IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

An Efficient Framework Using Learning Algorithm For Breast Cancer Diagnostics

P. Senthil Kumar[1], G. Buvaanyaa[2] and Dr. M. Gobi[3]

[1]Research Scholar, Ph.D. Category - B, R & D Centre
Bharathiar University, Coimbatore, Tamilnadu, India

[2]PhD Research Scholar, Department of computer Science
Chikkanna Government Arts College, Tiruppur, Tamilnadu, India

[3]Associate Professor
Department of Computer Science
Chikkannna Government Arts College, Tirupur, Tamilnadu, India

Abstract- Breast cancer affects more women than any other form of cancers. Breast cancer is diagnosed mostly by mammography. Medical data from CT scans, PET scans, and MRIs are among the most widely used types of information. The use of Deep learning and Machine Learning approaches has become essential for efficient and precise cancer prediction and detection since the work of analyzing this massive amount of data has gotten increasingly difficult. Clinically relevant information can be mined from medical photographs to better aid in illness diagnosis and early detection, which is the primary focus of medical image mining. Patients need careful symptom observation and a prediction automatic system that can identify the tumour as benign or malignant in order to receive effective treatment. While its primary function as a generic neural network based on convolution is to classify images with an image as input and a single label as output, in biomedical applications, it may also identify disease and locate its exact location. This issue can be solved by employing deep learning methods. A Deep learning with hybrid-based framework is suggested for tumour zone segmentation and prediction in mammography images. To assess the stage of breast cancer, the model uses a fully hybrid method that has been updated and broadened in design to operate with fewer training images and deliver more accurate tumour height and width segmentations.

Keywords- Accuracy, Gene expression analysis, Breast Cancer, Deep Learning, Classification, Prediction.

I. INTRODUCTION

When compared to other female-specific malignancies, breast cancer is the worst. Mammogram is the major diagnostic method for detecting breast cancer throughout the testing process. Breast cancer is currently listed as the 25th biggest cause of deaths globally [1]. About 50,000 women in India are diagnosed with this malignancy each year. Mammograms can detect three primary warning indicators indicative of cancer: masses, calcification, and architectural distortion. Therefore, it becomes even more crucial to discover and diagnose this malignancy as early as possible. Early detection of breast cancer is essential to increase the chance of a successful treatment outcome. In high-risk females, breast MRI has a high detection rate for even the smallest of cancer tumours. With a sensitivity of 97%, breast tumours are more accurately diagnosed. Early diagnosis can reduce the difficulty and severity associated with an illness. Screening can discover the disease at an early stage, before symptoms occur. The data mining methods of "clustering" and "classification" are particularly well-suited to the task of analyzing breast cancer photos. [2] The images are clustered into different disease categories using clustering, an unsupervised classification method. Choosing an efficient technique that is suited to the particular requirements of the task at issue is essential in medical diagnosis.

Medical data from CT scans, PET scans, and MRIs are among the most widely used types of information. The use of DM approaches has become essential for efficient and precise cancer prediction and detection since the work of analyzing this massive amount of data has gotten increasingly difficult. The primary purpose of medical image mining is to aid in the diagnosis and early identification of disease by extracting clinically useful information from medical images. Patients need careful symptom observation and a prediction automatic system that can identify the tumour as benign or malignant in order to receive effective treatment. In biological applications, convolutional neural networks may identify disease and determine its location, in addition to their basic role of image classification. Deep learning approaches can effectively address this issue.

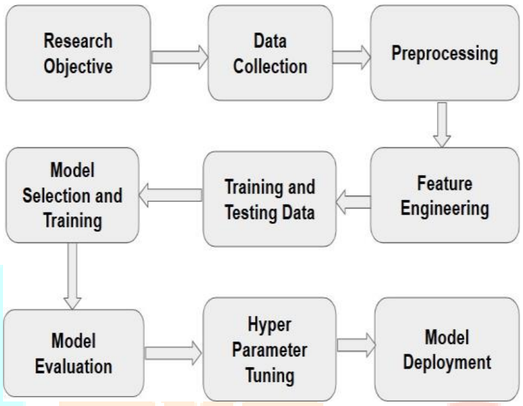


Fig. 1. Working Process

The evaluation of gene expression levels and the identification of associated diseases are essential components in the medical field. Accurately diagnosing the stage of breast cancer is vital, as it facilitates additional testing to ascertain whether the cancer has metastasized beyond the breast and to identify potential treatment options that may be effective. Fig 1 is explaining the details about working process.

The goals of this research are outlined as follows:

- To conduct an in-depth analysis of gene expression, the mechanisms of breast cancer, the use of machine learning algorithms, and the significance of accurately identifying breast cancer.
- To examine the current techniques utilized for determining the stages of breast cancer through the analysis of gene expression values.
- To pinpoint ideas that can enhance the efficiency of detecting breast cancer-related genes.
- To create and implement a novel algorithm derived from the recognized concepts to find out the various stages of breast cancer, along with the development of a software prototype.
- To evaluate the model within a healthcare setting.

II. MACHINE LEARNING & DEEP LEARNING IN MEDICAL IMAGE PROCESSING

In the previous few decades, there has been significant growth in creating sophisticated algorithms and effective preprocessing approaches in ML and DL. [1] Advancements in neural networks have led to the development of deep neural networks. In these circumstances, machine learning and deep learning have produced ways for more precisely diagnosing sickness in its early stages, reducing the frequency of readmissions in clinics and hospitals.

Deep learning tackles a broad range of issues in healthcare, such as personalized therapy recommendations, infection monitoring, and cancer detection. Thus, the adoption of artificial intelligence (AI) tools can facilitate the acquisition of new fidelity procedures and lower the expense of healthcare resulting from inaccurate diagnoses. In the field of medical imaging, DL has achieved tremendous progress, attaining remarkable outcomes in several tasks. There is still an obstacle in the form of the restricted availability of training information, especially in the healthcare domain where obtaining data can be expensive and governed by privacy laws [2]. Image mining, computer vision, and pattern recognition have all become more important aspects of medical image processing in shows figure 2.

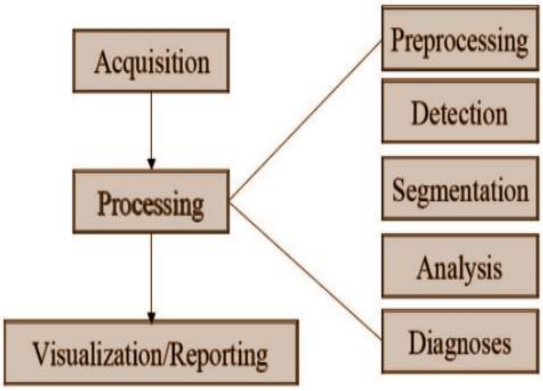


Fig. 2. ML types

III. BIOINFORMATICS

Bioinformatics represents a scientific discipline focused on the development of methodologies and software tools aimed at interpreting biological data. This field integrates various domains, including statistics, computer science, engineering and mathematics, to explore and derive insights from biological information. Techniques in bioinformatics, such as image and signal processing, facilitate the extraction of meaningful results from extensive datasets [3]. Within genetics and genomics, bioinformatics is instrumental in the sequencing and annotation of genomes, as well as in identifying mutations. It is crucial for analyzing and regulating protein and gene expression. Furthermore, bioinformatics tools enable the comparison of genetic and genomic datasets to uncover evolutionary patterns in molecular biology, thereby aiding in the ~~examination of networks and biological pathways, which are essential components of systematic biology. In~~ the field of structural biology, bioinformatics adds to the modeling and simulation of proteins, DNA, RNA, and the interactions between them, ultimately integrating information from biology to offer an in-depth comprehension of these relationships [4]. As a result, bioinformatics is an important field for the analysis and interpretation of a variety of data, such as nucleotides and sequences of amino acids, protein domains, and structures of proteins, and the total process of evaluating genetic data is known as biological computation.

Important subfields of informatics and biological computation include designing and implementing computer- aided techniques that efficiently extract, process, and manage different data types. Another focus is on creating new algorithms and statistical methods for analyzing interactions between different data types. The aim of bioinformatics is to gain better insights into processes in biology by developing computer methods that help. This includes pattern recognition, data mining, machine learning algorithms, and data visualization. The future of this field is likely to include sequence alignment, gene identification, genome assembly, pharmaceutical development and exploration, structure of proteins alignment, amino acid's structure prediction, analysis of gene expression, interactions between proteins, genome-wide association research, and modeling processes associated with evolution and division of cells, specifically mitosis.

IV. RELATED WORKS

Elbashir [5] presented a method of lightweight CNN architecture for breast cancer prediction. This method pre- processes gene expression data and transforms it into a 2D image. Then, the outlier removal was done using the Array- Array Intensity Correlation (AAIC) technique, and CNN was used for the classification process. By using the RNA-seq gene expression data, F-Score, Accuracy, Precision, sensitivity and specificity of 0.955, 98.76%, 100%, 91.43% and 100%, respectively, were obtained. However, applying CNNs to gene expression data increased the computational demands.

Jazayeri and Sajedi [6] proposed a Non-negative Matrix Factorization (NMF) and an Extreme Learning Machine (ELM) algorithm for classifying breast cancer. This method combined NMF with column splitting for dimension reduction, and ELM was used for the classification process. Experimented on the NCBI dataset, this model reduced the classification error rate by 2.7%, but it has problems handling feature redundancy, noises and irrelevant data.

Arya and Saha [7] suggested a two-stage stacked ensemble framework for predicting breast cancer, with CNN used for extracting the features in the first stage and a stacked ensemble model using these features for final classification in the second stage. Tested on a multi-model dataset and obtained a 90.2% accuracy and 0.93 AUC value. However, the CNN used in this model increased the complexity when stacked as an ensemble.

Lamba [8] presented a DNN-based classification for cancer in the breast. In this method, minority class balancing was performed using the SMOTE algorithm and BFS Best First Search was used for the selection of features and CFS before classifying using DNN. This model achieved 93% accuracy for GSE15852 datasets but has also suffered from overfitting issues due to a smaller sample size.

Cheng [9] developed a DNN-based breast cancer detection model and combined ensemble learning with Systems biology feature selection methods. This model obtained AUC values of 0.7677 and 0.7836 between genes and clinical features and a concordance index (CI) of 0.6683 for the METABRIC dataset.

Liu [10] proposed a hybrid DNN for predicting breast cancer based on multi-modal data that combines the gene model data with the image model data. The feature extraction network works based on weighted linear aggregation to improve the DNN performance in this method. This hybrid model obtained 88.07% accuracy for the TCGA-BRCA dataset but suffers from a high processing time of 40 minutes.

Mustafa [11] presented an ensemble model using multi-modal data and multiple neural networks for breast cancer survivability prediction. Here, CNN is used for clinical modalities. To handle data in multi-dimensional data and modalities in gene expression, LSTM is utilized and DNN is used for CNV effectively. This model obtained 98% accuracy, 99% F1-score, 98% precision, and 100% sensitivity for the METABRIC dataset, but the memory complexity is higher than other DL-based methods.

Comparing their performance using these results will be unfair since a method can work better for a dataset while underperforming for another dataset. The smaller sample size and high dimensionality of the gene expression datasets have significantly reduced the performance of ML and DL methods. Similarly, the complexity issues in DL-based methods are also a challenging concern.

V. METHODS AND MATERIALS

Machine learning is the core of computer aided diagnosis. Data analytics techniques such as machine learning teach computers to do what humans and animals naturally do: gain knowledge through experience. Algorithms that use machine learning learn information by analyzing data instead of relying on a predetermined equation as a model. With machine learning, computers are intended to learn automatically and adjust their actions accordingly without any human assistance. The ML types showing in the figure 3.

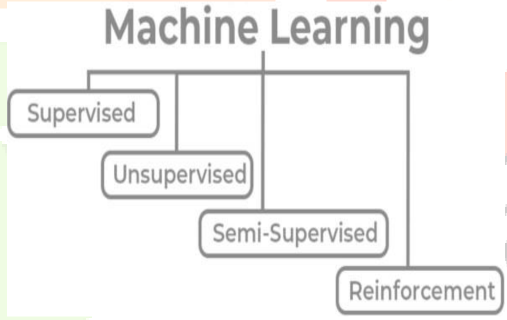


Fig. 3. ML types

A. Supervised Learning

Models are trained using supervised learning when they are given labelled data. This type of dataset consists of both inputs and outputs. Data is generally divided into 80:20 proportions in order to train the model - 80% for training, 20% for testing. Input data comes from training. Training data is the only source of model learning. In other words, the model is learning to build logic on its own. [13]

A model is ready for testing once it has been developed. The model will predict some value, which will be compared with actual output and the accuracy will be calculated. Data from the remaining 20% is fed into the test and the model has never seen it before. In general, supervised learning can be divided into two types.

Supervised learning is used in categorization to provide results that may be labelled in a strict manner (discrete values). Although the results of supervised learning are often presented in discrete form, the process itself may be seen as continuous. The following are a few examples of supervised learning in action: Random Forest, Gaussian Naive Bayes and Decision Trees.

B. Unsupervised Learning

This method consists of training a model on unclassified, unlabeled data and allowing it to work without explicit guidance. An unsorted set of data is grouped according to its similarity, patterns, and differences devoid of any prior data training. There are two possible methods to break down the topic of unsupervised learning: Cluster analysis is a method of organizing large amounts of data into manageable chunks based on their similar properties, such as the identification of repeat customers based on their purchasing patterns. An association rule learning problem consists of identifying rules that describe large parts of the data you provide, such as those that describe people who buy X also buy Y in addition.

C. Reinforcement Learning

In this method, actions are produced and errors or rewards are discovered through interacting with the environment. Reinforcement learning involves trial-and-error search and delayed rewards. Machines and software agents can automatically analyze a specific context to determine the best possible behavior to optimize their performance. [12] The reinforcement signal in the form of simple reward feedback that the agent needs to learn what is the best action. Here's a quick demonstration of how it's done.

There are two types of reinforcements - Positive & Negative

A Positive Reinforcement occurs when a particular behavior is reinforced by an event that encourages the behavior to become stronger and more frequent. Therefore, the behavior is positively affected.

Reinforcement learning has the following advantages:

- Increases performance
- Long-term sustainability

D. Deep Learning

Machine learning that is completely based on neural networks is called deep learning, and since neural networks mimic the human brain, it can be looked at as mimicking the human brain. Deep learning does not require explicit programming. Deep learning is not a new concept. [14] The company has been around for quite some time now. Today it's all the rage because earlier there wasn't that much processing power and data available. This exponential growth in computing power since the 1990s has allowed for the advent of DL and ML. The ML and DL process is illustrated in figure 4.

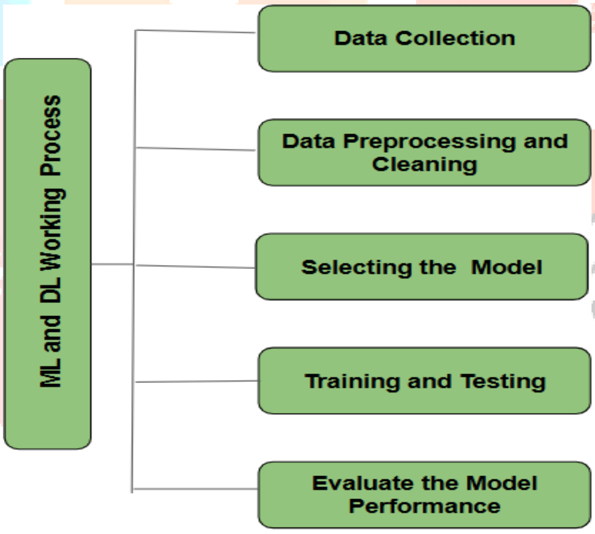

ML and DL Working Process
Data Collection
Data Preprocessing and Cleaning
Selecting the Model
Training and Testing
Evaluate the Model Performance

Fig. 4. ML process

The many DL architectures are listed below.

- The deep neural network typically contains multiple hidden layers (e.g., there are multiple input layers and hidden output layers. They are able to model and comprehend non-linear relationships.
- In the Deep Neural Networks class, Deep Belief Networks (DBNs) are included. Multi- layered belief networks exist. The following steps are required to build a Deep Belief Network
- A DBN is performed as follows:
- The Contrastive Divergence algorithm is applied to study a layer of features from visible units.
- The features of previously trained features are studied by considering their activations as visible units.
- After all hidden layers have been learned, the entire deep learning network is trained.
- Recurrent (conduct the same calculation for each element within a sequence) Neural Networks – Support parallel and sequential processing. The human brain (a large network of neurons connected

by feedback). They can remember important aspects of the input received, which makes them better at interpreting it.

E. Dataset Preparation:

The purpose of this study was to figure out whether breast cancer recurs. The BC-TCGA which contains totally 17814 genes is taken from the Mendeley dataset for the evaluation of the algorithm for breast cancer prediction. This data sets contains the samples of 590 different patients with 61 normal and 529 breast cancer patients and 17,814 genes for each sample.

F. Data Pre-processing:

In this step, unwanted data is removed such as features with missing values, data encoding techniques are applied in order to convert categorical variables into numeric data. Finally, each sample is labelled with Breast Cancer subtype based on presence or absence of particular hormone values.

G. ML Model Preparation & Training:

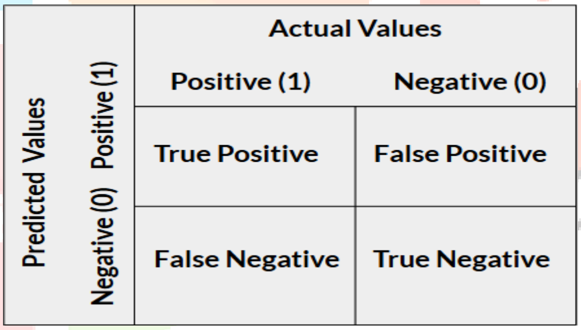
The dataset used in the research study is divided into training and testing dataset. Various ML and DL algorithms are applied on the dataset and performance is observed.

H. Result Validation & Explanation:

Obtained results are validated through Explainable AI, deployed application and suggestions taken from medical professionals.

VI. RESULTS AND DISCUSSION

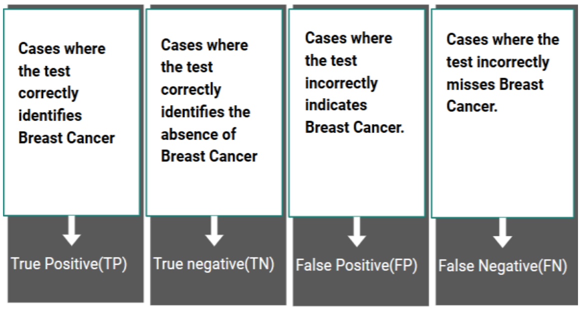
In the context of breast cancer prediction, TP, TN, FP, and FN are terms used to evaluate the performance of a classification model. These metrics are crucial for understanding the performance of a breast cancer prediction model. Several performance metrics, including sensitivity, specificity, and accuracy dependent on the confusion matrix (figures 5 and 6), are used to confirm the methodology's efficacy.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive	False Positive
	Negative (0)	False Negative	True Negative

Fig. 5. CM for results

The confusion matrix's components are as follows:



Cases where the test correctly identifies Breast Cancer → True Positive(TP)

Cases where the test correctly identifies the absence of Breast Cancer → True negative(TN)

Cases where the test incorrectly indicates Breast Cancer. → False Positive(FP)

Cases where the test incorrectly misses Breast Cancer. → False Negative(FN)

Fig. 6. Results Metrics

These metrics are crucial for understanding the performance of a breast cancer prediction model.

Now, will employ various classification algorithms for organizing these images with malignant or benign (normal or cancerous) categories. This work clearly states that,

Accuracy: The whole precision of the strategy estimates is measured by precision. It measures how many examples in the database were properly identified including true positives and true negatives.

Precision: Precision, also called as positive prognostic value, is the capability of the strategy for properly classifying the positive models (malignant cases). It is described as the relation of true positives to the entire predicted positive models.

High precision value refers that the model has a less rate of falsely classifying benign cases as malignant, which is crucial for avoiding unnecessary medical interventions

Recall: The capacity of the machine learning algorithm to correctly differentiate false results are positive (malignant cases) and the overall number of real positive specimens is measured by recall, which is referred to as sensitively or true optimistic rate. It is the proportion of real positive samples (true positives plus false negatives) to all actual positive samples.

F1- Score: The F1 score (also called the F-measure) is a performance metric used to evaluate a classification model, especially when the data is imbalanced (i.e., one class appears much more frequently than another).

Higher recall indicates that the developed model has a low rate of falsely classifying malignant cases as benign, which is critical for detecting all true positive cases.

Feedforward Neural Networks (FNN): The simplest form of neural networks where information moves in one direction - from input to output.

Graph Neural Networks (GNNs): Extend deep learning to graph-structured data (nodes + edges).

Hybrid RF-SVM classification Algorithm (HRFSVM): is a smart learning method that blends the simplicity of RBF (focusing on a particular spot in data) with extra tools for finding disease, ensuring faster and learning smarter.

Hybrid NB-KNN classification Algorithm (HNBKNN): An optimized approach in machine learning where specialized "experts" (models) each focus on finding a disease prediction, and their decisions are combined intelligently to get the perfect result.

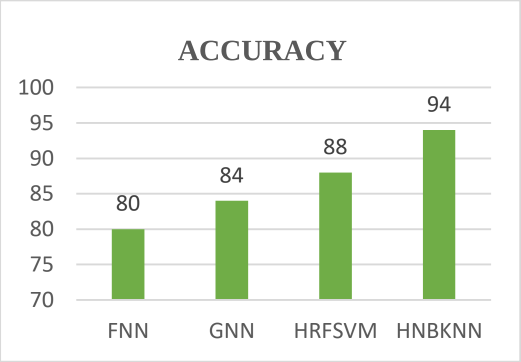


Fig. 7. Accuracy for Breast Cancer

HNBKNN algorithm gives 94 percentage accuracy. This value is higher than other methods results.

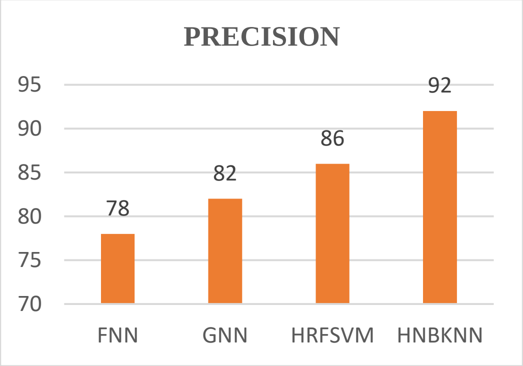


Fig. 8. Precision for Breast Cancer

HNBKNN algorithm gives 92 percentage precision. This value is higher than other methods results.

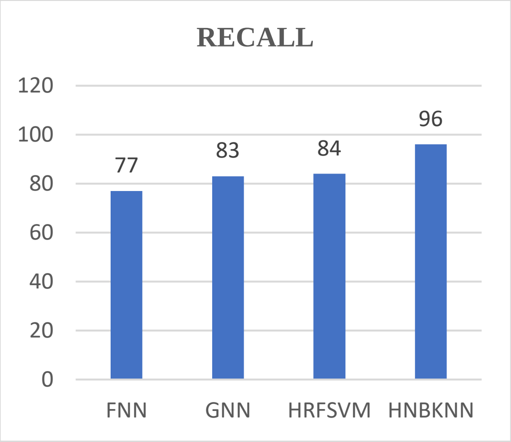


Fig. 9. Recall for Breast Cancer

HNBKNN algorithm gives 96 percentage recall. This value is higher than other methods results.

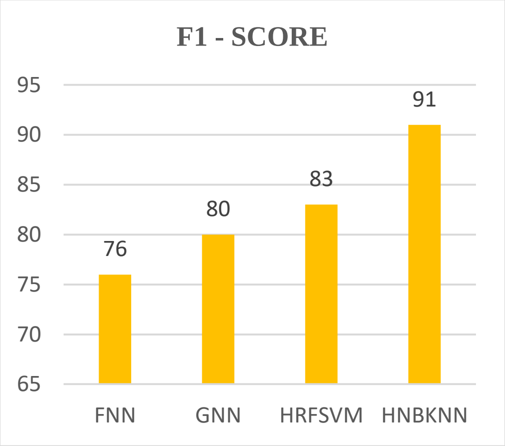


Fig. 10. F1-Score for Breast Cancer

HNBKNN algorithm gives 91 percentage f1-score. This value is higher than other methods results.

The results of accuracy display figure 7, The results of Precision displays figure 8, the results of Recall display figure 9 and the results of f1-score display figure 10. This research focused compare ML and DL techniques. Comparing ML and DL methods, hybrid method provides the good results.

VII. CONCLUSION AND FUTURE SCOPE

In this research work applied different ML and DL methods. The output of accuracy, recall, precision and f1-score results shows in the figures. This research focused compare ML and DL techniques. While comparing ML and DL methods, HNBKNN provides the good results. So, the proposed method is HNBKNN. The research work aims to contribute to society by proposing personalized, still affordable treatment options for breast cancer patients, which will surely reduce the number of deaths due to adverse effects after treatment, long waiting for the test results, etc. The increasing cases of deaths due to breast cancer in women shows the need of precision medicine approach for the treatment. Collaborating with medical product development industries and medical research institutes for creating cost-effective test kits for predicting breast cancer disease at an early stage shall be the targeted future for the extension of this research work. Moreover, the development of a Deep learning-based interface streamlines the collection of patient data, enabling the determination of molecular classification and subsequent treatment plan recommendations. Future directions may involve expanding the dataset, incorporating additional features, refining models, and conducting prospective studies to validate the system's recommendations in clinical practice. Treatment plans may vary from patient to patient.

REFERENCES

[1] Revathi.K, Karthikeyan.V.V, Priyanka.S, and Prakash.S.J, "Unveiling Genetic Disorders: Machine Learning and Deep Learning Approaches in Gene Expression Analysis", 2nd International Conference on Intelligent Cyber Physical Systems and Internet of Things, 2024, pp. 1315–13.

[2] Devi, N.R., Revathi, K., Lekhaa, T.R. and et al, "Medical Imaging Analysis Using Machine Learning By Design Thinking Approach", 4th IEEE International Conference on Smart Technologies in Computing, Electrical and Electronics, ICSTCEE 2023.

[3] D Vijaya Shree, BL Shiva Kumar, and et al, "The Role of Artificial Intelligence and Machine Learning Methodologies in Bioinformatics", 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Pages, 194-201, 2024.

[4] Shree. D. V, Shiva Kumar. B.L, Ganesan. V, Sivaraman. M and Sumitha. J, "Biological Data Analysis for Disease Prediction and Classification in Bioinformatics", 8th International Conference on Inventive Systems and Control, ICISC 2024, pp. 435–440, 2024.

[5] M.K. Elbashir, M. Ezz, M. Mohammed, and S. S. Saloum,"Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data". IEEE Access, vol. 7, pp. 185338- 185348, December 2019.

[6] N. Jazayeri, and H. Sajedi, "Breast cancer diagnosis based on genomic data and extreme learning machine". SN Applied Sciences, vol. 2, pp. 1-7, December 2020.

[7] N. Arya, and S. Saha, "Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model". IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19(2), pp. 1032-1041, August 2020.

[8] M. Lamba, G. Munjal, and Y. Gigras, "A hybrid gene selection model for molecular breast cancer classification using a deep neural network". International Journal of Applied Pattern Recognition, vol. 6(3), pp. 195-216, August 2021.

[9] L. H. Cheng, T. C. Hsu, and C. C. Lin, "Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction". Scientific Reports, vol. 11(1), pp. 14914, July 2021.

[10] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng, "A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multi-modal data". Irbm, vol. 43(1), pp. 62-74, Feburary 2022.

[11] E. Mustafa, E. K. Jadoon, S. Khaliq-uz-Zaman, M. A. Humayun, and M. Maray, "An Ensembled Framework for Human Breast Cancer Survivability Prediction Using Deep Learning". Diagnostics, vol. 13(10), pp. 1688, May 2023.

[12] Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika and et al, "Machine Learning Algorithms For Breast Cancer Prediction and Diagnosis", Procedia Computer Science, Volume 191, 2021, Pages 487-492.

[13] Sajib Kabiraj, Laboni Akter, M. Raihan and et al, "Prediction of Recurrence and Non-recurrence Events of Breast Cancer using Bagging Algorithm, " 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020.

[14] Irmawati, Ferda Ernawan, Mohammad Fakhreldin and Andi Saryoko, "Deep Learning Method Based for Breast Cancer Classification", International Conference on Information Technology Research and Innovation (ICITRI), 2023.

[15] G. Buvaanyaa, M. Gobi, “Crop Recommendation System Using Hybrid Classification Algorithm”, African Journal of Biological sciences”, Volume 6, Issue 14, Aug 2024, doi: 10.48047/AFJBS.6.14.2024.8913-8918.

[16] Dr. M. Gobi, R. Sridevi, “ECC Encryption and LSB Data Embedding Technique for Message Security”, Int. J. for Res. In Technological Studies, Vol.2, Issue 7, June 2015.

[17] M. Gobi, G. Buvaanyaa, “An Efficient Naïve Bayes Imputation Method for Missing Values”, IRJMETS, Vol.2, Issue 7, July-2020.

[18] G. Siva Brindha, Dr. M. Gobi, “Improving the Map Reduce Performance using Symmetric Key Algorithm”, International Journal of Science and Research (IJSR), Vol. 10, Issue 4, April 2021.

[19] Dr. S. Selvi, Dr. M. Gobi, “An Efficient Data Security Model using Hyper Elliptic Curve Cryptography and Stenography”, Int. J. Research and Development in Technology, Vol. 7, Issue 6, June – 2017.

[20] R. S. Vindan, M. Gobi, “Pinning based Energy aware Computation Offloading for Mobile Cloud Computing”, First Int. Con. Computational Science and Technology (ICCST), 2022.