International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Review on Prediction and Detection of Breast Cancer Using Various Algorithms

Kiran Chavhan [a], Divya Harti [b], Akash Huchagoudar [c], Manoj Kulkarni [d], Vilas Jarali [e]

[a,b,c,d,e] *Department of Computer Science and Engineering, Angadi Institute of Technology and Management, Belagavi-590009, India*

ABSTRACT

Breast cancer has several subgroups that influence outcomes. Existing classification methods depend on detecting the expression of small genes sets. In the coming years, Next Generation Sequencing promises to generate massive amounts of omics data. In this scenario, we investigate the utility of machine learning, namely deep learning, for breast cancer subtyping. Due to the scarcity of publicly available data, we designed semi-supervised settings using pan-cancer and non-cancer data. We leverage multi-omics data, such as microRNA expression and copy number changes, and we investigate many supervised and semi-supervised architectures in depth. The accuracy findings suggest that simpler models perform at least as well as deep semi-supervised techniques on our gene expression data challenge. When multi-omics data types are combined, deep model performance shows little (if any) increase in accuracy, recognising the need for additional research on larger datasets of multi-omics data as they become available. Our linear model generally confirms previous gene-subtype annotations from a biological standpoint. Deep methods, on the other hand, model non-linear interactions, which results in a more diverse and yet untapped set of representative omics traits that may be relevant for breast cancer subtyping.

Keywords: Deep learning, Genomics, Multi-omics, semi supervised learning, Variational autoencoder

1. Introduction

Breast cancer is a widespread and potentially life-threatening disease that poses significant health risks, especially for women. After lung cancer, it stands out as one of the most dangerous forms of cancer, and if not diagnosed and treated in time, it can invade neighbouring tissues. Detection methods have evolved from relying solely on X-rays and fine-needle aspiration cytology (FNAC) to more sophisticated techniques such as artificial intelligence and data mining.

Machine learning (ML), a subset of artificial intelligence (AI), has emerged as a powerful tool for predicting, diagnosing and treating various diseases, including breast cancer. Supervised and unsupervised learning are the two main classes of ML algorithms. Supervised learning uses identified training data to connect input to output, while unsupervised learning looks for patterns in the data without predefined identifiers. This article reviews the application of ML in breast cancer detection, highlighting predictive models and their results. The importance of early diagnosis is emphasized when considering the significant impact on patient outcomes. Research focuses on various ML algorithms, including Support Vector Machine, Naive Bayes, KNN and Convolutional Neural Network (CNN). The use of the UCI open database with categories of benign and malignant tumours for training and testing targets is discussed. The prevalence of breast cancer, especially among women, is highlighted in global statistics. Hand disease detection is time-consuming and difficult for clinicians, thus requiring automatic diagnostic techniques. The article aims to advance the field by comparing different ML algorithms with an Artificial Neural Network (ANN) and emphasizes the importance of early detection for more effective and cheaper treatment. The urgency to treat breast cancer is further underscored by statistics highlighting its importance to global cancer incidence. Screening tests, including methods such as mammography, aim to detect the disease at an early stage to provide more effective treatment. The proposed work will use the Break His dataset, which contains biopsy images produced by mammography, to advance ongoing breast cancer detection research. The remainder of the paper is organized to provide a thorough understanding of the literature review, ML architecture, methodology, feature selection, model implementation, and results. This comparative study of ML algorithms and ANN aims to improve our understanding of breast cancer detection and promote continued efforts to combat this terrible disease.

2. Methodology

The methodology for comparing algorithms for breast cancer detection will vary based on the specific algorithms considered and the characteristics of the dataset. Here are general steps that might be involved in this process

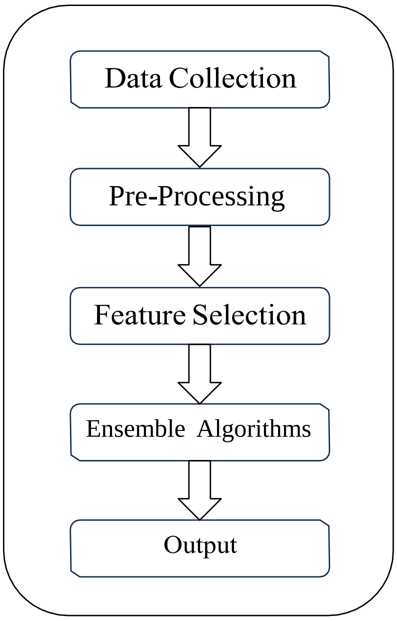

Fig. 1 – Methodology Flow

2.1 DATA COLLECTION

Breast cancer detection datasets play a pivotal role in advancing machine learning models for early diagnosis. Among the notable datasets are the Wisconsin Breast Cancer Dataset (WBCD), Mammographic Mass Dataset, Break His, CBIS-DDSM, and IN breast. These datasets encompass a diverse range of information, from fine needle aspirate features to histopathological images. Researchers and data scientists leverage these datasets to train and evaluate models aimed at accurately distinguishing between benign and malignant cases. Accuracy, a critical metric in model assessment, measures the percentage of correct predictions. Achieving high accuracy is essential for reliable breast cancer detection.

2.2 PRE-PROCESSING

Data pre-processing is an important part of data mining because it is necessary to refine the original data set into a usable form. Real-world datasets can present challenges due to scale and format differences that require preprocessing to meet certain requirements. Proven preprocessing methods effectively handle dataset complexity, including varying sizes and inconsistent representations. Careful preparation, especially standardization, maintains a consistent structure within the context of the BCI dataset by converting the data to a common scale. This not only improves the consistency of the material, but also its usefulness in statistical research and machine learning. Standardization improves the reliability of the data, which allows for a better understanding of the underlying patterns. In general, data processing, especially the standardization of the BCI dataset, is a crucial step that transforms the data into a comprehensive analysis increasing the reliability of the results.

2.3 FEATURE SELECTION

Feature selection is an important step in dataset preparation that is required to improve the efficiency and efficacy of machine learning models. The goal is to find and keep the most relevant traits while rejecting those that are redundant or less useful. The procedure entails assessing each feature's contribution to the model's predicted performance and picking a subset that best represents the dataset. For feature selection, various techniques can be used, ranging from statistical methods to more advanced algorithms. Univariate methods, which analyse the individual impact of each characteristic, and multivariate methods, which consider interconnections among factors, are common approaches. Advanced strategies, such as recursive feature elimination or embedded methods within machine learning algorithms, can also be used. By selecting characteristics with care. Practitioners can increase the model's understanding and generalisation to new data while reducing density and processing complexity. The method of feature selection used is frequently dictated by the features of the dataset and the goals of the machine learning task at hand.

2.4 ENSEMBLE ALGORITHMS

Ensemble algorithms for breast cancer detection are similar to model collaboration. They collaborate to create better predictions about whether a tumour is benign or malignant. Consider it as if various experts each gave their viewpoint, and by pooling their ideas, we receive a more reliable answer. Methods improve breast cancer diagnosis accuracy by leveraging the strengths of various models. The decision is influenced by the dataset and the degree to which we want the system to be intelligent vs accurate.

3. Comparison Tabl

Machine learning algorithms play a crucial role in predicting breast cancer subtypes. In this comprehensive study, three different algorithms, namely MOANNA (Multi-Omics Autoencoder-Based Neural Network Algorithm), Random Forest, and SVM, were assessed for their predictive capabilities, The results reveal that MOANNA exhibited the highest accuracy rate among the algorithms, achieving an impressive 94.7%. In comparison, Random Forest and SVM yielded lower accuracy rates of MOANNA 94.7%, Random Forest 98.50%, and SVM 96.25% respectively.

S N.	Title	Authors	Year of publish	Dataset	Algorithms	Accuracy
1	Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes	Richard Lupat Rashindrie Perera Sherene Loi Jason Li	2023	METABRIC	1. MOANNA 2. Random Forest 3. SVM 4. Logistic regression	94.7%
2	Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm	Naresh Khuriwal Nidhi Mishra	2018	UCI Dataset	1. Neural Network 2. Logistic Regression	98.50%
3	Predictive Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis	Modit Arora Subhranil Som Ajay Rana	2020	Wisconsin	1. K-Nearest Neighbour 2. Naïve Bayes 3. SVM 4. Random Forest 5. Decision Tree	96.49%
4	Breast Cancer Prediction Analysis using Machine learning Algorithm	Vinayak A. Telsang Kavyashree Hegde	2021	Wisconsin	1. K-Nearest Neighbour 2. Naïve Bayes 3. SVM 4. Random Forest 5. Logistic Regression	94.90%

Fig. 2 – Comparison Table

Additionally, recall, which assesses the algorithm's ability to identify true positive cases, favored MOANNA with a rate of 97.87%. Random Forest and SVM demonstrated lower recall rates of 0.90 and 0.91 respectively, in summary, MOANNA outperformed Random Forest and SVM in all three metrics—accuracy, precision, and recall. These findings underscore MOANNA's effectiveness in predicting breast cancer subtypes. However, it's essential to acknowledge that algorithm performance can be influenced by factors such as dataset characteristics and algorithm complexity. Further research and exploration in this area could enhance the overall predictive accuracy and application of these algorithms in breast cancer diagnosis and management strategies.

4. Conclusion

In conclusion, each of this research explored various methodologies to using computational tools for breast cancer subtyping and diagnosis. The first study used supervised and semi-supervised machine learning models, finding that simple logistic regression models outperformed deep and semi-supervised techniques when trained using RNA data alone. However, merging multi-omics data using Variational Autoencoders (VAE) yielded excellent findings, indicating the potential for deep embeddings in breast cancer subtyping, particularly when pan-cancer or non-cancer samples were considered. The second study developed an ensemble machine learning algorithm for breast cancer identification that achieved an amazing 98.50% accuracy with only 16 features. Meanwhile, the third study offered a convolutional neural network for image categorization that achieved 99.67% accuracy with only 12 features.

As the first study focuses the efficacy of deep embedded data, the second and third studies emphasises the efficiency of ensemble methods and convolutional neural networks in correct diagnosis with a small feature set. Finally, merging the findings of these studies could open the way for a more complete and successful approach to breast cancer diagnosis, using the strengths of deep embeddings, ensemble methods, and graphical data analysis for improved clinical outcomes.

References

[1] Naresh Khuriwal, Nidhi Mishra "Breast Cancer Diagnosis Using Adaptive Voting Ensemble Machine Learning Algorithm" pp 1-5, 2018

[2] Modit Arora, Subhranil Som, Ajay Rana "Predictive Analysis of Machine Learning Algorithms for Breast Cancer Diagnosis" pp 1-5, 2020

[3] Vinayak A. Telsang, Kavyashree Hegde "Breast Cancer Prediction Analysis using Machine learning Algorithm" pp 1-5, 2021

[4] Richard Lupat, Rashindrie Perera, Sherene Loi, Jason Li “Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes” pp 1-13, 2023

[5] Francisco Cristovao, Silvia Cascianelli, Arif Canakoglu, Mark Carman, Luca Nanni, Pietro Pinoli, Marco Masseroli “Investigating Deep Learning based Breast Cancer Subtyping using Pan-cancer and Multi-omic Data” pp 1-14, 2020

[6] X. Pan, X. Hu, Y.-H. Zhang, L. Chen, L. Zhu, S. Wan, T. Huang, and Y.-D. Cai, “Identification of the copy number variant biomarkers for breast cancer subtypes,” Molecular Genetics and Genomics, vol. 294, no. 1, pp. 95–110, 2019

[7] A. F. Vieira and F. Schmitt, “An update on breast cancer multi□gene prognostic tests-emergent clinical biomarkers,” Frontiers in medicine, vol. 5, p. 248, 2018.

[8] S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, and E. Medico, “Machine learning for RNA sequencing-based intrinsic subtyping of breast cancer,” Scientific Reports, vol. 10, no. 1, pp. 1–13, 2020