A Support Vector Machine and Decision Tree Based Breast Cancer Prediction

Tsehay Admassu Assegie, Sushma S. J.

Abstract: ***The first step in diagnosis of a breast cancer is the identification of the disease. Early detection of the breast cancer is significant to reduce the mortality rate due to breast cancer. Machine learning algorithms can be used in identification of the breast cancer. The supervised machine learning algorithms such as Support Vector Machine (SVM) and the Decision Tree are widely used in classification problems, such as the identification of breast cancer. In this study, a machine learning model is proposed by employing learning algorithms namely, the support vector machine and decision tree. The kaggle data repository consisting of 569 observations of malignant and benign observations is used to develop the proposed model. Finally, the model is evaluated using accuracy, confusion matrix precision and recall as metrics for evaluation of performance on the test set. The analysis result showed that, the support vector machine (SVM) has better accuracy and less number of misclassification rate and better precision than the decision tree algorithm. The average accuracy of the support vector machine (SVM) is 91.92 % and that of the decision tree classification model is 87.12 %.***

Keywords: breast cancer diagnosis, decision tree classification, SVM classification, machine learning.

I. INTRODUCTION

The Support Vector Machine (SVM) and the Decision tree are a supervised machine learning models used in classification, regression and outlier detection [1-2]. The SVM and Decision tree models are widely used in medical imaging, object recognition such as face detection, financial analysis, big data analysis, handwritten digits recognition and recommendation systems. One of the importance of the SVM and Decision tree algorithms is disease classification. Although, many supervised machine learning models such as SVM, Decision tree, Adaboost, Naive Bayes and neural network can be used in disease classification problems and diagnosis of different diseases, the accuracy and the complexity of training the models is different for each algorithm in classification as well as diagnosis of the diseases.

A Breast cancer is one of the most common cancers followed by cancers of the cervix [3]. And the breast cancer disease results in death if it is not threated early. Therefore, an early detection of the symptoms of the breast cancer significantly plays a great role in reduction of the mortality rate due to this disease. In the diagnosis of breast cancer the physicians use different features to identify the disease. The features used are mean texture values, radius of the breast and so on. In the diagnosis of the disease, the physicians use these features to classify the disease into possible set, and then a test is made on the samples.

As the number of patients with cancer disease is increasing, the physicians may be busy on the treatment. To solve the classification problem and assist the breast cancer diagnosis process using machine's capability, machine learning models can be used in the classification process in order to identify the symptoms of breast cancer as early as possible.

This study is devoted to build machine learning models that to assist the classification of breast cancer using the Support Vector Machine (SVM) and the Decision Tree classification algorithms and finally, compares the performance of these classification algorithms. Some of the problems that this paper is intended to address are the following:

1) What is the accuracy of the SVM and the decision tree algorithm on the breast cancer identification?
2) Can we develop a machine learning model for breast cancer identification with an acceptable level of accuracy?

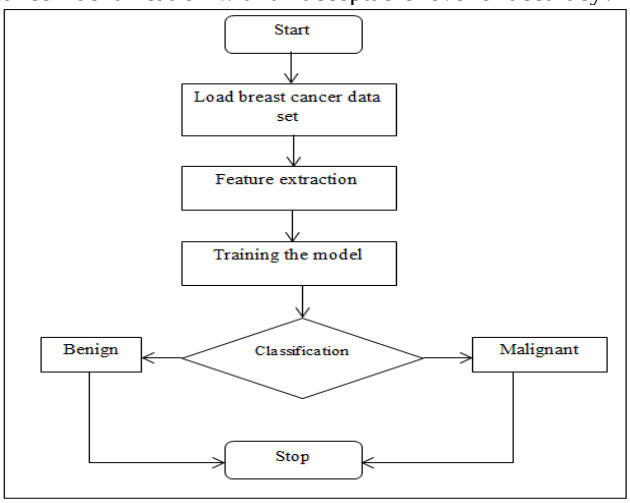

Fig. 1.flowchart of breast cancer diagnosis using SVM and Decision tree classification.

II. RELATED WORKS

The Support Vector Machine (SVM) and the Decision tree models are supervised machine learning models used in classification problems such disease classification and diagnosis. The basic SVM model is used in two class classification models [2]. The breast tumor is either malignant or benign and the kaggle dataset of breast cancer used in training and testing the models contain two classes. This implies that, SVM model can be used in breast cancer classification and diagnosis problem.

Revised Manuscript Received on February20, 2020.

Tsehay Admassu Assegie*, Department of Computing Technology, College of Engineering and Technology, Aksum University, Aksum, Ethiopia. E-mail: tsehayadmassu2006@gmail.com

Sushma S. J., Associate Professor, GSSSIETW, Mysuru, Karnataka, India. E-mail: enggsush@gmail.com

A comparative study on the Support Vector Machine (SVM) and Artificial Neural Network (ANN) [3] for breast cancer diagnosis shows that the on the Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifier has roughly similar accuracy in classification. But, the precision and recall of the Artificial Neural Network (ANN) is less than the Support Vector Machine (SVM).

In [4] a hybrid approach is used in feature extraction and mass classification of breast cancer. Another machine learning research on breast cancer diagnosis using the Support Vector Machine (SVM) and Artificial Neural Network was proposed in [5]. In the study, the authors used the SVM model for training the machine and the model had an accuracy of roughly 79 percent. As accuracy is one of the most important metric used in evaluating the performance of machine learning models and 79 percent is better accuracy, the authors achieved a better result in diagnosis of breast cancer.

The utilization of machine learning approaches is significantly increasing due to the need for computer-assisted diagnosis and disease classification. In this area computers are getting greater importance due to the advances in machine learning and the capability of machines in processing a huge amount of data with very limited amount of time to reach the goal [6]. The author studied breast cancer detection using the SVM algorithm. The dataset used in the study was Wisconsin Diagnostic Breast Cancer (WDBC) dataset.

As researched in [7], the number of deaths due to the breast cancer is 458,000 a year worldwide and this cancer is the most common type of cancer in females in developing and developed nation although, the mortality rate is less in developed nations compared to developing nations. An early diagnosis is important for detection of the breast cancer symptoms.

In a study [8] a method for breast cancer diagnosis using neural network was proposed. The authors used Breast Cancer Wisconsin dataset with 699 samples of which 458 were benign and 241 are malignant. The system was effective in classification of the sample into two classes the benign and malignant with accuracy of 96.5% on the random tests conducted on the model.

Mohammad Alifraheed [9] proposed a breast cancer identification using Artificial Neural Network based on mammograms images.

The machine learning models are widely used for prediction and early detection of breast cancer [10-18]. The authors compared Byes Navies, K-Nearest Neighbour (K-NN) and the Support Vector Machine (SVM) using the Wisconsin Breast Cancer dataset. Among the three models, the Support Vector Machine (SVM) performed better with accuracy and lower error rate. The metrics that the authors have used are only the accuracy, precision, sensitivity and specificity. In their study, the complexity of training the models and the other metrics used in measuring the performance of classification models such as recall and confusion matrix were not used to evaluate the performance of the models.

III. RESEARCH METHOD

This study is focuses on the comparing the predictive accuracy and training complexity of SVM and the decision tree on identification of breast cancer. The kaggle breast cancer dataset is used for training and testing the accuracy of each supervised machine learning algorithms namely, the SVM and the decision tree algorithm on breast cancer identification. In the dataset visualization, pre-processing and testing the accuracy of the models in diagnosis of breast cancer, the Python programming language is employed. In the dataset 357 observations are benign, cancer negative and the 212 observations are the malignant (cancerous) or breast positive. The accuracy and confusion matrix, recall, receiver operating characteristic (ROC) and precision are used as evaluation metrics to evaluate the performance of the proposed model on the kaggle breast cancer data repository.

A. Dataset Description

In this study, the kaggle dataset is used to create the supervised machine learning model for identitifcation of breast cancer using the Decision tree and the Support Vector Machine (SVM) classification algorithms. This dataset consists of the features of malignant and benign. Some of the features used in classification of breast cancer to the malignant and benign are breast radius mean, texture mean, perimeter mean, texture and so on. The dataset consists of 569 sample breast cancer instances and 30 labels or features. In the training, 75 percent of the dataset is used and 25 percent is used for testing the models. For pre-processing like, splitting the dataset into training and test set the sklearn train_test_split method is used to randomly split the entire 569 data samples into training set which consists of 75 percent of the dataset and 25 percent for testing.

IV. RESULT AND DISCUSSION

In this section, the accuracy of the Support Vector Machine (SVM) and the Decision tree models is explained and the results of the research shows that SVM is better in classification of breast cancer than decision tree. Figure 2 shows the accuracy of the Support Vector Machine (SVM) and the Decision tree in classification of breast cancer. A part form the accuracy, recall, precision and confusion matrix is used to evaluate the performance of the Support Vector Machine (SVM) and the Decision tree in classification.

A. Accuracy of the models

In this section, the accuracy of the Support Vector Machine (SVM) and the Decision tree models is explained and the results of the research shows that the Support Vector Machine (SVM) as better model in classification of breast cancer than the Decision tree. Figure 2 illustrates the accuracy of the Support Vector Machine (SVM) and the Decision tree in classification of breast cancer.

Table- I: Name of the Table that justify the values

Random test on the model	Accuracy in %
1	89.47
2	90.35
3	94.73
4	93.85
5	91.22

Table- II: Name of the Table that justify the values

Random test on the model	Accuracy in %
1	89.47
2	90.35
3	94.73
4	93.85
5	91.22

The result on random tests on the Support Vector Machine (SVM) and the Decision tree models shows that, the Support Vector Machine (SVM) as better model in breast cancer classification, as it has better accuracy than the Decision tree model.

As shown in figure 2, the average accuracy for Support Vector Machine (SVM) in breast cancer diagnosis is 91.92 % and that of the decision tree model is 87.12 %. Table I shows the accuracy of the Decision tree model and table II shows the accuracy of the Support Vector Machine (SVM) on five random tests on the test dataset.

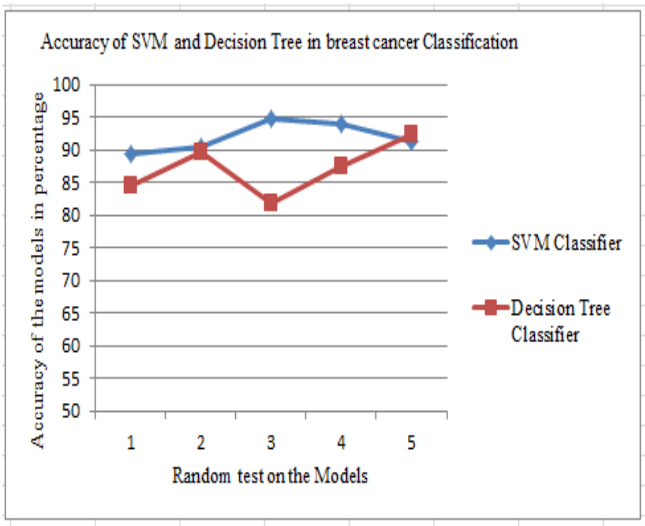

Fig. 2.Accuracy of SVM and Decision tree on breast cancer diagnosis

B. Precision Recall Analysis

In this section, the recall of the Support Vector Machine (SVM) and the Decision tree models is explained and the results of the research shows that Support Vector Machine (SVM) as better in classification of breast cancer than Decision tree. The recall of the Decision tree model is shown in table III.

Table- III: Name of the Table that justify the values

Random test on the model	Recall	
	Malignant	Benign
1	88	94
2	90	96
3	88	96
4	87	92
5	88	89

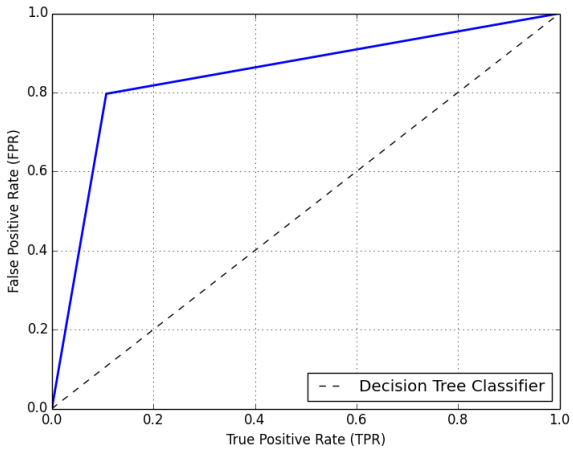

Fig. 3.Fig.4 ROC curve for Decision tree classifier on breast cancer diagnosis

Table- IV: Name of the Table that justify the values

Random test on the model	Recall	
	Malignant	Benign
1	81	92
2	87	90
3	87	86
4	88	89
5	88	92

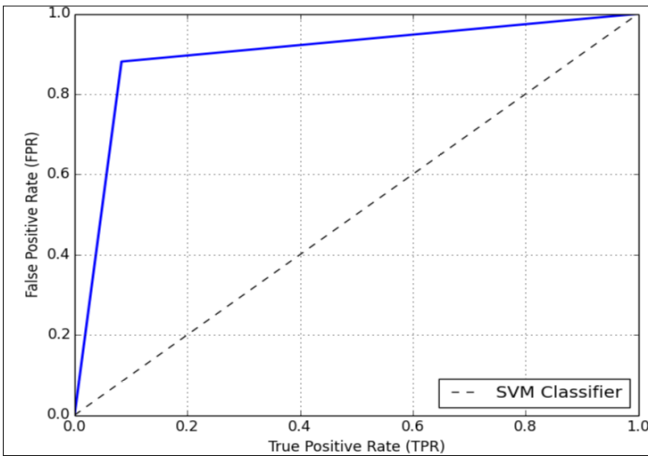

Fig. 4.Fig.4 ROC curve for SVM classifier in breast cancer diagnosis

The receiver operating characteristic (ROC) of SVM and decision tree is given in figure x and figure y respectively. By looking the ROC curve, the SVM classifier is has better recall rate than the decision tree classifier.

C. Confusion matrix analysis

Along with the accuracy, a confusion matrix is used as a metric to evaluate the performance of the Support Vector Machine (SVM) and the Decision tree models in the classification of breast cancer into either the malignant or the benign class. And the confusion matrix of Support Vector Machine (SVM) model is shown in figure 5 and that of Decision tree model is illustrated in figure 6.

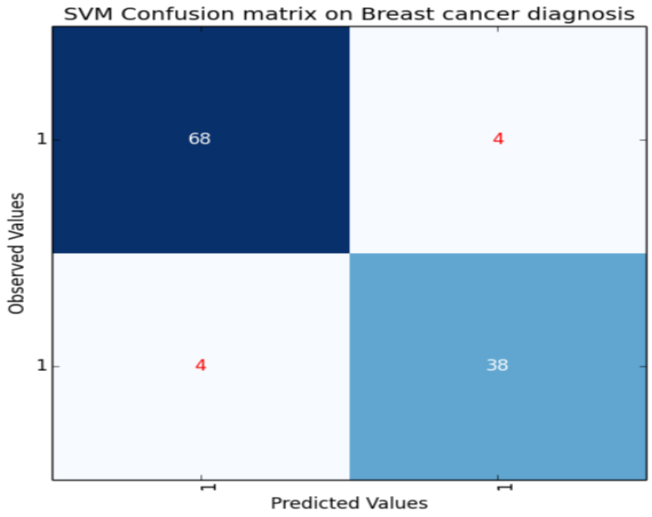

Fig. 5.Fig.5 Confusion Matrix of SVM in breast cancer diagnosis

In figure 5, the Support Vector Machine (SVM) confusion matrix is shown. The confusion matrix is very important metric to evaluate classifiers quality as well as performance. Because, the confusion matrix shows the number of misclassifications by the classification models and the confusion matrix also shows the true classifications.

As illustrated in figure 5, there are a total of 8 miss-classification or false positives and false negatives using the Support Vector Machine (SVM) model. This value is less than the false positives and false negatives in the Decision tree classification model. Form a total of 569 samples given in the dataset, 68 are classified as malignant and 38 are classified as benign.

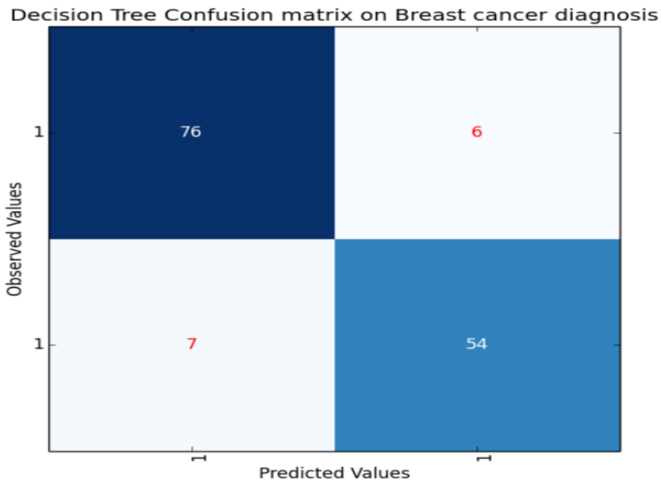

Fig. 6.Fig.6 Confusion Matrix of Decision Tree in breast cancer diagnosis

In figure 6, the Decision tree classifier model's confusion matrix is illustrated. As illustrated in figure 4, there are a total of 13 miss-classification or false positives and false negatives. And when compared to the Support Vector Machine (SVM) classifier the values of misclassification in the Decision tree classification model is greater than the misclassification in the Support Vector Machine classification model. The true positive and true negatives are 76 and 54 respectively using the Decision tree. This implies that Form a total of 569 samples given in the dataset, 76 are classified as malignant and 54 are classified as benign and 12 samples are misclassified.

D. Precision analysis

The precision of classification model is the measure of how precise the models are in the classification of breast cancer into the malignant and benign category. The precision is used as a metric to evaluate the performance of the Support Vector Machine (SVM) and the Decision tree models in the classification of breast cancer into either the malignant or the benign class. The precision of the Support Vector Machine (SVM) model is illustrated in table V and the precision of the decision tree classifier is illustrated in table V.

Table- V: Name of the Table that justify the values

Random test on the model	Precision	
	Malignant	Benign
1	0.83	0.94
2	0.92	0.89
3	0.92	0.87
4	0.79	0.91
5	0.85	0.90

Fig. 7.Fig.7 precision of the SVM model

Table- V: Name of the Table that justify the values

Random test on the model	Precision	
	Malignant	Benign
1	0.84	0.94
2	0.74	0.96
3	0.86	0.91
4	0.78	0.88
5	0.85	0.96

Fig. 8.Fig.8 precision of the SVM model

V. CONCLUSION

In this study a machine learning model is proposed by employing the most common supervised machine learning algorithms namely, the SVM and decision tree algorithm. The proposed model is developed and tested on a breast cancer data collected from the kaggle repository consisting of 569 observations. Finally, the proposed model is tested on the training dataset. The analysis on the performance of the proposed model shows that the support vector machine (SVM) has better accuracy on identification of breast cancer as compared to the decision tree algorithm. In addition to accuracy, the confusion matrix is used to evaluate the performance of SVM and the decision tree algorithm and the analysis of the results shows that SVM is better on identification of breast cancer with less number of false negatives or false positive.

REFERENCES

1. Vasanth P.C, Nataraj K.R, Facial Expression Recognition Using SVM Classifier, Indonesian Journal of Electrical Engineering and Informatics (IJEEI) Vol. 3, No. 1, March 2015.
2. Tsehay Admassu Assegie, Pramod Sekharan Nair, Handwritten digits recognition with decision tree classification: a machine learning approach, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 5, October 2019, pp. 4446~4451.
3. M. Atif, M. s. AISalhp, s. Devanesan, v. Masilamani, K. Farhat, D. Rabah, Spectral characterization of Breast Cancer, IEEE, 2014.
4. Mohamed A. Berbar , Hybrid methods for feature extraction for breast masses classification, Egyptian Informatics Journal 19 (2018) 63–73.
5. Madina Hamiane, Fatema Saeed, SVM Classification of MRI Brain Images for Computer Assisted Diagnosis, International Journal of Electrical and Computer Engineering (IJECE) Vol. 7, No. 5, October 2017.
6. Parameshwar R. Hegde, Manjunath M. Shenoy, B.H. ShekarComparison of Machine Learning Algorithms for Skin Disease Classification Using Color and Texture Features, IEEE, 2018.
7. Yulia Ery Kurniawati, Adhistya Erna Permanasari, Silmi Fauziati, Comparative Study on Data Mining Classification Methods for Cervical Cancer Prediction Using Pap Smear Results, IEEE, 2016.
8. Ashutosh Kumar Dubey, Umesh Gupta, Sonal Jain, Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data, International Journal on Advanced Sciences, Engineering and Information Technology, Vol.8 , 2018.
9. Ashraf Osman Ibrahim, Siti Mariyam Shamsuddin, Abdulrazak Yahya Saleh, Ali Ahmed , Mohd Arfian Ismail, Shahreen Kasim , Backpropagation Neural Network Based on Local Search Strategy and Enhanced Multi-objective Evolutionary Algorithm for Breast Cancer Diagnosis, International Journal on Advanced Sciences, Engineering and Information Technology, Vol.9 , 2019.
10. Mohammad Alifraheed , An approach for detection of the mass of breast cancer in mammogram images, Journal of Theoretical and Applied Information Technology 30th September 2018. Vol.96. No 18.
11. Hiba Asria , Hajar Mousannif,Hassan, Al Moatassime ,Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems, Procedia Computer Science 83 (2016) 1064 – 1069.
12. Nirmine Hammouch, Hassan Ammor, A confocal microwave imaging implementation for breast cancer detection, Indonesian Journal of Electrical Engineering and Informatics (IJEEI) Vol. 7, No. 2, June 2019, pp. 263~270.
13. Ashutosh Kumar Dubey, Umesh Gupta, Sonal Jain, Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data, International Journal on Advanced Sciences, Engineering and Information Technology, Vol.8 , 2018, No.1.
14. Mohammed Abdulrazaq Kahya, Classification enhancement of breast cancer histopathological image using penalized logistic regression, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 13, No. 1, January 2019, pp. 405~410.
15. Shofwatul Uyun, Lina Choridah, Feature Selection Mammogram based on Breast Cancer Mining, International Journal of Electrical and Computer Engineering (IJECE), Vol. 8, No. 1, February 2018, pp. 60~69.
16. Liu, Lie, Research on logistic regression algorithm of breast cancer diagnosis data by machine learning, 2018, IEEE.
17. Abien Fred M. Agarap, On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset, IEEE, Feb, 2019.
18. Tsehay Admassu Assegie, Pramod Sekharan Nair, the performance of different machine learning models on breast cancer prediction, international journal of scientific research and engineering, January 2020.

AUTHORS PROFILE

Tsehay Admassu Assegie, is BSc. In Computer Science, Dilla University, Ethiopia and MSc. In Computrer Science, Andhra University, Visakhaptnam, India. Lecturer department of computing technology, college of engineering and technology, Aksum University, Aksum, Ethiopia.

Sushma S. J., is Associate professor, GSSSIETW, MYSURU, KARNATAKA, India.