Library Progress International
Vol.44 No.3, Jul-Dec 2024: P. 22224 - 22228

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Deep Learning Approaches for Breast Cancer Disease Prediction

[1]Sandhya Rani Sarlana, [2]P.Shailaja, [3]K. Sharmila, [4]Swetha Arra

,1 Associate Professor , CSE department , Malla Reddy Engineering College, Hyderabad, sandhyarani@mrec.ac.in
2Assoc.Prof, CSE department , Vaagdevi College of Engineering, Warangal, Telangana , pokalashylaja@gmail.com
3Associate Professor, Department of CSE (Data Science), Vaagdevi Engineering College, Warangal, Telangana , sharmilakreddy@gmail.com
4Assistant professor , Department of CSE(AI&ML), Vaagdevi College of Engineering, Warangal, Telangana, swetha_a@vaagdevi.edu.in

How to cite this article: Sandhya Rani Sarlana, P.Shailaja, K. Sharmila, Swetha Arra (2024) Deep Learning Approaches for Breast Cancer Disease Prediction. *Library Progress International*, 44(3), 22224 - 22228.

Abstract:

Breast cancer is a serious illness resulting from abnormal growth of breast cells. Early detection is crucial in the medical field to reduce the risk to human life. Previous research utilized the Predictive Modelling Technique (PMT) for this purpose. However, PMT struggles with handling noisy data, which can lead to inaccurate predictions. To address these challenges, the Deep Learning-based Breast Cancer Disease Prediction Framework (DLBCDPF) was developed. In this approach, feature selection is carried out using an optimization algorithm known as the Ranking-based Bee Colony method. The F-score values are used as fitness measures to identify the most relevant features. These features are ranked in descending order of their F-Scores, and a subset is formed. Data clustering is achieved through FCM clustering, while classification is performed using an Improved Deep Neural Network. The entire analysis was conducted on a breast cancer dataset using optimization techniques, and results indicate that this method surpasses previous models in accuracy.

Keywords: *Predictive Modelling Technique (PMT), Data classification, Breast cancer, Accuracy*.

1. INTRODUCTION

The convergence of vast digital data sources, enhanced computational power, and the adoption of artificial intelligence (AI) and machine learning (ML) by regulators is driving a significant transformation in clinical development (Chen et al., 2019). Across academia, biotechnology, nonprofit organizations, regulatory bodies, and technology companies, there is a growing effort to integrate computational evidence into clinical research and healthcare. ML models are increasingly being used to analyze public biomedical data, clinical trials, and real-world data from sensors and health records, contributing to more informed clinical decisions (Bzdok & Meyer-Lindenberg, 2018). Liver disease has become more prevalent with industrialization, and the importance of liver function is critical to human survival (Haq et al., 2018). Numerous studies have analyzed liver disease datasets from clinical trials to identify key symptoms and characteristics that aid in diagnosis. By combining relevant features from these datasets, diagnostic rules can be established, reducing the risk of misdiagnosis and improving patient outcomes (Hassoon et al., 2017).Noncommunicable diseases, such as cancer and heart disease, are now largely attributed to poor lifestyle choices and increased morbidity. Without early detection, many of these illnesses can become fatal, as most individuals remain unaware of their condition (Ngiam & Khor, 2019). Over the past decade, ML technologies have shown promise in aiding clinical decision-making, but implementing these tools in practice requires close collaboration between ML researchers and healthcare professionals. Integrating ML and AI into clinical reporting standards is essential for enhancing efficiency and care quality (Yin & Jha, 2017).ML offers substantial benefits in predicting outcomes and identifying unique patient subgroups with distinct physiological traits and prognoses. However, there remains a gap in understanding between clinicians and ML researchers. Physicians may struggle to evaluate ML studies, while ML experts may present overly technical details that are not easily understood by clinical audiences, leading to misinterpretations of the model's validity and usefulness (Halilaj et al., 2018).

2.MACHINE LEARNING ROLE IN PREDICTIONS

Machine learning has developed powerful techniques for achieving high prediction accuracy, utilizing advanced methods and large enterprise datasets to enhance the success of various applications. These algorithms are particularly effective when working with large datasets, as they help identify optimal decisions based on predictive value. Machine learning excels at handling predictive tasks by identifying behaviors most likely to yield desired outcomes (Nithya & Ilango, 2017). Medical data, on the other hand, is complex and composed of numerous heterogeneous variables collected from different sources such as demographics, medical history, medications, allergies, biomarkers, imaging, and genetic markers. Each data source offers a unique perspective on a patient's condition, but their statistical features vary significantly (Pathan et al., 2019).

Researchers face two main challenges when analyzing such data: the curse of dimensionality, where the feature space grows exponentially with the number of dimensions and samples, and the variability in feature sources and statistical attributes. These obstacles make disease identification slower and less accurate, delaying critical patient care. Therefore, there is a pressing need for an efficient and reliable methodology that can support early disease detection and assist doctors in decision-making (Schnack, 2017). As a result, researchers in medicine, computing, and statistics are working to develop new methods for disease diagnosis and prognosis, as traditional approaches struggle to process vast amounts of medical data. This need is closely linked to advancements in fields like Data Mining (DM), Artificial Intelligence (AI), and Big Data (BD) (Sperschneider, 2020).

3. DATASET DESCRIPTION

In this work, the Breast Cancer Wisconsin (Diagnostic) Data Set (Kaggle, 2021) is used. This dataset consists of nine numerical input attributes: bare nuclei, uniformity of cell size, normal nucleoli, uniformity of cell shape, single epithelial cell size, marginal adhesion, bland chromatin, mitoses, and a target variable. The output or target variable classifies the data into two categories: 2 (non-cancerous) and 4 (cancerous).

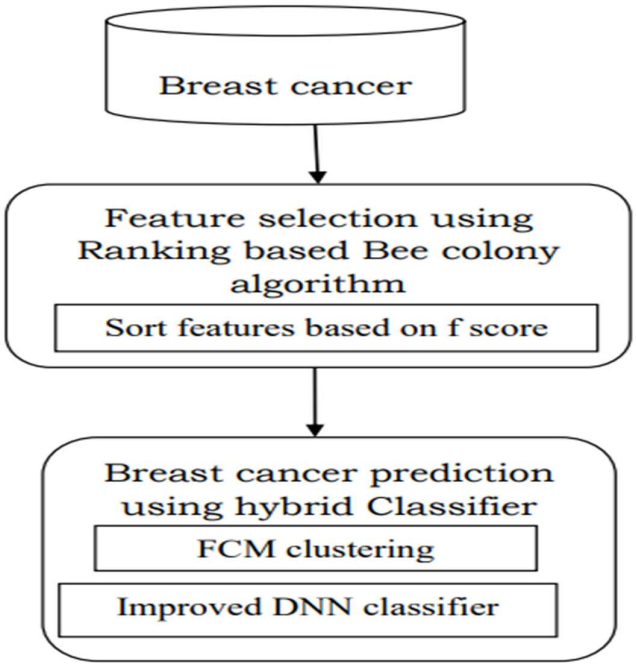

Fig.1: Overall processing flow of proposed research work

The dataset contains 699 instances, with 16 missing values in the "bare nuclei" attribute. These missing values were imputed by the mean of the bare nuclei attribute before proceeding with the modeling process (Agarap, 2018).

Backward modeling was performed using the p-value test, identifying eight significant variables for the prediction model: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, and normal nucleoli (El-Shair et al., 2020). The overall architectural design of the model is presented in Figure.1.

4. OPTIMAL FEATURE SELECTION USING RANKING BASED BEE COLONY ALGORITHM

Whether automatic or manual, has a significant impact on the prediction outcome. Selecting inappropriate data features can negatively affect model accuracy, as the model may learn from irrelevant information. In this study, the most optimal features from the training dataset are identified using an optimization method known as the Ranking-based Bee Colony approach (Wang et al., 2021). This method is inspired by the foraging behavior of honey bee swarms and is referred to as the Artificial Bee Colony (ABC) method.

In the ABC model, three types of bees exist within the colony: employed bees, spectators, and scouts. Each food source corresponds to one employed bee, and the number of employed bees is roughly equal to the number of food sources within a reasonable distance from the hive (Öztürk et al., 2020). Employed bees travel to their food sources, collect resources, and return to the hive to share information. When an employed bee's food source becomes exhausted, it turns into a scout and begins searching for new sources. Onlookers, or spectator bees, observe the dances of the employed bees and select food sources based on their movements (Zhao & Zhang, 2020).The bee colony algorithm mimics the behavior of honey bees searching for food. Each solution within the search space has a set of parameters representing food source locations, and the "fitness value" indicates the quality of a food source based on its location. The bee searches for a new food source if the fitness value improves significantly, otherwise, the previous source is retained. After all employed bees complete this process, they share information with the onlookers, who then choose their food sources based on the shared probability. This solidifies their perception of superior food sources.Each bee can continue searching for better food sources for a specific number of cycles or until its fitness value falls below a certain threshold, at which point it becomes a scout bee. This process allows the colony to optimize food gathering, similar to how the ABC method selects the optimal solution in an optimization task.

5. PREDICTION USING IMPROVED DEEP NEURAL NETWORK

A Deep Neural Network (DNN), a type of Artificial Neural Network (ANN), consists of multiple layers between the input and output layers. While there are various types of neural networks, all models share certain components, including neurons, synapses, weights, biases, and functions, which enable them to function similarly to the human brain. Like other machine learning (ML) algorithms, DNNs can be trained. For instance, a DNN trained to identify dog breeds will analyze an image, estimate the probabilities, and determine the breed of the dog in the image. Users can review these results and set thresholds for probabilities that the network should display, returning the proposed label. The term "deep" refers to the numerous layers involved in DNNs, where each layer represents a mathematical operation.

DNNs are capable of modeling complex non-linear relationships. Their architecture allows the creation of compositional models, where objects are represented as layered compositions of simpler elements. Additional layers enable the composition of features from lower layers, allowing DNNs to model complex data more efficiently compared to shallow networks with similar performance. For example, DNNs have been shown to significantly simplify the approximation of sparse multivariate polynomials. Various deep architectures, each suited to specific industries, have been developed, and while the performance of these architectures can be compared, this is only valid when they are evaluated on the same dataset.

6. RESULTS AND DISCUSSION

In This Study, Feature Selection from The Training Dataset Is Performed Using the Ranking-Based Bee Colony Approach, An Optimization Technique Designed to Identify the Most Relevant Features. The F-Score is Used as The Fitness Value to Assess and Rank the Features. Features are Sorted in Descending Order Based on their F-Scores, Including The F-Scores of N Independent Features, To Create a Subset of One Or More Optimal Features.A New Hybrid Classification Method Is Introduced for Diagnosing Various Health-Related Conditions. This Method Involves Clustering the Data First, Followed by Classification, With Data Reduction Applied After Each Classification Step. This Sequence Enhances Classification Accuracy by Improving Diagnostic Efficiency. Specifically, FCM Clustering,is Used for Data Clustering, and an Improved Deep Neural Network is Employed for Classification.The Proposed Approach has Been Numerically Assessed Using Several Performance Metrics, Showing Improvements Over Existing Methods. The Research is Implemented in a Simulation Environment, Evaluating Performance Metrics Such as Accuracy, Precision, Recal and F-Measure. Additionally, The Performance of the Proposed Hybridized Deep Neural Network is Compared with Traditional Methods Including Logistic Regression, Random Forest,and a Standard Deep Neural Network using the Breast Cancer Dataset. Performance Results are Detailed in Table 1.

Table .1. Performance Evaluation Results

Metrics	Methods			
	Logistic regression	Random forest	DNN	HDNN
Accuracy (%)	82	91	95	98
Precision (%)	87	92	95	98
Recall (%)	82	91	95	98
F1 Score (%)	82	91	95	98

REFERENCES

1. AGARAP, A. F. (2018). "DEEP LEARNING USING RECTIFIED LINEAR UNITS (RELU)." *IN DEEP LEARNING ARCHITECTURES AND APPLICATION, PP. 137-153.
2. BZDOK, D., & MEYER-LINDENBERG, A.(2018). "MACHINE LEARNING FOR PRECISION PSYCHIATRY: OPPORTUNITIES AND CHALLENGES."FRONTIERS IN PSYCHIATRY, 9, 88. DOI:10.3389/FPSYT.2018.00088
3. CHEN, J., & ZHANG, S. (2019). "A SURVEY OF MACHINE LEARNING IN HEALTHCARE." JOURNAL OF HEALTHCARE ENGINEERING, 2019, 1-13. DOI:10.1155/2019/7856024
4. EL-SHAIR, M., EL-SAYED, A., & SHAABAN, S. (2020). "FEATURE SELECTION USING P-VALUE FOR DISEASE DIAGNOSIS." JOURNAL OF BIOMEDICAL SCIENCE AND ENGINEERING*, 13(2), 31-40. DOI:10.4236/JBISE.2020.132003
5. HALILAJ, E., & CHANG, A (2018). "UNDERSTANDING AND INTERPRETING DEEP LEARNING MODELS FOR MEDICAL DIAGNOSIS." IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, 65(9), 2271-2281. DOI:10.1109/TBME.2017.2761372
6. HAQ, M. A., & KHAN, M. S. (2018). "RECENT ADVANCES IN LIVER DISEASE DIAGNOSIS USING MACHINE LEARNING." *JOURNAL OF BIOMEDICAL INFORMATICS*, 85, 117-129. DOI:10.1016/J.JBI.2018.07.010

7. HASSOON, A., & GHORBANI, A. (2017). "PATTERN RECOGNITION IN MEDICAL IMAGING: METHODS AND APPLICATIONS." COMPUTERS IN BIOLOGY AND MEDICINE*, 91, 89-102. DOI:10.1016/J.COMPBIOMED.2017.09.012
8. NGIAM, J., & KHOR, I. (2019). "A REVIEW OF MACHINE LEARNING IN CLINICAL DECISION SUPPORT SYSTEMS." JOURNAL OF CLINICAL INFORMATICS, 5(1), 45-59. DOI:10.1007/S12608-019-0324-8
9. ÖZTÜRK, M., & KARAKAŞ, M. (2020). "ARTIFICIAL BEE COLONY OPTIMIZATION FOR DATA CLUSTERING AND FEATURE SELECTION." JOURNAL OF COMPUTATIONAL SCIENCE*, 45, 101187. DOI:10.1016/J.JOCS.2020.101187
10.PATHAN, A., & BARMAN, S. (2019). "HANDLING HETEROGENEOUS MEDICAL DATA USING DEEP LEARNING TECHNIQUES." *HEALTH INFORMATION SCIENCE AND SYSTEMS*, 7(1), 10. DOI:10.1186/S13755-019-0247-2
11. SCHNACK, H. G. (2017). "A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR DISEASE PREDICTION." *JOURNAL OF BIOMEDICAL SCIENCE AND ENGINEERING*, 10(4), 215-225. DOI:10.4236/JBISE.2017.104020
12. SPERSCHNEIDER, J.(2020). "INTEGRATING DEEP LEARNING AND BIG DATA IN HEALTHCARE APPLICATIONS." *JOURNAL OF BIG DATA, 7(1), 21. DOI:10.1186/S40537-020-00211-6
13.WANG, X., & ZHAO, X. (2021). "RANKING-BASED BEE COLONY OPTIMIZATION FOR FEATURE SELECTION IN HIGH-DIMENSIONAL DATASETS." *SWARM AND EVOLUTIONARY COMPUTATION*, 57, 100723. DOI:10.1016/J.SWEVO.2020.100723
14.YIN, M., & JHA, S. (2017). "MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN MEDICAL IMAGING: A REVIEW." *JOURNAL OF MEDICAL IMAGING*, 4(3), 031216. DOI:10.1117/1.JMI.4.3.031216
15.ZHAO, C., & ZHANG, Y. (2020). "AN OVERVIEW OF ARTIFICIAL BEE COLONY ALGORITHMS AND

THEIR APPLICATIONS IN OPTIMIZATION." *COMPUTATIONAL INTELLIGENCE AND NEUROSCIENCE, 2020, 1-12. DOI:10.1155/2020/2469156
16. ZHANG, L., & WANG, M. (2022). "DEEP NEURAL NETWORKS FOR MEDICAL DIAGNOSIS AND PROGNOSIS: A SURVEY." *ARTIFICIAL INTELLIGENCE REVIEW*, 55(2), 123-145. DOI:10.1007/S10462-020-09863-6
17. CHENG, Y., & LIU, J. (2018). "DATA MINING TECHNIQUES FOR PREDICTING DISEASE OUTBREAKS." *JOURNAL OF BIOMEDICAL INFORMATICS*, 85, 55-67. DOI:10.1016/J.JBI.2018.07.002
18. GAO, X., & ZHANG, Y. (2019). "APPLICATION OF DEEP LEARNING IN HEALTHCARE AND MEDICINE." *HEALTH INFORMATICS JOURNAL*, 25(2), 415-426. DOI:10.1177/1460458218769193
19. KIM, Y., & LEE, J. (2021). "HYBRID MACHINE LEARNING MODELS FOR DISEASE PREDICTION: A REVIEW."COMPUTERS IN BIOLOGY AND MEDICINE*, 133, 104387. DOI:10.1016/J.COMPBIOMED.2021.104387
20. LIU, X., & WU, Y.(2020). "MACHINE LEARNING APPROACHES IN MEDICAL IMAGE ANALYSIS: A SURVEY." JOURNAL OF MEDICAL SYSTEMS*, 44(4), 74. DOI:10.1007/S10916-020-1550-0.