Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 11 Issue 2, February - 2024

Prediction Of Breast Cancer Using Support Vector Machine And Decision Tree Machine Learning Approaches

Chikezie Samuel Aneke[1]
Department Of Computer Engineering
University Of Uyo, Akwa Ibom State

Emmanuel Nsese Udoiwod[2]
Department Of Computer Engineering
University Of Uyo, Akwa Ibom State
emmanueludoiwod@uniuyo.edu.ng

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

***Abstract*— In this paper, prediction of breast cancer using support vector machine (SVM) and decision tree machine learning models was presented. The model training and validation are performed using the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset consists of 569 records of patients and 35 columns. Exploratory data analysis wais carried out on the diabetic patient dataset using the Pandas-Profiling library. The Seaborn library was used to show the Pearson Correlation of features in the dataset. The model training dataset was divided into 5 folds. Each fold was used as the validation set in 5 iterations. The results show that for the training, the SVM has F1 score with mean value of 98.505% while the decision tree model has F1 score with mean value of 99.334 %. On the other hand, for the validation dataset, the SVM has F1 score with mean value of 96.696 % while the decision tree model has F1 score with mean value of 91.6729 %.In addition, according to the results of the confusion matrix, the SVM has better performance as it had true (or correct) prediction with a higher value of 97.4 % while the decision tree has correct (true) prediction of 2.6%. Again, the SVM has better results for the untrue (or incorrect) prediction with a smaller value of 2.6 % while the decision tree has higher untrue (or incorrect) prediction of 5.3%. Essentially, the SVM model can predict the likelihood of breast cancer better that the decision tree model.**

Keywords*— *Breast Cancer Prediction, Support Vector Machine, Decision Tree, Machine Learning, Confusion Matrix

1. INTRODUCTION

In recent years, there has been increase in the incidence of breast cancer across the globe [1,2,3]. Accordingly, many non-governmental organizations, as well as government agencies are making more effort to address the issue. The efforts are geared towards creating awareness of breast cancer and getting people to know measures that can be used to detect the likelihood of occurrence of breast [4,5,6]. Also, information on life styles that can increase the chances of breast cancer are also publicized so as to encourage people to avoid such [7,8,9].

In addition, the medical practitioners and researchers have also applied some intelligent ways to diagnose and predict the likelihood of breast cancer in a patient based on medical historical data [10,11,12]. Such approach requires the use of intelligent algorithms which can be trained with the medical data records of breast cancer patients and hence enable such algorithms to predict with sufficient accuracy the likelihood of breast cancer in a person. Accordingly, in this work, support vector machine (SVM) and decision tree machine learning algorithms are employed to predict breast cancer [13,14,15,16]. A case study Breast Cancer Wisconsin (Diagnostic) dataset was used for the model training and validation [17,18,19]. The F1 score and the confusion matrix parameters were used to compare the prediction performance of the two machine learning models [20,21]. The essence of the study is to determine which of the two machine learning models is more suitable for breast cancer prediction.

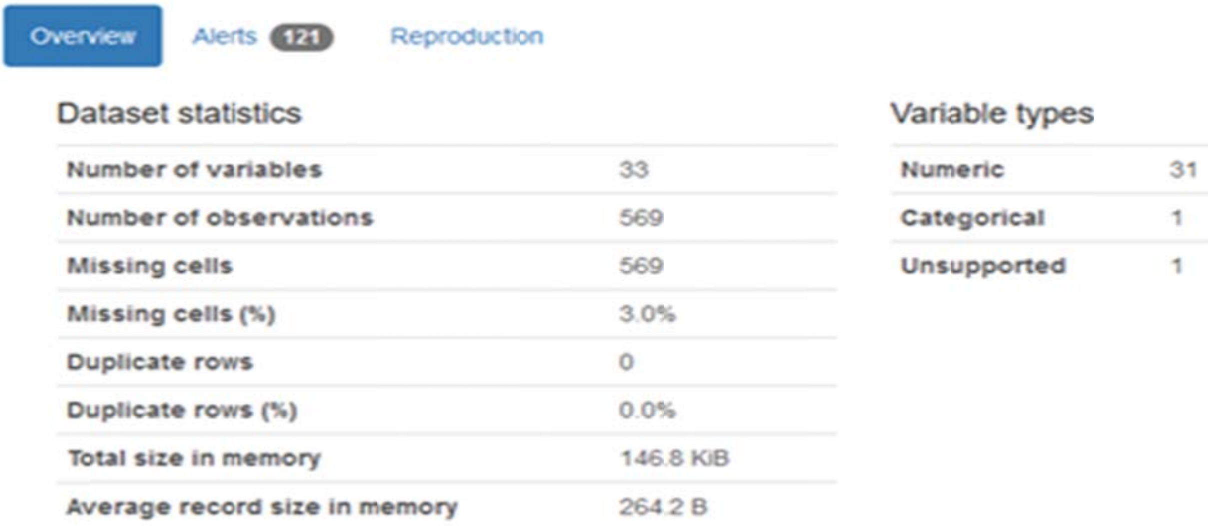
2. METHODOLOGY

In this paper, the focus in the application of support vector machine and decision tree machine learning models for the prediction of breast cancer. The model training and validation are performed using the Breast Cancer Wisconsin (Diagnostic) dataset. The dataset consists of 569 records of patients and 35 columns. The dataset metadata, referred to as features are presented in Table 1. In the features listed in Table 1, the column "Unnamed: 32" is irrelevant. There are null values. The column is removed during data cleaning. There are no missing values in the dataset. There are also no duplicate values.

Exploratory data analysis is carried out on the diabetic patient dataset using the Pandas-Profiling library [22, 23]. The screenshots shown in Figure 1 and Figure 2 show that there are 569 missing values which are from the 'Unnamed: 32' column. There are 31 numeric variables, 1 categorical variable which is the 'Diagnosis' column. The Unsupported variable is the 'Unnamed: 32' column'. The screenshot in Figure 2 shows that there are no missing values in any of the columns except the 'Unnamed: 32' column

Table 1: Features of diabetic patient dataset

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



Dataset statistics

Number of variables	33
Number of observations	569
Missing cells	569
Missing cells (%)	3.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	146.8 KiB
Average record size in memory	264.2 B

Variable types

Numeric	31
Categorical	1
Unsupported	1

Figure 1: Overview of the dataset

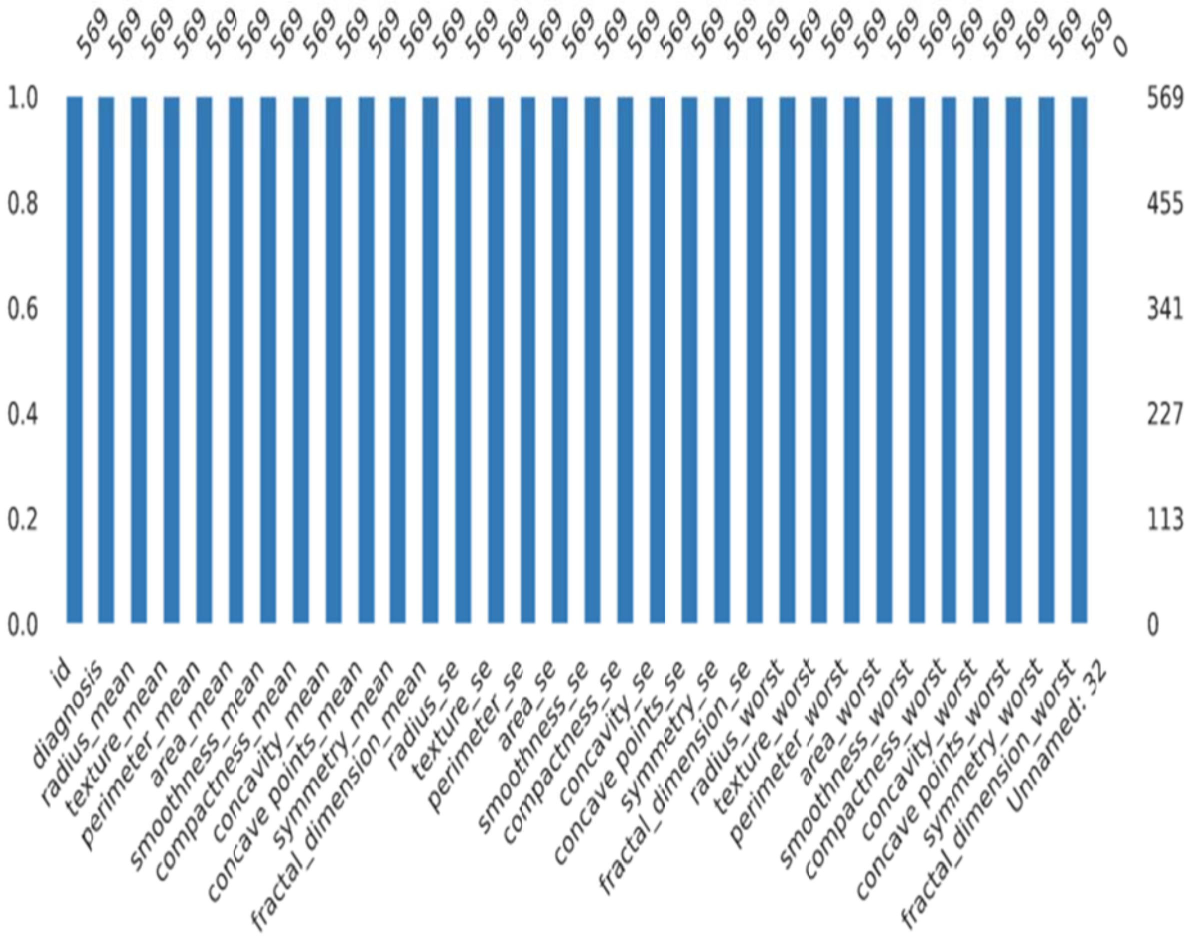


Figure 2: Visualization of nullity by column

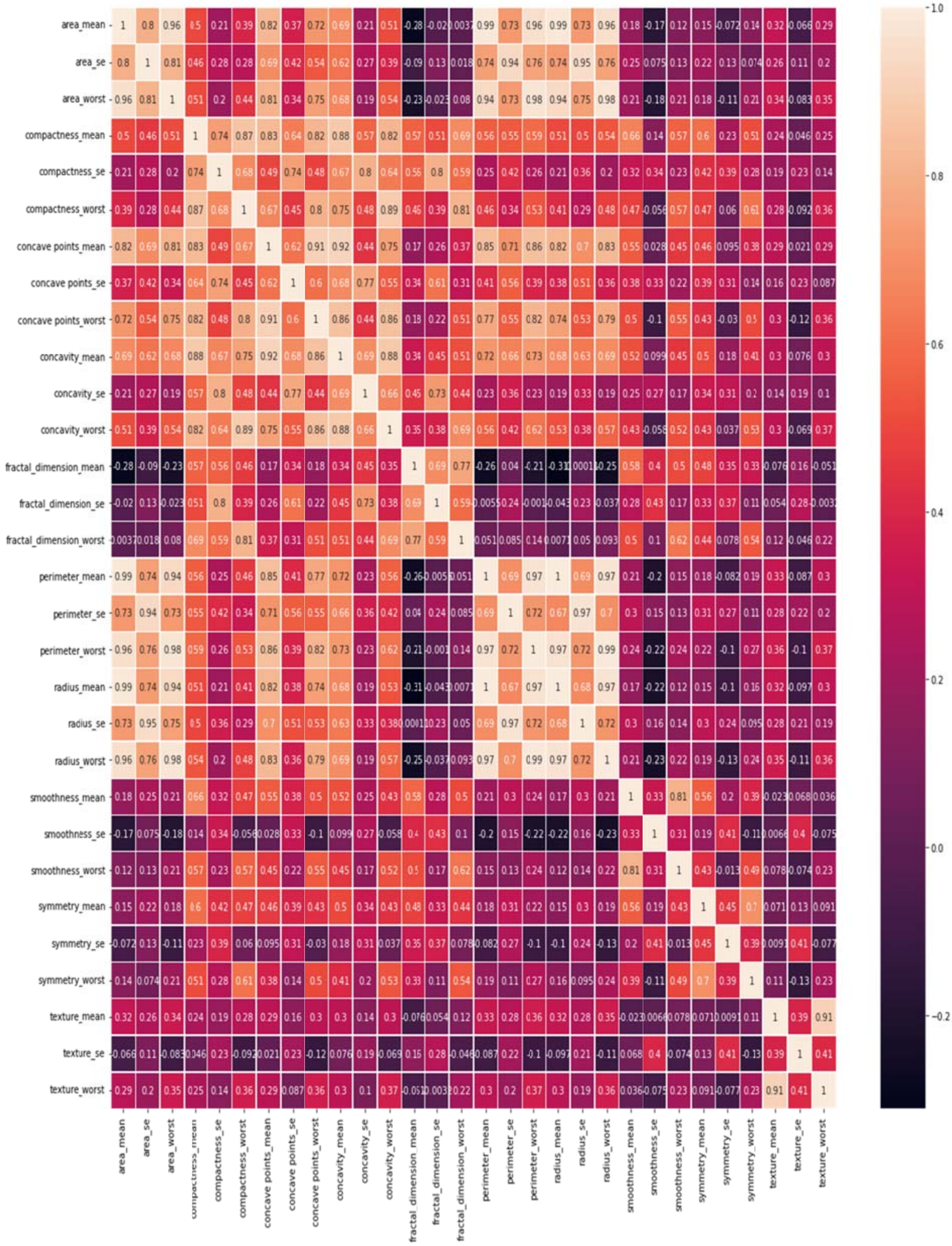


Figure 3 Pearson correlation of features in the dataset using seaborn

The Seaborn library is used to show the Pearson Correlation of features in the dataset [24,25] (Figure 3). Pearson Correlation is given as:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - [\sum y]^2]}} \qquad (1)$$

Where: n is the total number of observations, x is the first variable, y is the second variable and r is the pearson correlation value. From the Seaborn Heatmap, it can be seen that the *area_mean* column is highly positively correlated with the *area_se, area_worst, concave_point_mean, concave_points_worse, perimeter_mean, perimeter_se, perimeter_worse, radius_mean, radius_se, radius_worst.*

2.1 Model Training and validation

The model training dataset was divided into 5 folds. Each fold was used as the validation set in 5 iterations as shown in Figure 4. In the first iteration, the first part of the data is used for validation, while the other parts are used for training as illustrated in Figure 4. This process is repeated until all the folds of the dataset have been used for validation. Since it is a 5 folds technique, it means 20% of the dataset is used as the validation set while 80 % is for training.

Iteration 1	Test	Train	Train	Train	Train
Iteration 1	Train	Test	Train	Train	Train
Iteration 1	Train	Train	Test	Train	Train
Iteration 1	Train	Train	Train	Test	Train
Iteration 1	Train	Train	Train	Train	Test

Figure 4 : The screenshot of the 5-Fold Cross-Validation

2.2 Training of the Support Vector Machine (SVM) Model

The *SVC (Support Vector Classifier)* class is imported from the *svm* module of the Sci-kit learn library (see Figure 5). The regularization's intensity is inversely proportional to C. It means the higher the value of C, the lesser the regularization. Regularization is a technique used to reduce overfitting during training. Overfitting occurs when the machine learning model performs very well on the training set but performs poorly on the validation set. It means the model is not learning. It is just memorizing the training data. The parameter *C* is set to 3.7276. The *rbf* SVM kernel is used. The other SVM kernels available in the Sci-kit Learn library are *'poly', 'rbf', 'sigmoid',* and *'precomputed.'*

```
# Use the best parameters to train the logistic regression algorithm
from sklearn.svm import SVC

svm_model = SVC(C=3.727593720314938, kernel='rbf')
svm_train_val_result = cross_validation(svm_model, scaled_X_train, y_train, 5)
print(svm_train_val_result)
```

Figure 5: Training of the SVM Model

2.3 Training of the Decision Trees Model

DecisionTreeClassifier class was imported from the *tree* module of the Sci-kit Learn library. The *criterion* parameter is a function for determining the quality of a split. The *criterion* is *"entropy."* The *min_samples_split* represents the minimum amount of samples needed to separate an internal node in the decision tree. This parameter helps to avoid overfitting. The *min_samples_split* is 5. It means once we have 5 samples remaining, they should not be split again into various classes (as shown, Figure 6).

```
# Use the best parameters to train the Decision Tree algorithm
from sklearn.tree import DecisionTreeClassifier

dt_model = DecisionTreeClassifier(criterion="entropy",
                                  min_samples_split=5,
                                  random_state=0)

dt_train_val_result = cross_validation(dt_model, scaled_X_train, y_train, 5)
print(dt_train_val_result)
```

Figure 6: Training of the Decision Tree Model

3. RESULTS AND DISCUSSION

3.1 Training and Validation Results

The case study dataset is imbalanced as such, in this work the training and validation results, are focus on the F1 score metric. This is because accuracy is not effective metric on a dataset with imbalanced classes. However, accuracy, precision, recall, and f1-score metrics are used to evaluate the test set. The F1 scores for the training dataset across the 5 folds are as presented in Table 2 and Figure 7 while the F1 scores for the validation dataset across the 5 folds are as presented in Table 3 and Figure 7. The results show that for the training dataset (Table 2 and Figure 7), the SVM has F1 score with mean value of 98.505% while the decision tree model has F1 score with mean value of 99.334 %. On the other hand, for the validation dataset (Table 3 and Figure 8), the SVM has F1 score with mean value of 96.696 % while the decision tree model has F1 score with mean value of 91.6729 %. Essentially, the decision tree has higher (and hence better) F1 score in the training dataset than the SVM. However, the reverse is the case on the testing dataset where the SVM has higher (and hence better) F1 score in the training dataset than the decision tree. As such, other performance parameters available in confusion matrix are used to determine the model that is better for application in breast cancer prediction.

Table 2: The F1 scores for the training dataset across the 5 folds

	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
SVM F1 scores	98.5075	98.1273	99.2593	97.7444	98.8848	98.5046
Decision Tree F1 scores	99.2593	99.6337	99.2701	100.0000	98.5075	99.3341

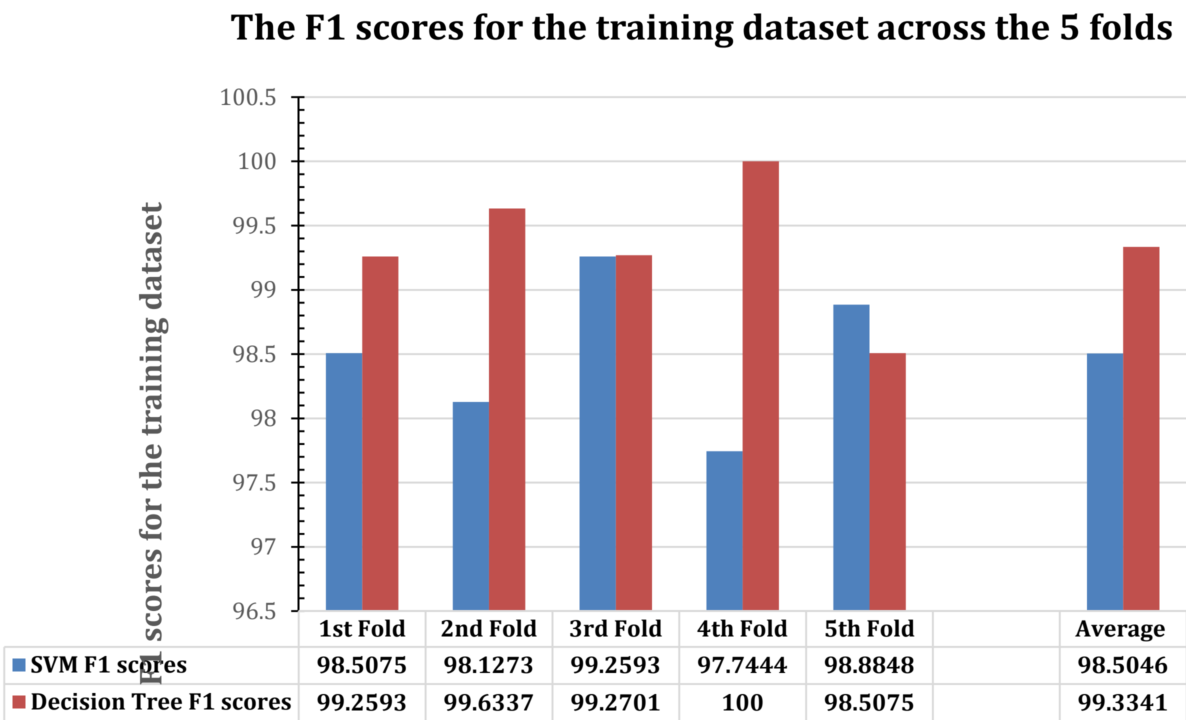


Figure 7 The F1 scores for the training dataset across the 5 folds

Table 3: The F1 scores for the validation dataset across the 5 folds

	1st Fold	2nd Fold	3rd Fold	4th Fold	5th Fold	Average
SVM F1 scores	95.3846	100.0000	95.3846	97.0588	95.6522	96.6960
Decision Tree F1 scores	95.5224	92.9577	89.8550	95.5224	84.5070	91.6729

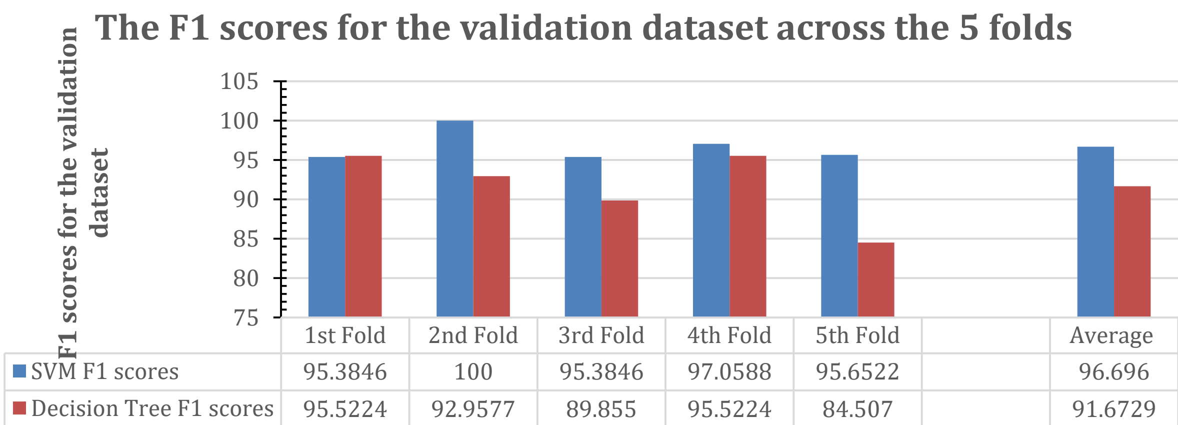


Figure 8 The F1 scores for the validation dataset across the 5 folds

The results on the confusion matrix are presented in Figure 9, Figure 10 and Figure 11 and they show that the number of true positives, false positives, true negatives, and false negatives. According to the confusion matrix results. Also, the statistics of true or correct prediction is presented in Figure 12 while the statistics of false or incorrect prediction is presented in Figure 13. According to the results, the SVM has better results for the true (or correct) prediction with a higher value of 97.4 % while the decision tree has correct (true) prediction of 2.6%. Again, the SVM has better results for the untrue (or incorrect) prediction with a smaller value of 2.6 % while the decision tree has higher untrue (or incorrect) prediction of 5.3%. Essentially, the SVM model can predict the likelihood of breast cancer better that the decision tree model.

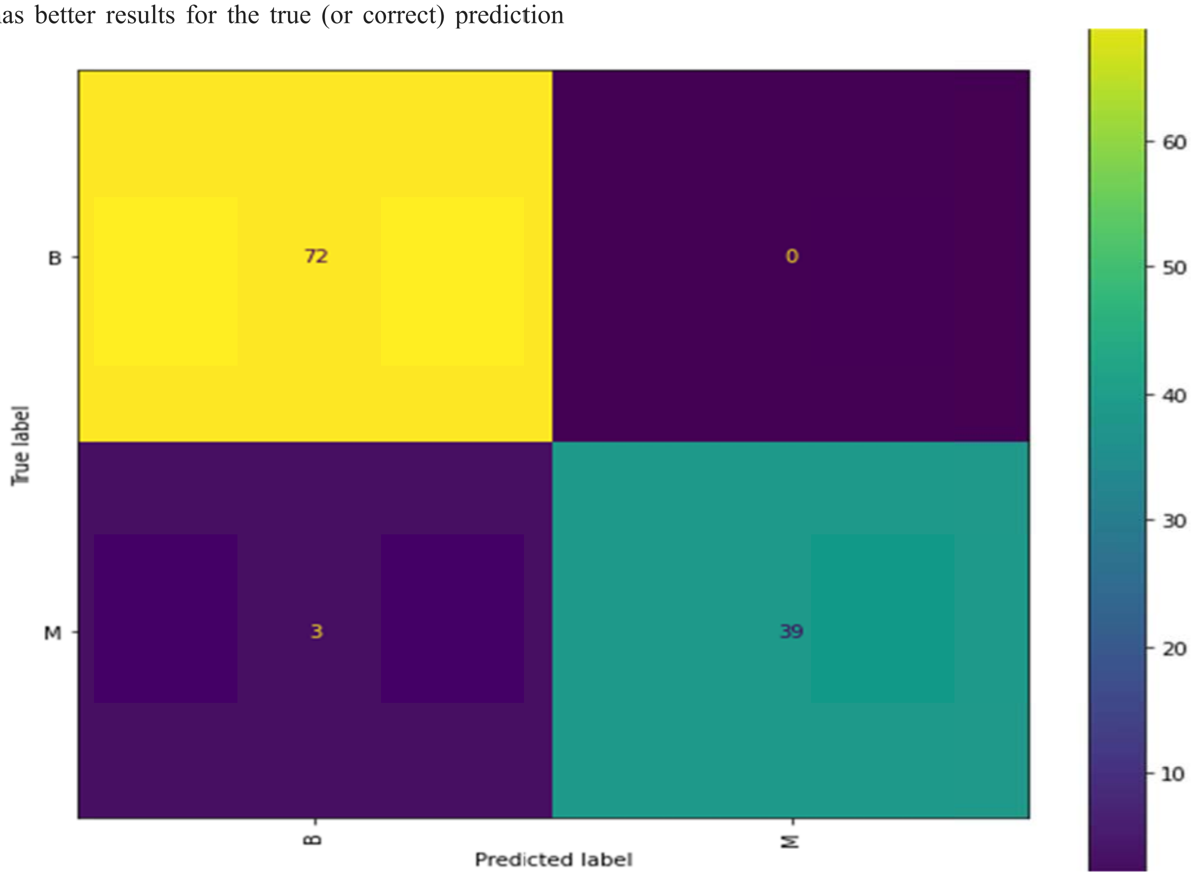


Figure 9: Confusion matrix heat map for the SVM Model

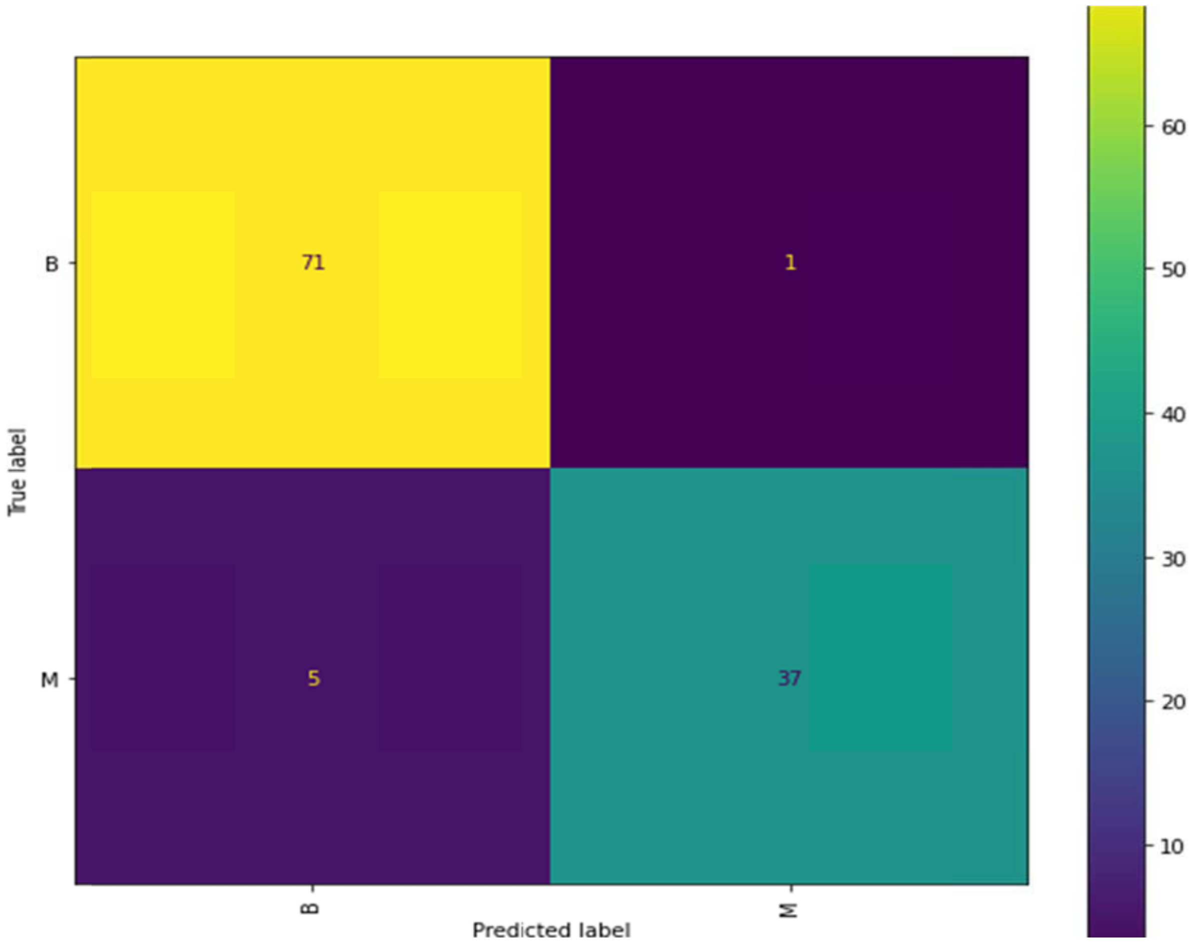


Figure 10: Confusion matrix heat map for Decision Tree Model

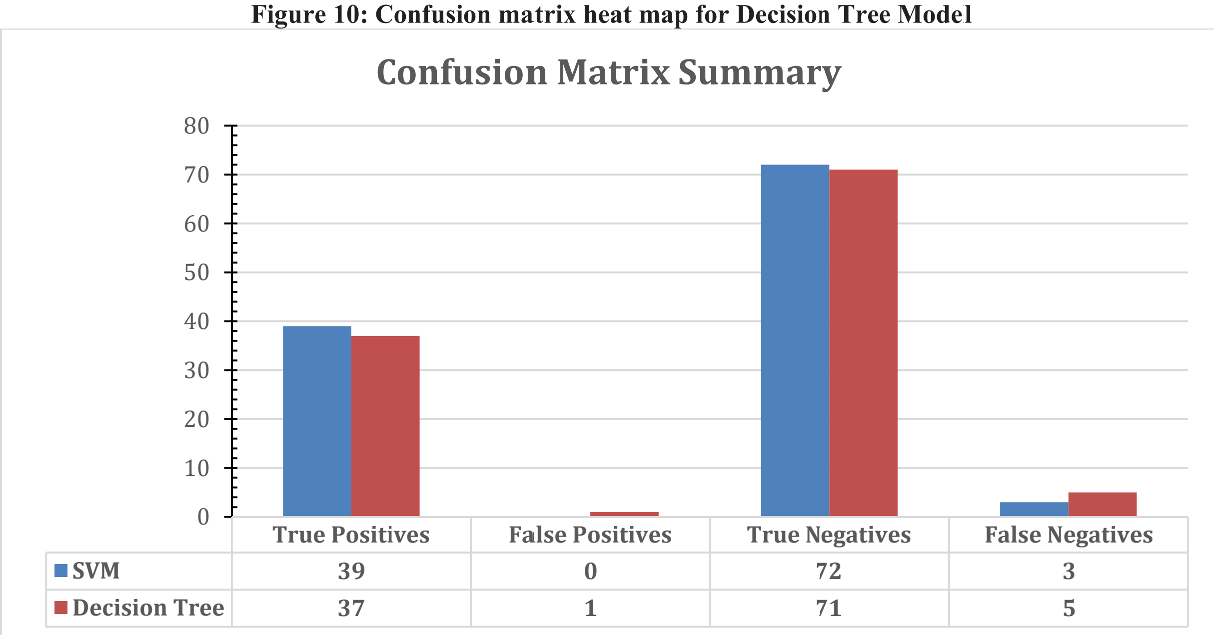


Figure 11 The summary of the confusion matrix for the two models

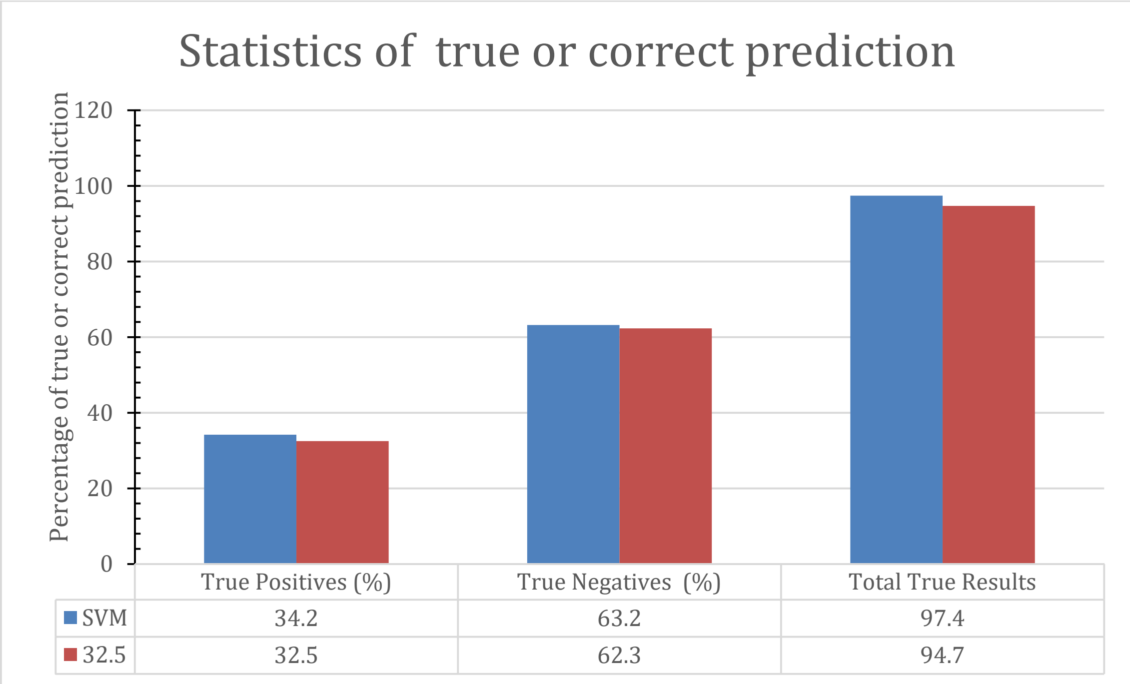


Figure 12 Statistics of true or correct prediction

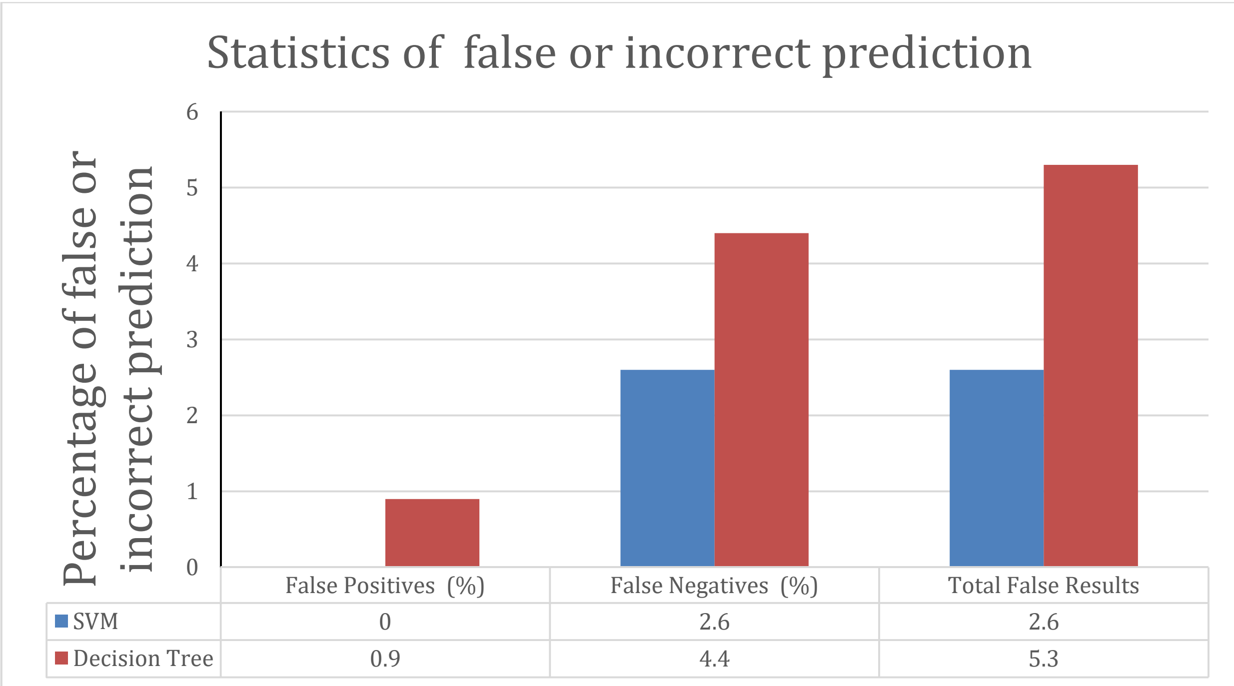


Figure 13 Statistics of true or correct prediction

4. CONCLUSION

The focus of this paper is on the support vector machine (SVM) and decision tree machine learning models which are trained for prediction of breast cancer. The cancer patients' dataset was acquired and the 5-fold technique was employed in splitting the dataset into training and validation set. The models were iteratively trained based on the 5-fold approach and the F1 scores were obtained for each model for each of the five folds. Also confusion matrix results were obtained for the two models. The results showed that the SVM model performed better than the decision tree in making correct predictions of breast cancer incidence in the patience.

REFERENCES

1. Lima, S. M., Kehm, R. D., & Terry, M. B. (2021). Global breast cancer incidence and mortality trends by region, age-groups, and fertility patterns. *EClinicalMedicine*, *38*.
2. Kashyap, D., Pal, D., Sharma, R., Garg, V. K., Goel, N., Koundal, D., ... & Belay, A. (2022). Global increase in breast cancer incidence: risk factors and preventive measures. *BioMed research international*, *2022*.

3. Lei, S., Zheng, R., Zhang, S., Wang, S., Chen, R., Sun, K., ... & Wei, W. (2021). Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Communications*, *41*(11), 1183-1194.
4. Osei-Afriyie, S., Addae, A. K., Oppong, S., Amu, H., Ampofo, E., & Osei, E. (2021). Breast cancer awareness, risk factors and screening practices among future health professionals in Ghana: A cross-sectional study. *PloS one*, *16*(6), e0253373.
5. Gupta, R., Gupta, S., Mehrotra, R., & Sodhani, P. (2020). Risk factors of breast cancer and breast self-examination in early detection: systematic review of awareness among Indian women in community and health care professionals. *Journal of public health*, *42*(1), 118-131.
6. World Health Organization. (2023). *Global breast cancer initiative implementation framework: assessing, strengthening and scaling-up of services for the early detection and management of breast cancer*. World Health Organization.
7. Jia, T., Liu, Y., Fan, Y., Wang, L., & Jiang, E. (2022). Association of healthy diet and physical activity with breast cancer: lifestyle interventions and oncology education. *Frontiers in public health*, *10*, 797794.
8. Rock, C. L., Thomson, C., Gansler, T., Gapstur, S. M., McCullough, M. L., Patel, A. V., ... & Doyle, C. (2020). American Cancer Society guideline for diet and physical activity for cancer prevention. *CA: a cancer journal for clinicians*, *70*(4), 245-271.
9. Newman, L. A. (2022). Breast cancer screening in low and middle-income countries. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *83*, 15-23.
10. Casal-Guisande, M., Comesaña-Campos, A., Dutra, I., Cerqueiro-Pequeño, J., & Bouza-Rodríguez, J. B. (2022). Design and development of an intelligent clinical decision support system applied to the evaluation of breast cancer risk. *Journal of personalized medicine*, *12*(2), 169.
11. Ogundokun, R. O., Misra, S., Douglas, M., Damaševičius, R., & Maskeliūnas, R. (2022). Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. *Future Internet*, *14*(5), 153.
12. Singh, L. K., Khanna, M., & Singh, R. (2023). Artificial intelligence based medical decision support system for early and accurate breast cancer prediction. *Advances in Engineering Software*, *175*, 103338.
13. Tsehay Admassu Assegie, S. S. (2020). A support vector machine and decision tree based breast cancer prediction. *International Journal of Engineering and Advanced Technology (IJEAT), ISSN*, 2249-8958.
14. Dinesh, P., Vickram, A. S., & Kalyanasundaram, P. (2024, May). Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy. In *AIP Conference Proceedings* (Vol. 2853, No. 1). AIP Publishing.
15. Naji, M. A., El Filali, S., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, *191*, 487-492.
16. Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., ... & Rhee, J. (2022). Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, *11*(9), 136.
17. Hossin, M. M., Shamrat, F. J. M., Bhuiyan, M. R., Hira, R. A., Khan, T., & Molla, S. (2023). Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. *Bulletin of Electrical Engineering and Informatics*, *12*(4), 2446-2456.
18. Mohammad, W. T., Teete, R., Al-Aaraj, H., Rubbai, Y. S. Y., & Arabyat, M. M. (2022). Diagnosis of breast cancer pathology on the wisconsin dataset with the help of data mining classification and clustering techniques. *Applied Bionics and Biomechanics*, *2022*.
19. Kadhim, R. R., & Kamil, M. Y. (2022). Comparison of breast cancer classification models on Wisconsin dataset. *Int J Reconfigurable & Embedded Syst ISSN*, *2089*(4864), 4864.
20. 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.
21. Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).
22. Harrison, M., & Petrou, T. (2020). *Pandas 1. x Cookbook: Practical recipes for scientific computing, time series analysis, and exploratory data analysis using Python*. Packt Publishing Ltd.
23. Gupta, S., & Sedamkar, R. R. (2020). Machine learning for healthcare: Introduction. *Machine learning with health care perspective: Machine learning and healthcare*, 1-25.
24. Vencer, L. V. T., Bansa, H., & Caballero, A. R. (2023, May). Data and Sentiment Analysis of Monkeypox Tweets using Natural Language Toolkit (NLTK). In *2023 8th International Conference on Business and Industrial Research (ICBIR)* (pp. 392-396). IEEE.
25. Dasari, K. B., & Devarakonda, N. (2022, September). SynFlood DDoS attack detection with SVM kernels using uncorrelated feature subsets selected by Pearson, spearman and Kendall correlation methods. In *2022 second international conference on computer science, engineering and applications (ICCSEA)* (pp. 1-6). IEEE.