UMYU Scientifica, Vol. 5 NO. 1, March 2026, Pp 163 – 178
https://doi.org/10.56919/usci.2651.015
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Breast Cancer Survival Rate Prediction Using Machine Learning Algorithm

Joshua Bature Hassan[1], Oparaji Nkem Anastasia[2], Alabi Orobosade Adewunmi[3], Esther Odunayo Oduntan[4], Olusegun Adeosun[3] & Daniel Dauda Wisdom[5]*

[1]Department of Computer Sciences Federal University Oye Ekiti, Ekiti State, Nigeria
[2]Faculty of Nursing, Chrisland University Abeokuta, Ogun State, Nigeria
[3]Depertment of Computer Science, Federal University of Agriculture Abeokuta, Ogun State, Nigeria
[4]Department of Computer Science, Federal University of Technology Ilaro, Ogun State, Nigeria
[5]Department of Cybersecurity Data Science, College of Computing Sciences, Federal University of Agriculture Abeokuta, Ogun State, Nigeria

ABSTRACT

Breast cancer is a formidable foe impacting numerous lives, emphasizing the critical need for accurate survival predictions to guide personalized treatment decisions. This research employed a machine learning algorithms, specifically decision tree models, for breast cancer survival prediction. The study seeks a model that deftly handles the intricacies of missing data and resonates meaningfully with the healthcare community. The study used a decision tree model, envisioning a precision virtuoso adeptly hitting the right notes, achieving a score of 0.73 for those overcoming breast cancer, a complex disease, and an admirable 0.87 for those navigating a more arduous journey. With a recall score of 0.90 for survivors, assuring that the majority are acknowledged. Simultaneously, with a score of 0.68 for those facing a more challenging path, it reveals a nuanced understanding of individuals who may not traverse it successfully. Delving into the nuanced realm of these metrics, the F1 score emerges as a meticulously choreographed dance, resonating at 0.81 for survivors and 0.76 for those facing formidable odds. This narrative isn't confined to numerical precision; it's a symphony of a model predicting with nuance. Stepping back, the overall accuracy of 0.79 was not merely a numerical outcome but the model's outstanding performance on the healthcare stage. Transcending precision, crafting a tool that converses in the language of healthcare professionals, facilitating nuanced decision-making. This research journey extended beyond numerical precision; it's an exploration into unraveling complexities and crafting a tool that transcends sterile lab origins. Envision this decision tree model not as an isolated entity but as a collaborative partner in diverse clinical environments. It's an evolving creation, refined based on real-world feedback, a symphony in harmony with the experiences of healthcare practitioners navigating the delicate terrain of breast cancer treatment. The model not only interprets data but comprehends the human stories behind it and aspires not merely to numerical accuracy but to a profound, positive impact on patient outcomes in the complex, real-world arena.

ARTICLE HISTORY

Received December 05, 2025
Accepted March 01, 2026
Published March 25, 2026

KEYWORDS

Breast_cancer, Decision_tree_model, survival_prediction, clinical_environments & heathcare_community

INTRODUCTION

Breast cancer is the most common cancer in women worldwide. It is a complex disease that requires comprehensive assessment and personalized treatment. The survival rate of breast cancer patients depends on various factors such as age, tumor size, histologic type, and stage of the disease. Early detection and accurate prediction of survival rates can significantly improve patient outcomes by enabling timely, tailored treatment. Machine learning techniques have emerged as powerful tools for predicting disease outcomes in clinical practice. In recent years, decision tree algorithms have gained popularity in medical research due to their ability to handle large and heterogeneous datasets, intuitive visualization of decision rules, and interpretability. Decision tree models can identify relevant factors and create decision rules that provide insight into the underlying mechanisms of disease progression and treatment response. Causing many deaths in the current situation.

Due to changes in food and lifestyle, the number of cancer cases in women is increasing day by day. It is the second most common cause of death in women in the world. (Meteb *et al* 2004), This uses concepts of Deep learning (DL) and Machine learning (ML) to predict breast cancer based on the data obtained. This cancer is characterized by the abnormal growth of fatty and fibrous tissues, and

Correspondence: Daniel Dauda Wisdom. Department of Cybersecurity Data Science, COLCOMPS, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria. ✉ danieldw@funaab.edu.ng.
How to cite: Hassan, J. B., Anastasia, O. N., Adewunmi, A. O., Oduntan, E. O., Adeosun, O., & Wisdom, D. D. (2026). Breast Cancer Survival Rate Prediction Using Machine Learning Algorithm. *UMYU Scientifica*, 5(1), 163 – 178. https://doi.org/10.56919/usci.2651.015

its different phases are caused by cancer cells spreading throughout the tissue (Ionescu, *et al.,* 2022). This is one of the most common cancers that affects women, but other types of cancer and those who are affected by them can be treated greatly, according to a government survey, when compared to breast cancer (Matyas *et al* 1965). The various phases of breast cancer are identified through proper treatment and detailed assessment (Jain *et al.* 2016).

If we do not provide proper therapy to our patients, it will result in their death. A number of methods for establishing an accurate diagnosis of breast cancer have been presented (Jiang *et al* 2017). Because the dataset contains a variety of distinct report attributes, machine learning may be easily applied to the dataset for prediction (Ragab, 2019). Even with technology that is not fully automatically designed to produce the output (Gouda and Abdelhalim 2012). Hence, we propose a fully automatic classification and prediction of breast cancer based on the dataset (Yildirim, 2023). Using a deep learning technique. This learning technique is recognized as the best method for predicting and classifying image datasets.

Earlier methods for classifying data were used, despite their lower accuracy, because they could be used for proper categorization and prediction. Deep learning algorithms and numerical dataset machine learning techniques are used to extract features and hidden features. The convolution value is obtained from the stride function, which extracts features from images of different sizes. CNN is one that produces proper output for the dataset we used in this study (Cortes and Vapnik 1995). Typical cancer screening procedures are grounded on the "gold-standard", which consists of three tests: clinical evaluation, radiological imaging, and
This traditional technique, based on regression, detects the presence of cancer, whereas new ML techniques and algorithms are built on model creation. During its training and testing stages, the model is intended to forecast unknown data and provide satisfactory predictions (Yildirim, 2023). Preprocessing, feature selection or extraction, and classification are the three major methodologies used in machine learning (Yildirim, 2023). The feature extraction part of the machine learning method is crucial for cancer diagnosis and prediction. This process may differentiate between benign and malignant tumors (Cristianini *et al.,* 2000). The "gold-standard" method for detecting cancer previously consisted of three parts: clinical evaluation, radiological imaging, and pathology testing (Gönen *et al* 2011).
The proposed technique indicates cancer presence based on regression, whereas newer algorithms are available. Model designed for predicting new data and should give good results in its training and testing phases (Ferroni and Roselli 2017). Here are 3 main steps: preprocessing features, extraction, and classification.
Figure 1 shows the types of breast cancers. The "gold standard" method for detecting cancer previously consisted of three parts: clinical evaluation, radiological imaging, and pathology testing. (Gönen *et al* 2011). The proposed technique indicates cancer presence based on regression, whereas newer algorithms are available. Model designed for predicting new data and should give good results in its training and testing phases (Ferroni and Roselli 2017). Here are 3 main steps: preprocessing features, extraction, and classification. Breast cancer is a prevalent and potentially life-threatening disease that affects millions of individuals worldwide. Predicting the survival rate of breast cancer patients plays a crucial role in guiding treatment decisions and improving patient outcomes. Traditional statistical models and machine learning algorithms have been used to forecast survival rates, with decision trees a prominent choice due to their interpretability and ability to handle both categorical and continuous variables. However, despite the advantages of decision tree-based models, several challenges remain to be addressed in breast cancer survival rate prediction. These challenges include: **Limited predictive accuracy:** Decision trees tend to overfit or underfit, resulting in limited predictive accuracy. Improving the model's ability to generalize and capture complex relationships among prognostic factors is essential for improving survival rate predictions. **Feature selection and variable importance:** Identifying the most relevant features or variables that significantly impact breast cancer survival rates is crucial for developing an efficient decision tree model. The identification of these factors can aid in better understanding the disease and guide healthcare practitioners in prioritizing interventions. **Handling missing data:** Medical datasets often contain missing values, which can adversely affect the performance of decision tree models. Developing effective strategies to handle missing data when building decision trees is necessary to ensure robust, accurate survival rate predictions. **Interpretable decision rules:** One of the primary advantages of decision tree models is their interpretability. However, complex decision trees can become challenging to interpret, hindering their utility in clinical settings. Ensuring that the decision tree model produces understandable and actionable rules is essential for healthcare professionals to make informed decisions. Thus, this study develops an improved decision tree-based model for breast cancer survival rate prediction that addresses the aforementioned challenges. The objective is to create a robust and interpretable model that accurately predicts survival rates, effectively selects relevant features, handles missing data, and generates understandable decision rules, ultimately assisting healthcare practitioners in making informed treatment decisions for breast cancer patients.

BREAST CANCER OVERVIEW

2.1 Breast Cancer

Cancer is a destructive disease, and breast cancer is one of the most common cancers affecting the health of women globally (Sharma and Dave 2010). Breast cancer is a type of cancer that develops in the breast tissue. It occurs when abnormal cells in the breast multiply and grow uncontrollably, forming a tumor. These tumors can invade surrounding tissues and spread to other parts of the body if left untreated. Breast cancer is the most common cancer among women worldwide and can also occur in men, although it is rare. According to the American Cancer

Society (ACS), breast cancer starts in the cells that line the ducts or lobules of the breast.

Ductal carcinoma, which begins in the milk ducts, is the most common type of breast cancer. Lobular carcinoma, which starts in the lobules, is another type. Other less common types of breast cancer include inflammatory breast cancer and Paget's disease of the breast (Islami *et al.,* 2022). *Breast cancer* is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin in different parts of the breast. A breast is made up of three main parts: lobules, ducts, and connective tissue.

The lobules are the glands that produce milk. The ducts are tubes that carry milk to the nipple. Connective tissue (which consists of fibrous and fatty tissue) surrounds and supports everything. Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized (islami *et al.,* 2022).

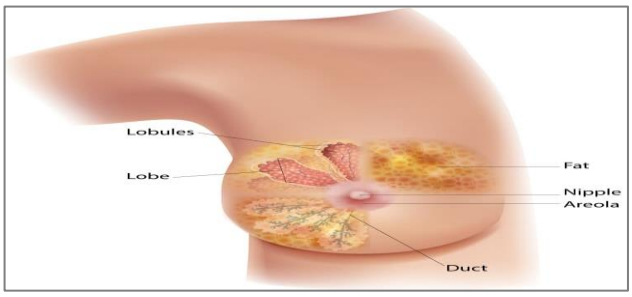

Figure 1: Breast Cells (Healthline.com)

2.2 Types of Breast Cancer

The most common types of breast cancer are:

1. ***Invasive ductal carcinoma:*** The cancer cells begin in the ducts and then grow outside the ducts into other parts of the breast tissue. Invasive cancer cells can also spread, or metastasize, to other parts of the body.

2. ***Invasive lobular carcinoma:*** Cancer cells begin in the lobules and spread from the lobules to nearby breast tissue. These invasive cancer cells can also spread to other parts of the body.

2.3 Symptoms of Breast Cancer

Different people have different symptoms of breast cancer. Some of these symptoms are:

1. New lump in the breast or underarm (armpit).

2. Thickening or swelling of part of the breast.

3. Irritation or dimpling of breast skin.

4. Redness or flaky skin in the nipple area or the breast.

5. Pulling in of the nipple or pain in the nipple area.

6. Nipple discharge other than breast milk, including blood.

7. Any change in the size or the shape of the breast.

8. Pain in any area of the breast.

2.4 Causes of Breast Cancer

Breast cancer develops when abnormal cells in your breast divide and multiply. But experts don't know exactly what causes this process to begin in the first place. However, research indicates that there are several risk factors that may increase your chances of developing breast cancer. These include:

Age: Being 55 or older increases your risk for breast cancer.

Sex: Women are much more likely to develop breast cancer than men.

Family history and genetics: If you have parents, siblings, children, or other close relatives who have been diagnosed with breast cancer, you're more likely to develop the disease at some point in your life. About 5% to 10% of breast cancers are due to single abnormal genes that are passed down from parents to children, and that can be discovered by genetic testing.

Smoking: Tobacco use has been linked to many different types of cancer, including breast cancer.

Alcohol use: Research indicates that drinking alcohol can increase your risk for certain types of breast cancer.

Obesity: Having obesity can increase your risk of breast cancer and breast cancer recurrence.

Radiation exposure: If you've had prior radiation therapy — especially to your head, neck or chest — you're more likely to develop breast cancer.

Hormone replacement therapy: People who use hormone replacement therapy (HRT) have a higher risk of being diagnosed with breast cancer.

2.5 Machine Learning

In general, machine learning is a field of artificial intelligence that equips systems with the ability to learn from experience automatically without human intervention and aims to predict future outcomes as accurately as possible, utilizing various algorithmic models. Machine Learning is very different from conventional computational approaches, where systems are explicitly programmed to calculate or solve a problem. Machine learning involves using input data to train a model that learns patterns in the data and uses that knowledge to predict unknown outcomes. The application of machine learning is incredibly vast. It is used in various applications, such as spam filtering, weather and stock market prediction, medical diagnosis, cancer detection, autopilot, house price prediction, face detection, and many more.

Typically, machine learning falls into three categories: supervised, unsupervised, and reinforcement learning. This thesis concerns supervised learning, which we will discuss in the next section. For now, we can define supervised learning as the approach where the model is trained with both input and output labels. In contrast, unsupervised learning is where the dataset has input labels (i.e., a model is trained with unlabeled data), from which it learns different patterns and structures. Reinforcement learning deals with learning how to achieve a complex goal by maximizing along a specific dimension step by step (e.g., maximizing the points won in each round [sky]).

2.5.1 Supervised Learning

Supervised learning is a machine learning approach in which both the input and output labels are provided to the model for training. The supervised model uses labeled input and output data for training and extracts patterns from the input data. These extracted patterns are used to support future judgments. Supervised learning can be formally represented as follows:

$$Y = f(x) \quad (1)$$

Where x represents the input variables, Y denotes an output variable and f(X) is a mapping function. The goal is to approximate the mapping function such that, when an unseen input is given to the mapping function, it correctly predicts the output variable (Y). Furthermore, supervised learning has two sub-categories: classification and regression. In a classification problem, the output variable is a category (e.g., fraud or genuine, rainy or sunny). In a regression problem, the output variable is a real value, (e.g., the price of a house, temperature, etc.). This thesis only deals with the classification problem.

2.5.2 Classification

A classification problem in machine learning is the task of predicting the class label of a given data point. For example, breast cancer survival prediction can be identified as a classification problem. In this case, the goal is to predict the survival rate. Generally, there are three types of classification: binary classification, where there are two output labels, multi-class classification, where there are more than two output labels (e.g., classifying a set of images of flowers which may be Rose or Lilly or Sunflower) and multi-label classification, where the data samples are not mutually exclusive and each data samples are assigned a set of target labels (e.g., classifying a crab on the basis of the sex and color in which the output labels can be male/female and red/black). This thesis addresses the binary classification problem in which the output label is either normal or fraud.

2.5.3 Class Imbalance problem

Most real-world applications have an unbalanced class distribution, where one class label heavily dominates the counts of others. One of the best example to explain class imbalance problem is the fraud detection task, where the number of fraud class label is very low as compared to the normal class label. Most machine learning algorithms perform poorly when the class distribution is unbalanced (i.e., the predictive model tends to classify minority examples as majority examples). So some questions may arise such as:

(i) How to tackle the class imbalance problem?
(ii) Which machine learning algorithms should be applied in the presence of unbalanced class distribution?
(iii) What evaluation metrics should be used to assess the performance of a predictive model when the dataset is highly unbalanced? In the next sections, we will discuss the solutions to these questions.

2.5.4 Handling class imbalance problem

This section explains various approaches to addressing class imbalance. We can roughly classify the approaches into three categories: resampling, ensemble-based, and cost-sensitive learning. This thesis only deals with the resampling and ensemble-based approaches, which we will go through in detail in the next sections. Just to give a brief overview, cost-sensitive learning accounts for misclassification costs. For example, in medical diagnosis of cancer, the misclassification cost of missing a cancer is much higher than the cost of predicting that a healthy person has cancer. Hence, by considering the misclassification cost of minority class more heavily than that of majority class, the true positive rate of the model can be improved.

2.5.4.1 Resampling approach

Most predictive models perform worst in the presence of an unbalanced class distribution. Therefore, some data preprocessing must be performed before providing the data as input to the model. In the case of a class imbalance problem, such data preprocessing is performed at the data level, using a resampling approach. Basically, there are three resampling approaches: undersampling, oversampling, and hybrid.

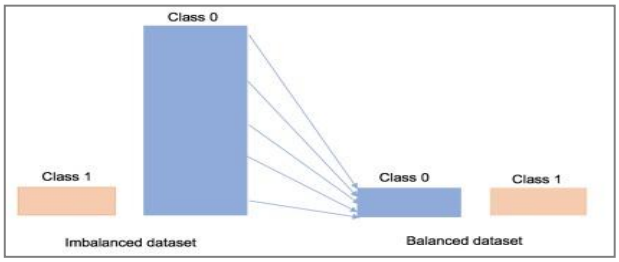

Figure 2: Under-sampling approach

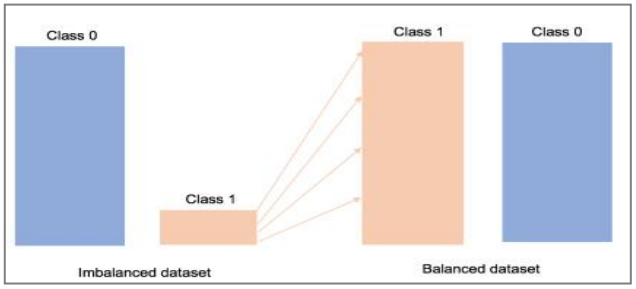

Figure 3: Oversampling approach

In the undersampling method, the majority class is reduced to balance the dataset, as shown in Figure 2. It is best to implement when the dataset is large, and reducing the majority samples can greatly improve runtime and reduce storage requirements. An oversampling method is the opposite of an undersampling method.

This method works with the minority class. It replicates the observations from the minority class to balance the ratio between the majority and minority samples, as shown in Figure 3. At last, a hybrid method applies both undersampling and oversampling for rebalancing. We will discuss some of the resampling approaches in the following subsections.

2.5.4.1.1 Random Underdamping

This method randomly eliminates the majority of samples to balance the dataset. This method is best used when the training data is large. Reducing the frequency of majority samples, it improves runtime and reduces storage requirements. However, the disadvantage of using such an approach is that some useful information may be eliminated in the process of eliminating the majority of samples. As a result, the classifier prediction may not be very accurate.

2.5.4.1.2 Tomek Link Removals

Tomek Link is a pair of examples of different classes that are each other's nearest neighbors. Given two samples E1 and E2 belonging to different classes, a pair (E1, E2) is a Tomek Link if there's not a sample E3 such that the distance between E1 and E3 is less than that of E1 and E2 or the distance between E2 and E3 is less than that of E1 and E2 [BPM04]. Removing Tomek links can be considered an undersampling approach, where the majority sample in the Tomek link is eliminated.

2.5.4.1.3 Random Oversampling

This method randomly replicates minority samples to balance the dataset. Unlike random undersampling, this approach does not lead to information loss. However, there's a high risk of overfitting the data, as it replicates the minority samples.

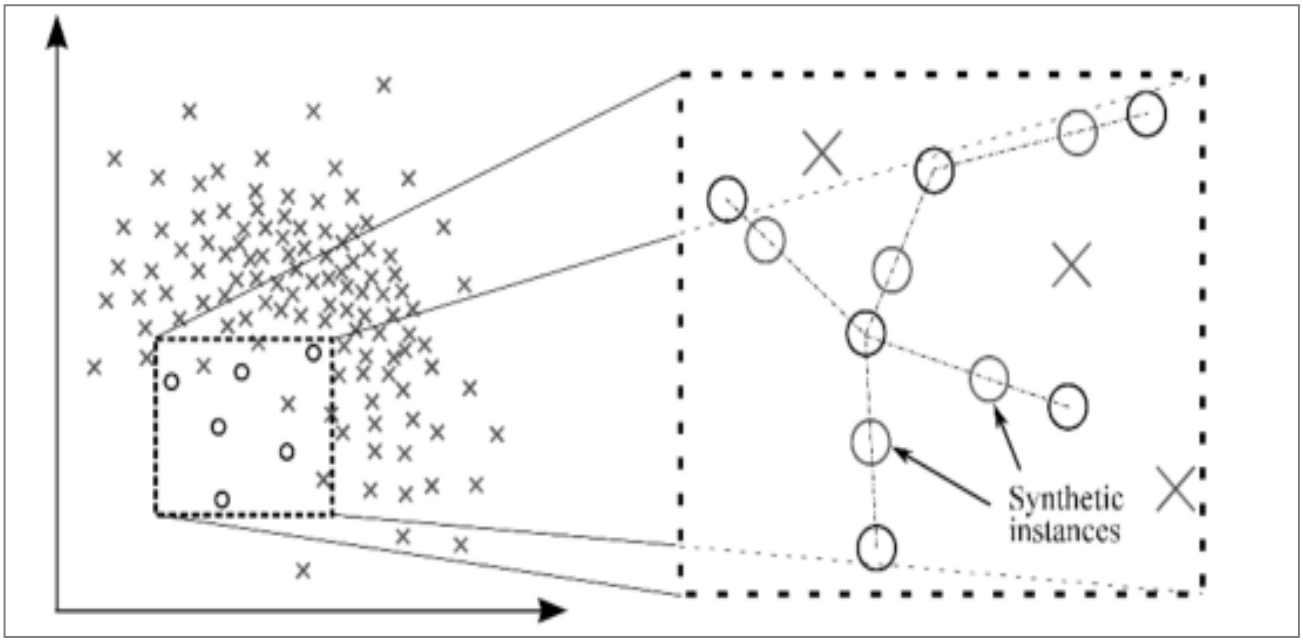

Figure 4: Generation of synthetic examples using SMOTE

2.5.4.1.4 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a popular technique for rebalancing datasets, developed by Chawla [BCHK02]. It aims to create new minority-class examples (synthetic instances) by interpolating between several nearest minority examples rather than by oversampling with replacement, as shown in Figure 4. As a result, it reduces overfitting to the training data. Depending upon the amount of oversampling required, nearest neighbors of minority examples are randomly selected.

2.5.4.1.5 Combination of SMOTE and Tomek Link removal

SMOTE is a powerful approach for balancing class distributions. However, when creating new synthetic minority examples, the minority class cluster might intrude into the majority class space. Providing such data to the model can lead to overfitting. Hence, to mitigate such a situation, both SMOTE and the Tomek Link removal approach can be applied to balance the class distribution. In this process, the original training dataset is first oversampled using SMOTE, and then Tomek Link removal is applied to the resulting dataset, producing a balanced dataset.

2.5.4.2 Ensemble approach

In the previous section, we discussed the data-level approach in which resampling techniques are used to balance the class distributions. In this section, we will discuss the algorithmic approach known as the ensemble approach. An ensemble approach involves modifying existing classification algorithms to address unbalanced

class distributions. In general, an ensemble approach is a learning algorithm that assembles a set of classifiers and applies them for classification by taking a vote on their predictions [Die00], as shown in Figure 5. Typically, there are two types of the ensemble approach: bagging and boosting.

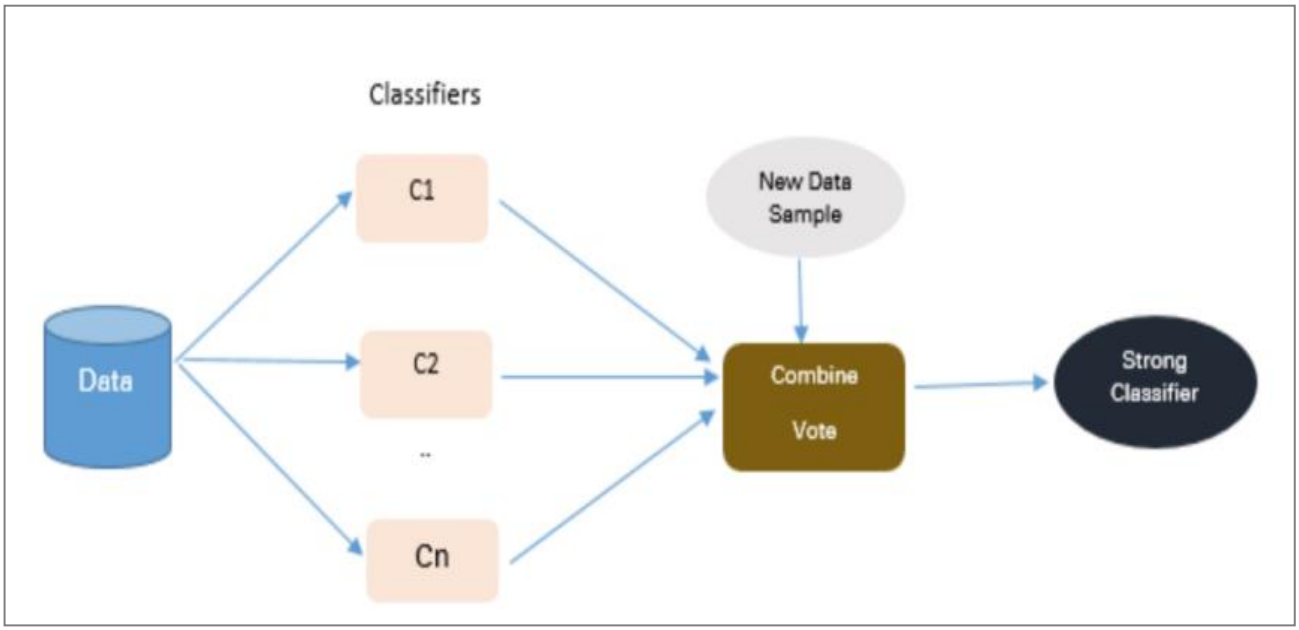

Figure 5: Ensemble approach

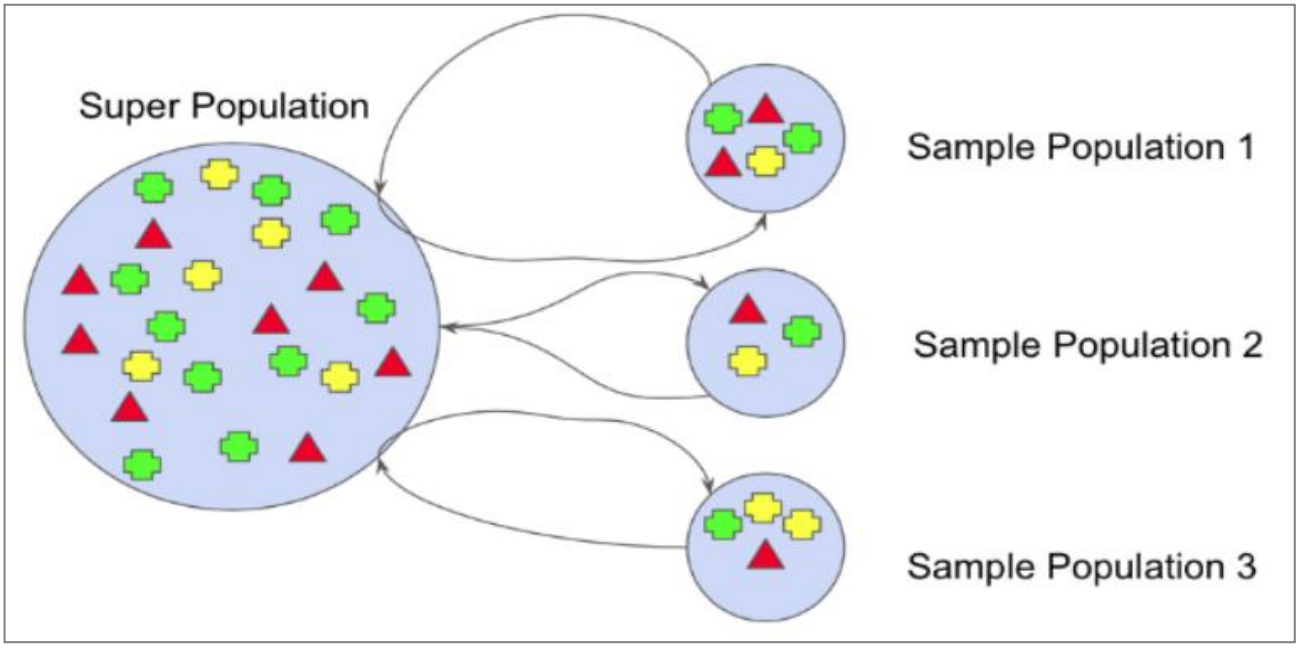

Figure 6: Bootstrapping

Before moving on to bagging, let's first talk about an essential concept in bootstrapping, which is used in bagging algorithms. In machine learning, bootstrapping is the process of sampling the training data randomly with replacement. Each bootstrap sample is generated so that each sample has different characteristics, as shown in Figure 6. When the models use these samples for training, they can learn various aspects of the data and can improve the prediction performance.

2.5.4.2.1 Bagging

Bagging, an abbreviation for Bootstrap Aggregation, is a simple yet very powerful ensemble technique. This method involves bootstrapping that generates new training samples from the original training set with replacement. These new training samples are called bootstrap training samples. Each bootstrap sample is used to train the individual model separately, which is then used for prediction. Finally, the predictions from all the bootstrapped models are aggregated by averaging the output (for regression) or voting (for classification). Figure 7. Gives a better picture of bagging. It helps to reduce overfitting and variance. Decision trees are typically used as base models in bagging. However, other types of methods can also be used. This thesis deals with the random forest algorithm as a bagging method.

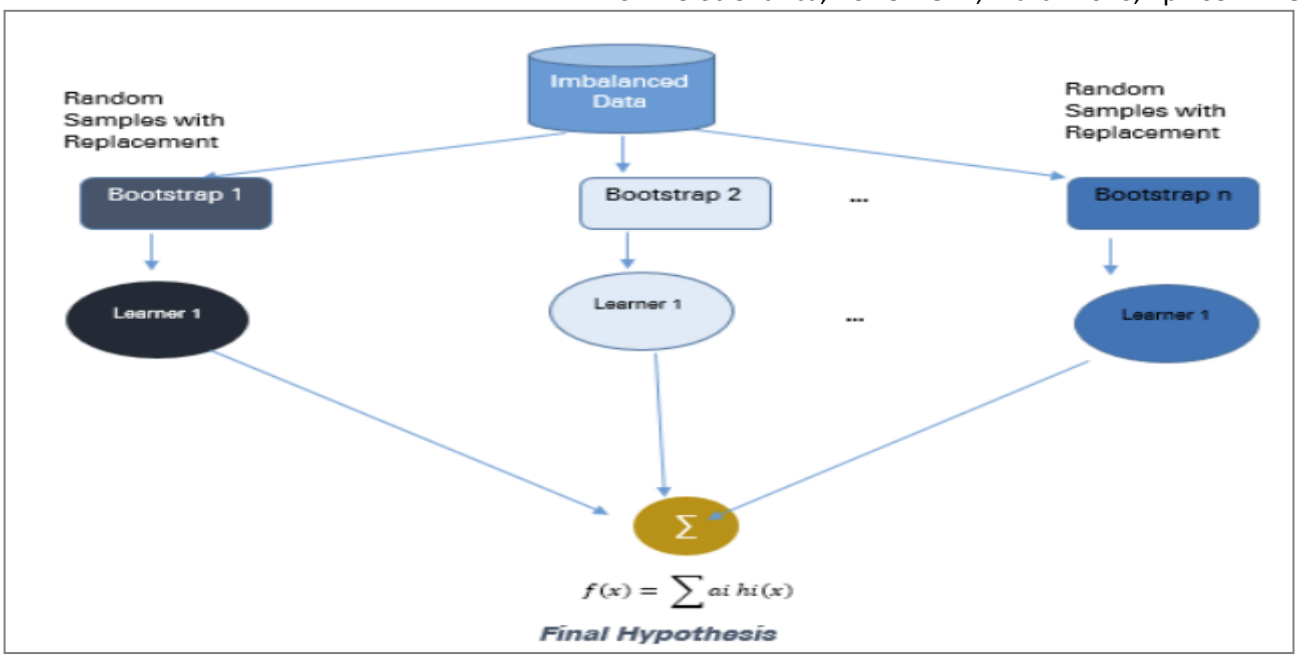

Figure 7: Overview of bagging

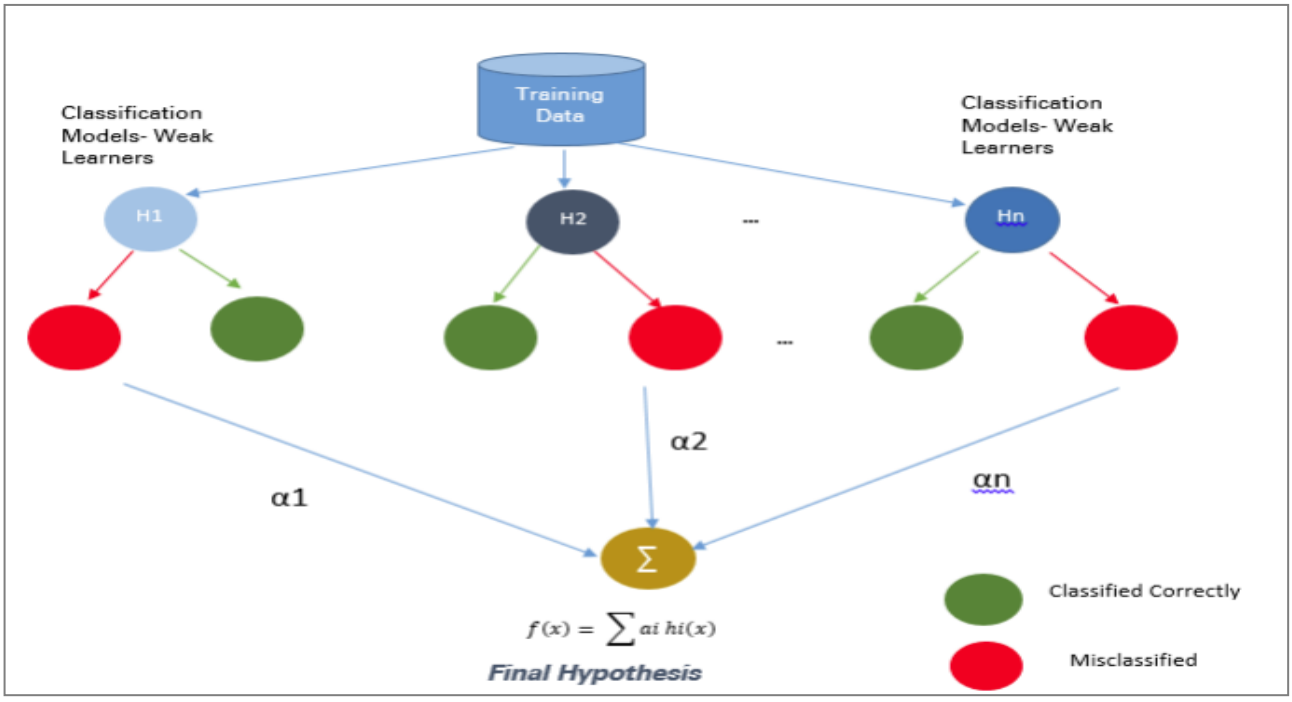

Figure 8: Overview of boosting

2.5.2.2 Boosting

Boosting is another very powerful ensemble technique. It involves combining weak learners, also called base learners, to create a strong learner that produces better results than an individual learner. Unlike bagging, in which each model runs in parallel and the outputs are combined at the end, boosting trains weak learners sequentially, so that each learner tries to correct its predecessor by assigning more weight to samples previously misclassified. Therefore, the future weak learner will focus more on the misclassified cases. Figure 8. Gives a better picture of boosting. It also uses bootstrapping, which helps it avoid overfitting and variance. There are many examples of boosting algorithms, such as AdaBoost, Gradient Boosting, and XGBoost. This work deals with XGBoost.

2.5.5 Selected Models

In this section, we will discuss the different models selected for the predictive analysis. Depending on the nature of the classification problem, we chose three very popular predictive models. They were logistic regression, random forest, and XGBoost.

2.6. Logistic regression

Logistic regression is the classical and best bicategorical algorithm, preferred for classification problems, especially bicategorical ones. The choice of algorithm is based on the principle of simplicity before complexity. Logistic regression is also an excellent choice because it is a well-established statistical method for predicting the outcome of a binomial or polynomial distribution. A multinomial

logistic regression algorithm can regenerate the model. A better classification algorithm is needed when the target field or data is a set field with two or more possible values.

The advantage of logistic regression is that it is faster to process and is suitable for bicategorical problems. It is also easier for a beginner to understand and see the weights of each feature. Then it is easier to update the model and incorporate new data to address different problems (Ionescu, *et al.,* 2022).

Furthermore, it has a disadvantage. There is a limit to the data and the adaptability of the scene. Not as adaptable as the decision tree algorithm. But this is an issue we can also determine in this project, based on the actual situation, whether logistic regression has a better ability to adapt to an extensive data set on survival rates (Osisanwo, *et al.,* 2017).

The main methods of logistic regression:

Objective: It is to look for some risk factor, then in the study, find a particular attribute to be either alive or survived

Prediction: Predicting the probability of survival under other independent variables, based on different algorithmic models.

Regression General Steps

Finding the h-function (i.e., the prediction function)

When constructing the predictive function h(x), the logistic function, also known as the sigmoid function, the first step is to build the predictive process, including the training data vector and the best parameters. The basic form of the function is shown in Figure 9.

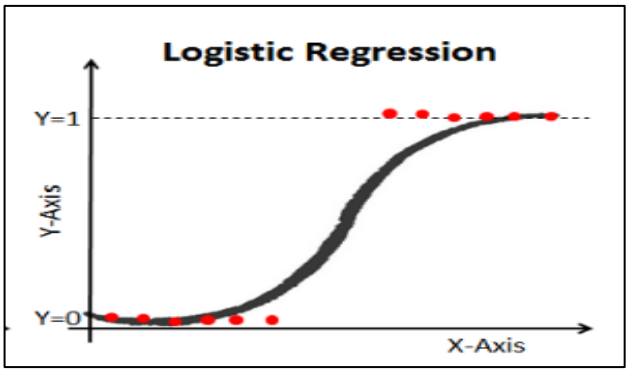

Figure 9: Architecture of logistic function

In constructing the J-function (loss function), the second step is to define the loss function-j. In general, there will be m samples, each with n characteristics. The Cost and J functions are as follows and are derived using maximum likelihood estimation (Cho *et al.,* 2014).

The Cost function Equation:

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & if\ y=1 \\ -log(1-h_\theta(x)) & if\ y=0 \end{cases} \quad (2)$$

The J Equation:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} cost\big(h_\theta(x_i), y_i\big)\big) = -\frac{1}{m}\big[\sum_{i=1}^{m}(y_i Log h_\theta(x_i) + (1-y_i)\log(1-h_\theta(x_i))\big] \quad (3)$$

Figures 2 and 3 indicate the J-function minimal and how to find the regression parameter (θ)

The final step in gradient descent is to solve for the minimum value of θ. The process of updating θ can then be summarized as follows:

Function of upating θ

$$\theta_j = \theta_{j-}\ \text{ѡ}\ \frac{1}{m}\sum_{1=1}^{m}(h_\theta(x_i) - y_i)\, xi \quad (4)$$

2.6.1 Decision tree (DT)

The use of decision trees is usually based on the known probabilities of various scenarios, and the decision tree is constructed to determine the probability that the expected net present value is greater than or equal to zero, thereby evaluating the risk of the training project (Scheidegger, *et al.,* 2021). Also, it judges the feasibility of the decision analysis method. Then we know that because this decision branch is drawn as a graph, much like the trunk of a tree, we name it a decision tree. Decision trees are a primary classification and regression method, and learning typically involves three steps: feature selection, decision tree generation, and decision tree pruning. In machine learning, a decision tree is a predictive model that represents a mapping between object properties and object values.

A classification tree (decision tree) is a very commonly used classification method. Similar to the dataset classification problem discussed in this paper, decision trees are often used to analyze data and make predictions. That is why we chose it for training the fraud detection system (Cho *et al.,* 2014).

That is a simple decision tree classification model: the red boxes are features.

There are two universal reasons for choosing Decision trees: Decision trees usually mimic human horizontal thinking, so it is easy to understand the data we provide and to make excellent interpretations. Decision trees allow you to see the logic of how the data is interpreted, unlike SVMs, NNs, and other similar black-box algorithms, where you do not see any internal information (Osisanwo, *et al.,* 2017). For example, as the Figure 10, we can see how the logic makes decisions. Plain and simple.

Then, what is a decision tree now? A decision tree is like a tree, where each node represents an element (attribute), each link (branch) represents a decision (rule), and each leaf represents a result (categorical or continuous value). The core of the entire decision tree is to create a tree like this for the entire dataset. And the decision tree process individual results (or minimizes errors) at each leaf on each plate.

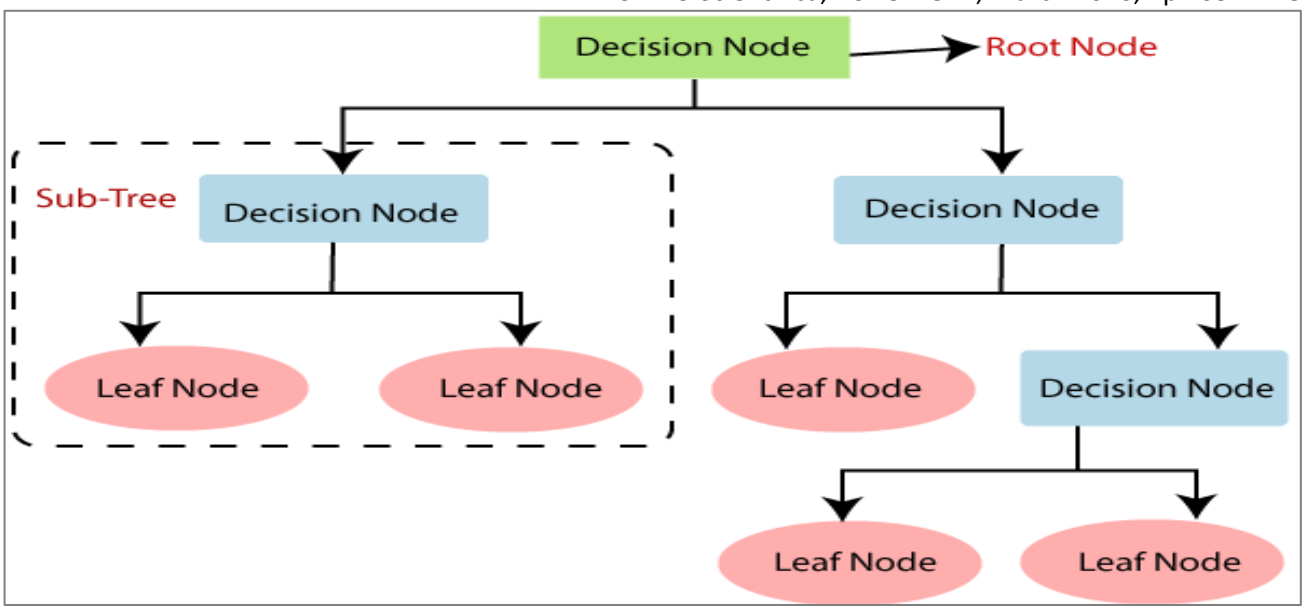

Figure 10: Architecture of a decision tree

Table 1. Treatment type received by all patients

No of Patients	No treatment	Chemotherapy	Amputation treatment	Hormonal therapy
87	Yes			
48		Yes		
101			Yes	
9				Yes
8		Yes		Yes
1		Yes	Yes	Yes

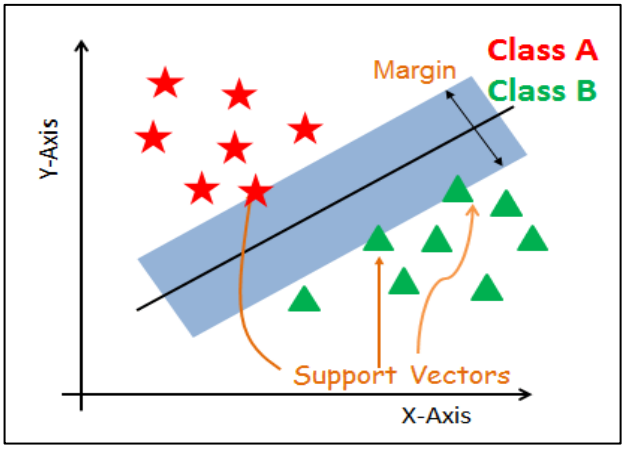

Figure 11: Diagram of support vectors

2.6.3 Support vector machine (SVM)

A Support Vector Machine (often abbreviated as SVM) is a supervised learning method most widely used in statistical classification and regression analysis. It is also the focus of this project. Support vector machines belong to a family of generalized linear classifiers, which are characterized by their ability to both minimize empirical errors and maximize geometric edge regions. Hence, support vector machines are also known as maximum edge region classifiers.

The core principle of the support vector machine is: mapping the vectors into a higher-dimensional space where a maximum-margin hyperplane is established. Two parallel hyperplanes are built on either side of the hyperplane that separates the data. Also, the separated hyperplanes maximize the distance between the two parallel hyperplanes (Cho *et al.,* 2014). It is assumed that the greater the space or gap between the parallel hyperplanes, the smaller the total error of the classifier. In this project, SVM is the supervised learning algorithm used for multi-class classification (Agarap, 2018).

Support Vectors

Support vectors are the data points, that are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to classifier construction. Figure 11 is a diagram of support vectors.

Hyperplane

A hyperplane is a decision plane that separates a set of objects with different class memberships.

Margin

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger between the classes, it is considered a good margin; a smaller margin is a bad margin.

2.6.6 Evaluation

This is the stage where machine learning algorithms are evaluated to determine how each performed. This step

shows us how accurate our results are and how efficiently each news headline is classified into its pre-defined class. Researchers in the field of literature have used numerous measures for this purpose, including accuracy, precision/recall, fallout, error, and more. A few of these measures are listed below;

1. **Precision** is defined as a fraction of the survival rate.
2. **Recall**: is the fraction of survival rates that are correctly predicted.
3. **True Positive**: means that survived patient is classified to its correct class.
4. **False Negative**: means that survived patient is classified to a wrong class.
5. **True Negative**: means that survived patient does not belong to that class and is misclassified.
6. **Accuracy**: the breast cancer survival rate is defined as the sum of true negatives and true positives.

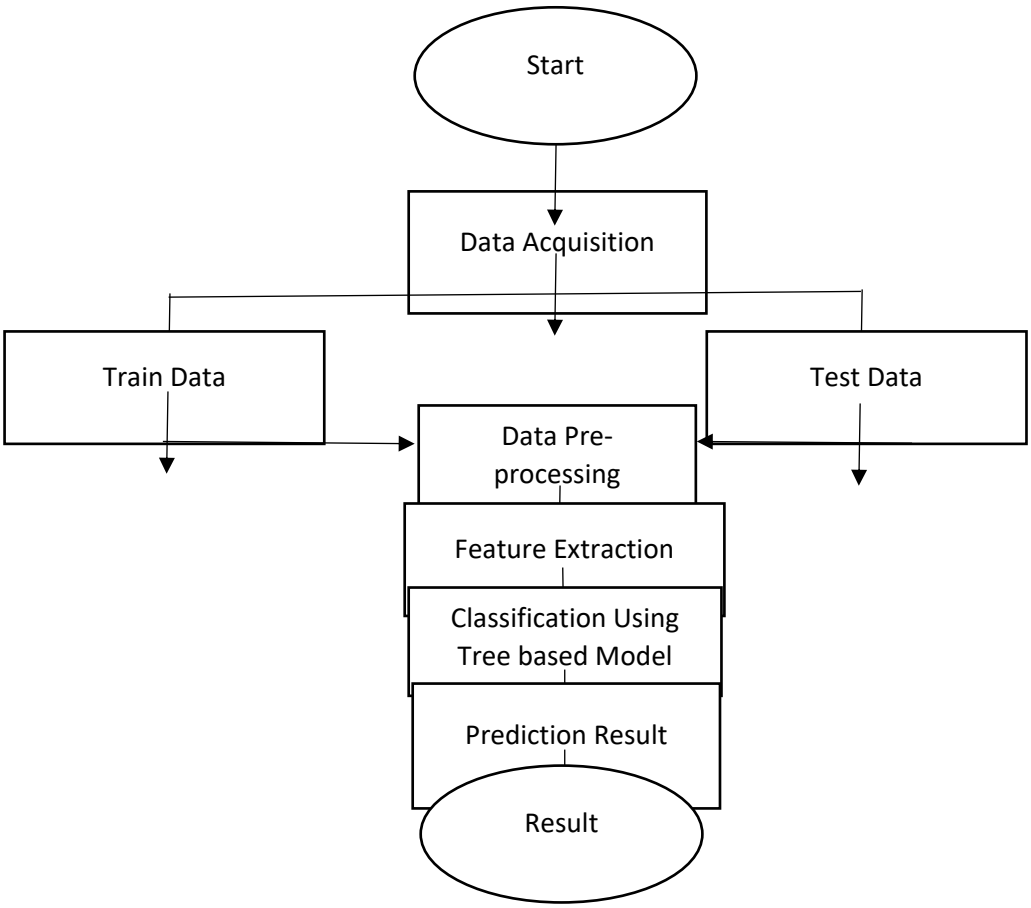

Figure 12: Proposed System Architecture of breast cancer survival rate prediction using tree data model

3.0 RELATED WORKS

The significance of detecting breast cancer at an early stage to enhance the chances of survival is widely recognized (Padmapriya et al., 2016). As a result, numerous researchers are actively involved in the exploration of breast cancer detection and prediction. Diverse machine learning techniques, including supervised and unsupervised, have been embraced in this field of algorithms (Dutta et al., 2019). In a study conducted by Ming et al. (2019), the focus was on personalized breast cancer prediction. They employed various machine learning algorithms and compared them to two existing statistical models: Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BODICEA) and Breast Cancer Risk Assessment Tool (BCRAT).

The findings demonstrated that machine learning techniques improved classification accuracy for both women with and without breast cancer, using the same independent variables as the statistical models. This finding supports the reliability and credibility of Machine Learning Algorithms (MLAs). A method called Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems (BCD-NFIS) was also introduced, which uses Fuzzy networks to reduce dependence on dataset features. As a result, this approach achieved an impressive accuracy of 98.24%, surpassing previous methodologies (Nallamala et al., 2019).

The occurrence of cancer, particularly mouth cancer, is attributed to various factors such as tobacco smoking and alcohol consumption. However, doctors still face uncertainty and confusion regarding the precise causes of these diseases, and their impacts are not clearly defined. Consequently, the Max-Min composition approach (Jothikumar et al., 2019) is noteworthy for addressing this matter.

In a comparative study (Shravya *et al.,* 2019), three machine learning algorithms (MLAs), namely Support Vector Machine (SVM), Logistic Regression (LOR), and

K-Nearest Neighbor (KNN), were evaluated alongside the utilization of the Dimensionality Reduction Technique. The findings revealed that SVM outperformed the other two classifiers, achieving a remarkable 92.7% accuracy. KNN came in second place with 92.23% accuracy, while Logistic Regression achieved 92.10%. (Dutta *et al.,* 2019) acknowledged the uncertain nature of features found in breast cancer datasets, leading to the utilization of a fuzzy inference system for predicting breast cancer.

This system was pitted against various machine learning techniques, including Decision Table, RBF-Network, Naive Bayes, and Random Tree. Among these methods, the fuzzy system achieved the highest performance, with an accuracy of 84.64%. The second-best performer was the Decision Table with an accuracy of 79.02%, while the Random Forest exhibited the poorest performance with an accuracy of 72.38%. Notably, the study did not include some popular ML algorithms such as SVM and KNN. Given the multitude of metrics used in decision-making during model construction, no single algorithm can be deemed superior to others. (Shajahaan *et al.,* 2013) examined the potential of decision trees for breast cancer prediction, comparing it with other supervised learning algorithms like ID3, Naïve Bayes, Random Forest, C4.5, and CART, employing the Wisconsin dataset. The evaluation focused on Precision and Recall as performance metrics.

The results demonstrated that the Random Forest outperformed the other supervised learning algorithms. However, given the problem domain's inherent nature, it is crucial to include additional ML metrics for a comprehensive assessment. In a study (Vijayarani *et al.,* 2011), the performance of three data mining algorithms (C4.5, RIPPER, and PART) was compared on two datasets: breast cancer and heart disease. The analysis focused on the number of rules generated. Another research (Spanhol *et al.,* 2016) involved examining over 7,000 breast cancer datasets obtained from 82 patients' histopathology images. Various classification algorithms were used, resulting in accuracy ranging from 80% to 85%.

In a similar vein, Agarap et al. (2018) compared six machine learning algorithms on the Wisconsin's diagnostic breast cancer dataset: Softmax Regression, Multilayer Perceptron (MLP), Linear Regression, Nearest Neighbor (NN) search, Support Vector Machine (SVM), and Gated Recurrent Unit SVM (GRU-SVM). MLP achieved the highest accuracy, reaching 99.04%.

Another approach to diagnosing breast cancer is the analysis of histopathological images of breast cancer tissues (Nuruddin *et al.,* 2019). In this method, machine learning models are used to automate the classification of benign and malignant tissue images. Several studies have reviewed machine learning algorithms (MLAs) for breast cancer prediction and classification (Eltalhi *et al.,* 2019). These papers have compared different algorithms to determine which is most effective. The authors of Bazila et al. (2018) discussed and compared the performance of Bayes classifiers, including Boosted Augmented Naive (BAN) Bayes, Tree Augmented Naive (TAN) Bayes, and Bayes Belief Network (BBN).

The study found that TAN with gradient boosting achieved the highest accuracy, sensitivity, and specificity. Apart from these metrics, the evaluation of algorithms also considers the Mean Absolute Error (MAE) and Total Time on Bench (TTB) as crucial performance indicators. A study was conducted by Chaurasia et al. (2014) to explore various data mining techniques for predicting breast cancer. The researchers used the Wisconsin dataset from UCL, consisting of 10 attributes and 699 instances. After removing sixteen instances with missing values, they were left with 683 instances. The study employed three supervised learning algorithms, namely IBK, BF Tree, and Sequential Minimal Optimization (SMO). The comparison revealed that SMO achieved the highest prediction accuracy at 96.2%. SMO also yielded a Kappa statistic (KS) of 0.92 and a lower mean absolute error (MAE) than the other algorithms.

In a separate study by Padmapriya et al. (2016), the objective was to assess the performance of different classification algorithms on mammogram images. The study employed three algorithms: J48, CART, and ADTree. The evaluation metrics used included specificity, Kappa statistics, and MAE. (Akinsola *et al.,* 2019) expressed the viewpoint that Multi-Criteria Decision Method (MCDM) approaches could be employed to identify optimal classification and regression models concerning supervised machine learning algorithms.

METHODOLOGY

4.1 Dataset Description

All data used in this project was gotten from Kaggle. This study's NKI dataset, the original dataset, was obtained from the data. World Website (https://data.world/datasets/breast-cancer). This dataset comprises gene expression profiles and clinical information for breast cancer patients aged 28-53. The original dataset used in this project has 15 columns and 4025 rows. One of which is a response variable (survived/not survived). Out of 1564 independent variables, 10 contain data on age, chemotherapy, hormonal treatment, amputation treatment, histological type, tumor diameter, number of nodes, cancer grade, lymphocytic infiltration, and gene expression profiling. Table 1 shows the treatment given to 272 patients.

4.2 Preparation of Data, Pre-processing

Raw data were verified for missing values within the dataset before being subjected to any analysis, and none were identified. Label 0 (not survived) and label 1 (survived) were used to categorize mortality. The standard scaler technique was used to scale the columns in gene expression profiling.

4.2.1 The XGB Technique to Determine the Significant Features

In machine learning, feature selection is extremely important. High-dimensional data increase the distance

between variables, making it more difficult to anticipate proper outcomes. The goal of feature selection is to eliminate non-essential characteristics in order to create models that are quicker, more stable, and more accurate. Irrelevant and undesired characteristics might impair the algorithms' capacity to accurately distinguish between labels, increasing the modeling error. The removal of features decreases the noise in the data and improves the performance of the models.

The XGB method was used to identify features, and the top 10 rated features were chosen for further research based on their significance score. The dataset now comprises 272 samples with the top 10 characteristics (195-label 0, 77-label 1). Because of its default use of the Gini index as an internal feature significance score and its reputation in machine learning contests, the XGB method was chosen. Individual trees are created by XGB utilizing several cores, and data are arranged to reduce lookup times.

4.3 Data splitting

For each experiment, we split the entire dataset into 70% training set and 30% test set. We used the training set for resampling, hyperparameter tuning, and model training, and the test set to evaluate the performance of the trained model. While splitting the data, we specified a random seed (a random number), ensuring the same split every time the program executed.

4.4 Data resampling

Data resampling is a technique commonly used in machine learning to address class imbalance, where one class of data samples is significantly more prevalent than the others. In the context of breast cancer survival rate prediction, data resampling can be beneficial when the dataset has an imbalanced distribution of survival outcomes (e.g., a small number of samples in the minority class, such as patients who did not survive).

There are two common approaches for data resampling: oversampling and undersampling. Let's discuss each of them:

1. **Oversampling**: Oversampling aims to increase the representation of the minority class by replicating or generating synthetic samples. The goal is to balance the class distribution and provide more training examples for the underrepresented class. Some popular oversampling techniques include:

Random Oversampling: This method randomly duplicates samples from the minority class until both classes are balanced. However, it may lead to overfitting if the minority class is overrepresented.

i. Synthetic Minority Over-sampling Technique (SMOTE): SMOTE creates synthetic samples by interpolating between the feature vectors of neighboring minority class samples. This approach generates new instances while maintaining the underlying distribution patterns of the minority class.

ii. Adaptive Synthetic (ADASYN): ADASYN is an extension of SMOTE that focuses on regions with higher densities of minority-class samples. It generates more synthetic samples for the difficult-to-learn instances and fewer for the easier ones.

2. **Undersampling**: Reduces the number of samples from the majority class to match the minority class's size. This approach aims to balance the class distribution by removing instances from the overrepresented class. Some common undersampling techniques include:

Random Undersampling: Randomly selects samples from the majority class until both classes are balanced. However, this approach may discard potentially useful data and result in information loss.

It's important to note that data resampling techniques should be used with caution and evaluated properly. Oversampling may introduce duplicate or synthetic samples that can bias the model, while undersampling may discard potentially valuable information. Additionally, resampling should be performed only on the training set, keeping the original test set intact for unbiased evaluation. A comprehensive analysis and experimentation with various resampling techniques, combined with appropriate evaluation metrics, can help determine the most suitable approach for predicting breast cancer survival rates.

4.5 Breast Cancer Survival Prediction

After feature extraction, the next phase is classification, an important step in which the aim is to classify a survival record as either surviving or not. In this study, a decision tree will be used as a tree-based model for survival classification into two categories (Figure 12).

Decision Tree Classier: Decision trees are constructed via an algorithmic approach, which identifies ways to split a dataset based on different conditions. It is one of the most widely used and practical methods for supervised learning (Osisanwo, *et al.,* 2017). The main aim of using Decision Trees is to create a training model that can predict the class or value of target variables by learning decision rules inferred from the training data (Cho *et al.,* 2014). The mathematical representation goes thus;

$$(x, Y) = (x_1, x_2, x_3, \ldots\ldots, x_k, Y) \quad (3.1)$$

The dependent variable Y is the target variable we are trying to understand, classify, or generalize. The vector x consists of the features x1, x2, x3, etc., used for that classification (Osisanwo, *et al.,* 2017).

4.6 Performance Evaluation Metrics

The F1 Score and Accuracy are the basic evaluation metrics considered in this work. These metrics helped determine the best-fit algorithm for news headline classification.

RESULTS AND IMPLEMENTATION

Breast Cancer is a complex and prevalent disease that affects millions of lives worldwide. Early detection and

accurate prediction of survival rates are crucial for improving patient outcomes and treatment plans. In the age of data-driven healthcare, predictive models have emerged as powerful tools to aid in this process. This project is very important in healthcare, where data science and machine learning techniques play a pivotal role. By harnessing the power of predictive models, medical professionals can better allocate resources, provide personalized treatments, and improve patient outcomes.

5.1 Confusion Matrix

The confusion matrix is a tool for evaluating the performance of a classification model. It provides a clear breakdown of true positive (TP), true negative (TN) false positive (FP), and false negative (FN) predictions. For our breast cancer survival prediction model, the confusion matrix result is given below (Figure 13):

i) True Positive (TP): Instances in which the model correctly predicted a positive outcome, indicating a breast cancer survivor.
ii) True Negative (TN): The model correctly identified instances in which the patient did not survive breast cancer.
iii) False Positives (FP): False alarms occurred when the model incorrectly predicted survival for patients who didn't actually survive.
iv) False Negatives (FN): Instances where the model failed to predict survival for patients who actually survived the disease.

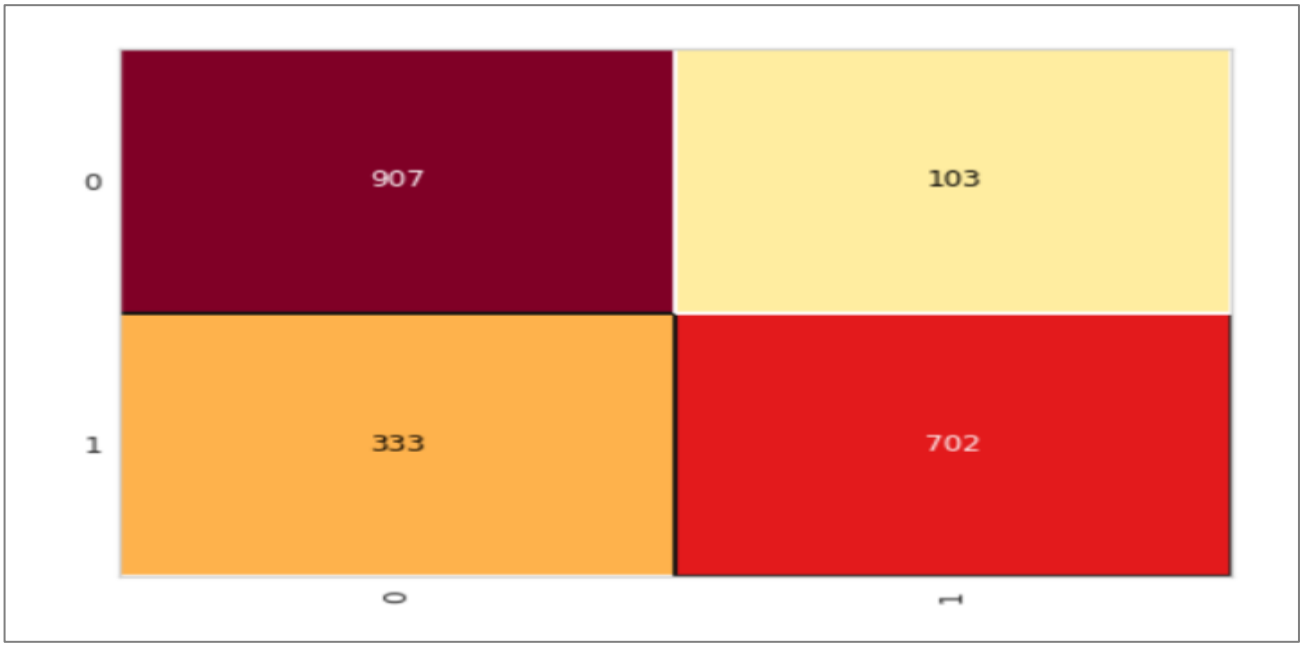

Figure 13: The confusion Matrix of the model

Table 2. The classification report performance of the model

Class / Metric	Precision	Recall	F1-score	Support
0	0.73	0.90	0.81	1010
1	0.87	0.68	0.76	1035
Accuracy			**0.79**	2045
Macro Avg	0.80	0.79	0.78	2045
Weighted Avg	0.80	0.79	0.78	2045

5.2 Classification Report Analysis:

The classification report provides a more comprehensive evaluation by considering metrics such as precision, recall, and F1-Score for each class. In this context (Table 2), the two classes are 'Survived' and 'Did not survive'.

i) Precision: The precision score indicates the proportion of correctly predicted positive instances (TP) out of all instances predicted as positive (TP + FP)
ii) Recall: this metric represents the percentage of true positive predictions (TP) correctly identified out of all actual positive instances (TP + FN)
iii) F1-score: The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of the model's accuracy.

In this study, we analyze breast cancer survival prediction using a machine learning model, focusing on the insights from the confusion matrix and classification report.

5.3 IMPLEMENTATION

The model was deployed to a website using streamlit python library. All the attributes are used to predict the attribute. Feature extraction is then performed on the input word, and it's fed into the model for sentiment prediction (Figure 14).

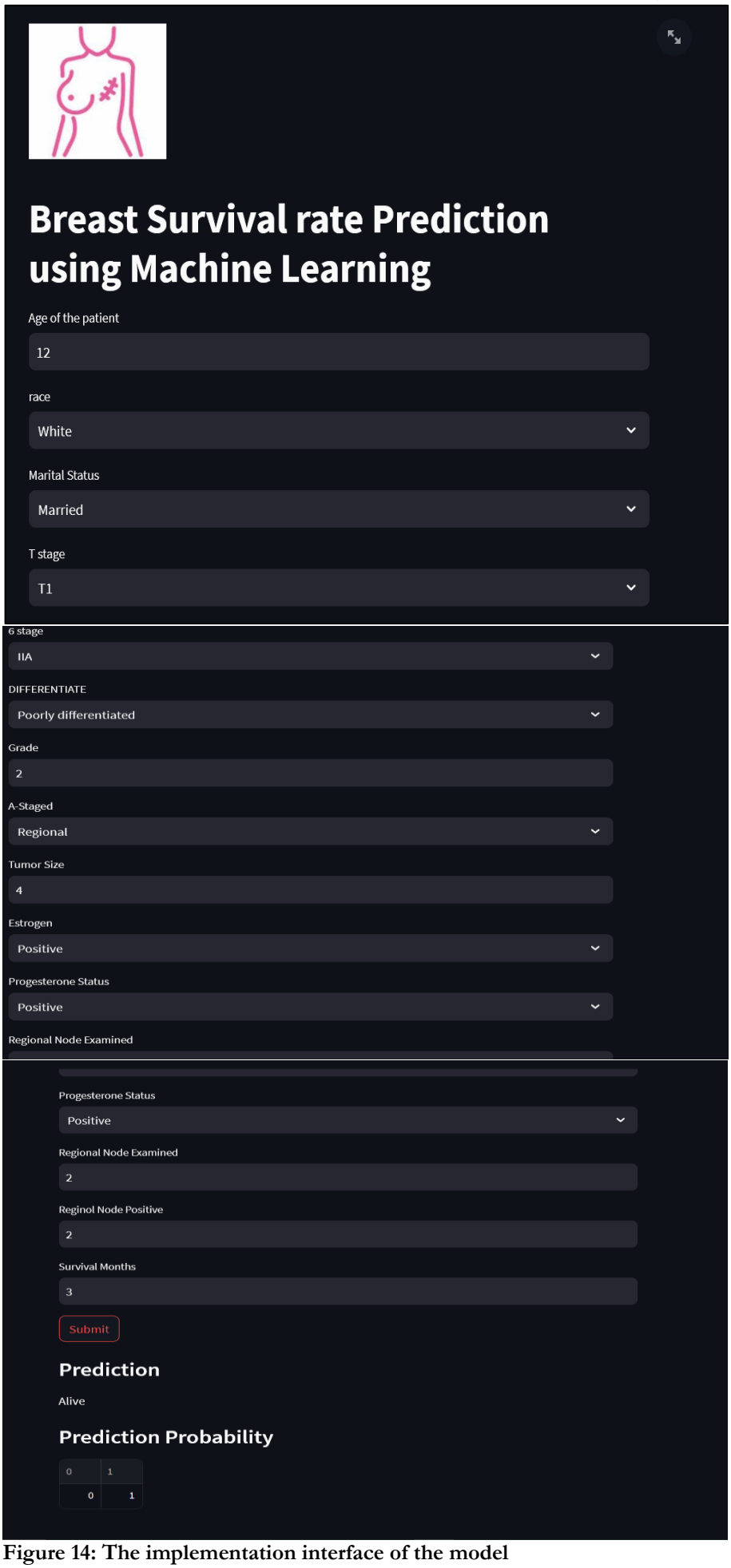

Figure 14: The implementation interface of the model

Below are the steps in using the website for prediction:

1. Prepare a single data word and predict its sentiment
2. Input the data into the respective fields

CONCLUSION

Breast cancer is a live threatening disease that could be treatable in the early stages of the condition before it spreads to other parts of the body. Thus, early prognosis is important to support the patients with the best possible care, medications, and therapies. A specific and precise prediction of mortality rates with a large data size can bring forth relevant information in terms of pervasiveness and public awareness. Different commercial organizations applied artificial intelligence (AI) and Machine learning (ML) algorithms along with various independent research groups to manifest the ability of computational intelligence and to assist medical experts in the diagnosis and prediction of cancer risk. One of the major hitches of computational intelligence in clinical medicine is the constraint of complex algorithms in handling multivariable data that strongly affect the decision-making process. AI and ML algorithms can be utilized in a smart way that optimizes the impact in a much-supervised way, with future challenges not limited to only one solution but require multitudinous ways of training, testing, and validating the models. It will provide more flexibility to the researcher and experts for the integration of these approaches into breast cancer prediction. The authors used different ML algorithms for breast cancer patients' datasets and compared their performance based on a ROC-AUC score. ML classifiers are also examined on the basis of accuracy, precision, true positives, true negatives, ROC-AUC, and F1-score. The advancement in basic concepts and technological progress in these ML approaches helped the researchers to make decisions and spread public awareness at the very early stage of breast cancer.

REFERENCE

Agarap, A. F. M. (2018). An application of machine learning algorithms on the Wisconsin diagnostic dataset. In *International Conference on Machine Learning and Soft Computing*.

Akinsola, J. E. T., Kuyoro, S. O., Awodele, O., & Kasali, F. A. (2019). Performance evaluation of supervised machine learning algorithms using multi-criteria decision making techniques. In *International Conference on Information Technology in Education and Development (ITED) Proceedings* (pp. 17–34).

Bazila Banu, A., & Ponniah, T. (2018). Comparison of bayes classifiers for breast cancer classification. *Asian Pacific Journal of Cancer Prevention, 19*(10), 2917–2920.

Chaurasia, V., & Pal, S. (2014). A novel approach for breast cancer detection using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering, 2*(1), 2456–2465.

Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724-1734). Association for Computational Linguistics. [Crossref]

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. [Crossref]

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press. [Crossref]

Dutta, S., Ghatak, S., Sarkar, A., Pal, R., Pal, R., & Roy, R. (2019). Cancer prediction based on fuzzy inference system. *Advances in Intelligent Systems and Computing, 851*, 127–136. [Crossref]

Eltalhi, S., & Kutrani, H. (2019). Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. *Journal of Dental and Medical Sciences, 18*(4), 85–94.

Ferroni, P., Roselli, M., Zanzotto, F. M., & Guadagni, F. (2018). Artificial intelligence for cancer-associated thrombosis risk assessment. *The Lancet Haematology, 5*(7), e288–e289. [Crossref]

Ferroni, P., Zanzotto, F. M., Scarpato, N., Riondino, S., Nanni, U., Roselli, M., & Guadagni, F. (2017). Risk assessment for venous thromboembolism in chemotherapy treated ambulatory cancer patients: A precision medicine approach. *Mediterranean Journal of Hematology and Infectious Diseases, 9*(1), Article e2017031. [Crossref]

Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research, 12*, 2211–2268.

Gouda, I. S., Abdelhalim, M., & Zeid, M. A. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC), 32*, 569.

Healthline. (n.d.). *Healthline*. [Link]

Ionescu, S., Nicolescu, A. C., Marincas, M., Madge, O. L., & Radu, P. (2022). An update on the general features of breast cancer in male patients-A literature review. *Diagnostics, 12*(7), Article 1554. [Crossref]

Islami, F., Guerra, C. E., Minihan, A., Yabroff, K. R., Fedewa, S. A., Sloan, K., Wiedt, T. L., Thomson, B., Siegel, R. L., Nargis, N., Winn, R. A., Jemal, A., & Ward, E. M. (2022). American Cancer Society's report on the status of cancer disparities in the United States, 2021. *CA: A Cancer Journal for Clinicians, 72*(2), 112-143. [Crossref]

Jain, A., & Levy, D. (2016). Breast mass classification using deep convolutional neural networks. In *30th Conference on Neural Information Processing Systems (NIPS 2016)* (pp. 1–6). Barcelona, Spain.

Jiang, F. (2017). Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th International Conference on Bioinformatics and Computational Biology* (pp. 1–6). [Crossref]

Jothikumar, R., Shanmugam, S. G., Nagarajan, M., Premkumar, S., & Asokan, A. (2019). Analyzes of mouth cancer using max-min composition in soft computing. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(3), 825–830. [Crossref]

Matyas, J. (1965). Random optimization. *Automation and Remote Control, 26*, 246–253.

Meteb, M. A. (2004). A hybrid deep learning model for breast cancer diagnosis based on transfer learning and pulse-coupled neural networks. [Unpublished manuscript or institutional publication].

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *Breast Cancer Research, 21*, Article 75. [Crossref]

Nallamala, S. H., Mishra, P., & Koneru, S. V. (2019). Qualitative metrics on breast cancer diagnosis with neuro fuzzy inference systems. *International Journal of Advanced Trends in Computer Science and Engineering, 8*(2), 259–264. [Crossref]

Nazeri, K., Aminpour, A., & Ebrahimi, M. (2018). Two-stage convolutional neural network for breast cancer histology image classification. In *International Conference Image Analysis and Recognition* (pp. 717–726). [Crossref]

Nuruddin Qaisar Bhuiyan, M., Shamsujjoha, M., Ripon, S. H., Proma, F. H., & Khan, F. (2019). Transfer learning and supervised classifier based prediction model for breast cancer. In *Handbook of Deep Learning in Biomedical Engineering* (pp. 1–20). Elsevier. [Crossref]

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology, 48*(3), 128–138. [Crossref]

Padmapriya, B., & Velmurugan, T. (2016). Classification algorithm based analysis of breast cancer data. *International Journal of Data Mining Techniques and Applications, 5*(1), 1–8. [Crossref]

Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ, 7*, e6201. [Crossref]

Scheidegger, F., Istrate, R., Mariani, G., Benini, L., Bekas, C., & Malossi, A. C. I. (2021). Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy. *The Visual Computer, 37*(6), 1593-1606. [Crossref]

Shajahaan, S. S., Shanthi, S., & Manochitra, V. (2013). Application of data mining techniques to model breast cancer data. *International Journal of Emerging Technology and Advanced Engineering, 3*(11), 1–8.

Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010). Various types and management of breast cancer. *Journal of Advanced Pharmaceutical Technology & Research, 1*(2), 109–126. [Crossref]

Shravya, C. H., Pravalika, K., & Subhani, S. (2019). Prediction of breast cancer using supervised machine learning techniques. *International Journal of Innovative Technology and Exploring Engineering, 8*(6), 1106–1110.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering, 63*(7), 1455–1462. [Crossref]

Vijayarani, S., & Divya, M. (2011). An efficient algorithm for generating classification rules. *International Journal of Computer Science and Technology, 2*(4), 512–515.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1992). *Breast cancer Wisconsin (diagnostic) data set*. UCI Machine Learning Repository. [Link]

Yildirim, A. E. (2023). *Prediction of breast cancer diagnosis with a recursive-self adaptive cohort intelligence feature selection and classification method* [Master's thesis, University of South Florida]. ProQuest Dissertations Publishing.

Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs, 2*(2), Article 13. [Crossref]