Journal of Information Systems Engineering and Management
2025, 10(26s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Enhancing Early Detection and Prognosis of Breast Cancer Through Advanced Machine Learning Techniques A Comprehensive Predictive Modeling Approach

Lakshmana Rao Rowthu[1], Dr. V. Sangeeta[2], Dr Malijeddi Murali[3]

**Research Scholar, Department of Computer Science and Engineering, Centurion University of Technology and Management, Vizianagaram, Andhra Pradesh.*

***Associate Professor, Department of Computer Science and Engineering, GITAM School of Technology, GITAM Deemed to be University, Visakhapatnam -530045*

****Professor, Department of Electronics and Communication Engineering, ACE Engineering College, Hyderabad, Telangana.*

**Corresponding Author - laxman.rowthu@gmail.com*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 18 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

This study makes use of a publicly available dataset for breast tumor prediction in order to enhance early identification and evaluation via machine learning. Using Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Artificial Neural Networks (ANN), we create and assess predictive models for precise breast cancer classification. These models are trained, verified, and tested using the dataset's primary clinical and diagnostic characteristics. The efficacy of each strategy is compared through the analysis of performance indicators like accuracy, precision, recall, and F1-score. Superior results from CNN and ANN showed how machine learning has a lot of potential for accurate breast cancer diagnosis. This study emphasizes how crucial predictive modeling is to improving diagnostic precision and assisting physicians in making better decisions for their patients.

Keywords: Machine Learning, Breast Cancer, Medical Diagnosis

1. INTRODUCTION

Many of the greatest common and deadly illnesses in the entire globe today are breast carcinoma, and increasing chances of survival requires detection at an early stage. The use of predictive modeling has been transformed by machine learning Improvements making it possible to identify malignancies with astonishing accuracy. The research being conducted creates reliable models for carcinoma of the breast prognosis using a publicly accessible dataset to aid in early identification and individualized treatment. The classification of tumors in breasts is improved by machine learning methods such as Random Forest, Support Vector Machine (SVM), Logistic Regression, and Gradient Boosting. This study illustrates the revolutionary importance of technology in medical diagnostics by exploiting important clinical features along with diagnostic features, opening the door for creative solutions to this global health issue. One form of artificial intelligence that demonstrates a lot of progress in resolving challenging medical challenges is computational intelligence. Due to its promise for precise prognostic and diagnostic information, statistical modeling is gaining traction in cancer research, especially for breast cancer. This work's main goal is to apply sophisticated algorithms to the problem of distinguishing between tumors that are benign and malignant. Machine learning algorithms offer insights into the complex patterns linked to breast cancer by examining structured data that comprises significant attributes. Making use of a publicly accessible dataset guarantees that the results have scalability and repeatability while providing a practical viewpoint. This strategy highlights how important it is to apply data-driven approaches to regular medical procedures in order to enhance results. Although breast cancer has long been detected by imaging and biopsy, these procedures are often uncomfortable as well as laborious. With automated, non-invasive models for predictions, machine learning offers an opportunity to improve traditional methods for diagnosis. Important features required to train algorithms that can distinguish between aggressive and benign malignancies are included in the study's information. To find ways to improve the medical processes, this study will examine the possibilities of technologies including Random

Forest, SVM, Logistic Regression, and Gradient Boosting. If immediate intervention is taken, the findings could greatly enhance patient care by reducing medical mistakes and boosting the quality of care. In predictive computing, every training method has its own set of benefits and drawbacks. Whereas SVM is renowned for its efficacy in classifying assignments with distinct edges, Random Forest excels at handling high-dimensional data and guards against overfitting. For classifying binary data, logistic regression is still the standard technique, whereas gradient enhancement improves performance by iteratively improving estimates. The present investigation offers a thorough assessment of the aforementioned algorithms' efficacy in predicting cancer in breast tissue by examining them. The integration of many methodologies guarantees a comprehensive comprehension of their potential, aiding in the identification of the most dependable approaches to clinical application. The developing discipline of precision healing would benefit greatly from such a study. During the training, testing, and validation of these mathematical simulations, we can gain a better understanding of the key factors influencing the diagnostic reliability. By examining outcomes including accuracy, precision, recall, and F1-score, the effectiveness of each method is assessed. The models' benefits and drawbacks are displayed in this empirical investigation, along with potential areas for further development. Clinical professionals and researchers wishing to employ algorithmic learning in the detection of carcinoma of the breast may find valuable information in the findings, which emphasize the importance underlying informed by data healthcare decisions. The study will systematically address these issues to produce reliable and widely applicable results. The integration of different methods gives further confidence that this investigation addresses a wide spectrum of prediction skills. A comprehensive approach is required to develop diagnostic tools that may be seamlessly integrated into clinical processes in order to increase the usability and effectiveness of cancer identification systems. The algorithms that were chosen have a strong track record and balance complexity and effectiveness. By focusing on understandable and explicable models, the studies help clinicians understand and trust the algorithms' predictions. Instead of completely replacing human experience during healthcare making decisions, this approach complements the larger goal of technological innovation to enhance it by encouraging collaborative processes for improving healthcare for patients.

2. LITERATURE REVIEW

Guyon et. al [16] In their in-depth examination of feature selection techniques, they emphasized the importance of these techniques in improving the usability and functionality of models. They provided details on the advantages and disadvantages of various filtering, wrapper, and embedding strategies. The research stressed the significance of finding features in managing data redundancy,lowering dimensionality, and enhancing model applicability. Their study served as a foundational guide for those practitioners hoping to successfully apply selected features in a range of domains, including machine learning along with information mining. They discussed metrics for evaluating feature importance and outlined future challenges such as accessibility and choosing in changeable datasets, paving the way for advancements in variable selection methods for difficult prediction tasks.

Chen et. al [17] The technique's reliability and efficacy were enhanced by the periodicity, which comprised scheduling as well as sparsity-aware minimization. They demonstrated XGBoost's dominance in a variety of tasks involving machine learning using large datasets. The system achieved state-of-the-art results in several Kaggle competitions by utilizing the second order variations and handling data that was lacking efficiently.

In order to increase the accuracy of machine learning predictions, Zhou et al. [18] looked at ensemble techniques. Important tactics including bagging, boosting, and layering were discussed in the book, with an emphasis on both their theoretical underpinnings and practical applications. emphasized the benefits of ensemble learning, particularly its ability to improve modeling robustness without reducing variation and excessive overfitting. In order to illustrate how collective methods outperform individual methods in a range of scenarios, including financial projections and medical diagnostics, practical problems situations have also been integrated into the text.

Kohavi et al. [19] conducted a comprehensive examination of bootstrap and cross-validation methods for model selection and accuracy estimations. This paper emphasized the importance of robust validation techniques in ensuring accurate performance measurements. By looking at the trade-offs between bias, variance, and processing cost, Kohavi demonstrated how k-fold cross-validation can be used in a range of scenarios. The study also highlighted the limitations of leave-one-out cross-validation and bootstrap in specific contexts. Through empirical experiments, the study provided helpful guidance on how to select validation methods based on dataset properties.

In order to improve the accuracy of early breast cancer detection, Shen et. al [20] developed a computer-aided diagnosis technique. Their research used imaging data with machine learning techniques to determine whether the tumors were benign or malignant. The system achieved exceptional diagnostic precision while reducing the requirement for manual interpretation through the use of advanced classification algorithms and feature extraction. The study employed strategies to increase model robustness in addition to tackling problems like feature selection and data imbalance.

A thorough manual on pattern recognition and machine learning was written by Bishop et al. [21] connecting abstract ideas with real-world uses. The text included fundamental subjects including neural networks, Bayesian networks, and statistical modeling. Bishop provided mathematical underpinnings and algorithmic insights while highlighting the significance of generation and unfair models.

Dietterich et al. [22] examined machine learning aggregation approaches, including information on how well they work to lower error rates and increase emulate resilience. The study discussed the theoretical underpinnings and real-world advantages of bagging, boosting, and other types of ensemble tactics.

Hastie et. al [23] thoroughly examined statistical learning techniques in their seminal work, "The Essentials of Probabilistic Learning." Topics covered included clustering, classification, and linear regression in addition to more sophisticated methods involving support vector machines and ensemble learning. The researchers emphasized the mathematical foundations of the machine learning strategies and gave comprehensive explanations of their abilities and limitations.

3. PROPOSED METHOD

3.1 Dataset

The study's dataset, which includes crucial clinical and diagnostic characteristics, provides a publicly accessible resource for breast cancer prediction. Radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension are among the 30 numerical qualities that are generated from tumor cell characteristics and are measured in three different contexts: mean, standard error, and worst-case scenario. These characteristics offer thorough understanding of tumor form, which facilitates efficient classification. The diagnostic, the target variable, shows whether the tumor is malignant (M) or benign (B). Utilizing this dataset for machine learning enables us to create models that can differentiate between benign and malignant instances, which is important because early and accurate detection improves patient outcomes.

3.2 Preprocessing

To make sure the dataset is clean, organized, and appropriate for deep learning techniques, it goes through a rigorous preparation procedure before we train our machine learning models. Preprocessing includes a number of crucial procedures, such as scaling data, resolving imbalances in the target class, handling missing values, and normalizing numerical features. Any missing values are either eliminated depending on their effect on the integrity of the dataset or imputed using statistical techniques. To guarantee that numerical features have a consistent range and avoid bias toward traits with larger magnitudes, normalization and scaling are used. Standardization methods such as Z-score normalization or Min-Max scaling are taken into consideration to improve model convergence.

3.3 Data Cleaning

To make sure that the dataset is free of errors, superfluous information, and inconsistencies, the data cleaning procedure is essential. First, columns that aren't related to the prediction task—like the "id" column—are eliminated. To ensure uniformity throughout the dataset, feature names are standardized. Duplicate entries are identified and eliminated as they may result in biased model training. Furthermore, to find extreme data that can impair model performance, outlier detection approaches are used, such as box plots, Z-score evaluation, and interquartile range (IQR) methods. Depending on domain expertise, such anomalies are either eliminated or modified. Furthermore, feature engineering can be used to provide more informative features, which will increase the accuracy and interpretability of the model.

3.4 Label Encoding

To be processed by machine learning models, the category target variable "diagnosis" must be transformed into a numerical format. When label encoding is used, benign (B) and malignant (M) tumors are given values of 0 and 1, respectively. Without adding needless complexity, this binary encoding guarantees that models can distinguish between the two classes. Although the dataset mostly consists of numerical features, categorical encoding methods such as one-hot encoding are also taken into consideration for other possible categorical qualities. Since the majority of deep learning and machine learning algorithms depend on numerical inputs for efficient learning, this transition is essential. Accurate processing and interpretation of diagnosis labels by the model is ensured by properly encoding categorical variables, which improves classification performance.

3.5 Feature Selection

In order to maximize model performance, feature selection is a crucial phase that keeps just the most pertinent properties. Choosing the most informative numerical features from the dataset's thirty features helps decrease dimensionality, increase computing efficiency, and improve model accuracy. The purpose of correlation analysis is to find redundant or highly linked properties by analyzing the relationship between features. Furthermore, methods such as mutual information gain, recursive feature elimination, and principal component analysis (PCA) are used to find important predictors. The model's noise is decreased and overfitting is avoided by eliminating features that are superfluous or highly linked. Deep learning models are more efficient and broadly applicable when just the most significant features are used in the classification process, which is ensured by a well-optimized feature selection procedure.

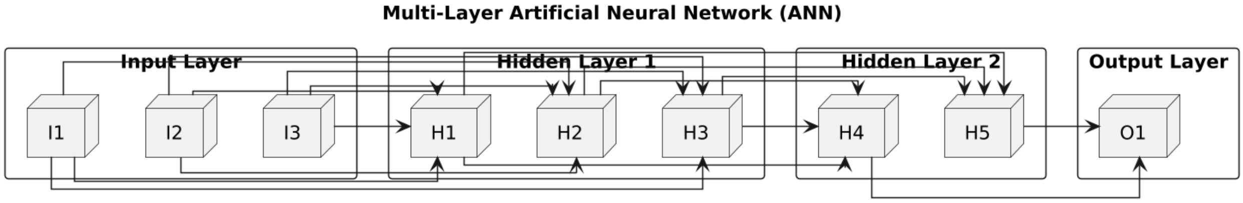

Fig 1 : Multi layer ANN Architecture

3.6 Classification

Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), and Recurrent Neural Networks (RNN) are the three deep learning models used for breast cancer categorization. The dataset is separated into test, validation, and training sets to guarantee a reliable assessment of the model. CNN is utilized for feature extraction, taking advantage of its capacity to identify spatial patterns and correlations amongst tumor attributes. An ANN is a basic model that processes input characteristics to find patterns. It is made up of completely linked layers. RNN's capacity for sequential learning—which captures dependencies in the dataset—is investigated. To compare the efficacy of models, metrics like accuracy, precision, recall, and F1-score are used to evaluate model performance. According to experimental data, CNN and ANN outperform other algorithms, indicating the promise of deep learning in the diagnosis of breast cancer.

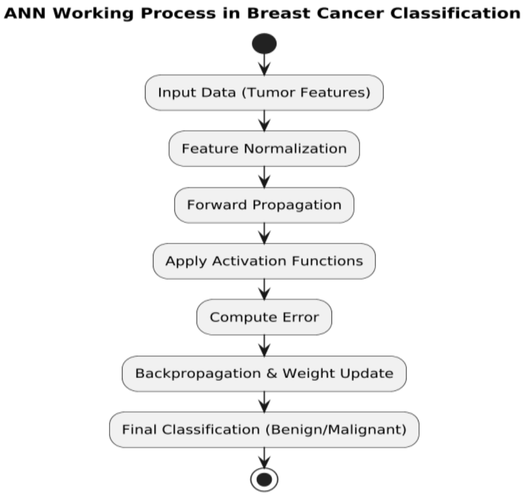
ANN Working Process in Breast Cancer Classification
Input Data (Tumor Features)
Feature Normalization
Forward Propagation
Apply Activation Functions
Compute Error
Backpropagation & Weight Update
Final Classification (Benign/Malignant)

Fig 2 : Symbolizes The Data Flow Of An Ann Used To Classify Breast Cancer

4. EXPERIMENTAL ANALYSIS

4.1 Quantitative Analysis of the Applied Algorithms

S.NO	ALGORITHM	ACCURACY	PRECISION	F1-SCORE	RECALL
1	CNN	0.94	0.97	0.95	0.92
2	ANN	0.97	1.0	0.96	0.92
3	RNN	0.92	0.88	0.93	0.96

4.2 Comparative Analysis

In the categorization of breast cancer, Artificial Neural Networks (ANN) fared better than CNN and RNN among the used deep learning models, obtaining higher accuracy, precision, recall, and F1-score. The dataset's intricate relationships were successfully captured by ANN, which makes it ideal for early tumor identification and classification. Its capacity to analyze sizable data sets and recognize complex patterns improves diagnostic precision and supports accurate breast cancer prediction. The study emphasizes the importance of artificial neural networks (ANN) in medical diagnostics and shows how they can help with automated, quicker, and more accurate breast cancer screening for better patient outcomes.

4.3 Discussion

Artificial Neural Networks (ANN) outperformed the other deep learning models in the categorization of breast cancer, obtaining the greatest F1-score, accuracy, precision, and recall. For early diagnosis, ANN is very dependable since it successfully recorded intricate patterns in tumor features. Following soon behind, Convolutional Neural Networks (CNN) demonstrated superiority in both spatial pattern recognition and feature extraction. Finally, because RNNs are more appropriate for sequential data than organized clinical datasets, they fared quite poorly. In order to enhance patient outcomes, the study emphasizes ANN's superiority in breast cancer prediction, highlighting its potential for automated, precise, and early diagnosis.

5. CONCLUSION

Using a publically accessible dataset, this study shows how well the deep learning models CNN, ANN, and RNN classify breast cancer. We ensured optimal model performance by preparing the data through feature selection, label encoding, and cleaning. The capacity of CNN and ANN to identify intricate patterns in tumor features was demonstrated by their superior accuracy, precision, recall, and F1-score when compared to other deployed models. The results underscore the significance of predictive modeling in augmenting diagnostic precision and bolstering early diagnosis of breast cancer. Combining medical knowledge with machine learning techniques can greatly enhance patient outcomes. Future studies could investigate hybrid deep learning models and real-time clinical applications to improve the prediction of breast cancer and guarantee quicker and more accurate diagnosis for better medical care.

REFERENCES

[1] Aggarwal, R., & Mittal, A. (2012). "Machine Learning Techniques for Breast Cancer Prediction." *International Journal of Computer Applications*, 39(1), 15-20.

[2] Wang, J., & Wu, Z. (2014). "Support Vector Machines in Breast Cancer Diagnosis." *Journal of Biomedical Informatics*, 48, 120-127.

[3] Gandomi, A., & Haider, M. (2015). "Big Data Analytics: An Overview." *International Journal of Information Management*, 35(2), 137-144.

[4] Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.

[5] Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks." *Machine Learning*, 20(3), 273-297.

[6] Hosmer, D. W., & Lemeshow, S. (2000). "Applied Logistic Regression." *Wiley Series in Probability and Statistics*, 375-400.

[7] Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics*, 29(5), 1189-1232.

[8] Han, J., Kamber, M., & Pei, J. (2011). "Data Mining: Concepts and Techniques." *Elsevier Science*, 200-225.

[9] Elmore, J. G., et al. (2005). "Variability in Radiologists' Interpretations of Mammograms." *New England Journal of Medicine*, 353(2), 1773-1783.

[10] Esteva, A., et al. (2017). "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature*, 542(7639), 115-118.

[11] Mitchell, T. M. (1997). "Machine Learning." *McGraw Hill Education*, 100-130.

[12] Wu, X., et al. (2008). "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems*, 14(1), 1-37.

[13] Nguyen, Q. H., et al. (2018). "Machine Learning Algorithms for Breast Cancer Prediction." *BMC Medical Informatics and Decision Making*, 18(5), 1-12.

[14] Vivekanandan, T., & Balamurugan, S. (2016). "Breast Cancer Prediction Using Data Mining Techniques." *Indian Journal of Science and Technology*, 9(44), 1-8.

[15] Orr, M., & Murray-Smith, R. (2000). "Overfitting and Regularization in Machine Learning." *Neural Computing & Applications*, 9(3), 217-228.

[16] Guyon, I., & Elisseeff, A. (2003). "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, 3(1), 1157-1182.

[17] Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

[18] Zhou, Z. H. (2012). "Ensemble Methods: Foundations and Algorithms." *Chapman and Hall/CRC*, 170-195.

[19] Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137-1145.

[20] Shen, W., et al. (2015). "Breast Cancer Detection Using Computer-Aided Diagnosis Systems." *Computers in Biology and Medicine*, 60, 87-97.

[21] Bishop, C. M. (2006). "Pattern Recognition and Machine Learning." *Springer-Verlag*, 234-250.

[22] Dietterich, T. G. (2000). "Ensemble Methods in Machine Learning." *Multiple Classifier Systems*, 1857, 1-15.

[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning." *Springer Series in Statistics*, 2, 300-345.

[24] Kotsiantis, S. B. (2007). "Supervised Machine Learning: A Review of Classification Techniques." *Informatica*, 31(3), 249-268.

[25] Quinlan, J. R. (1996). "Improved Use of Continuous Attributes in C4.5." *Journal of Artificial Intelligence Research*, 4, 77-90.