Comparative Performance Evaluation Of Machine Learning Algorithms For Breast Cancer Prediction

Emmanuel Nsese Udoiwod[1]
Department Of Computer Engineering
University Of Uyo, Akwa Ibom State
emmanueludoiwod@uniuyo.edu.ng

Ahiara, W . C. [2]
Department of Computer Engineering
Michael Okpara University of Agriculture Umudike, Abia State
ahiara.wilson@mouau.edu.ng

Umoren Mfonobong A.[3]
Department of Electrical /Electronic Engineering,
University of Uyo, Nigeria
mfon4gigis@yahoo.com

***Abstract*— In this paper, comparative performance evaluation of machine learning algorithms for breast cancer prediction is presented. The machine learning algorithms are decision tree, random forest, K-nearest neighbors, logistic regression and the support vector machine. The Wisconsin breast cancer (WBC) dataset available online in Kaggle repository is employed. The WBC Dataset consists of 569 patients' records with 35 columns per record. The acquired dataset was preprocessed; notably irrelevant features in the data records were removed. Also, feature scaling and 5-fold data splitting were conducted. Each of the five machine learning models were iteratively trained and also validated using each of the fivefold data divided into 80 % training and 20 % validation dataset. The model results show that for the training dataset, the Decision tree algorithm has the best F1 score of 99.331% followed by the Logistic Regression model with F1 score of 98.733 % while the KNN model has the worst F1 score of 96.003 %. On the other hand, for the validation dataset, the SVM algorithm has the best F1 score of 96.696% followed by the Logistic Regression model with F1 score of 96.678 % while the Decision tree model has the worst F1 score of 91.673 %. In addition, the confusion matrix results show that the logistic regression model gave the lowest number of false predictions and the highest number of true or correct predictions. As such, among the five machine learning models studied, the logistic regression model has the best prediction performance and hence is recommended for breast cancer prediction.**

Keywords*— Decision Tree, Breast Cancer Prediction, Random Forest, Wisconsin Breast Cancer Dataset, Machine Learning Algorithms, K-Nearest Neighbors , K-Nearest Neighbors, Logistic Regression, Support Vector Machine

1. INTRODUCTION

Across the globe, cancer has been identified as one of the worst diseases which claims millions of lives every year [1,2,3]. Among the different kinds of cancer, breast cancer is the one most prevalent among women [4,5,6]. There were about 2.26 million new instances of breast cancer worldwide in 2020 [7], making it the most prevalent kind of cancer. Additionally, it is the most prevalent form of cancer in women in both developed and developing nations, which is a significant issue for public health [8,9,10].

In any case, early detection and treatment has been identified as a more suitable way to address the rising challenges posed by breast cancer [11,12,13,14,15]. In this wise, medical experts are adopting approaches that can be used to detect breast cancer at the early stage of its development or to predict the likelihood of its occurrence in a patient so as to take proactive measures to avert it occurrence. Notably, machine learning solutions are employed in recent years to assist medical expert to analyze medical records of patients and thereby predict the likelihood of breast cancer [16,17,18,19,20]. Accordingly, this work is focused on the application of five different machine learning models to predict breast cancer based on a case study dataset of medical records of breast cancer patients. The machine learning algorithms considered in this study includes decision tree, random forest, K-nearest neighbors, logistic regression and the support vector machine [21,22,23]. The prediction performance of the five models are evaluated through metrics like the F1 score, true

positive, true negative , false positive and false negative predictions [24,25]. The overall essence of the study is to identify the machine learning model that is most suitable for breast caner prediction.

2. METHODOLOGY

The performance of the five machine learning models in predicting breast cancer is considered in this work. The five models are decision tree, random forest, K nearest neighbors, logistic regression and the support vector machine. The Wisconsin breast cancer (WBC) dataset available in Kaggle repository is employed in the study [26,27,28]. The WBC dataset consists of patients' 569 records with 35 columns per record, as shown in Table 1.

Table 1 The columns in the WBC data records

S/N	Features	Count	Data Type
0	id	569 non-null	int64
1	Diagnosis	569 non-null	object
2	radius mean	569 non-null	float64
3	texture mean	569 non-null	float64
4	perimeter mean	569 non-null	float64
5	area mean	569 non-null	float64
6	smoothness mean	569 non-null	float64
7	compactness mean	569 non-null	float64
8	concavity mean	569 non-null	float64
9	concave points mean	569 non-null	float64
10	symmetry mean	569 non-null	float64
11	fractal_dimension_mean	569 non-null	float64
12	radius_se	569 non-null	float64
13	texture_se	569 non-null	float64
14	perimeter_se	569 non-null	float64
15	area_se	569 non-null	float64
16	smoothness_se	569 non-null	float64
17	compactness_se	569 non-null	float64
18	concavity_se	569 non-null	float64
19	concave points_se	569 non-null	float64
20	symmetry_se	569 non-null	float64
21	fractal_dimension_se	569 non-null	float64
22	radius_worst	569 non-null	float64
23	texture_worst	569 non-null	float64
24	perimeter_worst	569 non-null	float64
25	area_worst	569 non-null	float64
26	smoothness_worst	569 non-null	float64
27	compactness_worst	569 non-null	float64
28	concavity_worst	569 non-null	float64
29	concave points_worst	569 non-null	float64
30	symmetry_worst	569 non-null	float64
31	fractal_dimension_worst	569 non-null	float64
32	Unnamed: 32	0 non-null	float64

2.1 Data Cleaning

The acquired dataset was preprocessed and the first step of data pre-processing (cleaning) is to remove irrelevant features. The *'Unnamed:32'* and *'id'* columns (in Table 1) are removed. The next step is to separate the features and the target variable. The "*diagnosis*" column the *target variable* and was employed to help with the prediction.

In the target variable, 'M' represents malignant cancer, while 'B' represents benign cancer (as shown in the screenshot of Figure 1). These values are converted to numbers before the training of the machine learning algorithms started. Label encoding was also performed to convert the categorical variables in the target column to numbers. The number 0 represents the benign cancer class while the number 1 represents the malignant cancer class.

The dataset is split into the training (which is 80% of the data) and test set (which is 20% of the data). Specifically, there are 455 records in the training set (as shown in Figure 2) and 114 records in the test set (as shown in Figure 3). Specifically, the 80 5 by 20 % data split is used in a 5-fold data splitting technique which is used to train and validate the each of the five models.

The last step in the data cleaning process is feature scaling. The method of feature scaling used is standardization. Standardization is applied to the training set and test set to keep all the features on the same scale (as

shown in Figure 4). It also helps to speed up the training process. The formula for standardization is as follows:

$$x = [x - \text{mean}(x)]/\text{standard deviation}(x) \quad (1)$$

```
Matrix of features
[[1.799e+01 1.038e+01 1.228e+02 ... 2.654e-01 4.601e-01 1.189e-01]
 [2.057e+01 1.777e+01 1.329e+02 ... 1.860e-01 2.750e-01 8.902e-02]
 [1.969e+01 2.125e+01 1.300e+02 ... 2.430e-01 3.613e-01 8.758e-02]
 ...
 [1.660e+01 2.808e+01 1.083e+02 ... 1.418e-01 2.218e-01 7.820e-02]
 [2.060e+01 2.933e+01 1.401e+02 ... 2.650e-01 4.087e-01 1.240e-01]
 [7.760e+00 2.454e+01 4.792e+01 ... 0.000e+00 2.871e-01 7.039e-02]]
Target Variable
['M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M'
 'M' 'B' 'B' 'B' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M'
 'M' 'B' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M'
 'M' 'B' 'M' 'M' 'B' 'B' 'B' 'B' 'M' 'B' 'M' 'M' 'B' 'B' 'B' 'B' 'M' 'B'
 'M' 'M' 'B' 'M' 'B' 'M' 'M' 'B' 'B' 'B' 'M' 'M' 'B' 'M' 'M' 'M' 'B' 'B'
 'B' 'M' 'B' 'B' 'M' 'M' 'B' 'B' 'B' 'M' 'M' 'B' 'B' 'B' 'B' 'M' 'B' 'B'
 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'M' 'M' 'B' 'B' 'B'
 'M' 'M' 'B' 'M' 'B' 'M' 'M' 'B' 'M' 'M' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'B'
 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'M'
 'M' 'B' 'M' 'B' 'B' 'M' 'M' 'B' 'B' 'M' 'M' 'B' 'B' 'B' 'B' 'M' 'B' 'B'
 'M' 'M' 'M' 'B' 'M' 'B' 'M' 'B' 'B' 'B' 'M' 'B' 'B' 'M' 'M' 'B' 'M' 'M'
 'M' 'M' 'B' 'M' 'M' 'M' 'B' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'M' 'M' 'M'
 'B' 'B' 'M' 'M' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'B' 'B' 'M'
 'B' 'B' 'M' 'M' 'B' 'M' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'B'
 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'M' 'B' 'B' 'B' 'B'
 'B' 'B' 'M' 'B' 'M' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'M' 'B' 'B' 'B' 'B'
 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'B' 'B' 'B'
 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'M' 'B' 'M'
 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'B' 'B' 'B' 'M' 'B' 'M' 'B' 'M' 'B' 'B'
 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'M' 'B' 'B' 'B' 'B' 'B' 'B'
 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'B' 'M' 'M' 'M' 'B' 'M' 'M' 'B' 'B' 'B' 'B'
 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'M' 'B' 'B' 'M' 'M' 'B' 'B'
 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B'
 'M' 'B' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'B'
 'M' 'M' 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'M' 'B' 'B' 'M' 'B' 'M' 'B' 'B' 'M'
 'B' 'M' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'B' 'M' 'M' 'B' 'B' 'B' 'B' 'B' 'B']
```

Figure 1 Matrix of features and target variables separated

```
Matrix of features (Training set)
[[1.602e+01 2.324e+01 1.027e+02 ... 9.975e-02 2.948e-01 8.452e-02]
 [1.232e+01 1.239e+01 7.885e+01 ... 9.391e-02 2.827e-01 6.771e-02]
 [1.285e+01 2.137e+01 8.263e+01 ... 5.601e-02 2.488e-01 8.151e-02]
 ...
 [1.300e+01 2.513e+01 8.261e+01 ... 5.921e-02 2.306e-01 6.291e-02]
 [1.420e+01 2.053e+01 9.241e+01 ... 1.339e-01 2.534e-01 7.858e-02]
 [1.706e+01 2.100e+01 1.118e+02 ... 1.827e-01 2.623e-01 7.599e-02]]
-------------------------------------------------
Target Variable (Training set)
[1 0 0 1 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 1 1 0 0 0
 1 0 0 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 1 0 0 0 1 0 0 1 0 0 0 1 0
 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 1 1 1
 0 0 0 0 0 1 0 0 1 1 0 1 0 1 0 0 1 0 0 1 1 0 0 0 0 1 0 1 0 0 1 1 0 1 1 1 0
 1 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 0 0 1 1
 0 0 1 0 1 0 1 0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 0 0 1 0 1 0 1
 1 1 0 0 1 0 0 1 1 0 0 1 1 1 0 1 0 0 0 0 0 0 0 1 1 0 1 0 0 1 1 0 0 0 1 1 0
 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 1 1 1 0 1 1 1 0
 0 0 0 1 1 1 0 0 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0
 1 0 0 1 1 1 0 1 1 0 0 0 0 1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 1
 0 1 0 1 0 0 1 1 0 0 1 0 1 1 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0 0 1 0 1
 0 1 0 1 1 1 0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1
 1 0 0 0 0 0 0 0 0 0 1]
```

Figure 2 The cut section of the screenshot to the records in the training set

```
Matrix of features (Test set)
[[1.141e+01 1.082e+01 7.334e+01 ... 8.958e-02 3.016e-01 8.523e-02]
 [2.094e+01 2.356e+01 1.389e+02 ... 2.105e-01 3.126e-01 7.849e-02]
 [1.617e+01 1.607e+01 1.063e+02 ... 1.251e-01 3.153e-01 8.960e-02]
 ...
 [1.453e+01 1.934e+01 9.425e+01 ... 9.594e-02 2.471e-01 7.463e-02]
 [1.512e+01 1.668e+01 9.878e+01 ... 1.252e-01 3.415e-01 9.740e-02]
 [1.607e+01 1.965e+01 1.041e+02 ... 1.520e-01 2.650e-01 6.387e-02]]
-------------------------------------------------
Target Variable (Test set)
[0 1 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 0 0
 0 1 0 0 0 1 0 0 1 0 1 0 0 1 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 1 0 1 0 1 1 0
 0 0 0 1 0 1 1 1 0 0 0 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0
 0 1 1]
```

Figure 3 The cut section of the screenshot to the records in the test set

```
Scaled Matrix of features (Training set)
[[ 0.51855873  0.89182579  0.4246317  ... -0.23574392  0.05456632
   0.02183673]
 [-0.51636409 -1.63971029 -0.54134872 ... -0.32320788 -0.13757624
  -0.90440164]
 [-0.36811839  0.45551496 -0.38824993 ... -0.89082504 -0.675893
  -0.14401559]
 ...
 [-0.32616206  1.33280304 -0.38905998 ... -0.84289958 -0.96490082
  -1.16888375]
 [ 0.00948859  0.25952507  0.00786279 ...  0.27571059 -0.60284707
  -0.30545988]
 [ 0.80945595  0.36918608  0.79320285 ...  1.00657384 -0.46151907
  -0.44817001]]

Scaled Matrix of features (Testing set)
[[-0.77089916 -2.00602473 -0.76451652 ... -0.38805702  0.16254727
   0.06095804]
 [ 1.89472636  0.96648861  1.89081582 ...  1.42292627  0.33722232
  -0.31041892]
 [ 0.56051506 -0.78108791  0.57044007 ...  0.14391558  0.38009711
   0.30174696]
 ...
 [ 0.10179251 -0.01812727  0.08238707 ... -0.29280517 -0.70288824
  -0.52310661]
 [ 0.26682074 -0.63876192  0.26586259 ...  0.14541325  0.79614133
   0.73153038]
 [ 0.53254417  0.05420233  0.48133495 ...  0.54678897 -0.41864428
  -1.11598733]]
```

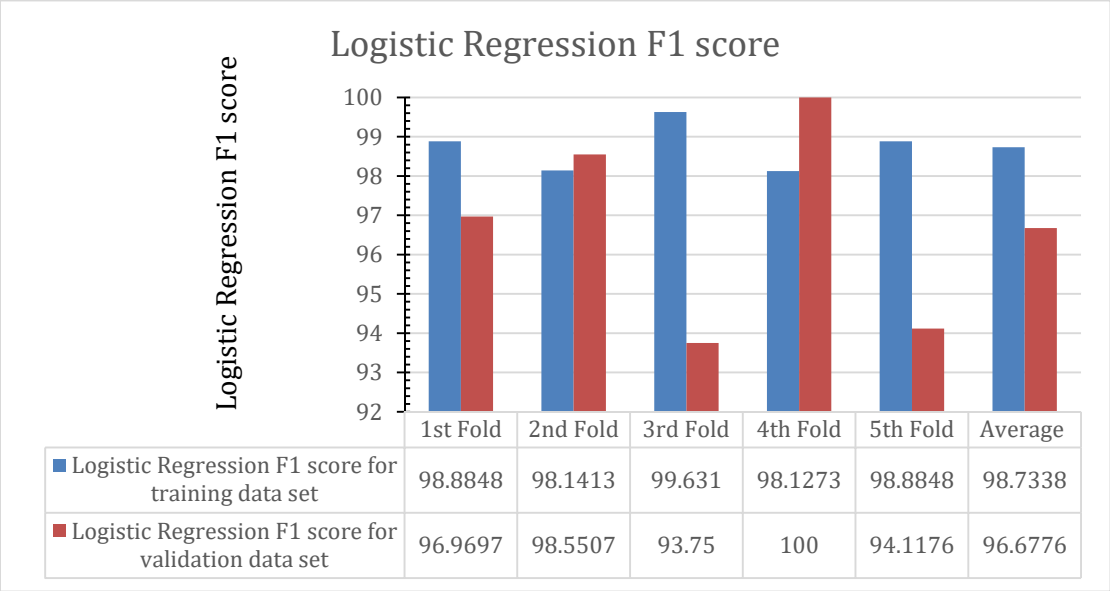
Figure 4 Matrix of features of training and test set after feature scaling

3. RESULTS AND DISCUSSION

The fivefold training and validation results of the five models are shown in Figure 5 for the Logistic Regression model, Figure 6 for the KNN F1 score, Figure 7 for the SVM F1 score, Figure 8 for the Decision Tree F1 score and Figure 9 for the Random Forest F1 score. In addition, the mean F1 score obtained for the five models are shown in Figure 10 for the mean F1 scores obtained from the training dataset and Figure 11 for the mean F1 scores obtained from the validation dataset. The results show that for the training dataset, the Decision tree algorithm has the best F1 score of 99.331% followed by the Logistic Regression model with F1 score of 98.733 % while the KNN model has the worst F1 score of 96.003 %. On the other hand, for the validation dataset, the SVM algorithm has the best F1 score of 96.696% followed by the Logistic Regression model with F1 score of 96.678 % while the Decision tree model has the worst F1 score of 91.673 %.

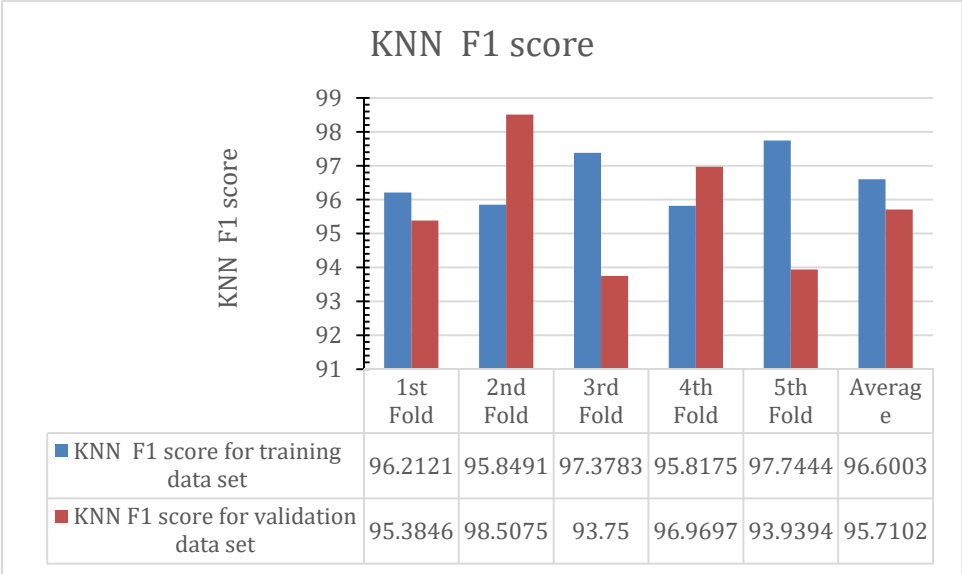
The summary of confusion matrix result is presented in Figure 12 and it shows that True and False Prediction Performance of each of the five models. In this study, the class of interest is the 'Malignant' class. Hence, according to the summary of confusion matrix result presented in Figure 12, the logistic regression model correctly classified 40 patients with malignant cancer. It misclassified 1 patient as having malignant cancer, whereas the patient's cancer was benign. It correctly classified 71 patients as having benign cancer. It misclassified 2 patients as having benign cancer, whereas they actually have malignant cancer. The KNN model correctly classified 38 patients with malignant cancer. It misclassified 1 patient as having malignant cancer, whereas the patient's cancer was benign. It correctly classified 71 patients as having benign cancer. It misclassified 4 patients as having benign cancer, whereas they actually have malignant cancer. The SVM model correctly classified 39 patients with malignant cancer. It misclassified 0 patients as having malignant cancer, whereas the patient's cancer was benign. It correctly classified 72 patients as having benign cancer. It misclassified 3 patients as having benign cancer, whereas they actually have malignant cancer. The Decision Tree model correctly classified 37 patients with malignant cancer. It misclassified 1 patient as having malignant cancer, whereas the patient's cancer was benign. It correctly classified 71 patients as having benign cancer. It misclassified 5 patients as having benign cancer, whereas they actually have malignant cancer. The Random Forest model correctly classified 38 patients with malignant cancer. It misclassified 0 patient as having malignant cancer, whereas the patient's cancer was benign. It correctly classified 72 patients as having benign cancer. It misclassified 4 patients as having benign cancer, whereas they actually have malignant cancer. The logistic regression model gave the lowest number of false negatives. This is

very important in the Healthcare industry. As such, among the five machine learning models studied, the logistic regression model has the best prediction performance for breast cancer.



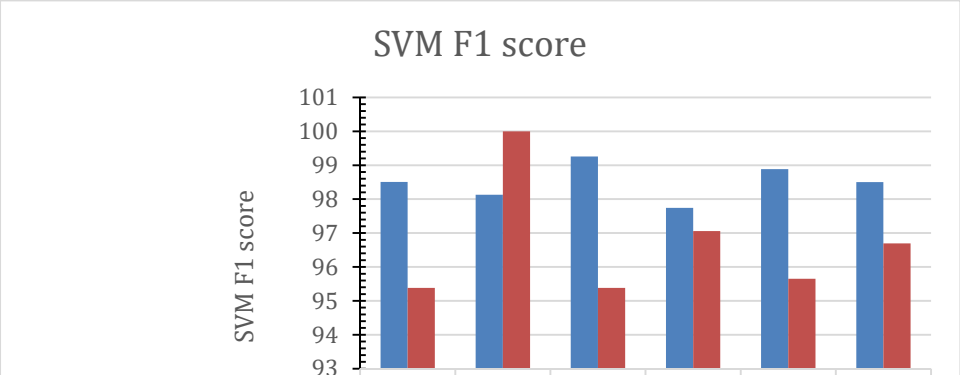
	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
Logistic Regression F1 score for training data set	98.8848	98.1413	99.631	98.1273	98.8848	98.7338
Logistic Regression F1 score for validation data set	96.9697	98.5507	93.75	100	94.1176	96.6776

Figure 5 The Logistic Regression F1 score



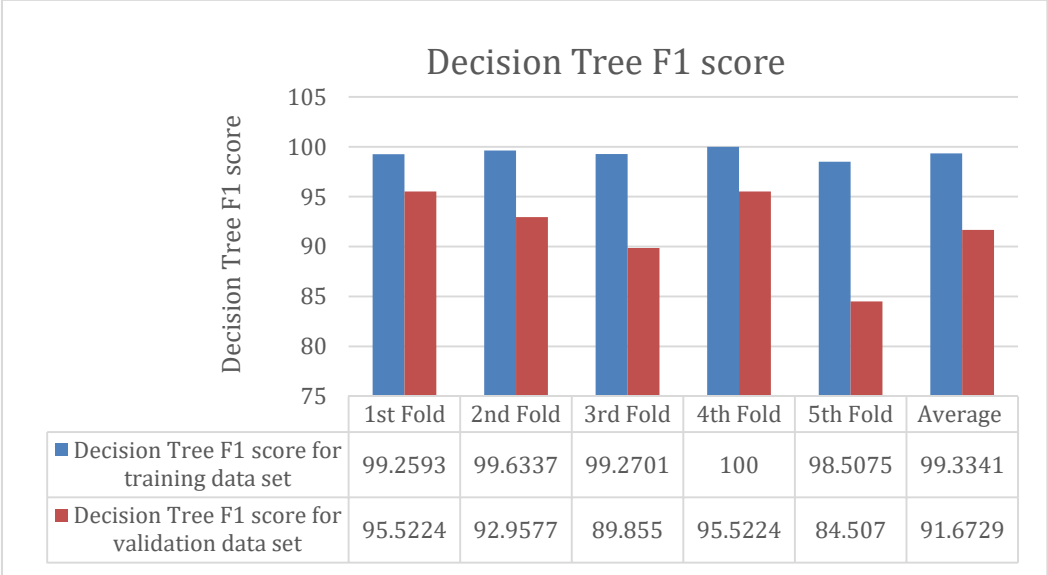
	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
KNN F1 score for training data set	96.2121	95.8491	97.3783	95.8175	97.7444	96.6003
KNN F1 score for validation data set	95.3846	98.5075	93.75	96.9697	93.9394	95.7102

Figure 6 The KNN F1 score



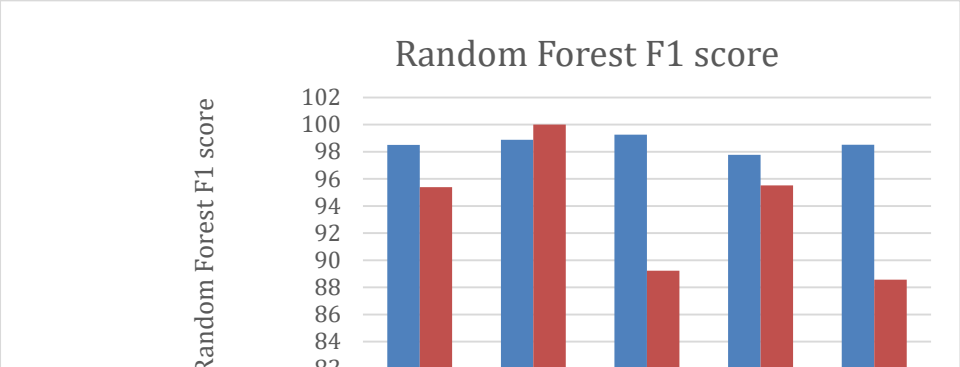
	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
SVM F1 score for training data set	98.5075	98.1273	99.2593	97.7444	98.8848	98.5046
SVM F1 score for validation data set	95.3846	100	95.3846	97.0588	95.6522	96.696

Figure 7 The SVM F1 score



	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
Decision Tree F1 score for training data set	99.2593	99.6337	99.2701	100	98.5075	99.3341
Decision Tree F1 score for validation data set	95.5224	92.9577	89.855	95.5224	84.507	91.6729

Figure 8 The Decision Tree F1 score



	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold
Random Forest F1 score for training data set	98.5075	98.8848	99.2647	97.7778	98.5185
Random Forest F1 score for validation data set	95.3846	100	89.2308	95.5224	88.5714

Figure 9 The Random Forest F1 score

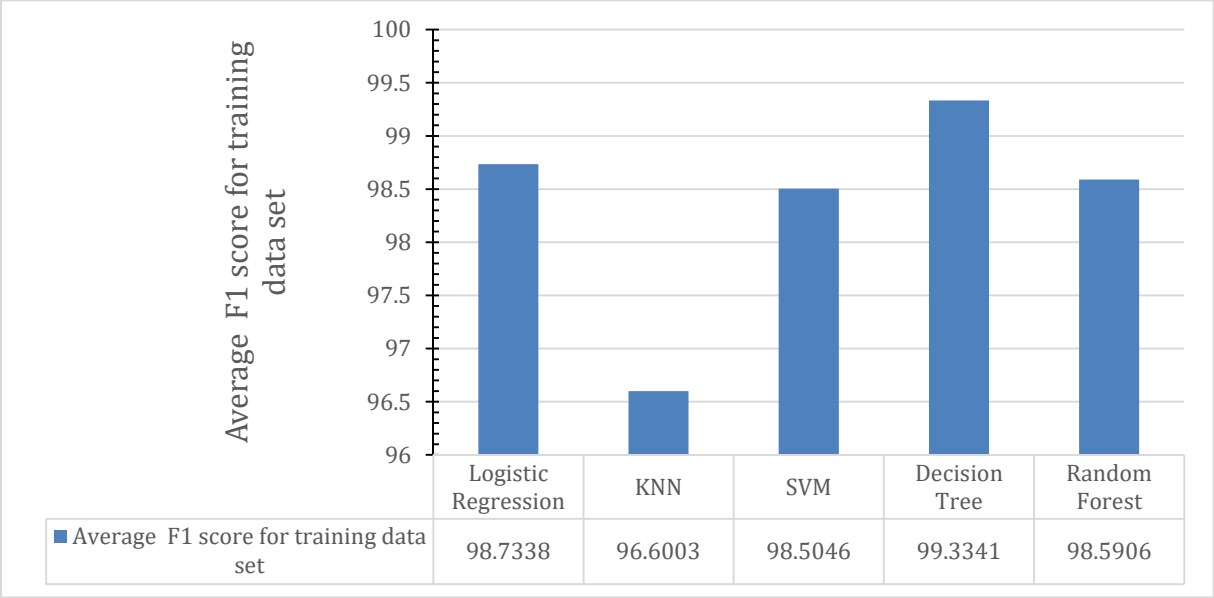


Figure 10 The mean F1 scores for the five machine learning models based on the training dataset

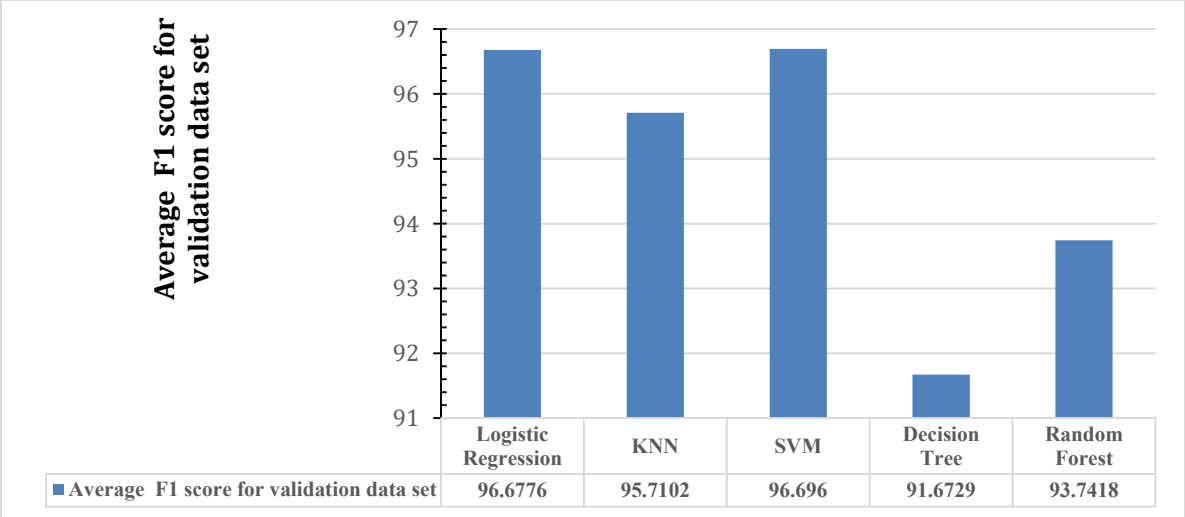


Figure 11 The mean F1 scores for the five machine learning models based on the validation dataset

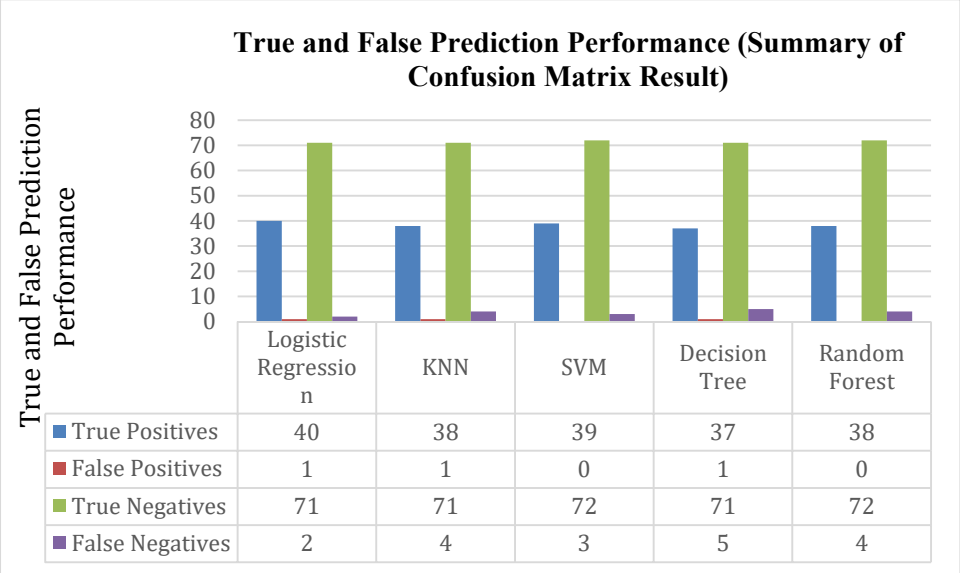


Figure 12 Summary of Confusion Matrix Result

4. CONCLUSION

The breast cancer prediction capability of five machine learning models based on a given case study breast cancer patients dataset is presented. The machine learning models considered are decision tree, random forest, K nearest neighbors, logistic regression, and the support vector machine. The models are individually trained and validated iteratively using 5-fold dataset splitting technique. The results show that the logistic regression model has the best prediction performance as it has the highest number of true or correct predictions and the lowest number of false or incorrect predictions. On the other hand, the decision tree model has the least prediction performance as it has the lowest number of true or correct predictions and the highest number of false or incorrect predictions.

REFERENCES

1. Chhikara, B. S., & Parang, K. (2023). Global Cancer Statistics 2022: the trends projection analysis. *Chemical Biology Letters*, *10*(1), 451-451.
2. Alexander, S. M., Retnakumar, R. J., Chouhan, D., Devi, T. N. B., Dharmaseelan, S., Devadas, K., ... & Chattopadhyay, S. (2021). Helicobacter pylori in human stomach: the inconsistencies in clinical outcomes and the probable causes. *Frontiers in microbiology*, *12*, 713955.
3. World Health Organization. (2022). Invisible numbers: the true extent of noncommunicable diseases and what to do about them.
4. Smolarz, B., Nowak, A. Z., & Romanowicz, H. (2022). Breast cancer—epidemiology, classification, pathogenesis and treatment (review of literature). *Cancers*, *14*(10), 2569.
5. Iacoviello, L., Bonaccio, M., de Gaetano, G., & Donati, M. B. (2021, July). Epidemiology of breast cancer, a paradigm of the "common soil" hypothesis. In *Seminars in cancer biology* (Vol. 72, pp. 4-10). Academic Press.
6. Cuthrell, K. M., & Tzenios, N. (2023). Breast Cancer: Updated and Deep Insights. *International Research Journal of Oncology*, *6*(1), 104-118.
7. Luo, C., Li, N., Lu, B., Cai, J., Lu, M., Zhang, Y., ... & Dai, M. (2022). Global and regional trends in incidence and mortality of female breast cancer and associated factors at national level in 2000 to 2019. *Chinese medical journal*, *135*(01), 42-51.
8. Oglat, A. A., Hasan, H., Mhanna, H. Y. A., & Akhdar, H. F. (2024). Study of North Jordanian women's knowledge of breast cancer causes and medical imaging screening advantages. *Informatics in Medicine Unlocked*, *47*, 101490.
9. Aizaz, M., Khan, M., Khan, F. I., Munir, A., Ahmad, S., & Obeagu, E. I. (2023). Burden of breast cancer: developing countries perspective. *International Journal of Innovative and Applied Research*, *11*(1), 31-37.
10. Adetifa, F. A., & Ojikutu, R. K. (2009). Prevalence and trends in breast cancer in Lagos State, Nigeria. *African Research Review*, *3*(5).
11. Black, E., & Richmond, R. (2019). Improving early detection of breast cancer in sub-Saharan Africa: why mammography may not be the way forward. *Globalization and health*, *15*, 1-11.
12. Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., ... & Anderson, B. O. (2020). Breast cancer early detection: A phased approach to implementation. *Cancer*, *126*, 2379-2393.
13. Birnbaum, J. K., Duggan, C., Anderson, B. O., & Etzioni, R. (2018). Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. *The Lancet Global Health*, *6*(8), e885-e893.
14. Birnbaum, J. K., Duggan, C., Anderson, B. O., & Etzioni, R. (2018). Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study. *The Lancet Global Health*, *6*(8), e885-e893.
15. Harford, J. B. (2011). Breast-cancer early detection in low-income and middle-income countries: do what you can versus one size fits all. *The lancet oncology*, *12*(3), 306-312.
16. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.
17. Akselrod-Ballin, A., Chorev, M., Shoshan, Y., Spiro, A., Hazan, A., Melamed, R., ... & Guindy, M. (2019). Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, *292*(2), 331-342.
18. Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F., & Roselli, M. (2019). Breast cancer prognosis using a machine learning approach. *Cancers*, *11*(3), 328.
19. Khan, M. M., Islam, S., Sarkar, S., Ayaz, F. I., Kabir, M. M., Tazin, T., ... & Almalki, F. A. (2022). Machine learning based comparative analysis for breast cancer prediction. *Journal of Healthcare Engineering*, *2022*.
20. Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., & Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, *49*(3), 1-40.
21. Saberioon, M., Císař, P., Labbé, L., Souček, P., Pelissier, P., & Kerneis, T. (2018). Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (Oncorhynchus mykiss) classification using image-based features. *Sensors*, *18*(4), 1027.
22. Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. *Journal of Data Analysis and Information Processing*, *8*(4), 341-357.
23. Thanh Noi, P., & Kappas, M. (2017). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover

classification using Sentinel-2 imagery. *Sensors*, *18*(1), 18.

24. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1-13.
25. Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.
26. Awan, M. Z., Arif, M. S., Abideen, M. Z. U., & Abodayeh, K. (2024). Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach. *Indonesian Journal of Electrical Engineering and Computer Science*, *34*(3), 2032-2044.
27. Roy, B. R., Pal, M., Das, S., & Huq, A. (2020, November). Comparative Study of Machine Learning Approaches on Diagnosing Breast Cancer for Two Different Dataset. In *2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT)* (pp. 29-34). IEEE.
28. Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of breast cancer using machine learning techniques-a survey. *Procedia Computer Science*, *218*, 1434-1443.