Breast Cancer Prediction Using Machine Learning

Nasheed Hossain Serajee, Saad Bin Mannan, Rawad Abdulghafor, Sharyar Wani, Adamu Abubakar, Akeem Olowolayemo

Department of Computer Science, Kulliyyah of Information & Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia

nasheed09@gmail.com, saad.mannan37@gmail.com, rawad@iium.edu.my, sharyarwani@iium.edu.my, akeem@iium.edu.my

Abstract— One of the most common cancers is breast cancer that occurs in women and it contributes greatly to the number of deaths that occur worldwide. Breast cancer is caused due to the presence of cancerous lumps inside the breast. A breast lump is a mass that develops in the breast. The lumps can be of various sizes and textures. The lumps found inside the breasts can be either cancerous or non-cancerous. If the lump is cancerous, then no diagnosis needs to be carried out. If the lump is found to be cancerous, then further diagnosis will be carried out to check whether the cancer has affected the rest of the body. The tests that are used for diagnosis are MRI, mammogram, ultrasound, and biopsy. Breast cancer is responsible for death of women from cancer. It is accountable for 16 percent of the overall deaths caused by cancer in the world. In this paper, we are going to predict whether lumps present in the breast are cancerous. To achieve this, we are going to make use of four algorithms which are Support Vector Machines (SVM), K-Nearest Neighbour (KNN). Random Forest and Naïve Bayes. We will compare the efficiency of the machine learning algorithms based on classification metrics and deduce the best one for this research.

Keywords— machine learning, breast cancer, dataset, algorithm

I. Introduction

Databases that contain medical data are increasing profusely. The databases comprise of data that includes examinations, measurements, tests, prescriptions etc. This huge data makes it difficult to discover useful information using traditional methods. Hence, the use of new methods and tools for learning valuable information from these data has replaced the use of traditional methods. Analysing this data with new analytical approaches to discover exciting patterns and unseen information is more demanding now.

The origin of breast cancer is inside the breast. The cancer appears when the cell growth is out of proportion. The cancer cells produce a tumour that is normally detected by performing x-ray. The tumour cells can also be felt as lumps inside the breast. They can be either cancerous or non-cancerous. When the tumour is cancerous, the cells will grow in an abnormal manner and form lumps inside the breast. When the cells do not spread to the rest of the body, then they are considered as non-cancerous. Fibro adenoma is the most common kind of cancerous breast tumour that needs to be surgically removed to stop the growth of the cancerous cells. Apart from that, no other treatment is necessary [1].

In a cancerous tumour, the malignant cells will disseminate to the rest of the body if the proper treatment is not carried out. For example, if a cancerous tumour inside the breast doesn't receive treatment, it may then spread to the muscles beneath the breast. It is also possible for it to develop into the skin that covers the breasts. At certain times, the cancer may spread to other organs of the body. They have the ability to enter the bloodstream or lymphatic system and spread throughout the body. When malignant cells encounter a new environment, they may detach, resulting in the formation of a new tumour. Secondary or metastasis refers to the new tumour. Breast cancer develops when cells in the ducts and lobules of the breast become malignant. If the breast cancer is detected at an early stage, it can be treated. But if the cancer spreads to different parts of the body, it becomes untreatable. Although, the cancer can be controlled for a long time but eventually it will lead to the patient's demise. When breast cancer occurs, it can cause various symptoms. According to [2], the symptoms include are breast pain, nipple pain, swelling of the breasts, irritation, changes in color of the breast or nipple, retraction of the nipples, nipple discharge, changes in breast size, size or appearance of the breast, sore on the breast or nipple, rough and continuous cough, changed appetite, night sweats, weight gain or loss, blood found in the urine or stool, pain after meals, weakness and tiredness. Breast cancer is the most common cancer among women all over the world. Every year, it affects roughly 2.1 million women and is the leading cause of cancer-related fatalities among women. According to [3], an estimation of 627000 deaths in women resulted from breast cancer, i.e., 15% of overall of deaths caused by cancer in women.

II. RELATED WORKS

Suleyman et al. [4] investigated a machine learning system for diagnosing patients with breast cancer. The Cancer Genome Atlas (TCGA) study assessed the entire exome sequencing data from 358 breast cancer patients with similar ethnicities. They employed scores from the non-negative matrix factorization approach to divide the patients into subgroups. They compared the three subgroups and uncovered 358 genes in the late-stage category that have significantly higher mutation rates. The functional characterization of these genes revealed that there are several essential functional gene families in the late state rich subset of patients that have a high mutational load. Finally, after completing all of the study, the researchers created a supervised classification model to predict the patients' stage. The classification model developed in this study might provide a reasonable prediction of a cancer patient's stage based on the profiles of their mutations.

The application of supervised machine learning algorithms to predict breast cancer was examined by Shravya et al. [5]. They used the UCI repository to obtain data for their investigation. They made use of Logistic Regression, Support Vector Machine (SVM), and K Nearest Neighbour on their dataset (KNN). The algorithms' accuracy, precision, sensitivity, specificity, and false positive rate, as well as their efficiency, were calculated and compared. They came to the conclusion that SVM was the best algorithm for predictive analysis, with a 92.7 percent accuracy.

Wenbin et al. [6] showed how machine learning algorithms may be utilized to predict breast cancer and for prognosis. They used different types of supervised algorithms on multiple breast cancer benchmark datasets. They observed the accuracy of every algorithm on different datasets. On WBCD dataset algorithms accuracy ranges from 94.36% to 99.90%. The scores are as follows: ANN scored 99.68%, SVM scored 99.10%, Decision Trees scored 99.9% and K-NN scored 98.70.

Afshar et.al [7] showed multiple methods that can be used for classifying breast cancer by using machine learning. They showed statistics of the datasets that have been used mostly for research purposes and then pointed out the algorithms that were used by the previous researchers. According to them SVM had the highest accuracy rate

III. METHODOLOGY

This chapter includes the methodology of conducting the research. It also contains the description of the algorithms that will be used for this research.

A. Structure of Research Methodology

The research papers related to ours will be thoroughly studied that will help us to compare our proposed topic with similar existent ones. Next, we will study the dataset and search for any errors in it. Later we will choose the algorithms that we feel are convenient for our research and compare them based on certain evaluation metrics. Eventually, the best algorithm will be determined for this research.

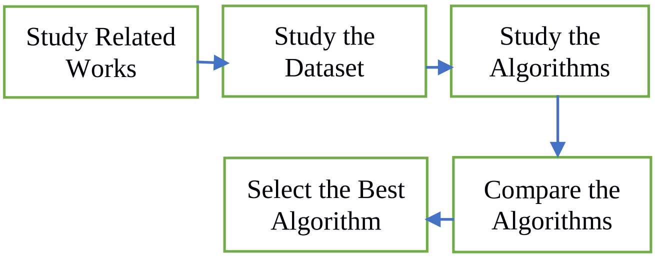


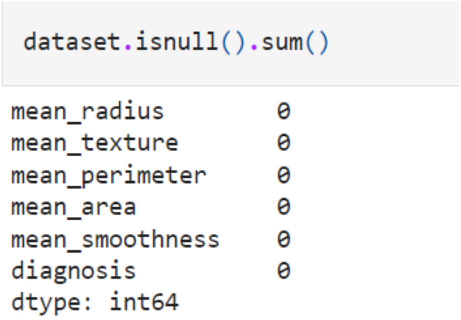
Fig. 1 Structure of Research Methodology

B. Selecting Best Algorithms

The algorithms mentioned above are all suited for classification but our aim is to find the best one among them. They can be compared by finding out the confusion matrix, precision and recall, receiver operating characteristic (ROC) area, accuracy, and other metrics. The algorithm with the most accurate and precise values will be declared the best.

C. Data Collection and Pre-Processing

The dataset that we are going to use in our research has been collected from kaggle.com. The name of the dataset is Breast Cancer Wisconsin Dataset. The dataset contains 6 features and over 5000 instances. Fig 2 shows that there are no missing values in the dataset and we concluded that the dataset needs no data modification. Hence, we will proceed with the dataset for further analysis.



```
dataset.isnull().sum()

mean_radius        0
mean_texture       0
mean_perimeter     0
mean_area          0
mean_smoothness    0
diagnosis          0
dtype: int64
```

Fig. 2 Missing values in the dataset

D. Machine Learning Algorithms

In our research, we are going to make use of four algorithms and compare the results. The algorithms are k-Nearest Neighbour (k-NN), Support Vector Machines (SVM), Random Forest and Naive Bayes Classifier. The dataset that we are going to use for our research is labelled which means that all the rows of data are being organized and have a distinct column name. According to the nature of our dataset, we will carry out Supervised Learning. In Supervised Learning, algorithms take a known set of input and output data and train machine learning models to predict the new output data based on unseen input data.

K-Nearest Neighbours(k-NN)

k-Nearest Neighbours (k-NN) is a type of supervised algorithm that is extensively used for classification and regression prediction problems. It is also identified as lazy learning algorithm because of not having a specialized training phase and makes use of all the data for training during classification. k-NN is also a non-parametric learning algorithm. When introduced to an unknown test data and after it has already "learned" from training data, it gives a numerical prediction based on a "similarity" measure, usually a distance function. In other words, when given a test data, using the similarity function it finds out to which training data it is most similar to and then gives a prediction [8]. The similarity function, usually the Euclidean distance function [8]:

$$\sqrt{(x - a)^2 + (y - b)^2} \text{ ---- eq} \qquad (1)$$

This equation is used to find out the minimum distance of the points to the k value. Apart from Euclidean distance, Manhattan Distance [8] can be used.

$$|x1 - x2| + |y1 - y2| \qquad (2)$$

The following steps describes how the k-NN algorithm [8] works:

Step 1 – The appropriate dataset is chosen and separated into two groups: training and testing.

Step 2 – the value of k, the number of nearest neighbours, is chosen. This value of k has to be an integer.

Step 3 – for each point in the testing set, the steps mentioned below are carried out:

- The Euclidean distance function is used to calculate the distance between testing data and each row of training data.
- The values are then sorted in ascending order.
- The top k rows will then be selected from the values that are sorted.
- The algorithm will then assign a class to the test point based on the rows' most recurrent class.

Support Vector Machines (SVM)

The support vector machine is used to find a hyperplane in an N-dimensional space (N — the number of features) that categorizes the data points in a specific way. Hyperplanes are decision boundaries that aid in the categorization of data points. The dimension of the hyperplane is determined by the number of columns (features) in the dataset. Many different hyperplanes can be used to separate the two types of data points. The best plane with the greatest margin, i.e. the greatest distance between the data points of the two classes [9], is chosen.

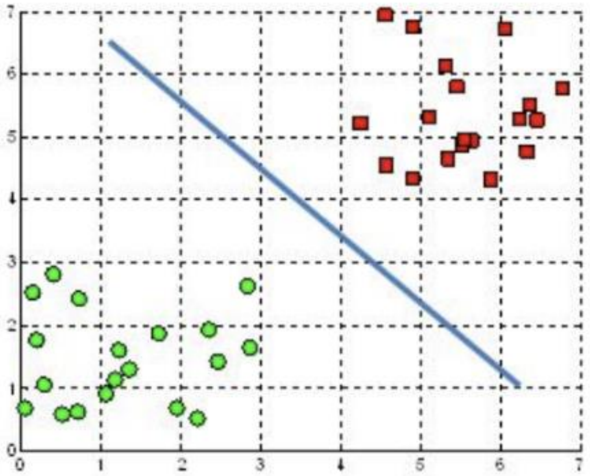


Fig. 3 Hyperplane in 2D

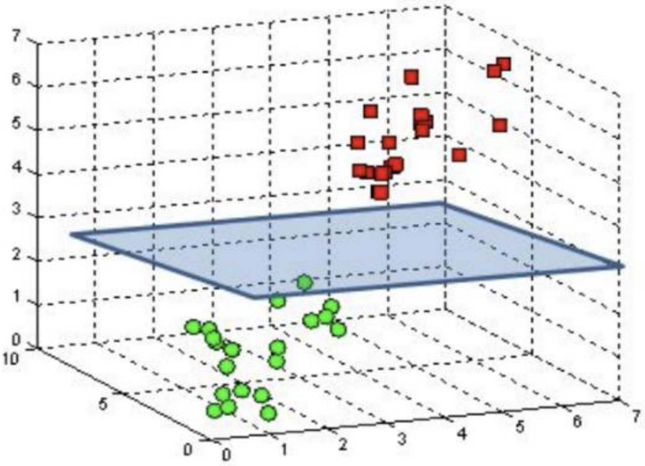


Fig. 4 Hyperplane in 3D

In the diagrams [12] above, the blue line separates the data into two different classified groups. This blue line is the classifier. Our dataset contains non-linear data and hence we must make use of a nonlinear SVM. The steps [9] for non-linear classification are:

- Define Φ – A mapping equation which is found using the kernel trick.
- Transform the points using Φ which will separate the data points.
- Plot the new data points.
- Choose the supporting vectors from the graph.
- Form quadratic equations using the support vectors.
- Find the value of gradient, y-intercept from the equations.
- Draw the line that separates the data points.

Naive Bayes Classifier

The Naive Bayes Classifier algorithm is a probabilistic algorithm used in classification problems. The algorithm used Bayes Theorem to predict the outcome. The Bayes Theorem [10] is described below:

$$P(A|B) = P(B|A) * P(A) / P(B) \qquad (3)$$

The variable B is the target variable, which represents a probability based on given conditions. The variable A represents the parameters/features.

The steps in Naive Bayes Classification [10] are:

- Calculate the prior probability from the dataset that contains different classes.
- Find Likelihood probability for every attribute of the data.
- Insert the values in the Bayes Formula (see eq3) and compute posterior probability.
- Find the class with higher probability.
- The types of Naive Bayes Classifier are:
- Multinomial Naive Bayes: Used for classification problem that uses labelled data.
- Bernoulli Naive Bayes: Used for classification problem that uses Boolean variables.
- Gaussian Naive Bayes: Used for classification that uses continuous data.

Random Forest

Random Forest is a supervised machine learning approach that uses decision trees to solve classification and regression problems. The Random Forest algorithm works as follows [11]:

- Take a random sample of 'n' records from the dataset.
- Create a decision tree with the 'n' records.
- Repeat the first two stages to get the ideal number of trees.
- Each tree in the forest predicts the category to which the new record belongs in classification tasks.

IV. EXPERIMENTATION

Building the Machine Learning Model

All of the four algorithms are used to build the model and eventually compare the results to select the best one among them. The steps towards building the machine learning model are described below:

- Data Collection – As mentioned before, the dataset that we are going to use in our research has been collected from kaggle.com. The name of the dataset is Breast Cancer Wisconsin Dataset. The dataset contains 6 features and over 5000 instances.
- Pre-Processing the Data – In this process, we will clean and prepare the data for training. Processes like cleaning the data (remove duplicate values, correct errors, input/remove missing values, normalization), randomizing the data, visualizing the data to help detect relevant relationships and at last dividing the data into training and testing sets.
- Train the Model - Creating the machine learning model with all of the algorithms proposed for this research. The training data will be fit into the classifier and the classifier will be used for prediction.
- Making predictions on new data – After successfully building our model, it will be ready to test on unseen data.
- Evaluate the Model - Metrics like as classification accuracy and the f1 score will be used to evaluate the model's efficiency and performance.

V. RESULTS

This section contains the results of the efficiency of the algorithms. The same machine learning model was created using the four different algorithms and the results are shown below:

TABLE I
EFFICIENCY OF DIFFERENT ALGORITHMS

Algorithm	ROC area	Classification Accuracy	F1 score	Precision	Recall
k-NN	0.925	0.889	0.887	0.889	0.889
SVM	0.980	0.932	0.932	0.933	0.932
Random Forest	0.972	0.925	0.924	0.925	0.925
Naive Bayes	0.962	0.909	0.909	0.909	0.909

In Table I, Support Vector Machine (SVM) has scored the highest percentage in all metrics used. The classification accuracy obtained by SVM (93.2%) is better than KNN, Random Forest and Naive Bayes that have an accuracy lower than SVM. The ROC curve is a metric that may be used to better understand the capabilities of a machine learning system. The percentage of accurately predicted positive observations among all observations is known as recall. By taking all the metrics into consideration we can conclude that SVM is the most suitable algorithm for this research.

V. CONCLUSION

There are a variety of machine learning methods that may be used to analyse medical data. The goal is to create a model that is both effective and accurate. On the Wisconsin Breast Cancer dataset, we used four key algorithms: SVM, Naive Bayes, k-NN, and Random Forest. We compared the efficacy of these algorithms using certain metrics like classification accuracy, precision, f1 score, recall and ROC area to find the best algorithm. SVM has the highest percentages in all of the above metrics and has proven to be the best algorithm for this research.

REFERENCES

[1] *"What Is Breast Cancer? | Breast Cancer Definition", Cancer.org, 2019. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.*

[2] *“Breast Cancer Prediction Using Data Mining Method”. [Online]. Available: https://www.researchgate.net/publication/319688741_Breast_Cancer_Prediction_Using_Data_Mining_Method.*

[3] *“Breast cancer", World Health Organization, 2019. [Online]. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.*

[4] *Suleyman Vural, Xiaosheng Wang and Chittibabu Guda, “Classification of breast cancer patients using somatic mutation profiles and machine learning approaches,” in The International Conference on Intelligent Biology and Medicine (ICIBM), Indianapolis, USA, 2015.*

[5] *Mogana Darshini Ganggayah, Nur Aishah Taib, Yip Cheng Har, Pietro Lio and Sarinder Kaur Dhillon, “Predicting factors for survival of breast cancer patients using machine learning techniques,” in BMC Medical Informatics and Decision Making, 2019.*

[6] *Hiba Asria, Hajar Mousannifb, Hassan Al Moatassimec, and Thomas Noeld, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” in The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS), 2016, pp 1064-1069.*

[7] *M. Tahmoores, A. Afshar, B. Bashari Rad, K. B. Nowshath and M. A. Bamiah, “Early Detection of Breast Cancer Using Machine Learning Techniques,” in Journal of Telecommunication, Electronic and Computer Engineering, vol. 10, 2018*

[8] *KNN algorithm - Finding nearest neighbors. (n.d.). RxJS, ggplot2, Python Data Persistence, Caffe2, PyBrain, Python Data Access, H2O, Colab, Theano, Flutter, KNime, Mean.js, Weka, Solidity. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.html*

[9] *Gandhi, R. (2018, July 5). Support vector machine — Introduction to machine learning algorithms Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47?gi=884a326d1cd2*

[10] *“Naive Bayes Classifier - Towards Data Science”. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c.*

[11] *“Random Forest Algorithm with Python and Scikit-Learn”. [Online]. Available: https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/.*

[12] *Kumar, N., 2021. Introduction to Support Vector Machines (SVMs). [online] MarkTechPost. Available at: <https://www.marktechpost.com/2021/03/25/introduction-to-support-vector-machines-svms/> [Accessed 28 November 2021].*