Support Vector Machine Classifier for Prediction of Breast Malignancy Using Wisconsin Breast Cancer Dataset

Reddy Anuradha
Assistant Professor, Department of CSE,
Malla Reddy Institute of Technology & Science,
Hyderabad, Telangana
anuradhareddy.anu@gmail.com

Abstract - Cancer is the world's second largest cause of death. In 2018, 9.6 million people died from cancer. In any medical sickness, breast cancer is one of the most delicate and endemic diseases. This is one of the primary causes of female death in the world. Breast cancer kills one out of every eleven women around the world. "Early detection equals improved odds of survival," says a well-known cancer adage. As a result, early detection is essential for successfully preventing breast cancer and lowering morality. Breast Cancer is a type of cancer that affects one of the most significant issues that humanity has faced in recent decades has been diagnosis and prediction. Cancer detection that is accurate can save millions of lives. Effective technologies for diagnosing malignant breasts aid healthcare providers in diagnosing and treating patients in a fast and accurate manner. Experiments were carried out in this study to categorize breast cancer as benign or malignant using the Wisconsin Diagnosis Breast Cancer (WDBC) database. Support Vector Machine is a supervised learning technique (SVM). The SVM classifier's classification performance is evaluated. Experiments demonstrate that the SVM model has a fantastic performance, with a classification accuracy of 96.09 percent on the testing subset.

Keywords- Wisconsin Breast Cancer Breast cancer, Mammography, Artificial intelligence, support vector machine, Wisconsin Breast Cancer dataset

I. INTRODUCTION

Breast cancer is the most frequent cancer among women throughout the world. Breast cancer is remains the most common and second leading cause of death in women, despite years of technical and scientific research. The traditional approach for diagnosing breast cancer was X-ray mammography. Sharp needle aspiration cytology (FNAC) is employed since there are so many different ways to interpret mammography [1]. FNAC's average accurate identification accuracy is only 90 percent. As a result, it is critical to develop alternate approaches for detecting breast cancer. To reduce the amount of incorrect diagnoses, data mining tools are a suitable alternative [2].

Early detection of cancer can enhance life expectancy by up to 98 percent. Figure 1 shows the various types of malignancies, with breast cancer taking the lead with 24 percent.

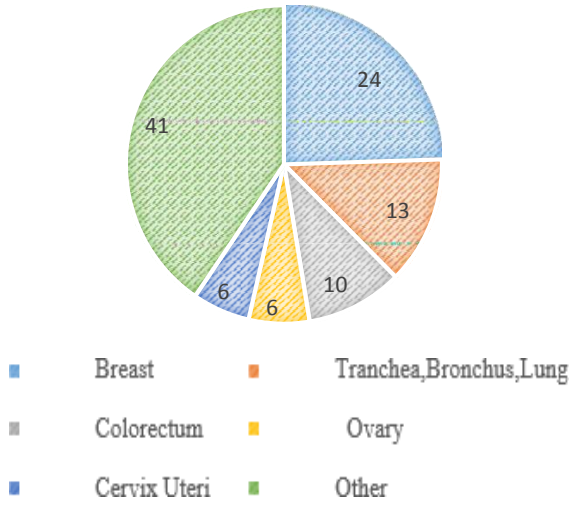

Fig. 1. various types of malignancies

Artificial intelligence (AI) can be used to detect and diagnose breast cancer more accurately, as well as to avoid overtreatment. Nonetheless, merging Artificial Intelligence (AI) and Machine Learning (ML) approaches allows for better prediction and decision-making accuracy [3,4]. For example, evaluating whether or not a patient need surgery based on the biopsy results for detecting breast cancer [5].

Breast cancer is the most frequent malignancy among women, according to a World Health Organization (WHO) assessment. Around 5% of Indian women are at risk of breast cancer, compared to 12.5 percent in Europe and the United States. Breast cancer is usually easy to detect if particular symptoms appear. Some women with breast cancer, on the other hand, experience no symptoms. As a result, early detection of breast cancer is critical.

Early identification of breast cancer aids in early diagnosis and therapy, as early diagnosis and treatment are critical for long-term survival. Breast cancer can be detected, diagnosed, and treated early enough to preserve a patient's life [6, 7].

II. LITERATURE REVIEW

Many studies have been undertaken on the application of Machine Learning (ML) to the detection of breast cancer. For a

greater rate of accuracy, detection and diagnosis can be done using a variety of methods or a combination of algorithms.

S. Gc, R. Kasaudhan [8] Worked on removing characteristics such as volatility, range, and compactness. The performance was assessed using SVM classification. Their labour production had the highest variance (95%), range (94%), and compactness (86%). SVM can be deemed a good approach for detecting breast cancer based on their results.

Durai et al.[9] Data Mining was used to recognise diseases such as breast cancer. He employed LRC and compared it to BFI, ID3, J48, and SVM, among other approaches. The output reveals that LRC is the most precise, with a precision of 99.25 percent.

Hafizah et al.[10] SVM and ANN were compared using various breast cancer data sets, including WBCD, BUPA JNC, Data, and Ovarian. Despite the fact that both methods have great performance, SVM was found to be superior to ANN in the research.

Tsirogiannis [11] on medical databases, bagging techniques such as decision trees, neural networks, and SVM were used. Bagging approaches, according to the research, are more accurate.

Avramov and Si [12] worked on feature extraction and the performance impact of selecting. They used five classification models and three methods of correlation selection (PCA, T-Test Significance, and Random feature selection) (LR, DT, KNN, LSVM, and CSVM). Stacking the logistic, SVM, and CSVM improve accuracy to 98.56 percent, which is the best result.

Azar and El-Said [13] I worked on six distinct SVM techniques. Against see which approach works best in terms of accuracy, sensitivity, specificity, and ROC, he compared ST-SVM to LPSVM, LSVM, SSVM, PSVM, and NSVM. With accuracy of 97.1429 percent, sensitivity of 98.2456 percent, specificity of 95.082 percent, and ROC of 99.38 percent, LPSVM emerged as the winner. As a result, LPSVM provides the best performance and accuracy.

Angeline [14] examined the performance of Nave Bayes, Decision Tree (C4.5), K-Nearest Neighbor, and Support Vector Machine in predicting the primary location of cancer in Wisconsin Breast Cancer (WBC). According to the findings, SVM outperforms other methods.

Mehmet Fatih Akay[15] On the problem of identifying breast cancer, a proposed medical decision-making system based on SVM paired with feature selection was used. Based on the findings, SVM-based models have shown to be quite effective in classifying breast cancer.

Leena Vig[16] Random Forest classifiers, Artificial Neural Networks, Nave Bayes, and Support Vector Machines were used in the investigation. The results reveal that ANNs, Random Forests, and SVMs can provide models with high accuracy, sensitivity, and specificity, while Nave Bayes does not.

III. PROCEDURE

We follow a few steps to diagnose breast cancer.

A. Data Collection & Preparation

Wisconsin Breast Cancer (WBC) is a breast cancer dataset taken from the UCI machine learning repository dataset [17]. This dataset contains 569 cases that are classified as benign or malignant, with 357 cases (62.74 percent) being benign and 212 cases (37.25 percent) being malignant. The data is divided into two classes, B and M, with B denoting the benign and M denoting the malignant. Breast cancer is the most common condition in medical diagnosis, and its prevalence is rising every year. Except for sample code number and class, the dataset comprises 32 features: radius mean, texture mean, area mean, smoothness, compactness, and concavity [18, 19]. The benign cases are classified as positive because they have little impact on the body, while the malignant instances are classified as negative since they are cancerous cells that have a negative impact on the body in our study. In the data set, there are 16 missing feature values. The mean is used to fill in the gaps in the missing features. Finally, to ensure proper data propagation, the data set is randomized.

B. Indicators of Performance Measurement

Several performance measures are used to assess the effectiveness of machine learning algorithms. To evaluate the parameter, a confusion matrix consisting of TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) is created for the actual and projected class. The meanings of the terms are listed below.

TP (True Positive) = Identified correctly; FP (False Positive) = Correctly Rejected; TN (True Negative) = Incorrectly Identified; FN (False Negative) = Rejected Incorrectly

IV. ALGORITHM

Support vector machines supervised learning algorithms are employed in our research.

A. SVM

The Support Vector Machine is a classification model that uses supervised learning and has excellent classification performance [20]. Each data item is represented in the SVM algorithm model as n-dimensional coordinates [21]. Where n is the total number of categorization features. In data point coordinates, each feature's value is stated. The SVM includes decision hyperplanes that use maximum margin to divide distinct classes of data points [22]. Support vectors are data points found near hyperplanes. The classification process creates non-linear decision boundaries and categorizes data points that aren't represented in vector space.

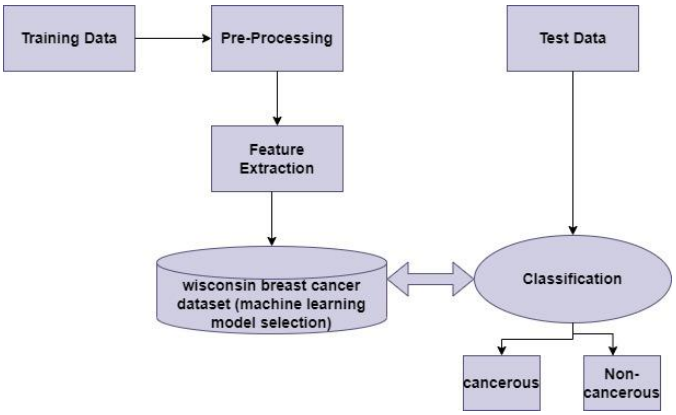

Fig. 2. Process framework of the model

Data used to train the model on the basis of example classes is referred to as training data [23].

Pre-processing: Using the describe technique in machine learning, calculate the mean and standard deviation, then delete the id and class.

Extraction of Features: The process of removing unnecessary characteristics and training the model on useful features is known as feature extraction.

WBCD is a data set. Our model was trained using the Wisconsin Breast Cancer Dataset. It is updated once a year in order to be more useful in the training model.

After classification, the SVM (Support Vector Machine) method model of Machine Learning can be used to predict the outcome [24]. Categorize a diseased raw data as benign or malignant more quickly.

V. CONCLUSION

We concentrated on breast cancer in this study because it is a very severe disease that kills many women around the world. In the field of Medicare and Biomedical, breast cancer prognosis is quite important. The goal of this work was to create a classifier that could predict the most serious malignancy, breast cancer. In this paper, we developed a collaborative strategy for diagnosing this disease and providing information on the patient's condition. The breast cancer model is described as a classification job in this article, and the Support Vector Machine (SVM) approach is used to classify breast cancer as benign or malignant. SVM produces accuracy and precision as a result. To summarize the created method, the first step is to collect patient data in the form of a text or csv file. Remove the features that aren't relevant, such as id and other. Finally, for classification, the SVM classifier is utilized, which trains models to categorize cancer patients based on their diagnosis. The model's usefulness is demonstrated by the experimental results. On test subsets, SVM achieves a classification accuracy of 96.09 percent.

REFERENCES

[1] Patan, R., Ghantasala, G. P., Sekaran, R., Gupta, D., & Ramachandran, M. (2020). Smart healthcare and quality of service in IoT using grey filter convolutional based cyber physical system. *Sustainable Cities and Society*, *59*, 102141

[2] Bhowmik, C., Ghantasala, G. P., & AnuRadha, R. (2021). A Comparison of Various Data Mining Algorithms to Distinguish Mammogram Calcification Using Computer-Aided Testing Tools. In *Proceedings of the Second International Conference on Information Management and Machine Intelligence* (pp. 537-546). Springer, Singapore

[3] Ghantasala, G. P., Kallam, S., Kumari, N. V., & Patan, R. (2020, March). Texture Recognization and Image Smoothing for Microcalcification and Mass Detection in Abnormal Region. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-6). IEEE

[4] Sreehari, E., & Ghantasala, P. G. (2019). Climate Changes Prediction Using Simple Linear Regression. *Journal of Computational and Theoretical Nanoscience*, *16*(2), 655-658

[5] Ghantasala, G. P., & Kumari, N. V. (2021). Breast Cancer Treatment Using Automated Robot Support Technology For Mri Breast Biopsy. *INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS*, *1*(2), 235-242

[6] Ghantasala, G. P., & Kumari, N. V. (2021). Identification of Normal and Abnormal Mammographic Images Using Deep Neural Network. *Asian Journal For Convergence In Technology (AJCT)*, *7*(1), 71-74

[7] Chandana, P., Ghantasala, G. P., Jeny, J. R. V., Sekaran, K., Deepika, N., Nam, Y., & Kadry, S. (2020). An effective identification of crop diseases using faster region based convolutional neural network and expert systems. *International Journal of Electrical and Computer Engineering (IJECE)*, *10*(6), 6531-6540

[8] S. Gc, R. Kasaudhan, T. K. Heo, and H.D. Choi, "Variability Measurement for Breast Cancer Classification of Mammographic Masses," in *Proceedings of the 2015 Conference on research in adaptive and convergent systems (RACS)*, Prague, Czech Republic, 2015, pp. 177–182

[9] S. G. Durai, S. H. Ganesh, and A. J. Christy, "Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer," In Computing and Communication Technologies (WCCCT), 2017 World Congress on 2018

[10] S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, "Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study," J. Teknol, vol. 65, pp. 73–81, 2018

[11] Tsirogiannis, G. L., et al. "Classification of medical data with a robust multi-level combination scheme." Neural Networks, 2004. Proceedings. 2004 IEEE International JointConference on. Vol. 3. IEEE, (2018)

[12] T. K. Avramov and D. Si, "Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis," Proc. Int. Conf. Comput. Data Anal. -ICCDA '17, pp. 69–74, 2018

[13] A. T. Azar, and S. A. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition," Neural Comput. Appl., vol. 24,no. 5, pp. 1163–1177, 2018

[14] Christobel, Angeline, and Y. Sivaprakasam. "An empirical comparison of data mining classification methods." International Journal of Computer Information Systems 3.2(2011): 24-28

[15] Mehmet Fatih Akay, "Support Vector Machines Combined With Feature Selection For Breast Cancer Diagnosis", ExpertSystems with Applications 36, 3240–3247, 2018

[16] LeenaVig , "Comparative Analysis of Different Classifiers for the Wisconsin Breast Cancer Dataset", Open Access Library Journal, Volume 1 | e660,2018

[17] G S Pradeep Ghantasala, D. Nageswara Rao, Mandal K (2021) MACHINE LEARNING ALGORITHMS BASED BREAST CANCER PREDICTION MODEL. Journal of Cardiovascular Disease Research, 12 (4), 50-56. doi:10.31838/jcdr.2021.12.04.04

[18] Ghantasala, G. P., Kumari, N. V., & Patan, R. (2021). Cancer prediction and diagnosis hinged on HCML in IOMT environment. In *Machine Learning and the Internet of Medical Things in Healthcare* (pp. 179-207). Academic Press

[19] Kishore, D. R., Syeda, N., Suneetha, D., Kumari, C. S., & Ghantasala, G. P. (2021). Multi Scale Image Fusion through Laplacian Pyramid and Deep Learning on Thermal Images. *Annals of the Romanian Society for Cell Biology*, 3728-3734

[20] Kumari, N. V., & Ghantasala, G. P. (2020). Support Vector Machine Based Supervised Machine Learning Algorithm for Finding ROC and LDA Region. *Journal of Operating Systems Development & Trends*, *7*(1), 26-33

[21] Ghantasala, G. P., Tanuja, B., Teja, G. S., & Abhilash, A. S. (2020). Feature Extraction and Evaluation of Colon Cancer using PCA, LDA and Gene Expression. *Forest*, *10*(98), 99

[22] G. S. Pradeep Ghantasala, Nalli Vinaya Kumari. Mammographic CADe and CADx for Identifying Microcalcification Using Support Vector Machine. Journal of Communication Engineering & Systems. 2020; 10(2): 9–16p

[23] Ghantasala, G. P., Reddy, A., Peyyala, S., & Rao, D. N. (2021). Breast Cancer Prediction In Virtue Of Big Data Analytics. *INTERNATIONAL JOURNAL OF EDUCATION, SOCIAL SCIENCES AND LINGUISTICS*, *1*(1), 130-136

[24] Ghantasala, G. P., Reddy, A. R., & Arvindhan, M. Prediction of Coronavirus (COVID-19) Disease Health Monitoring with Clinical Support System and Its Objectives. In *Machine Learning and Analytics in Healthcare Systems* (pp. 237-260). CRC Press