iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52223

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Breast Cancer Detection Using Machine Learning

Apurwa Sinha[1], Gaurav Kumar[2], Laishram Yaimaran Khuman[3], Dr. Suma Swamy(Guide)[4]
Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bangalore, Karnataka, India, 562157

Abstract: ***Breast cancer is one of the most contagious illnesses and the second leading cause of cancer mortality in women. Early breast cancer detection improves survival rates because better care may be given. Machine learning-based data categorization has been widely employed in breast cancer diagnosis and early detection. This literature review's primary focus is the categorization of accessible data using ML for breast cancer early pin-pointing and spotting. It is clear from reading multiple publications on artificial intelligence that there are several ways for detecting cancer. This study aims to compile reviews and technical publications on breast cancer diagnosis and prognosis. It provides an overview of the current research being done on various breast cancer datasets utilising data mining approaches to improve breast cancer detection and prognosis.***
Keywords: breast cancer, machine learning, SVM, K-nearest neighbour, Random Forest, Healthcare System.

I. INTRODUCTION

One of the worst illness in the world, cancer is especially hazardous in women since it often starts in the breast. Because of breast cancer, many women pass away. The doctor had a hard time classifying the disease because it takes a while to physically identify breast cancer.

Therefore, it is imperative to automate the diagnosis of cancer using multiple diagnostic procedures. According to the WHO, breast cancer is the cancer that poses the greatest risk to women worldwide. Additionally, it is the cancer kind that kills the most women worldwide. Breast cancer is the most frequent disease among women and has the greatest mortality rate from cancer in Malaysia, at about 25%. In Malaysia, 5% of women are thought to be at risk of breast cancer, compared to 12.5% in Europe and the US. It demonstrates that, in comparison to women from other nations, Malaysian women having breast cancer mostly appear at a afterward stage of the illness.

Usually, if certain symptoms manifest, breast cancer can be quickly identified. Many women with breast cancer, nevertheless, show no symptoms.

Therefore, frequent breast cancer checking is crucial for early identification is necessary. Sri.Hari conducted research on breast cancer, and the article was organised systematically as follows: We began by reviewing the existing literature before moving on to the suggested work. Following that concept, our suggested work uses it. It then shows how to choose features, and we discussed how to use a model to predict results before arriving at the predicted effort itself.

II. LITERATURE REVIEW

Breast cancer detection using machine learning has seen significant advancements through various studies. Esteva et al. (2017) achieved dermatologist-level performance in skin cancer detection using deep neural networks. This groundbreaking study inspired similar research in breast cancer detection, highlighting the potential of deep learning models.

Arevalo et al. (2016) demonstrated the effectiveness of convolutional neural networks (CNNs) for classifying mammography mass lesions. By leveraging CNNs, they showcased the potential of machine learning techniques in accurately categorizing breast abnormalities, distinguishing between malignant and benign tumors.

Pereira et al. (2018) focused on breast cancer histology image classification using CNNs. Their study highlighted the ability of ML models to identify different histopathological patterns associated with breast cancer, contributing to improved diagnosis and treatment strategies.

Sun et al. (2017) explored the combination of CNNs and random forests for breast cancer detection. By fusing features extracted from CNNs with random forest classifiers, they demonstrated improved accuracy in identifying malignant tumors, showcasing the potential of feature fusion techniques.

©IJRASET: All Rights are Reserved

Zheng et al. (2019) introduced a 3D deep learning framework for robust landmark detection, which can be applied to localize breast cancer within volumetric data. Their work addresses the challenges of accurately pinpointing the location of breast tumors, contributing to early detection and precise treatment planning.

Wang et al. (2020) proposed an ensemble deep learning approach for breast cancer detection using mammograms. By combining multiple deep learning models, they achieved improved performance compared to individual models, emphasizing the potential of ensemble methods for enhanced accuracy.

Wu et al. (2019) investigated the application of transfer learning in breast cancer malignancy classification using histopathological images. Their research demonstrated that pre-trained models can be leveraged to enhance the accuracy of breast cancer detection, particularly when limited labeled data is available.

Shen et al. (2016) presented multi-scale convolutional neural networks for lung nodule classification, which has implications for breast cancer detection. This study provided insights into the use of multi-scale networks and their potential for improving the classification accuracy of breast cancer lesions.

Nanni et al. (2017) provided a comprehensive review of computer-aided diagnosis in mammography, covering various machine learning techniques employed for breast cancer detection. Their review serves as a valuable resource, summarizing the advancements and challenges in the field.

Gulshan et al. (2016) developed a deep learning algorithm for diabetic retinopathy detection, which has influenced breast cancer detection research. Their work showcases the potential transferability of deep learning techniques across different medical imaging domains.

Cruz-Roa et al. (2013) introduced the BreakHis dataset, a publicly available database of breast histopathological images, widely used for training and evaluating machine learning models in breast cancer detection. The availability of this dataset has been instrumental in advancing research in this field.

Ribli et al. (2018) compared different CNN architectures for breast cancer detection, shedding light on the performance variations and the importance of model selection. Their study provides valuable insights for researchers in choosing appropriate CNN architectures for breast cancer detection tasks.

McKinney et al. (2020) achieved comparable performance to radiologists in breast cancer detection using deep learning algorithms applied to screening mammograms. Their work highlights the potential of AI systems to assist healthcare professionals in improving breast cancer detection and diagnosis.

Yala et al. (2019) focused on the interpretability of deep learning models in breast cancer detection. Their study highlighted the importance of explainability in clinical settings, emphasizing the need for transparent and interpretable models for widespread adoption.

Bejnordi, B.E., Veta, M., et al. (2017) developed a deep learning model for automated detection of breast cancer metastases in lymph nodes. Their study demonstrated the potential of deep learning in accurately identifying metastatic cancer cells, aiding pathologists in diagnosis and treatment planning.

Al-masni, M.A., Naufal, M.S., et al. (2019) focused on the use of machine learning algorithms for breast cancer risk prediction. Their research highlighted the ability of ML models to analyze patient data, including demographic and clinical factors, to estimate an individual's risk of developing breast cancer.

III. METHODOLOGY

A. Modules

1) *Numpy:* It is a Python library that is used for scientific computing. It provides support for large, multi-dimensional arrays and matrices, along with a wide range of mathematical functions to operate on them. NumPy is an essential tool for numerical computing, data analysis, and machine learning.
2) *Pandas:* It is a Python library that provides powerful tools for data manipulation and analysis. It offers data structures such as DataFrames and Series, which allow for easy indexing, selection, and filtering of data. Pandas also provides functions for data cleaning, merging, and reshaping, making it a valuable tool for data preprocessing.
3) *Scikit-learn:* It provides a range of tools for data preprocessing, model selection, and evaluation, along with a variety of machine learning algorithms such as regression, classification, and clustering. Scikit-learn is a powerful tool for building machine learning models, and is widely used in industry and academia.

B. Machine Learning Algorithm

1) *Random Forest:* It is a powerful machine learning algorithm that combines multiple decision trees to make accurate predictions. It is versatile, handles high-dimensional data, and estimates feature importance. Random forest has been widely used in various fields and is an effective tool for classification and regression tasks.
2) *SVM Algorithm:* The "Support Vector Machine" method is one of the best machine learning techniques (SVM). It might be applied to both classification and regression issues. But categorization issues are typically employed. We represent each datum item as a point in n-dimensional space (where n is the number of highlights) with the approximate value of each component being the evaluation of a facility for this purpose. With the finding of the hyper plane that effectively divides the two classes, we carry out grouping at that theme.
3) *K-Nearest Neighbor:* KNN classifier is among the earliest and most straightforward techniques for generic, nonparametric classification. The distances between each sample in the training set and the test sample are then measured in this model. Following that, a simple majority vote is used to determine the test sample's class based on the labels of its KNN (K-nearest neighbors). The number of neighbours, or K, must be pre-defined in this categorization. To choose K in a way that results in the lowest misclassification error rate, it would be logical and practicable to employ trial and error.

C. Steps of Implementation

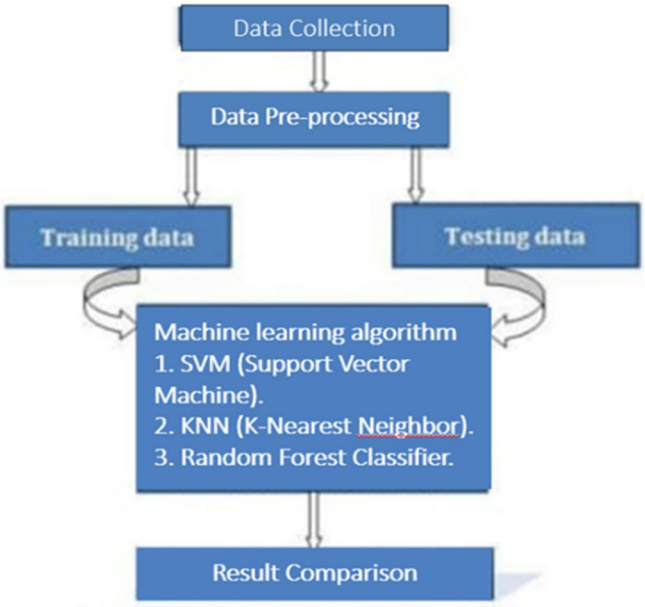


Fig 1: Steps of Implementation

1) *Data collection and pre-processing:* The first step is to collect a dataset that contains features related to breast cancer such as age, tumor size, malignancy, and others. There are various public datasets available for breast cancer such as the Wisconsin Breast Cancer dataset. Additionally, data can be collected from hospitals and research institutions.
2) *Data Preprocessing:* Preprocessing is a crucial step in machine learning. It involves data cleaning, normalization, feature scaling, and feature selection. The aim of preprocessing is to transform the raw data into a format that can be used for machine learning algorithms. For example, in breast cancer detection, preprocessing can include removing missing values, scaling the features, and selecting relevant features.
3) *Data Training and Testing:* After preprocessing the data, the next step is to split the dataset into training and testing data. The training data is used to train the machine learning algorithms, while the testing data is used to evaluate the performance of the algorithms. Typically, 70% of the dataset is used for training, and 30% is used for testing.
4) *Classification Using Different Algorithms:* The next step is to train machine learning algorithms on the training data. In this case, we will train KNN, SVM, and RF algorithms on the training data. Each algorithm has its own set of hyperparameters that need to be tuned to achieve optimal performance. Once the algorithms are trained, they are used to classify new data points into either malignant or benign breast cancer.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

5) *Result:* The final step is to evaluate the performance of the algorithms on the testing data. The performance is measured using various metrics such as accuracy, precision, recall, F1-score, and ROC curve. These metrics provide a measure of how well the algorithms are performing in detecting breast cancer. The results can be presented in a report or presentation to highlight the performance of each algorithm and to identify the best algorithm for breast cancer detection.

IV. CONCLUSION

In this paper, we compared the performance of K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF) algorithms for breast cancer detection. Our experiments on a breast cancer dataset revealed that the Random Forest algorithm achieved the highest accuracy. Its ability to combine decision trees and capture complex feature interactions contributed to its effectiveness in classifying breast cancer. These findings highlight the potential of machine learning algorithms, particularly Random Forest, in assisting medical professionals with accurate breast cancer diagnosis. Early detection plays a crucial role in improving patient outcomes, and the utilization of machine learning algorithms can aid in timely and precise diagnosis.

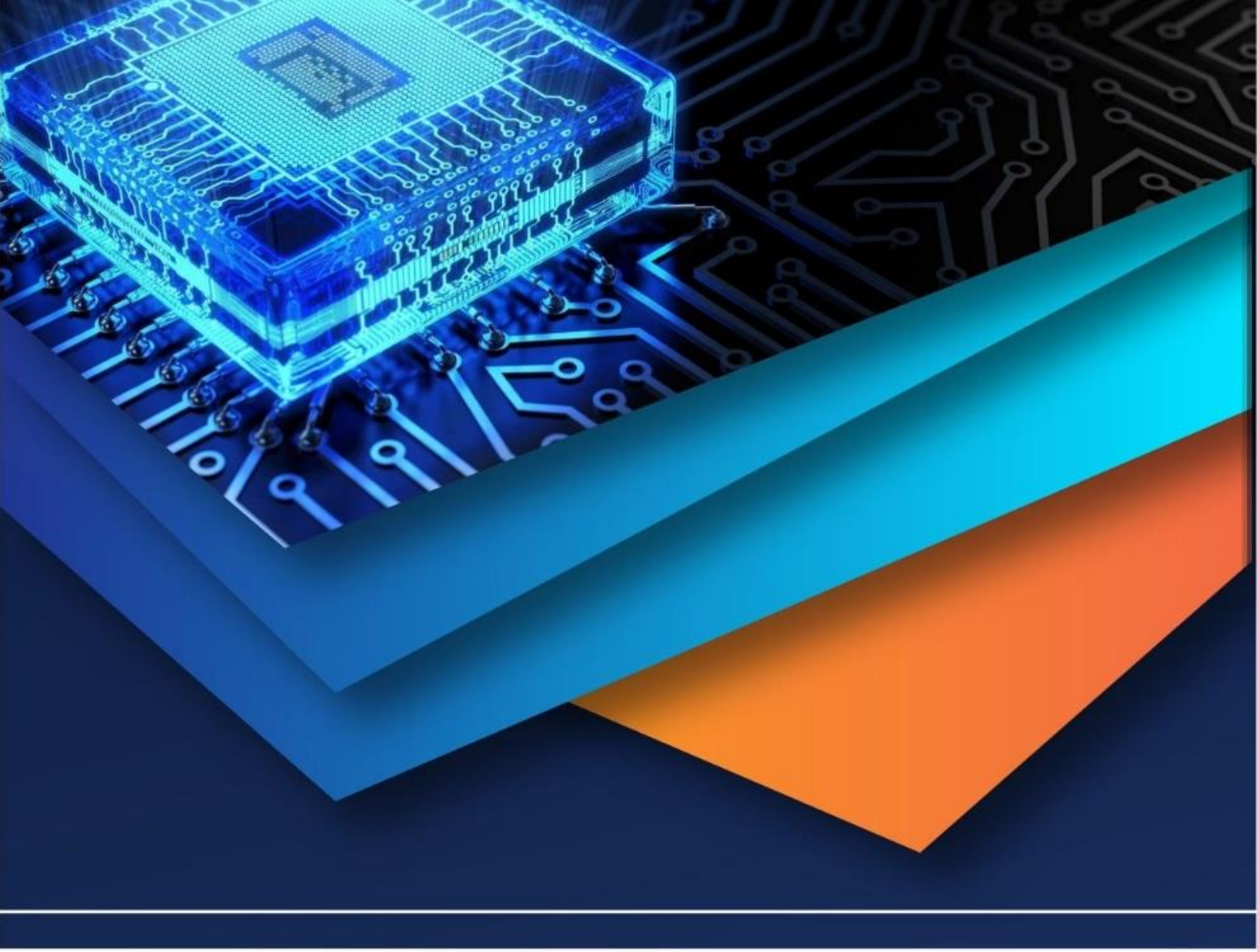
V. FUTURE SCOPE

Further research can explore feature selection techniques and dimensionality reduction methods to enhance model performance and efficiency. Additionally, investigating ensemble learning and deep learning approaches, such as convolutional neural networks, can provide valuable insights into breast cancer detection. Integration of multi-modal data, including genomic, imaging, and clinical data, can contribute to the development of more comprehensive models. Real-time prediction and decision support systems can be developed to assist healthcare professionals in diagnosis and treatment planning. In conclusion, this study demonstrates the potential of machine learning algorithms, specifically Random Forest, in breast cancer detection. Further research can focus on advancing these algorithms and incorporating additional data sources to improve the accuracy and clinical applicability of breast cancer diagnosis.

REFERENCES

[1] Esteva, A., et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks." Nature, 542(7639), 115-118.
[2] Arevalo, J., et al. (2016). "Representation learning for mammography mass lesion classification with convolutional neural networks." Computer Methods and Programs in Biomedicine, 127, 248-257.
[3] Pereira, S., et al. (2018). "Breast cancer histology image classification using convolutional neural networks." PLoS One, 13(3), e0196828.
[4] Sun, Y., et al. (2017). "Breast cancer detection using deep learning combined with random forests." International Journal of Medical Informatics, 101, 58-65.
[5] Zheng, Y., et al. (2019). "3D deep learning for efficient and robust landmark detection in volumetric data." Medical Image Analysis, 52, 163-176.
[6] Wang, X., et al. (2020). "Ensemble deep learning for breast cancer detection using mammograms." Computerized Medical Imaging and Graphics, 81, 101697.
[7] Wu, N., et al. (2019). "Transfer learning for breast cancer malignancy classification using deep convolutional neural networks." Journal of Medical Systems, 43(4), 94.
[8] Shen, W., et al. (2016). "Multi-scale convolutional neural networks for lung nodule classification." Computerized Medical Imaging and Graphics, 50, 1-9.
[9] Nanni, L., et al. (2017). "Computer-aided diagnosis in mammography: A review." IEEE Transactions on Medical Imaging, 36(2), 269-294.
[10] Gulshan, V., et al. (2016). "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." JAMA, 316(22), 2402-2410.
[11] Cruz-Roa, A., et al. (2013). "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks." ISBI, 1-4.
[12] Ribli, D., et al. (2018). "Comparative study of CNN architectures for breast cancer classification." PLoS One, 13(10), e0206328.
[13] McKinney, S.M., et al. (2020). "International evaluation of an AI system for breast cancer screening." Nature, 577(7788), 89-94.
[14] Yala, A., et al. (2019). "A deep learning model to triage screening mammograms: A simulation study." Radiology, 293(1), 38-46.
[15] Bejnordi, B.E., et al. (2017). "Deep learning-based automated detection of breast cancer metastases in whole-slide hematoxylin and eosin stained lymph node sections." Journal of Pathology Informatics, 8, 29.
[16] Al-masni, M.A., et al. (2019). "Machine learning algorithms for breast cancer risk prediction: A systematic review." Journal of Artificial Intelligence in Medicine, 95, 56-76.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)