].

Breast cancer diagnosis is categorised into three types: biopsy, mammography, and physical examination. Among these diagnostic methods, mammography is the most common type, but professional radiologists must interpret the tests. However, one shortcoming is that different radiologists have different inferences for the same mammogram, resulting in multiple interpretations [5]. Moreover, the accuracy rate of mammography is 65% to 78%. A biopsy is performed to measure breast cancer malignancy when mammography distinguishes a tumour. It is imperative to mention that the accuracy rate of biopsy is almost 100%, but it is time-consuming, painful, aggressive, and costly. Due to these problems, doctors may find it difficult to determine whether a tumour is malignant or benign. For these reasons, machine-learning methods can play a significant role in diagnosis [2].

In recent years, machine-learning (ML) algorithms have been used in healthcare systems, mostly for the diagnosis of breast cancer [6]. In the past, a patient's diagnostic accuracy is depended on the physician's expertise. This experience of a physician is built

over many years of observations of a patient's symptoms. Still, the accuracy cannot be reliable. With the arrival of computing techniques, acquiring and storing data has become easier. Intelligent healthcare systems are thus reliable and valuable domain. These systems can help physicians, and physicians diagnose patients with accurate and meaningful benchmarks. Moreover, these advances can help individuals plan their future health conditions. In this way, machine-learning methods can control the difficult manual work of healthcare professionals [7,8].

Computer-aided breast cancer detection techniques generally classify patients into two classes: benign class (non-cancer patients) and malignant class (patients with cancer). Various intelligent techniques have been introduced to classify data, where some techniques include feature selection approaches, and others perform classification without feature selection [9]. In [10], authors introduced a novel data mining method to accurately predict breast cancer (BC). The study aimed to develop an automated Expert System (ES) to offer an effective diagnosis of breast cancer. Therefore, the authors implemented Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) to examine breast cancer data. The study used Wisconsin Breast Cancer. In their first experiment, they tested the SVM technique using multiple values. They observed that adjusting regularisation parameters could significantly enhance the performance of conventional SVMs employed for breast cancer detection. The accuracy rate in the first experiment was 99.71%. In the next experiment, they conducted a novel breast cancer method based on two ensemble methods. They named their model CWV-BANN-SVM as they combined boosting ANN and two SVM algorithms. The study used well-known metrics, such as F1 score, AUC, Accuracy, FNR, FPR, and Gini. Their model reached an accuracy of 100%. A more recent study [11] developed a novel ensemble-based framework called Meta-Health Stack to envisage breast cancer efficiently. The novel framework Meta-Health Stack is comprised of two parts: feature selection and classification. In the first section, the Extra Trees classifier was used in their framework to extract the most appropriate features and to combine the attributes acquired from Information Gain, Pearson's Correlation, and Variance Inflation Factor to detect hidden patterns of the tumour. In the second section, the study combined three methods, Voting, Bagging, and Boosting, with the same weights through the stacking method. The findings of their study suggest that the proposed approach performed well when checked on the breast cancer dataset. The proposed approach reached a precision of 98% and resulted in a 97% F1 score when checked on the Wisconsin Breast Cancer (WDBC) dataset. The study offers worthy contributions to the breast cancer domain as this method considers various factors including tumour characteristics, medical history, and genetic testing to develop personalised treatment plans for patients. Moreover, the study utilises a stack of technologies that includes machine learning, patient data analysis, and genetics. By doing so the study aimed to overcome the shortcomings of conventional techniques to diagnose breast cancer. On the other hand, the study has a few limitations as well. For instance, the study considered only one case study, which limits the generalisation of the findings. In addition, the addition of multiple methods and technologies can increase the overall costs associated with breast cancer care. This may limit the accessibility of this proposed approach to a certain group of patients. Moreover, it is unclear whether the method will have long-term advantages for breast cancer patients or not. Machine-learning methods can learn from previous data and enhance data accuracy, thus leading to improved prediction and early detection. This is particularly crucial for diagnosing breast cancer, as early detection can increase the chances of successful treatment. For the above reasons, we agree that machine-learning techniques play an important role in breast cancer classification and early detection. This study presents a detailed review and comparison of the application of six popular machine-learning models in the field of breast cancer diagnosis. These models are Logistic Regression, Random Forest, Decision Tree, K-Neighbors, MLP, and XGBoost. It is imperative to mention that a number of classification approaches used in previous studies achieved high classification accuracy. The introduction of novel approaches is important to provide more options to the original breast cancer

datasets. Moreover, researchers argue that different classification approaches have specific advantages and shortcomings. Hence, the introduction of novel approaches can further enhance the efficiency of existing approaches as well.

The main contributions of this study are summarised below:

- This study proposed the use of ML algorithms in the breast cancer domain. The study compared six popular ML algorithms: Logistic Regression, Random Forest, Decision Tree, K-Neighbors, MLP, and XGBoost, using the Wisconsin Diagnostic Breast Cancer dataset.
- The study conducted a quantitative comparison of six classification methods.

2. Previous Works

In this section, the study reviews the existing literature on the classification of breast cancer data domain. Most of the reviewed works focused on the classification techniques, while some focused on the feature selection phase.

The study in [12] compares classification algorithms for breast cancer diagnosis. The study used several deep learning algorithms to detect breast cancer and classify breast cancer types with activation functions: Rectifier, Tanh, Exprectifier, and Maxout. Moreover, machine-learning algorithms, such as Support Vector Machine, Decision Tree, Naïve Bayes, Vote (SVM, DT, and NB), AdaBoost, and Random Forest, were compared for breast cancer based on tumour cells. The study used the Wisconsin Breast Cancer dataset and Rapidminer, a machine-learning tool. The findings show that a high accuracy of 96.99% was achieved with deep learning by the Exprectifier activation function. The high accuracy rate indicates that it is a promising method to classify various types of breast cancer datasets accurately. Moreover, the study explored the robustness of their technique noise and variations and noted that deep learning methods are highly resilient and can classify the cells accurately. The findings indicate that machine-learning methods, specifically those utilising the Exprectifier activation function, are able to revolutionise the diagnosis and treatment of breast cancer. In addition, this study offers a deep insight into the application of deep learning methods for breast cancer classification. However, the study has a few limitations, which cannot be ignored. For instance, detailed information about the framework and configuration of the techniques used in their study is missing. This would have helped readers to understand the technical aspects of this study.

Similarly, Ref. [2] introduced exploratory data methods and proposed four predictive methods to enhance breast cancer diagnosis. The study delved deep into four-layered data exploratory techniques to identify the feature classification of enhanced into benign class and malignant class. The Breast Cancer Coimbra Dataset (BCCD) and Wisconsin Diagnostic Breast Cancer (WDBC) datasets were used to check the proposed classifiers' performance and methods' performance. Moreover, the study applied performance metrics such as K-fold cross-validation and confusion matrices to check each classifier's efficiency and training time. The findings show that exploratory data techniques improved the overall performance as SVM attained 99.3%, Logistic Regression with 98.06%, and KNN achieved 97.35% accuracy with the WDBC dataset. The implementation process and results can help physicians adopt an effective method to understand and prognose breast cancer tumours. The high accuracy rates of the proposed approach have the potential to reduce false negatives and false positives, thus leading to advanced patient outcomes. Moreover, the findings of this study show that the proposed data exploratory technique outperforms conventional methods to diagnose breast cancer. The model can also be used for breast cancer screening in asymptomatic women, which can facilitate early detection and treatment. It is imperative to mention that additional research is required to validate the approach in much larger and more diverse datasets. Moreover, the study is unable to provide the precise reason for malignant features, which requires a domain expert.

In [13], a combination of multiple classifiers was presented. The study investigates the utilization of various classifiers in breast cancer diagnosis on three benchmark datasets. These classifiers include Multi-Layer Perception (MLP), J48 Decision Tree, Naïve Bayes (NB),

K-Nearest Neighbor, and Sequential Minimal Optimization (SMO). Different combinations were used to determine the best combination of these classifiers on WDBC, WBCD, and WPBC benchmark databases using confusion matrix and classification accuracy. The study evaluated these classifiers based on classification accuracy and confusion matrix, by employing a 10-fold cross-validation technique. The study also introduced a fusion at the classification level to point out the most appropriate multi-classifier method for each dataset. The findings of this study showed that the combination of the J48 Decision Tree and MLP with PCA feature selection yields superior outcomes than other classifiers. In the WDBC dataset, the study finds that using single classifiers (SMO) or fusing SMO with MLP or IBK is better than other classifiers. Finally, the fusion of MLP, J48, SMO, and IBK is superior to other classifiers in the WPBC dataset.

The study [14] compared six machine-learning frameworks, i.e., Linear Regression, GRU-SVM, Support Vector Machine, Nearest Neighbor, Softmax Regression, and Multi-Layer Perceptron. The study examined these algorithms' classification accuracy, sensitivity, and specificity on the Wisconsin Breast Cancer (WDBC) dataset. The WDBC dataset comprises features that were figured from digitalised images. Moreover, the study partitioned 70% for the training phase and 30% for the testing process, respectively. The findings of their study show that the machine-learning frameworks in the dataset performed well, as all of them exceeded 90% test accuracy. The MLP framework stood out among the compared frameworks with 99.04% test accuracy. Nevertheless, all the machine-learning approaches performed exceptionally well with accuracy exceeding 90%. The L2-SVM algorithm used in the study showed superiority over the results from a previous study that used SVM with Gaussian Radial Basis Function (RBF) as its kernel for classification. The previous study had a test accuracy of 89.28%, while the L2-SVM in this study had a test accuracy of about 96.09%. However, the L2-SVM was based on a higher training data of 10% compared to 70% in this study. The GRU-SVM algorithm had a mid-level performance with a test accuracy of 93.75%. The study confirms that all the approaches displayed better performance on the binary classification of breast cancer. Nonetheless, to further substantiate the results of the study, a cross-validation technique such as k-fold cross-validation should be employed to provide a more accurate measure of model prediction performance and assist in determining the most optimal hyper-parameters for the ML algorithms. Overall, the study demonstrates the effectiveness of machine-learning algorithms in breast cancer diagnosis.

The study in [15] explains that computer-aided detection methods based on machine learning give accurate breast cancer diagnoses. The study compared several algorithms with the help of various techniques, such as data mining methods, ensemble methods, and blood analysis. The compared algorithms are Random Forest, Naïve Bayes, Support Vector Machine, Artificial Neural Network, Decision Tree, and Nearest Neighbor on the WDBC dataset. The objective of the study was to choose the best-performing algorithm as the backend for their website. The purpose of the website is to classify cancer as malignant or benign. The proposed system involves a step-by-step process that starts with the patient booking an appointment using the website. The patient meets the doctor for the appointment and undergoes a breast mammogram or an ultrasound. The doctor then performs a manual check of the patient and detects lumps through an ultrasound. If lumps are detected, a biopsy is performed, and the digitised image of the Fine Needle Aspirate forms the features of the dataset. The numbers obtained from the biopsy will be provided to the system by the doctor, and the model will detect whether it is a benign or malignant cancer. According to the study, the purpose of this proposed system is to offer a consistent and effective technique to detect breast cancer, which can increase the accuracy of diagnosis and reduce the possibility of misdiagnosis. However, it is important to mention that the proposed method can be further improved by considering innovative features.

3. Materials and Methods

According to [13], ML approaches in the healthcare sector are gaining much attention due to the efficiency of these algorithms in prediction and classification systems, more importantly, assisting healthcare practitioners in their decisions. Other than improving patients' health-related issues, ML algorithms assist in enhancing medical studies and reducing the cost of medicines.

According to a report by Cancer Research UK, the survival rate of breast cancer is up to 100% if detected at its initial stage. However, the survival rate can be as low as 15% if detected in the latest stage [8]. More recently, machine-learning algorithms have played a key role in the diagnosis of breast cancer by utilising classification methods to spot adult women with breast cancer, discriminate malignant from benign tumours, and forecast prognosis [16]. Moreover, classification accuracy can help medical practitioners to prescribe the most effective treatment regime. In addition, machine learning is a type of Artificial Intelligence (AI) that utilises a range of optimisation, statistical, and probabilistic tools to enhance performance from new data and past incidences, exclusive of explicitly programmed commands [8]. In addition, machine-learning approaches have the ability to deal with large, high-dimensional complex data and can extract important features, which cannot be extracted using conventional statistical tools [17]. The use of machine-learning (ML) algorithms and data science in the health sector shows prolific results as such frameworks significantly assist medical practitioners [18]. The increasing trend of breast cancer cases has allowed scientists to use data that have great use in furthering clinical research. This also comes with machine learning and data science applications in this breast cancer domain [19]. Recent studies emphasised the significance of machine learning as researchers introduced the utilisation of ML algorithms to classify breast cancer using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

Researchers argue that classification is a complex optimisation problem. Researchers applied various machine-learning methods to solve the classification problem. Researchers strive to find the most efficient framework to attain the most accurate classification outcome, though the data quality can also influence the classification outcome. Moreover, the rare occurrence of data instances will also impact the number of algorithm applications. In the past, most machine-learning frameworks were tested in open-source databases. More recently, benchmark datasets have arisen in the literature. In the breast cancer domain, the Wisconsin Breast Cancer Diagnostic (WBCD) is a more commonly used breast cancer benchmark dataset.

3.1. Machine-Learning Methods Used in This Study

In this subsection, the study briefly describes the ML classification algorithms used in this research. These algorithms are Logistic Regression, Random Forest, Decision Tree, K-Neighbors, Multi-Layer Perception (MLP), and XGBoost.

3.1.1. Random Forest

Random Forest (RF) classifier is a category of ensemble learning method. It is a well-known supervised learning model that is utilised to sort out various classification issues [20,21]. Moreover, RF is an efficient ensemble that recognises non-liners data patterns. RF can handle categorical and numerical data effectively [22]. In addition, the RF method can handle issues, such as over-fitting. It is one of the most powerful methods for classification, recognition of patterns, etc.

3.1.2. Logistic Regression (LR)

Logistic regression, in terms of statistics, is used to solve binary classification issues to model events and classes probabilistically. It is a statistical method used to model binary classification challenges using logistic functions [5]. One of the assumptions of LR is that the data follow the linear function. It uses a sigmoid function to model the data [23].

3.1.3. Decision Tree (DT)

The Decision Tree method conceives things just like humans, thus making it easier and more popular to understand the inputs with a reasonable interpretation of the problems [24]. In this ML framework, a decision tree signifies a tree, and its nodes denote the traits. Moreover, the decision tree links denote a decision rule, and the leaf nodes signify an output class [25]. The total size denotes the number of nodes of the tree [26].

3.1.4. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor is one of the simplest classification methods. In this method, the training samples are referred to as Nearest Neighbors [27]. Moreover, the class labels of the test sample of the K-Neighbors decide the classification of the test sample. The value of k is important and must be sensibly chosen if the k value is too small, then the classifier may suffer the over-fitting issue due to noise in training data. Moreover, when the k value is too large, the issue of misclassification may occur as a classifier [28].

3.1.5. The Multi-Layer Perceptron (MLP)

The MLP algorithm is a feed-forward back-propagation network, a popular Neural Network (NN) method. It is a popular supervised learning algorithm that consists of input and output layers and single or multiple hidden layers which extract important information during learning and assign modifiable weighting coefficients to the components of the input layers [29,30].

3.1.6. Extreme Gradient Boosting (XGBoost)

The XGBoost is a high-scalability decision tree ensemble based on gradient boosting. XGBoost minimises the loss function to attain an additive expansion of the objective function [30]. The XGBoost algorithm has shown great classification results. It is one of the most effective algorithms for data classification.

3.2. Experimental Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is a benchmark dataset, publicly available in UCI machine-learning Repository [31] that contains details about breast cancer tumours. This dataset was originally collected by William H. Wolberg at the University of Wisconsin Hospital, Madison, in the early 1990s. Since then, several classification methods have been applied to analyse this dataset. The machine-learning frameworks were trained on breast cancer detection using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The dataset contains 569 instances: 357 benign and 212 malignant, each representing a tumour sample. There are 30 features, which are numerical measures of the characteristics of the cell nuclei present in the sample, including mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, and their standard errors, and worst values. The target variable is the diagnosis, which can be either benign (non-cancerous) or malignant (cancerous), indicated by the values 0 and 1, respectively.

4. Experimental Setup

This section deliberates on the dataset description, data analysis, confusion, and evaluation matrices for this research work.

Data Analysing and Preprocessing

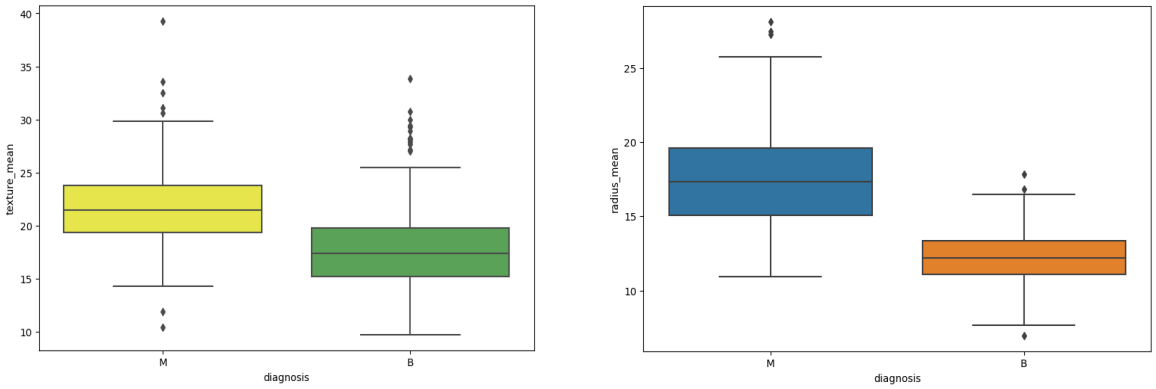
Data analysis is a process of inspecting, cleansing, transforming, and modelling data to discover useful information, inform conclusions, and support decision-making. This step is important to improve the classification accuracy. This step also involves useless columns. In our study, we found that two columns, 'Unnamed: 32' and 'id', contained irrelevant and redundant information. These columns were subsequently removed from our dataset in

order to clean our dataset and improve overall data quality and ensure greater accuracy. In addition, another important step in the data analysis is analysing each variable separately.

Analysing Our Target Feature

In this study, our target variable has only two classes: M and B. Here, M represents malignant cases, and on the other hand, B represents benign cases, respectively. It is imperative to mention that the dataset used in this study is imbalanced, i.e., there are more benign cases in the dataset than malignant cases. Therefore, we kept the same ratio while splitting our dataset into training and testing sets. Upon counting the unique values in the 'diagnosis' column, we found that there are 357 instances of benign diagnoses and 212 instances of malignant diagnoses. This indicates that there are more cases of benign diagnoses in the dataset than malignant ones. Understanding the distribution of diagnoses in the dataset is important as it can provide insights into the prevalence and severity of the condition being studied. In this case, the data suggest that most of the cases in the dataset are benign, but there is still a significant number of malignant cases that also need to be considered.

As we can see from the summary statistics, malignant tumours are larger in size compared to benign tumours. Furthermore, most benign tumours have a smaller radius than malignant tumours. This information is further supported by the boxplots shown in Figure 1a,b. The mean radius of malignant cancer cells is greater than the mean radius of benign cancer cells, indicating that malignant cancer cells are indeed larger than benign ones. Furthermore, the variance and standard deviation of malignant cancer cells are higher than that of benign cancer cells, implying that their size can vary significantly. This further emphasises the fact that malignant tumours are larger than benign tumours. It is clear from these findings that size plays an important role in classifying a cancer cell as either malignant or benign. The density plot allows us to reveal feature distributions. As shown in Figure 2, the distribution of all numerical features is consistent. All numerical variables have a clear leftward skew.



(**a**) (**b**)

Figure 1. This figure shows texture mean and radius mean for both cases: (**a**) box plot for radius mean and (**b**) box plot for texture mean.

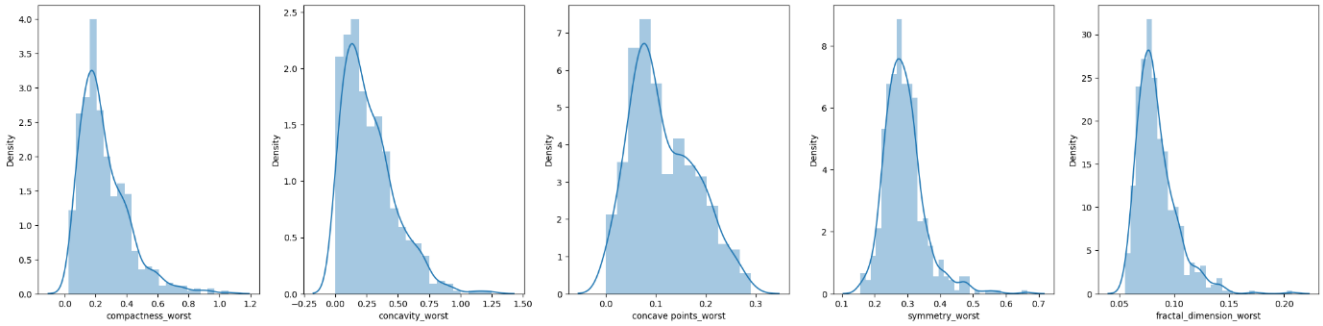


Figure 2. *Cont.*

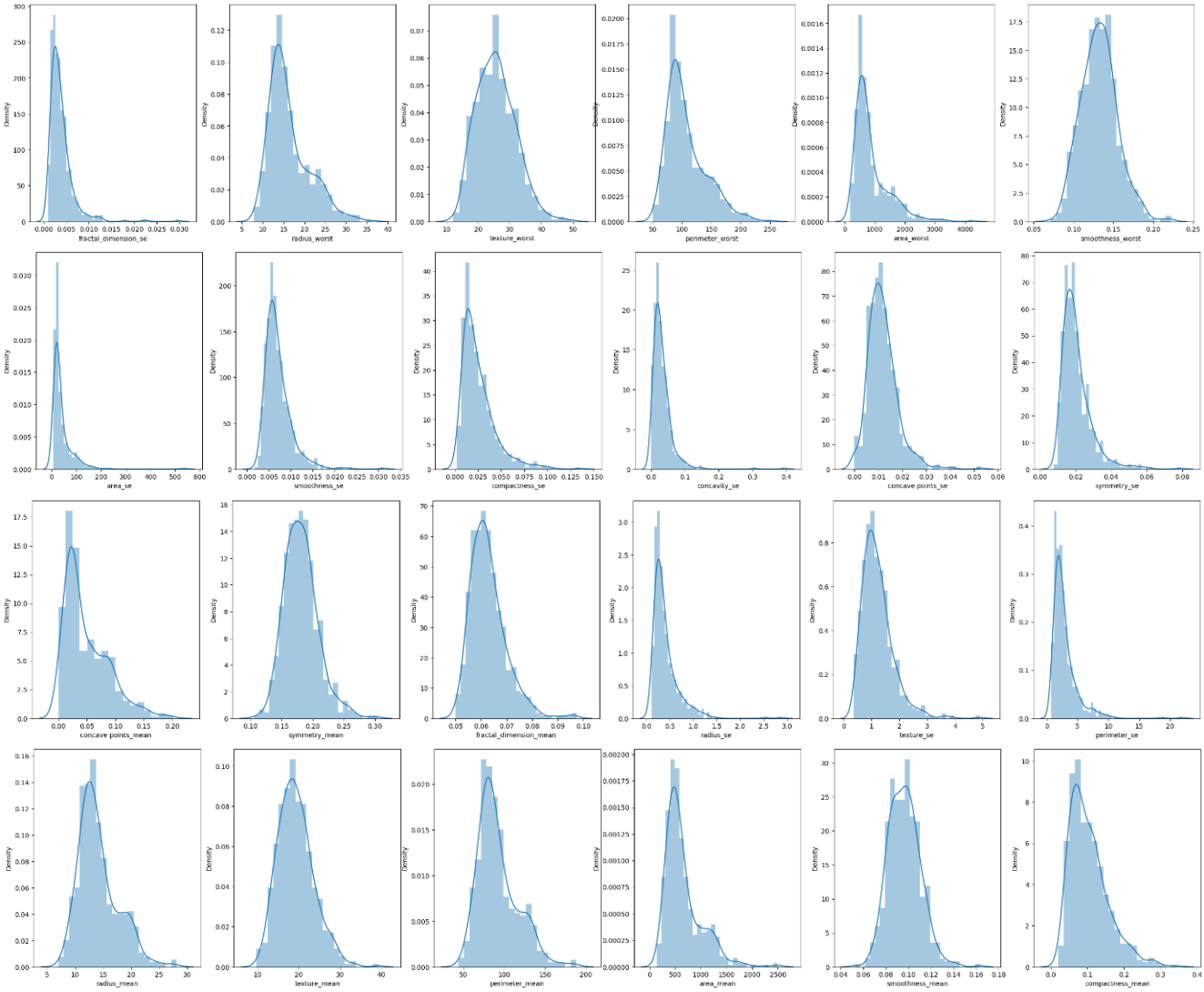


Figure 2. Distribution of all numerical features.

Figure 3 presents a correlation matrix for all features, also known as a heatmap matrix. The correlation coefficient can vary from −1 to 1. Moreover, the correlation value nearer to 1 shows that the features are highly correlated and indicates that all features positively depend on each other. On the contrary, a correlation value closer to 0 signifies that the features are not dependent on one another and that the correlation is perfect. Correlation measures the strength of the relationship between variables. In our dataset, only a few

columns have a negative correlation with the 'diagnosis' column, while around half of the columns have a correlation of over 50% with the 'diagnosis' column.

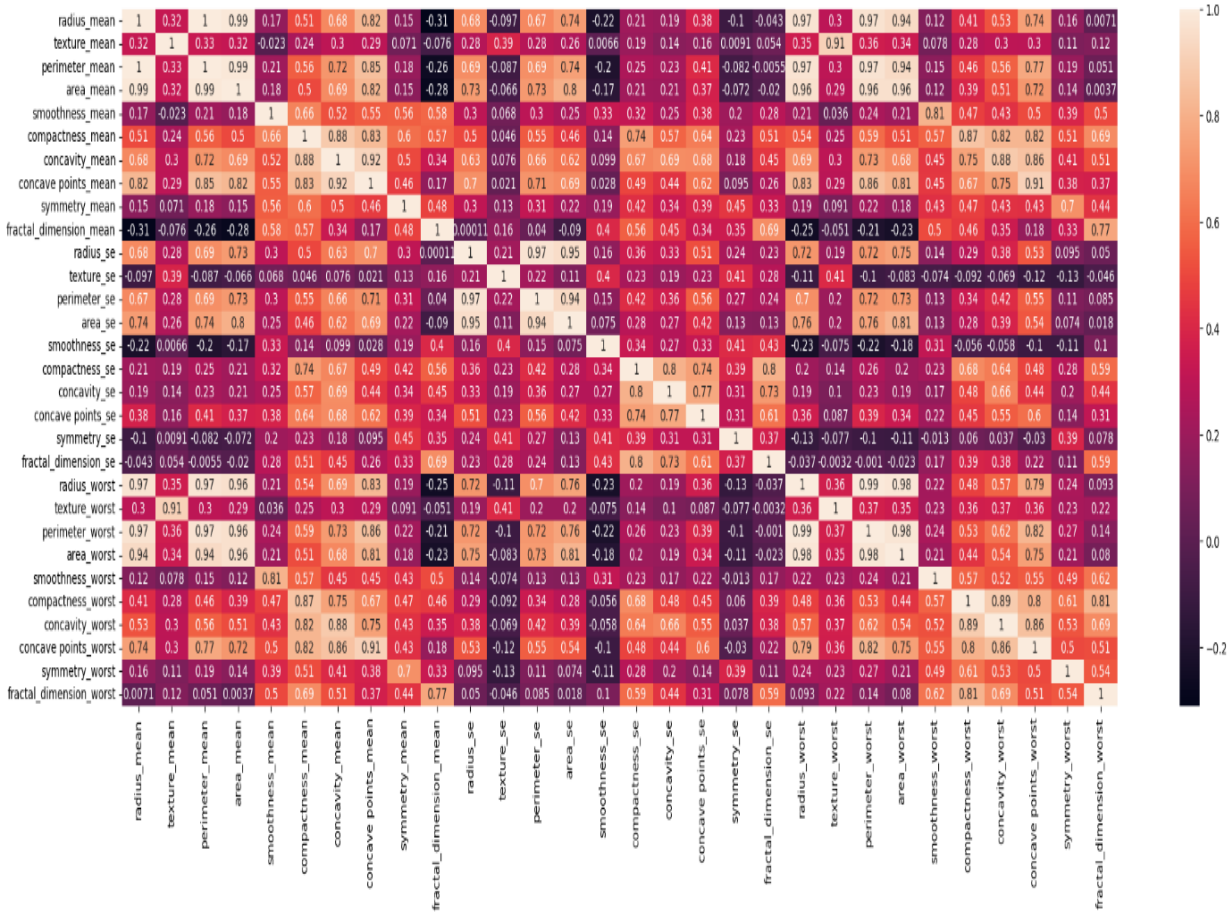


Figure 3. Heatmap matrix for all selected features.

5. Experimental Results

In this section, the findings are analysed for several classification methods which are used in this study. Our study used six machine-learning techniques for effective feature extraction and classification of breast cancer diagnosis. These methods are Logistic Regression, Random Forest, Decision Tree, K-Neighbors Classifier, Multi-Layer Perception (MLP) Classifier, and XGBoost. The findings are analysed using the confusion matrix. The dataset is divided into two parts: 80% for the training phase and 20% for the testing process.

5.1. Model Building and Performance Evaluation

5.1.1. Logistic Regression (LR)

The performance analysis of Logistic Regression is provided in Table 1. It is clear that the LR method gave an accuracy of 0.98%. Furthermore, the method achieved a precision of 0.96% for 0 and 0.97% for 1. Similarly, the experimental findings show a recall rate of 0.99% for 0 and 0.93% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 1. LR method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.96	0.99	0.97	72
1	0.97	0.93	0.95	42
Accuracy			0.98	114
Macro Avg.	0.98	0.98	0.98	114
Weighted Avg.	0.98	0.98	0.98	114

Table 1 presents 72 benign tumours, and the LR algorithm predicted 71 correctly. Out of 72 benign tumours, the algorithm considers that 1 case is malignant, but it is actually benign. In addition, there were 42 malignant tumour data, and the LR algorithm predicted 39 correctly. Moreover, the LR algorithm considers that three cases are benign, but actually, these cases are malignant.

5.1.2. Decision Tree

The performance analysis of the Decision Tree method is given in Table 2. The Decision Tree method gave 0.98% accuracy. It is clear that the Decision Tree method achieved a precision of 0.99% for 0 and 0.98% for 1. Similarly, the experimental findings show a recall rate of 0.99% for 0 and 0.98% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 2. DC method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.99	0.99	0.99	72
1	0.98	0.98	0.98	42
Accuracy			0.98	114
Marco Avg.	0.98	0.98	0.98	114
Weighted Avg.	0.98	0.98	0.98	114

Table 2 presents 72 benign tumours, and the DT algorithm predicted 71 correctly. Out of 72 benign tumours, the algorithm considers that 1 case is malignant, but it is actually benign. In addition, there were 42 malignant tumour data, and the DT algorithm predicted 41 correctly. Moreover, the LR algorithm considers that one case is benign, but the case is actually malignant.

5.1.3. Random Forest

The performance analysis of the Random Forest method is given in Table 3. The Random Forest method gave 97% accuracy. It is clear that the Random Forest method achieved a precision of 0.96% for 0 and 1.0% for 1. Similarly, the experimental findings show a recall rate of 1.00% for 0 and 0.93% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 3. RF method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.96	1.00	0.98	72
1	1.00	0.93	0.96	42
Accuracy			0.97	114
Macro Avg.	0.98	0.96	0.97	114
Weighted Avg.	0.97	0.97	0.97	114

In the above table, there were 72 benign tumours, and the RF algorithm predicted all of them correctly. In addition, there were 42 malignant tumour data, and the RF algorithm

predicted 39 correctly. Moreover, the LR algorithm considers that three cases are benign, but these cases are actually malignant.

5.1.4. K-Neighbors

The performance analysis of the K-Neighbors is given in Table 4. The K-Neighbor Classifier gave 0.89% accuracy. It is clear that the K-Neighbor Classifier achieved a precision of 0.86% for 0 and 1.00% for 1. Similarly, the experimental findings show a recall rate of 1.00% for 0 and 0.71% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 4. KN method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.86	1.00	0.92	72
1	1.00	0.71	0.83	42
Accuracy			0.89	114
Macro Avg.	0.93	0.86	0.88	114
Weighted Avg.	0.91	0.89	0.89	114

In the above table, there were 72 benign tumours, and the K-Neighbors Classifier predicted all of them correctly. In addition, there were 42 malignant tumour data, and the RF algorithm predicted 30 correctly. Moreover, the LR algorithm considers that 12 cases are benign, but these cases are malignant.

5.1.5. MLP

The performance analysis of the MLP is given in Table 5. The MLP classifier gave 0.92% accuracy. It is clear that the MLP Classifier achieved a precision of 0.90% for 0 and 0.97% for 1. Similarly, the experimental findings show a recall rate of 0.99% for 0 and 0.81% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 5. MLP method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.90	0.99	0.94	72
1	0.97	0.81	0.88	42
Accuracy			0.92	114
Marco Avg.	0.94	0.90	0.91	114
Weighted Avg.	0.93	0.92	0.92	114

The above table showed 72 benign tumours, and the MLP Classifier predicted 71 correctly. Out of 72 benign tumours, the algorithm considers that 1 case is malignant, but actually, it is benign. In addition, there were 42 malignant tumour data, and the MLP Classifier predicted 34 correctly. Moreover, the MLP Classifier considers that eight cases are benign, but these cases are actually malignant.

5.1.6. XGBoost

The performance analysis of the XGBoost is given in Table 6. The XGBoost Classifier gave 0.94% accuracy. It is clear that the XGBoost Classifier achieved a precision of 0.92% for 0 and 0.97% for 1. Similarly, the experimental findings show a recall rate of 0.99% for 0 and 0.86% for 1. The table also presents the accuracy, macro average, weighted average, and support scores.

Table 6. XGBoost method output based on precision, recall, and F1 score.

	Precision	Recall	F1 Score	Support
0	0.92	0.99	0.95	72
1	0.97	0.86	0.91	42
Accuracy			0.94	114
Macro Avg.	0.95	0.92	0.93	114
Weighted Avg.	0.94	0.94	0.94	114

The above table showed 72 benign tumours and the MLP Classifier predicted 71 correctly. Out of 72 benign tumours, the algorithm considers that 1 case is malignant, but it is benign. In addition, there were 42 malignant tumour data, and the XGBoost Classifier predicted 36 correctly. Moreover, the MLP Classifier considers that six cases are benign, but these cases are actually malignant.

5.2. Comparative Analysis of Classifiers

In the previous subsection, the study presented a performance analysis of individual classifiers. In this subsection, the study presents a comparative analysis of all the methods and classifiers on accuracy, precision, and F1 score. The following table depicts the performance analysis of these methods.

Table 7 presents the classification performance of each algorithm based on accuracy, precision, recall, and F1 score. The K-Neighbor Classifier gave the lowest accuracy rate when compared with other methods. Moreover, the decision tree method gave the highest accuracy rate. The table also depicts each method's recall, precision, and F1 scores. Overall, the decision tree method performed well with respect to all confusion matrices and other scores.

Table 7. Classification performance of each algorithm.

ML Algorithms	Confusion Matrix		Accuracy	Precision	Recall	F1 Score
	0	**1**				
Logistic Regression			0.96%			
0	71	1		0.96%	0.99%	0.97%
1	3	39		0.97%	0.93%	0.95%
Decision Tree			**0.98%**			
0	71	1		**0.99%**	0.99%	**0.99%**
1	1	41		0.98%	**0.98%**	**0.98%**
Random Forest			0.97%			
0	72	0		0.96%	**1.00%**	0.98%
1	3	39		**1.00%**	0.93%	0.96%
K-Neighbor Classifier			0.89%			
0	72	0		0.86%	1.00%	0.92%
1	12	30		**1.00%**	0.71%	0.83%
MLP Classifier			0.92%			
0	71	1		0.90%	0.99%	0.94%
1	8	34		0.97%	0.81%	0.88%
XGBoost			0.94%			
0	71	1		0.92%	0.99%	0.95%
1	6	36		0.97%	0.86%	0.91%

The bold numbers indicate the best performance of the methods.

6. Discussion

In this research, we used the WDBC dataset to examine the best machine-learning classification algorithm for effective feature extraction and classification of breast cancer diagnosis. For the purposes mentioned above, we analysed the performance of six machine-

learning techniques for effective feature engineering and classification of breast cancer diagnosis. These methods are Logistic Regression, Random Forest, Decision Tree, K-Neighbors, Multi-Layer Perception (MLP), and XGBoost. Our study suggests that the Decision Tree method was the most effective and successful method, with an accuracy value of 0.98 when we analysed it according to the settings of this study. The Random Forest method remained the second most effective and successful method, with an accuracy value of 0.97. The Random Forest was followed by the Logistic Regression method with an accuracy value of 0.96. This is followed by the XGBoost with an accuracy value of 0.94. In addition, the MLP achieved an accuracy value of 0.92%. Moreover, the study also confirmed that K-Neighbor achieved the lowest accuracy value of 0.89.

The findings of our study were mostly analysed by considering the accuracy value. However, the study also utilised cross-validation methods. These cross-validation methods are precision, recall, and F1 score. These methods were used to check the crucial values of TP, FP, TN, and FN to deal with the predicted and actual classes. They presented the precision, F1 score, and recall values to examine the performances of these ML classification algorithms. The findings show that the Decision Tree method performed better than other methods in terms of these values. This shows that the Decision Tree method successfully identified the tumour cases and classified the cancerous features as malignant.

In a 2017 study [32], the WBCD dataset was analysed using a voting classifier, an ensemble technique. This ensemble approach combines multiple models with a strategy that considers the varying predicted reliability of each classifier across different output classes. This technique combined the strengths of Support Vector Machines (SVMs), Naive Bayes, and J48 classifiers to achieve a highly impressive accuracy rate of 97.13%. Notably, this accuracy rate outperformed each classifier used in the technique. These findings offer promising insights into the potential of ensemble techniques to improve classification accuracy across various datasets. However, our study achieved better accuracy rates than this study.

Similarly, in [33], the study used four machine-learning approaches, SVM, KNN, Naïve Bayes, and Decision Tree, and evaluated their performance on two datasets. The models were trained using features selected at various threshold levels and validated using independent gene expression datasets. The results of this study indicated that the Support Vector Machine algorithm outperformed the other three algorithms in accurately classifying breast cancer into triple-negative and non-triple-negative types. The SVM method achieved an accuracy level of 73%. The study concludes that ML algorithms can be used as an effective tool for identifying the two types of breast cancer. However, compared to our study, their study achieved inferior accuracy rates.

In ref. [34], the study used the WDBC dataset to predict breast cancer accurately. The study implemented multiple machine-learning algorithms: SVM, Logistic Regression, KNN, Decision Tree, Naïve Bayes, and Random Forest. The study calculated and compared these algorithms' accuracy to determine the most suitable one. Notably, both Random Forest and Support Vector Machine classifiers outperformed other classifiers with an accuracy rate of 96.5%. The study was able to achieve higher accuracy rates for each method. However, in terms of better accuracy rates, our method outperformed the method used by their study. The findings of our study highlight the importance of feature engineering techniques on datasets to enhance prediction accuracy.

In ref. [35], the study conducted a comparison between various machine-learning approaches, such as Decision Tree, Support Vector Machine (SVM), Naïve Bayes, and KNN, on the WBDC dataset. The study's objective was to check these methods' precision, accuracy, sensitivity, and specificity to check their efficiency and effectiveness in classifying data. According to their study, the SVM approach outperformed the other algorithms with a remarkable 97.13% accuracy and the lowest error rate. However, our findings yielded insightful results when compared with the findings of [34]. However, our study outperformed in terms of better accuracy. Our objective of achieving better accuracy rates in breast cancer prediction was met when compared with the method used in this paper.

The higher the accuracy, the more reliable the algorithm makes predictions. Our study, therefore, provides valuable insights into the best machine-learning algorithm for the Wisconsin Breast Cancer dataset. Overall, the findings of this study demonstrate the importance of choosing the right algorithm for a particular dataset.

Compared to previous studies, our study gave a better performance in terms of accuracy. The objective of our study was achieved when compared with other methods in the literature in terms of a better accuracy rate in breast cancer prediction. Table 8 presents the result comparison of our study with previous studies.

Table 8. Summary of comparison of results involving the utilization of machine-learning algorithms.

References	Sampling Strategies	Highest Classification Accuracy
[32]	10-fold cross-validation	97.13%
[33]	Feature selection at different thresholds	73%
[34]	75−25 training–testing	96.5%
[35]	10-fold cross-validation	97.13%
Our study	80–20 training–testing	98%

7. Conclusions

Early detection of breast cancer is of the utmost importance, as breast cancer is one of the major causes of mortality in women. However, early detection of breast cancer can play a significant role in averting a high death rate. Recently, with the advent of advanced machine-learning classifiers, the process of detecting breast cancer tumours at an early stage has become more accurate and efficient. These classifiers use various algorithms to analyse data and identify abnormalities that may indicate the presence of breast cancer. These methods have not only improved the accuracy of diagnosis but also reduced the need for invasive procedures. Therefore, modern machine-learning techniques can play a great role in detecting breast cancer. In this study, we explain several machine-learning methods and their scope in breast cancer diagnosis. This study combined classifiers with feature selection for breast cancer diagnosis. We applied six classification algorithms to the WDBC dataset to check the classification accuracy of these algorithms. The findings of our study show that the Decision Tree model was the best-performing one, achieving an average accuracy of 98.64%.

Moving forward, there are several avenues for further research and development in the field of machine learning for breast cancer diagnosis. One potential area of focus is the integration of multiple machine-learning algorithms to improve the accuracy and reliability of breast cancer detection. In addition, further investigation into feature selection methods could help to identify the most relevant features for breast cancer diagnosis, ultimately leading to more efficient and accurate diagnoses. Additionally, exploring the use of deep learning techniques, such as convolutional neural networks, could potentially lead to even higher accuracy rates in the detection of breast cancer. Furthermore, it is important to consider the potential implications and ethical considerations of integrating machine-learning tools into clinical settings. Continued research into the impact of machine learning on patient outcomes and healthcare delivery will be essential in ensuring that these tools are used in responsible and effective ways. Overall, the findings of this study highlight the tremendous potential of machine learning in breast cancer diagnosis and underscore the need for continued research and development in this field. With further investigation and refinement, machine learning could ultimately improve the accuracy and efficiency of breast cancer diagnosis, leading to better outcomes for patients and improved public health.

Author Contributions: Conceptualization, E.S. and S.P.; methodology, E.S. and S.P.; software, E.S.; validation, E.S.; formal analysis, E.S.; investigation, E.S. and S.P.; resources, S.P.; data curation, E.S.; writing—original draft preparation, E.S.; writing—review and editing, S.P.; visualization, E.S.;

supervision, S.P.; project administration, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Bournemouth University, United Kingdom.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Breast Cancer Wisconsin (Diagnostic) Dataset | Kaggle (Breast Cancer Wisconsin (Diagnostic) Data Set).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chaurasia, V.; Pal, S.; Tiwari, B. Prediction of Benign and Malignant Breast Cancer Using Data Mining Techniques. *J. Algorithm Comput. Technol.* **2018**, *12*, 119–126. [CrossRef]
2. Rasool, A.; Bunterngchit, C.; Tiejian, L.; Islam, M.d.R.; Qu, Q.; Jiang, Q. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3211. [CrossRef] [PubMed]
3. IARC. *IARC Biennial Report 2020–2021*; IARC: Lyon, France, 2021.
4. WHO Editors World Health Organization (WHO). 12 July 2023. Breast Cancer. Available online: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 26 July 2023).
5. Khandezamin, Z.; Naderan, M.; Rashti, M.J. Detection and Classification of Breast Cancer Using Logistic Regression Feature Selection and GMDH Classifier. *J. Biomed. Inform.* **2020**, *111*, 103591. [CrossRef]
6. Karabatak, M. A New Classifier for Breast Cancer Detection Based on Naïve Bayesian. *Measurement* **2015**, *72*, 32–36. [CrossRef]
7. Meesad, P.; Yen, G.G. Combined Numerical and Linguistic Knowledge Representation and Its Application to Medical Diagnosis. *IEEE Trans. Syst. Man Cybern. Part. A Syst. Hum.* **2003**, *33*, 206–222. [CrossRef]
8. Yue, W.; Wang, Z.; Chen, H.; Payne, A.; Liu, X. Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs* **2018**, *2*, 13. [CrossRef]
9. Golatkar, A.; Anand, D.; Sethi, A. *Classification of Breast Cancer Histology Using Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 837–844.
10. Abdar, M.; Makarenkov, V. CWV-BANN-SVM Ensemble Learning Classifier for an Accurate Diagnosis of Breast Cancer. *Measurement* **2019**, *146*, 557–570. [CrossRef]
11. Samieinasab, M.; Torabzadeh, S.A.; Behnam, A.; Aghsami, A.; Jolai, F. Meta-Health Stack: A New Approach for Breast Cancer Prediction. *Healthc. Anal.* **2022**, *2*, 100010. [CrossRef]
12. Mekha, P.; Teeyasuksaet, N. Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells. In Proceedings of the 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), Nan, Thailand, 30 January–2 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 343–346.
13. Salama, G.I.; Abdelhalim, M. Magdy Abd-elghany Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. *Int. J. Comput. Inf. Technol.* **2012**, *1*, 36–43.
14. Agarap, A.F.M. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. In Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, Phu Quoc Island, Vietnam, 2–4 February 2018; ACM: New York, NY, USA, 2018; pp. 5–9.
15. Rane, N.; Sunny, J.; Kanade, R.; Devi, S. Breast Cancer Classification and Prediction Using Machine Learning. *Int. J. Eng. Res. Technol.* **2020**, *9*, 576–580. [CrossRef]
16. Suh, Y.J.; Jung, J.; Cho, B.-J. Automated Breast Cancer Detection in Digital Mammograms of Various Densities via Deep Learning. *J. Pers. Med.* **2020**, *10*, 211. [CrossRef] [PubMed]
17. Najafabadi, M.M.; Villanustre, F.; Khoshgoftaar, T.M.; Seliya, N.; Wald, R.; Muharemagic, E. Deep Learning Applications and Challenges in Big Data Analytics. *J. Big Data* **2015**, *2*, 1. [CrossRef]
18. Verma, V.K.; Verma, S. Machine Learning Applications in Healthcare Sector: An Overview. *Mater. Today Proc.* **2022**, *57*, 2144–2147. [CrossRef]
19. Bazazeh, D.; Shubair, R. Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis. In Proceedings of the 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, United Arab Emirates, 6–8 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–4.
20. Huljanah, M.; Rustam, Z.; Utama, S.; Siswantining, T. Feature Selection Using Random Forest Classifier for Predicting Prostate Cancer. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *546*, 052031. [CrossRef]
21. Jayaraj, D.; Sathiamoorthy, S. Random Forest Based Classification Model for Lung Cancer Prediction on Computer Tomography Images. In Proceedings of the 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 27–29 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 100–104.

22. Ghiasi, M.M.; Zendehboudi, S. Application of Decision Tree-Based Ensemble Learning in the Classification of Breast Cancer. *Comput. Biol. Med.* **2021**, *128*, 104089. [CrossRef]
23. Liu, L. Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning. In Proceedings of the 2018 International Conference on Robots & Intelligent System (ICRIS), Amsterdam, The Netherlands, 21–23 February 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 157–160.
24. Olanow, C.W.; Watts, R.L.; Koller, W.C. An Algorithm (Decision Tree) for the Management of Parkinson's Disease (2001): Treatment. *Neurology* **2001**, *56*, S1–S88. [CrossRef]
25. Pandya, R.; Pandya, J. C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *Int. J. Comput. Appl.* **2015**, *117*, 18–21. [CrossRef]
26. Tsang, S.; Kao, B.; Yip, K.Y.; Ho, W.-S.; Lee, S.D. Decision Trees for Uncertain Data. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 64–78. [CrossRef]
27. Al-Hadidi, M.R.; Alarabeyyat, A.; Alhanahnah, M. Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm. In Proceedings of the 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, UK, 31 August–1 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 35–39.
28. MurtiRawat, R.; Panchal, S.; Singh, V.K.; Panchal, Y. Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 534–540.
29. Desai, M.; Shah, M. An Anatomization on Breast Cancer Detection and Diagnosis Employing Multi-Layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN). *Clin. eHealth* **2021**, *4*, 1–11. [CrossRef]
30. Mahesh, T.R.; Vinoth Kumar, V.; Muthukumaran, V.; Shashikala, H.K.; Swapna, B.; Guluwadi, S. Performance Analysis of XGBoost Ensemble Methods for Survivability with the Classification of Breast Cancer. *J. Sens.* **2022**, *2022*, 4649510. [CrossRef]
31. William, H.; Wolberg, W.; Street, N.; Olvi, L. Mangasarian. In *UCI Machine Learning Repository*; School of Information and Computer Science, University of California: Irvine, CA, USA, 1995; Available online: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (accessed on 6 June 2023).
32. Kumar, U.K.; Nikhil, M.B.S.; Sumangali, K. Prediction of breast cancer using voting classifier technique. In Proceedings of the 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Chennai, India, 2–4 August 2017. [CrossRef]
33. Wu, J.; Hicks, C. Breast cancer type classification using machine learning. *J. Pers. Med.* **2021**, *11*, 61. [CrossRef] [PubMed]
34. Ara, S.; Das, A.; Dey, A. Malignant and benign breast cancer classification using machine learning algorithms. In Proceedings of the 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 5–7 April 2021; pp. 97–101.
35. Asri, H.; Mousannif, H.; Moatassime, H.A.; Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Comput. Sci.* **2016**, *83*, 1064–1069. [CrossRef]