Wisconsin Diagnosis Breast Cancer Dataset: Classification Analysis of Feature Selection

Yihang Fang 20363451
School of Computer Science
University of Nottingham
China
Email: alyyf20@exmail.nottingham.ac.uk

Ziyi Zhao 20353100
School of Computer Science
University of Nottingham
China
Email: alyzz40@nottingham.ac.uk

***Abstract*— This study compares Random Forest, Logistic Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN), and XGBOOST on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset by calculating their classification test accuracy, sensitivity, and specificity. The dataset was divided using a K-fold cross-check in the following way: four copies of training data and one copy of test data for testing and all the classifiers' parameters were assigned. Results show that the best classifier was the Logistic Regression model with an average auc value close to 1, followed by the Random Forest models, while XGBOOST was the poorest model.**

Index Terms—Classification, XGBOOST, ANN, SVM, Random Forest, Logistic Regression

I. Introduction

This Cancer is one of the biggest human problems in the developing world, mostly affecting women [1], [2]. This study used Wisconsin Diagnostic Breast Cancer dataset for research analysis to explore the idea that no single arbitrary variable can be used alone as a criterion to distinguish between benign or malignant. In this study, the purpose of modelling is to address the following research questions:

1) Are the feature variables in the dataset effective for distinguishing benign from malignant conditions?
2) Which categorization model is most effective, and do the scoring criteria influence the results?
3) Identifying combinations of nine cellular characteristics that accurately predict malignancy.

In order to address the aforementioned concerns, this study analyses five classification algorithms: Random Forest, Logistic Regression , SVM, ANN and XGBOOST. Finally, a summary is presented by analysing their classification test accuracy, sensitivity, specificity and Auc values.

II. Literature Review

Previous study, as indicated in Table 1, has recognised the relevance of the same research issue by suggesting the application of machine learning (ML) techniques to identify breast tumors using the Wisconsin Diagnosis of Breast Cancer dataset[1], [3], which resulted in statistically significant findings.

TABLE I.
UMMARY OF STUDIES

Study	Methods Used	Results
Alshouili et al. 2019 [4]	Decision Jungle and Decision Tree	Classification accuracy rate was 96.5 %.
Bhardwaj et.al 2015 [5]	Artificia Neural network	Classification average accuracy rate were 97.73%, for 50–50 training–testing partitions.
Shen et al. 2014 [6]	Feature selection method and SVM	Classification accuracy rate was 92 %.
Jeyasingh and Veluchamy 2017 [7]	Feature selection, Random Forest and Modified Bat Algorithm (MBA)	Classification accuracy rate was 96.85%
Fitrah Umami and Sarno 2020 [8]	Generalized Linear Model, Logistic Regression, and Gradient Boosted Decision Tree	Classification accuracy rate of Generalized Linear Model was 99.4%
Hiba et al. 2016 [9]	SVM, Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree	Classification accuracy rate of SVM was 97.13%
Sinha et al. 2019 [10]	Decision Tree, Naive Bayes, SVM, and K-Nearest Neighbors (KNN)	Support Vector Machine (SVM) gives the highest accuracy of 97.3%
Mugdil et al. 2019 [11]	Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Logistic Regression , and Support Vector Machine (SVM)	Classification accuracy rate of K-Nearest Neighbors (KNN) was 98.0%

III.METHODOLOGY

As seen in Fig. 1, the proposed model is broken down into numerous parts, each of which is addressed in detail below.

Fig. 1. Data processing flow chart

A. *Data Set Information*

The Wisconsin Diagnosis Breast Cancer Dataset consists of 699 observations on 11 variables, one of which is a character variable, nine of which are ordered or nominal variables, and one target class. Each variable, except for the first variable Id, was transformed into 11 raw numeric attributes with values ranging from 0 to 10. We transformed the last variable into a factor variable for subsequent classification. The distribution of the data in the sample set in Figure 2 shows that benign cancers (2) 457 class and malignant cancer (4) 241 class.

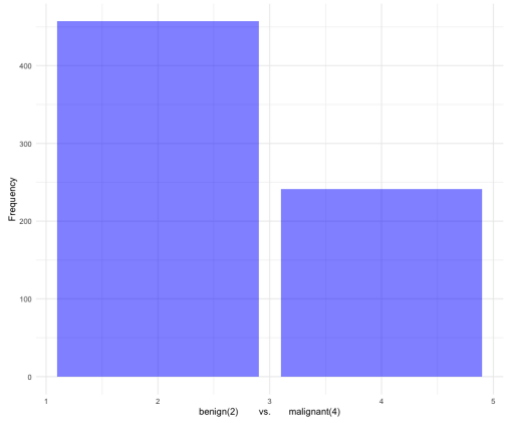

Fig. 2. Sample Distribution

B. *Dataset Preprocessing*

Initial analysis of the global observation data revealed that the variable Bare.nuclei contained 16 missing values, as depicted in Figure 3. There are two options for handling the values of these variables. Due to missing values comprise just 2.29 percent of the entire data, therefore the first way is to delete them. Another appropriate way is to replace them with the median or mean values. After conducting experiments, it was determined that removing the missing values is better than the other methods, since the data had more observations, and for simplicity, we chose to delete these 16 observations.

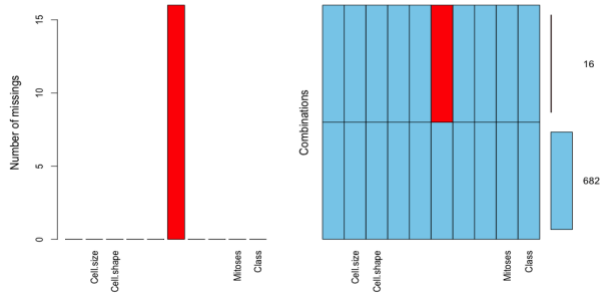

Fig. 3. Data set overview diagram

After addressing the missing values, we examine the distribution of the data characteristics and notice the presence of several outliers. Figure 4 illustrates a comparison of the processed data distribution. In addition, the dataset was normalized using Equation 1 such that the data adhered to a typical normal distribution, hence removing the influence of distribution discrepancies on the model's performance.

$$z = \frac{X-\mu}{\sigma} \tag{1}$$

where X is the feature to be standardized, μ is the mean value of the feature, and σ is the standard deviation of the feature [3].

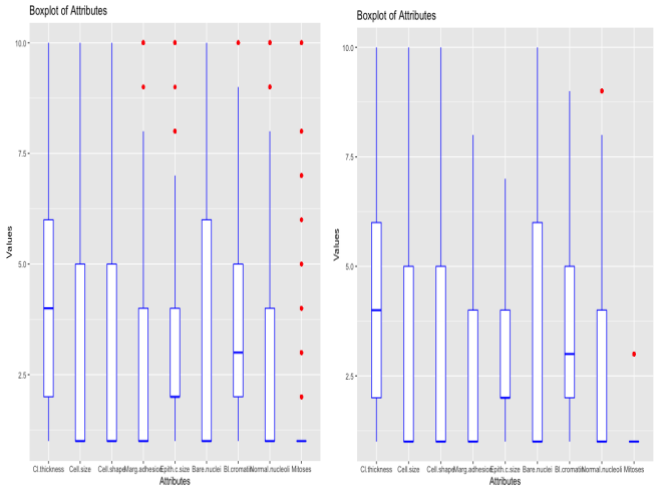

Fig. 4. Data boxplot before vs. after

It was important to ascertain whether or not the statistics were balanced [1]. As can be observed, there are roughly two times as many benign as malignant tumours in this dataset. The following step will be to do a variable correlation analysis in

order to visualise the relationship between attributes; a graph of this analysis is given in Figure 5.

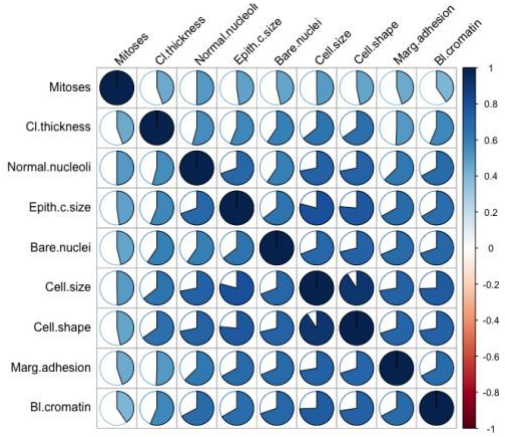

Fig. 5. Heatmap of the dataset features

C. Feature Extraction

This is a critical phase in the research process. Feature selection is a frequently used strategy for identifying meaningful data by deleting irrelevant and duplicate data [6]. In this study, we suggest that the most important attributes be taken into account. The first method is to select all features. And the other method is recursive feature elimination. Select th e most useful feature at first. Then keep traversing the remainin g features until the most useful N features are found. After feat ure selection, five features Bare.nuclei, Cl.thickness, Cell.size, Cell.shape, Bl.cromatin are selected. Results are shown in Tab le 2. In particular, except Random Forest, feature selection imp roves the auc-value on the result of other methods. (Hint: there is no difference in Logistic Regression).

TABLE II.

RESULTS OF FEAUTURE SELECTION

Classifier	**Average Auc (before feature selection)**	**Average Auc (after feature selection)**
Random Forest	0.9740	0.9634
SVM (radial)	0.9654	0.9693
ANN	0.9649	0.9696
XGBOOST	0.9559	0.9622

D. Proposed Method

This paper presents a comparison of five algorithms: Random Forest, Logistic Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN) and XGBOOST on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to enhance the classification accuracy. K-fold cross-validation is a type of cross-validation in which the whole data set is divided into k copies (each essentially equal), of which k-1 copies are used as training data. In this study, we set k=5, i.e. four copies of training data and one copy of test data for testing. This method selects the most useful features each time.

- *Random Forest*

Random forest is a group learning model for regression and classification problems [7]. Intuitively, a random forest is a set of several decision trees. Each decision tree is a classifier, so for an input sample, N trees will give N classifications. The random forest takes all of the classification votes and picks the category with the most votes as the final output.

As a consequence, employing a Random Forest rather than a decision tree is more logical in certain cases. There are four stages to the algorithm:

1) Assume that there are N samples in the training set. Randomly selecting N samples using put-back, then using the resultant N samples to train a decision tree.
2) If there are M features, choose m at random from M features at each segmentation, where m < M. Then choose the best segmentation variable (based on the information gain and Gini coefficient) among these m variables.
3) Repeat step 2 until no splitting is possible.
4) Forecasting through aggregated decision trees.

Figure 6 showed the Receiver Operating Characteristic (ROC) of Random Forest model, which means as the area under the ROC curve grows, or as the AUC approaches 1, the more effective the model becomes. The results show that the model works well.

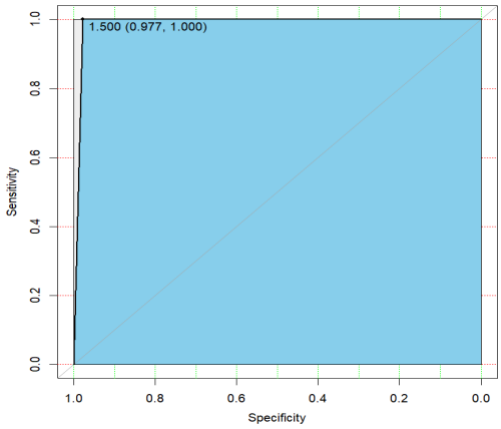

Fig. 6. ROC of sample Random Forest

- *Logistic Regression*

Logistic regression (Eq.2) is a supervised classification approach that produces binary output and is used to predict the output of a categorical dependent variable given a set of attributes or input. [2].

$$h_\theta(x) = g(\theta^T X + b) \tag{2}$$

Equation 3 is a sigmoid function that applies on $(\theta^T X + b)$, which is suitable for classification.

$$g(z) = \frac{1}{1+e^{-z}} \tag{3}$$

Sigmoid function maps a linear function between [0,1]. Therefore, selecting the appropriate threshold can be able to

classify the sample. The mean squared error (MSE) (Eq.4) is usually used for measuring the loss of the model.

$$J_\theta = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2 \quad (4)$$

Where y is the true class in the dataset. To find the best parameter θ, stochastic gradient descent minimized the loss of J_θ. According to Figure 7, while the sensitivity value of Random Forest is higher than Logistic Regression. Overall, Random Forest model is more stable than the Logistic Regression model.

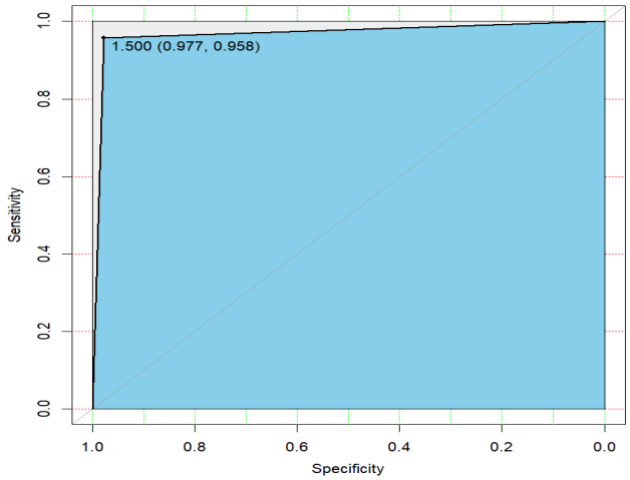

Fig. 7. ROC of sample Logistic Regression

- *Support Vector Machine (SVM)*

The support vector machine is a widely used machine learning tool. The algorithm's purpose is to find an N-dimensional hyperplane that classifies the data points [1]. The majority of this method is spent determining which plane optimizes the margin, which is largely used for binary classification [3]. The main task of SVM is finding a hyperplane (Eq.5) to maximize the margin.

$$\omega^T X + b = 0 \quad (5)$$

For this study, 4 different kernel functions (sigmoid, linear, polynomial, and radial) (Eq. 6) are used.

$$K(x_1, x_2) = \varphi(x_1)^T \varphi(x_2) \quad (6)$$

Then optimize the problem to find the best hyperplane for support vector machines. Where C is the penalty parameter, δ^i is the cost function.

$$min\frac{1}{2}\left|\left|\omega^2\right|\right| + C\sum_{i=1}^{N}\delta^i \quad (7)$$

$$\delta^i \geq 0 \quad (8)$$

$$y_i(\omega^T X_i + b) >= 1\text{-}\delta^i \quad (9)$$

The results of the various functions are summarized in Table 3, and the graphs for four different functions of SVM are shown in Figure 8. It is clear that the sigmoid function performs best, with an accuracy of 96.78 percent and an average AUC value of 0.9694.

TABLE III.

ESULTS OF DIFFERENT FUNCTION RUNS

Function	Acc (%)	Sec (%)	Spec (%)	Average Auc
radial	96.64	96.23	96.85	0.9654
linear	96.78	96.21	97.08	0.9665
polynomial	95.75	90.35	98.65	0.9450
sigmoid	96.78	97.48	96.40	0.9694

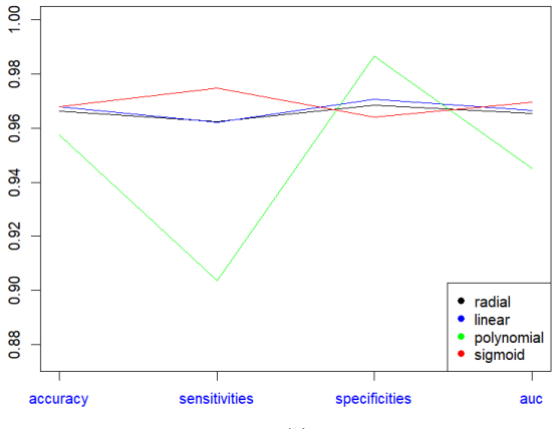

Fig. 8. Accuracy, sensitivities, specificities, and auc of four functions of SVM

- *Artificial Neural Network (ANN)*

Artificial neural networks (ANN), a widely popular machine learning approach, are inspired by the biological neural networks seen in the human brain [5]. There are three layers in an ANN: an input layer, an output layer, and a hidden input layer. The output of the input layer is sent into the hidden layer as its input. The activation function, which transforms linear input into nonlinear input, is used to move data to the next layer. The whole process calculates the weight of each variable and each layer and calculates the error, then traverses each layer by backpropagation to measure the error contribution of each connection, and finally adjusts the weights and biases to reduce the error.

$$h_\theta(x) = \theta^T X + b \quad (10)$$

$$g(z) = \frac{1}{1+e^{-z}} \quad (11)$$

For this study, the activation function showed in Eq.11. The input for the input layer showed in Eq.10. The Figure 9 describes the model for this study which shows that there is only 1 hidden layer in the model.

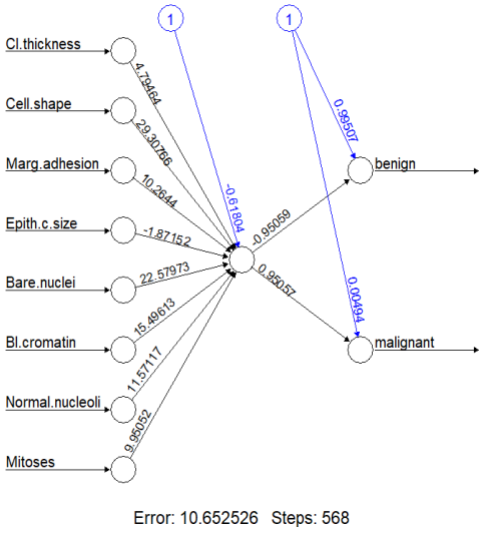

Fig. 9. ANN Analysis result

- *XGBOOST*

XGBoost is a decision tree-based boosting machine learning technique. It is a distributed gradient augmentation library that has been optimized to be efficient, adaptable, and portable. Each decision tree predicts the residual between the true value and the previous residual of the sum of the predicted values of all decision trees. The predicted values of all decision trees are the final result obtained by summing up all the predicted values.

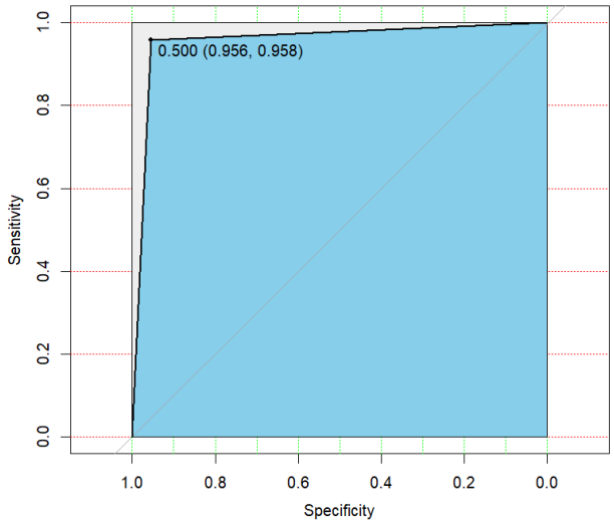

Fig. 10. ROC of sample XGBoost

According to Figure 10, the area under the ROC curve of the XGBoost model is smaller than both the Random Forest and Logistic Regression models, which represents a lower accuracy rate.

IV. Result and Discussion

A. Comparative Analysis

Many studies have used the confusion matrix in Table 4 below to evaluate the model [8], [10]. It is commonly used as a visualization tool that shows classification accuracy, sensitivity, and specificity. This is used to compare outcomes of actual and predicted classes.

TABLE IV.
ONFUSION MATRIX FOR BREAST CANCER DATASETS

Actual \ Predict	**Benign**	**Malignant**
Benign	True Benign (TB)	False Malignant (FM)
Malignant	False Benign (FB)	True Malignant (TM)

True Benign: Probability (+) that an individual is in the Benign stage.

False Benign: Negative probability of a benign stage (-).

True Malignant: Probability (-) that an individual is in the Malignant stage.

False Malignant: Probability of a positive malignant phase test (+).

Accuracy: Actual classifications divided by total classifications in the dataset.

$$\text{Accuracy} = \frac{TB+TM}{TB+TM+FB+FM} \quad (12)$$

$$Sensitivity = \frac{TB}{TP+FM} \quad (13)$$

$$Specificity = \frac{TM}{TM+FB} \quad (14)$$

B. Performance Analysis

The results of classifier algorithms are shown in Table 5. On this dataset, all five algorithms achieved an accuracy of 96 percent or higher, with Random Forest achieving the best accuracy (97.36 percent) and SVM coming in second (96.78 percent). The best classifier was a Logistic Regression model with an average auc value close to 1, followed by the Random Forest models, while XGBOOST was the poorest model. On the basis of the method we used for the simple classification, the classification outcomes are extremely close.

TABLE V.
ERFORMANCE TABLE OF CLASSIFIER ALGORITHMS

Classifier	**Acc (%)**	**Sec (%)**	**Spec (%)**	**Average Auc**
Random Forest	97.36	97.50	97.29	0.9740
Logistic Regression	96.48	94.55	97.53	0.9944

SVM (sigmoid)	96.78	97.48	96.40	0.9694
ANN	96.20	97.48	95.50	0.9649
XGBOOST	96.05	94.11	97.08	0.9559

V.CONCLUSION AND FUTURE WORK

Breast cancer is one of the most dangerous diseases affecting women. In this study, we use different machine learning algorithms to make predictions about people's diagnosis. This study demonstrates that Logistic Regression, SVM, and Random Forest are the most appropriate categorization methods. In addition, removing missing values is the best way of data pre-processing. There also have some limitations, such as different methods of feature selection on datasets; future research might try other methods of feature selection. Alternatively, our k-fold cross-validation grouping is small due to the size of our dataset. If massive amounts of data are encountered in future applications, we can select n times k-fold to increase the model's precision.

CONTRIBUTIONS

In the early stages of the project, Yihang Fang was in charge of implementing three algorithms: Artificial Neural Network (ANN), XGBOOST, and Random Forest, while Ziyi Zhao was in charge of Logistic Regression, Support Vector Machine (SVM), and Decision Tree. In the middle of the project, each coder checked and refined the code, and after deliberation, the application of Decision Tree was eventually removed since we felt it duplicated the performance of Random Forests and so was not displayed in the study. Later in the project, the paper was co-authored, with equal contributions from both authors.

REFERENCES

[1] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare (Switzerland)*, vol. 8, no. 2, 2020, doi: 10.3390/healthcare8020111.

[2] A. K. Dubey, U. Gupta, and S. Jain, "Breast cancer statistics and prediction methodology: A systematic review and analysis," *Asian Pacific Journal of Cancer Prevention*, vol. 16, no. 10, pp. 4237–4245, 2015, doi: 10.7314/APJCP.2015.16.10.4237.

[3] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," *Proceedings of the 2nd international conference on machine learning and soft computing*, vol. 47, no. 5 S, pp. 5–9, 2018, doi: 10.1002/1097-0142(19810301)47:5+<1164::AID-CNCR2820471318>3.0.CO;2-K.

[4] K. Alshouiliy, A. Shivanna, S. Ray, A. Alghamdi, and D. P. Agrawal, "Analysis and Prediction of Breast Cancer using AzureML Platform," *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2019*, pp. 212–218, 2019, doi: 10.1109/IEMCON.2019.8936294.

[5] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4611–4620, 2015, doi: 10.1016/j.eswa.2015.01.065.

[6] R. Shen, Y. Yang, and F. Shao, "Intelligent breast cancer prediction model using data mining techniques," *Proceedings - 2014 6th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2014*, vol. 1, pp. 384–387, 2014, doi: 10.1109/IHMSC.2014.100.

[7] S. Jeyasingh and M. Veluchamy, "Modified bat algorithm for feature selection with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset," *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 5, pp. 1257–1264, 2017, doi: 10.22034/APJCP.2017.18.5.1257.

[8] R. Fitrah Umami and R. Sarno, "Analysis of classification algorithm for Wisconsin diagnosis breast cancer data study," *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, pp. 464–469, 2020, doi: 10.1109/iSemantic50169.2020.9234295.

[9] H. Asri, H. Mousannif, H. al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Computer Science*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[10] A. Sinha, B. Sahoo, S. Swarup Rautaray, and M. Pandey, "An Optimized Model for Breast Cancer Prediction Using Frequent Itemsets Mining," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 5, pp. 11–18, 2019, doi: 10.5815/ijieeb.2019.05.02.

[11] P. Mudgil, "Breast cancer prediction algorithms analysis," *International Journal of Advance Research*, vol. 5, no. 3, pp. 424–427, 2019, [Online]. Available: www.IJARIIT.com