Machine Learning for Human Intelligence
2024, Vol. 2, Issue 1

Research Article

A Hybrid Method for Breast Cancer Classification Utilizing Feature Fusion

Khadija Bibi[1,⋆] and Faheem Naveed[2]

[1,2]Department of Computer Science, COMSATS University, Islamabad, Pakistan.; Email: khadijaqadir54@gmail.com, faheemnaveed95@gmail.com
⋆Corresponding author: Khadija Bibi (khadijaqadir54@gmail.com)

Article History

Academic Editor:
Dr. Muhammad Sajid
Submitted: January 11, 2024
Revised: February 25, 2024
Accepted: March 1, 2024

Keywords:
Mammary carcinoma; Computational algorithms; Breast neoplasm; Categorization

Abstract

Breast cancer is a prevalent form of malignancy, particularly among women. Estimates indicate that one in nine women receives a diagnosis of breast cancer. The insufficiency of adequate facilities is causing delays in breast cancer diagnosis, hence elevating the prospective mortality rate. A variety of statistical techniques and machine learning algorithms are frequently utilized in research to enhance the accuracy of breast cancer detection. Machine learning (ML) has yielded significant outcomes for physicians, and the healthcare sector is employing ML-based models for the detection of breast cancer in women. This facilitates the analysis of healthcare data and employs conventional computer-aided detection (CAD) to evaluate breast cancer. Machine learning has been integrated into clinical practice, enabling physicians to assess the ML model for early breast cancer detection. This research utilizes various machine learning methods to categorize cancer as malignant or benign. MLP signifies a more efficient and accurate methodology for breast cancer categorization. The Matthews correlation coefficient for the MLP is 0.89%, whereas the accuracy score for the random forest is 0.94%.

1 Introduction

The 2018 survey indicates that the prevalence of breast cancer among 2.4 million women can be inferred from the statistic that one in four women is afflicted by cancer [1], [2]. Breast cancer incidence is lower in Asian countries compared to Western countries. However, the incidence of breast cancer has risen in Asian countries over time as well [2]. The advancement of clinical innovations enables health professionals to promote more effective medical prevention systems. E-medical services frameworks are beneficial in numerous therapeutic domains. The applications of biological images in computer vision are expanding because to their significance in providing radiologists with critical data for enhanced treatment of problems. Unique clinical imaging modalities, including High-Frequency Electromagnetic Radiation (X-ray), MRIs, Ultrasound, and computed tomography (CT) scans, influence the diagnostic and therapeutic processes for patients. The proliferation of malignant cell clusters within or next to the brain leads to the formation of a brain tumor. The abrupt demise of cancerous cells in the brain affects the patient's health. The analysis, diagnosis, and treatment of cerebral pictures with modern medical imaging techniques are essential areas of research for physicians, radiologists, and clinical specialists.

Recent estimates indicate that one in nine women in Pakistan is afflicted by cancer, as shown in the patient registration methods. Pakistan ranks #1 among Asian countries for the increasing incidence of cancer in women with advancing age. The absence of adequate facilities is regrettably causing delays

in breast cancer diagnoses, resulting in a heightened potential mortality rate [3]. Employing a scientific approach to breast cancer initially could substantially enhance the survival rate of Pakistani women [4]. In Pakistan, there exists no repository or database pertaining to any disease, including breast cancer; data is solely sourced from hospitals, who provide information on the number of new cancer patients and their yearly mortality rate [6, 1].

The incidence of breast cancer among women in Western countries has risen among those aged 50 and older. Younger women in developing countries have a 47% higher incidence of breast cancer compared to older women. Breast cancer is typically observed in women aged 40 to 60 within the Asian demographic. Breast cancer patients in Asian nations, including India, Korea, and Japan, are often aged between 40 to 49 or 50 to 55 years [7].

Breast cancer is a disease originating from uncontrolled cell proliferation in breast tissue. It manifests anyplace in the body because to the proliferation of our cells. It impedes the body's functionality. It primarily originates from a tumor or tumor proliferation. Occasionally, a mass referred to as a tumor is present, although not all such masses are malignant. A biopsy is the procedure in which a portion of the mass is excised to determine the presence of malignancy. Non-cancerous lumps are referred to be benign. Masses that disseminate cancer perilously across the body are termed malignant [8]. A range of imaging modalities and methods were employed for the management and prognosis of a brain tumor [3]. In image processing techniques, segmentation is a crucial phase that facilitates the elimination of aberrant brain regions from MRIs. Tumor area segmentation is a critical process for predicting, managing, and evaluating therapy outcomes in cancer. MRI encompasses many sequence techniques for segmentation, including TI, T1c, T2-weighted, and FLAIR methods. MRIs possess distinct properties, including image textures, local histograms, and structure tensor eigenvalues utilized in segmentation.

1.1 Risk Factors

A significant number of women get breast cancer due to unidentified risk factors. Aside from being a woman, this constitutes a significant risk factor for developing cancer. A prior biopsy elevates your risk of breast cancer. An increased age correlates with a higher likelihood of developing breast cancer. Breast cancer is predominantly observed in younger individuals if there is a familial history involving a sister, mother, or daughter diagnosed with the disease [11].

Additional risk factors for breast cancer include increased alcohol consumption, tobacco use, and the use of various medicines [12]. Women diagnosed with colon, ovarian, or endometrial cancer exhibit an elevated risk of developing breast cancer. Upon the onset of your symptoms, consult a qualified physician or confide in a trusted friend or family member.

1.2 Cancer Survivors

A cancer survivor is a someone who continues to live after a cancer diagnosis. Each cancer survivor possesses unique characteristics, concerns, and obstacles [13]. Over the past 55 years, there has been a rising ratio of cancer patients. A 1971 survey indicated that 3 million individuals have cancer. Today, that figure has increased from 3 million to 15.5 million [14]. Approximately 67% of individuals survive each year. Seventeen percent of all cancer survivors were diagnosed two decades ago. Forty-seven percent of survivors are aged 70 or older [15].

1.3 Machine Learning in Healthcare

A process that extracts valuable information from extensive datasets, utilizing functions and data mining techniques, aids in identifying various diseases [16]. Statistics, databases, fuzzy sets, neural networks, and data warehouses contribute to the diagnosis of different types of cancer, such as lung cancer, leukemia, and prostate cancer [17]. The identification of conventional cancer methodologies relies on "the gold standard." Data mining involves three types of tests: pathology tests, clinical examinations, and radiological imaging [18].

The primary three stages of machine learning to be applied to available datasets are data preparation, feature selection or extraction, and classification [19]. A significant method that aids in predicting

cancer signs is known as feature extraction. It elucidates cancer by assessing patient data and categorizing tumors as benign or malignant [20].

2 Literature Review

In recent years, deep learning frameworks, methodologies, and algorithms have demonstrated a significant advancement in the intelligent extrapolation of subtypes and breast cancer forecasts.

The researchers employed the dataset designated as TCGA-BRCA as a mockup set for the subtype. It also examined and forecasted the molecular mechanisms underlying breast cancer [31]. They constructed a hybrid deep learning model and projected multimodal data. They integrated the genetic modality data of patients with the modality data that included photos.

Diverse designs have been proposed for brain tumor segmentation, which seeks to improve the system's accuracy. The tumor segmentation challenge can be accomplished by judiciously identifying both individual pixels and dense pixel clusters. This section will address the novel methodologies employed by researchers for brain tumor segmentation tasks. The authors in [31] employed a WRN that autonomously segments glioblastoma to eliminate characteristics from multi-modal brain tumor samples. Subsequently, the WRN functions yield a global representation at various levels through PPNet to adjust the recovery unit, wherein the original inputs are reintroduced into the network. The method generates pixel-level predictions that match the dimensions of the original inputs; however, it is constrained by issues such as overfitting and feature loss identified in the WRN Module. Increasing the number of layers exacerbates these problems, leading to a setting of 4 in the paper, although this does not consistently yield improved results for every image at that level. In [31], the authors suggested a contemporary cascaded U-Net for brain tumor segmentation, wherein the overall tumor is initially segmented, followed by the segmentation of the inner tumor parts. To establish unique frameworks for breast corpus classification, they implemented a structure composed of convolutional neurons featuring two convolutional layers [32]. Within their CNN framework, they demonstrated an increase from seventy-nine percent to eighty-six percent in testing compared to the outdated radionics frameworks.

In standard clinical practice, the deep learning model for evaluating mammographic breast density was presented to the radiologist during mammography interpretation [33]. The percentage of mammograms classified as dense by all radiologists diminished from 47.0 percent prior to the introduction of the deep learning model to 41.0 percent subsequent to its adoption.

This research presents a novel methodology that employs deep learning inside an ensemble framework, incorporating multiple distinct machine learning models [34]. They provided five distinct categorization models utilizing informative gene data obtained by differential organic phenomenon analysis [31], [35].

3 Materials and Methods

This article predicts breast cancer via various machine learning techniques. The dataset utilized for breast cancer prediction is sourced from Kaggle [38], where it is publicly accessible for research purposes. The collection comprises 569 instances and 6 characteristics.

Table 1: Characteristics of the Dataset

Characteristics	Explanation	Scope
Diagnosis	Diagnosis of breast cells (1 = Malignant, 0 = Benign)	Zero to one
Average Perimeter	Average dimensions of the core neoplasm	Zero to one
Average Area	-	Zero to one
Mean Smoothness	Average of local variation in radius length	Zero to one

Data pre-processing is a crucial phase prior to categorization. Data pre-processing encompasses

data cleaning, dimensionality reduction, transformation, normalization, and processing. Our data cleaning method involves imputing any missing values with the mean of the respective attributes. We utilize the Pearson correlation coefficient for feature selection in cancer cell detection as shown in Figure 1 [39], [40].

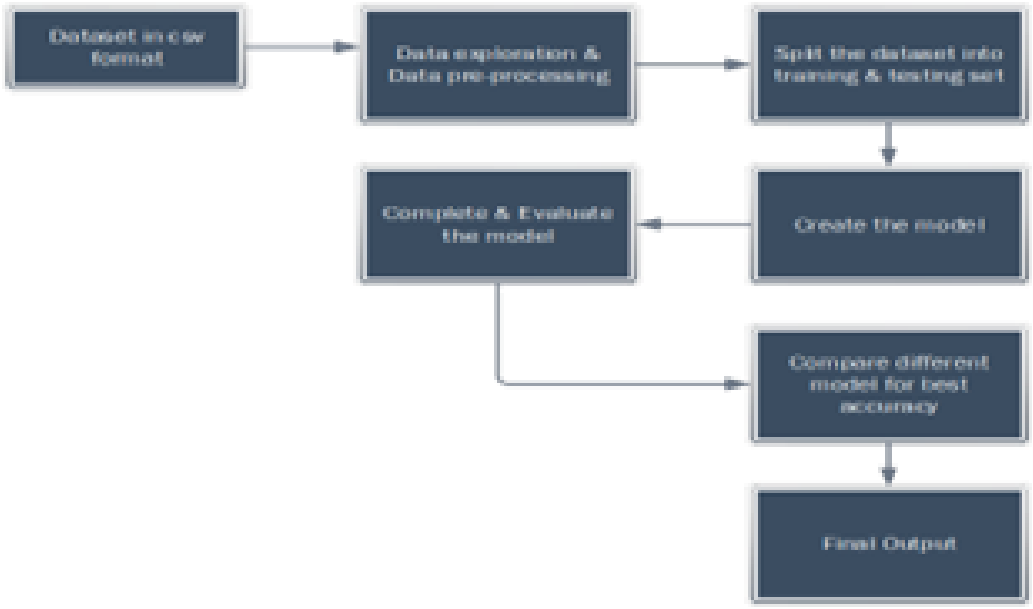

Figure 1: Proposed Methodology

3.1 Modeling Technique

This study employs six distinct categorization models: Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). The primary objective of employing these models is to yield optimal outcomes for cancer prediction. The confusion matrix employed to evaluate these performance metrics includes True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Utilizing Data Augmentation techniques to enhance the volume of picture training data by generating modified images from the original dataset. This strategy introduces changes in the images, enhancing the model's capacity to learn and generalize from previously unseen data more effectively [21]. The model is consequently generalized by including fluctuations in the training dataset, thereby reducing its susceptibility to overfitting. Various data augmentation techniques are employed, including horizontal flipping, a height shift of 5%, a rotation of 20%, a width shift of 5%, a shear of 5%, a zoom of 5%, and a fill mode configured to closest. This study use specific factors and formulae to assess performance. Accuracy (Acc) is the ratio of correctly classified instances (true positives and true negatives) to the total number of cases:

4 Experimental Results

This section discusses data analysis, data description, and data preprocessing. We have employed machine learning algorithms (ML) for the categorization of breast cancer as either malignant or benign. The fundamental processes of machine learning include dataset exploration, data preprocessing and cleaning, followed by partitioning the data and applying the model as shown in Figure 2.

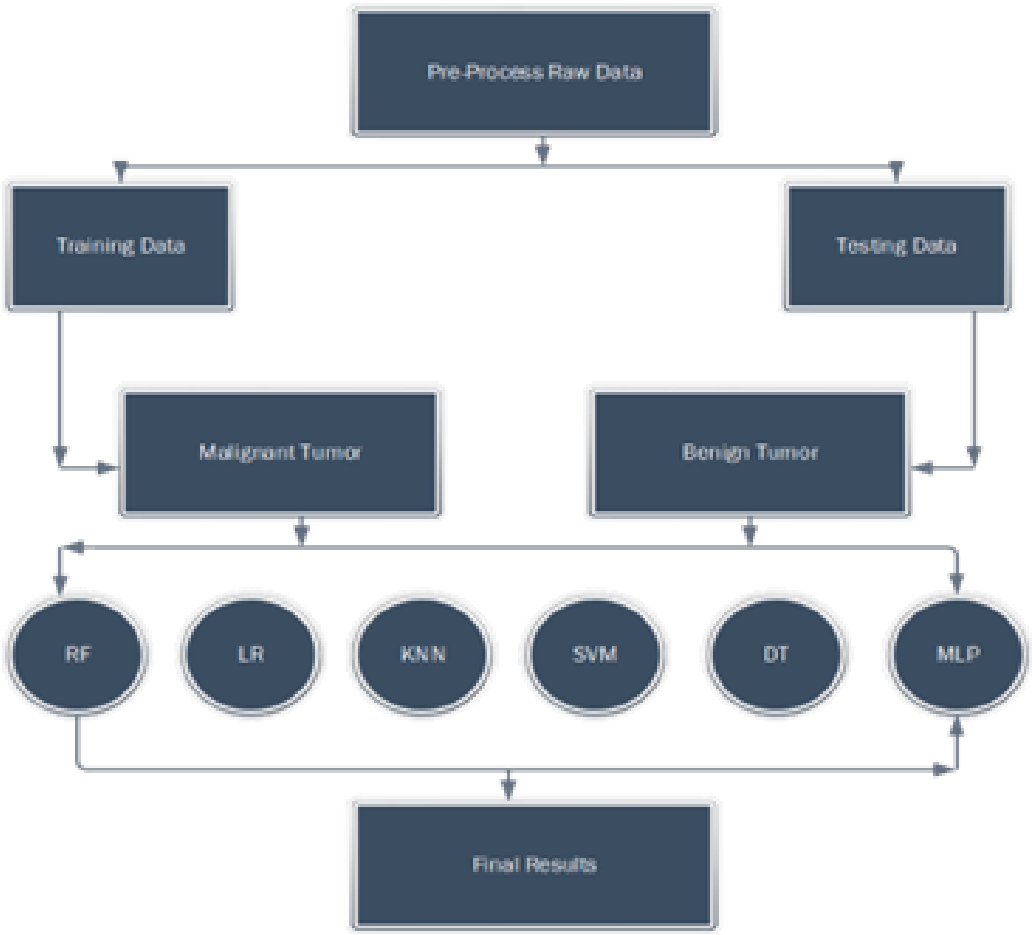

Figure 2: Process of Proposed Model

4.1 Classification Results

4.1.1 Random Forest

Initially, the Random Forest model was employed. Random forest is an ensemble classifier that use decision tree techniques [47]. The Matthews correlation coefficient for the random forest model is 0.80, and the accuracy score is 0.90.

4.1.2 Logistic Regression

A logistic regression model was employed. Despite its name suggesting a focus on regression, this method is really employed for classification purposes [46]. The Matthews correlation coefficient for logistic regression is 0.87%, while the accuracy score for random forest is 0.93%.

4.1.3 K-Nearest Neighbors (KNN)

The KNN model was employed thirdly. This algorithm is crucial for classification and regression, commonly employed in data mining. KNN is effective for small datasets, which contributes to its ease of implementation [49]. The Matthews correlation coefficient for KNN is 0.87%, while the accuracy score for the random forest is 0.93%.

4.1.4 Decision Tree

A decision tree model was employed. Decision trees [42] are utilized in two distinct data mining techniques: categorization and prediction. It serves to visually delineate rules that are easily interpretable

and comprehensible [43]. The Matthews correlation coefficient for the decision tree is 0.75%, whereas the accuracy score for the random forest is 0.87%.

4.1.5 Multi-Layer Perceptron (MLP)

The MLP model was utilized lastly. The Matthews correlation coefficient for the MLP is 0.89%, whereas the accuracy score for the random forest is 0.94% as givein in Table 2.

Table 2: Classification Report for Multi-Layer Perceptron

	Precision	Recall	F1-score	Support
0	0.94	0.94	0.94	48
1	0.95	0.95	0.95	66
micro avg	0.95	0.95	0.95	114
macro avg	0.95	0.95	0.95	114
weighted avg	0.95	0.95	0.95	114

In a decision tree, the accuracy is 0.87%, whereas the MLP achieves an accuracy of 0.94%. Based on this accuracy rate, we propose that MLP is optimal for classifying breast tumors as shown in Figure 3.

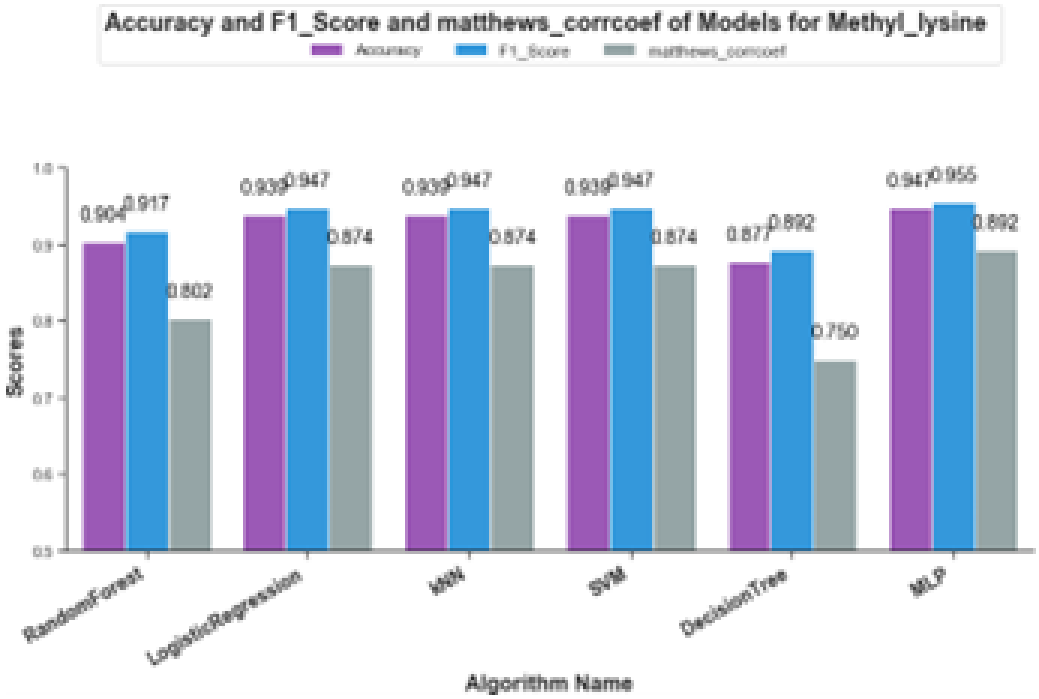

Figure 3: Best Accuracy Algorithm

5 Conclusions

A multitude of approaches in data mining and machine learning exist for the study of medical data. The principal objective in data mining and machine learning is the development of an enhanced classifier for medical science. This research included six algorithms: logistic regression, random forest, K-nearest neighbors (KNN), support vector machine (SVM), decision tree, and multilayer perceptron (MLP).

The MLP algorithm achieves the maximum accuracy at 0.94%. The decision tree exhibits the lowest accuracy at 0.83%. The primary objective of employing machine learning algorithms is to facilitate tumor detection, whereas in the realm of medical science, the equivalent processes require more time

and financial resources. Machine learning approaches function as a clinical helper for novice doctors and physicians in diagnosing breast cancer. MLP has demonstrated superiority over all other approaches in predicting breast cancer. The further application of MLP may yield significant advantages in cancer prediction. This research concludes that machine learning approaches can automatically and more accurately detect diseases.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., et al. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6), 394–424.

[2] Vos, T., Barber, R. M., Bell, B., Biryukov, S., et al. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: A systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 386(9995), 743–800.

[3] Sohail, S., & Alam, S. N. (2007). Breast cancer in Pakistan - awareness and early detection. *Journal of College of Physicians and Surgeons Pakistan*, 17(12), 711–712.

[4] Karachi, I., Yasmeen, P. F., & Zaheer, S. (2014). Functional time series models for estimating future age-specific breast cancer incidence rates for women. *Journal of Health Sciences*, 2, 213–221.

[5] Karadeli, E., Erbay, G., Parlakgumus, A., & Koc, Z. (2018). Utility of diffusion-weighted magnetic resonance imaging with multiple b-values in the evaluation of pancreatic malignant and benign lesions and pancreatitis. *Journal of the College of Physicians and Surgeons Pakistan*, 28(2), 103–109.

[6] Hanif, M., Sabeen, B., Maqbool, A., Ahmed, A., Nadeem, F., et al. (2015). Breast cancer: Incidence (thirteen-year data analysis) and one-year clinicopathological data of patients in a tertiary care cancer hospital. *International Journal of Biology and Biotechnology*, 12(3), 373–379.

[7] Park, E. H., Min, S. Y., Kim, Z., Yoon, C. S., Jung, K. W., et al. (2017). Fundamental statistics of breast cancer in Korea in 2014: The decade-long overall survival progress. *Journal of Breast Cancer*, 20(1), 1–11.

[8] CDC Prevention. (2021). What is breast cancer? Retrieved from https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm

[9] Healthline. (2021). Breast cancer: Symptoms, stages, types, and more. Retrieved from https://www.health-line.com/health/breast-cancer#types

[10] CDC Prevention. (2021). What constitutes breast cancer screening? Retrieved from https://www.cdc.gov/cancer/breast/basic_information/screening.htm

[11] Qaseem, A., Lin, J. S., Mustafa, R. A., Horwitch, C. A., & Wilt, T. J. (2019). Screening for breast cancer in women at average risk: A guidance statement from the American College of Physicians. *Annals of Internal Medicine*, 170(8), 547–560.

[12] CDC Prevention. (2021). What are the risk factors for breast cancer? Retrieved from https://www.cdc.gov/cancer/breast/basic_information/risk_factors.htm

[13] CancerNet. (2021). Breast cancer: Survivorship. Retrieved from https://www.cancer.net/cancer-types/breast-cancer/survivorship

[14] Mariotto, A. B., Etzioni, R., Hurlbert, M., Penberthy, L., & Mayer, M. (2017). Estimation of the prevalence of women with metastatic breast cancer in the United States. *Cancer Epidemiology, Biomarkers & Prevention*, 26(6), 809–815.

[15] Sung, H., DeSantis, C. E., Fedewa, S. A., Kantelhardt, E. J., & Jemal, A. (2019). Subtypes of breast cancer among Black women and other Black women in the United States. *Cancer*, 125(19), 3401–3411.

[16] Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, 12(4), 1243–1257.

[17] Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems*, 26(1), 100–112.

[18] Shahbazl, M., Faruq, S., Shaheen, M., & Masood, S. A. (2012). Cancer diagnosis utilizing data mining technology. *Life Science Journal*, 9(1), 308–313.

[19] Vaka, A. R., Soni, B., & Reddy, S. (2020). Detection of breast cancer utilizing machine learning. *ICT Express*, 6(4), 320–324.

[20] Sandri, V., Gonçalves, I. L., Neves, G. M. D., & Paraboni, M. L. R. (2020). Diagnostic significance of C-reactive protein and hematological parameters in acute toxoplasmosis. *Journal of Parasitic Diseases*, 44(4), 785–793.

[21] Eltalhi, S., & Kutrani, H. (2019). Breast cancer diagnosis and prediction utilizing machine learning and data mining methodologies: A review. *IOSR Journal of Dental and Medical Sciences*, 18(4), 85–94.

[22] Wang, J. M., Qian, C. L., Che, C. H., & He, H. T. (2010). Research on the methodology of network traffic classification utilizing machine learning. In *Fifth Annual ChinaGrid Conference* (pp. 262–266). IEEE.

[23] Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

[24] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 249–256).

[25] Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

[26] Meads, C., Ahmed, I., & Riley, R. D. (2012). A systematic review of breast cancer incidence risk prediction models accompanied by meta-analysis on their performance. *Breast Cancer Research and Treatment*, 132(2), 365–377.

[27] Wu, Y., Abbey, C. K., Chen, X., Liu, J., Page, D. C., et al. (2015). Formulating a utility decision framework for the assessment of predictive models in breast cancer risk assessment. *Journal of Medical Imaging*, 2(4), 041005.

[28] Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast cancer detection utilizing machine learning algorithms. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 114–118). IEEE.

[29] Wadhwa, G., Mathur, M., Todorova, M., Omondiagbe, D. A., Veeramani, S., et al. (2019). Machine learning classification techniques for breast cancer diagnosis. In *IOP Conference Series: Materials Science and Engineering* (Vol. 495, No. 1, p. 012033). IOP Publishing.

[30] Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Employing machine learning algorithms for the prediction of breast cancer risk and diagnosis. *Procedia Computer Science*, 83, 1064–1069.

[31] Zhang, X., He, D., Zheng, Y., Huo, H., Li, S., et al. (2020). Deep learning-based analysis of breast cancer utilizing an advanced ensemble classifier and linear discriminant analysis. *IEEE Access*, 8, 120208–120217.

[32] Guo, X., Liu, Z., Sun, C., Zhang, L., Wang, Y., et al. (2020). Deep learning radiomics of ultrasonography: Identifying the risk of axillary non-sentinel lymph node involvement in primary breast cancer. *EBioMedicine*, 60, 103018.

[33] Pang, T., Hsiu, J., Wong, D., Lin, W., & Seng, C. (2020). Deep learning radiomics in breast cancer across various modalities: An overview and future directions. *Expert Systems with Applications*, 158, 113501.

[34] Li, L., Feng, Q., & Wang, X. (2020). PreMSIm: An R package for predicting microsatellite instability based on expression profiling. *Computational and Structural Biotechnology Journal*, 18, 668–675.

[35] Yang, X., Wu, L., Ye, W., Zhao, K., Wang, Y., et al. (2019). Deep learning signature based on staging CT for preoperative prediction of sentinel lymph node. *Academic Radiology*, 4, 1–8.

[36] Shayma'a, A. H., Sayed, M. S., Abdalla, M. I., & Rashwan, M. A. (2020). Classification of breast cancer masses utilizing deep convolutional neural networks and transfer learning. *Multimedia Tools and Applications*, 79(41), 30735–30768.

[37] Gu, D., Su, K., & Zhao, H. (2020). A case-based ensemble learning system for elucidating breast cancer recurrence prediction. *Artificial Intelligence in Medicine*, 107, 101858.

[38] Kaggle. (2021). Breast cancer prediction dataset. Retrieved from https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset

[39] Tuggener, L., Amirian, M., Rombach, K., Lörwald, S., & Varlet, A. (2019). Automated machine learning in practice: State of the art and recent results. In *2019 6th Swiss Conference on Data Science* (pp. 31–36). IEEE.

[40] Mahesh, B. (2020). Machine learning algorithms: A review. *International Journal of Science and Research (IJSR)*, 9, 381–386.

[41] Wang, L. (2005). *Support vector machines: Theory and applications* (Vol. 177). Springer Science & Business Media.

[42] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms for classification in data mining. *International Journal of Science and Research*, 5(4), 2094–2097.

[43] Yang, Y., Li, J., & Yang, Y. (2015). The research of the fast SVM classifier method. In *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (pp. 121–124). IEEE.

[44] Schwamm, L. H., Holloway, R. G., Amarenco, P., Audebert, H. J., Bakas, T., et al. (2009). A review of the evidence for the utilization of telemedicine within stroke care systems: A scientific statement from the American Heart Association/American Stroke Association. *Stroke*, 40(7), 2616–2634.

[45] Ibrahim, A. A., Hashad, A. I., & Shawky, N. E. M. (2017). A comparative analysis of open-source data mining tools for breast cancer classification. In *Handbook of Research on Machine Learning Innovations and Trends* (pp. 636–651). IGI Global.

[46] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1), 3–14.

[47] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[48] Ayodele, T. O. (2010). Categories of machine learning algorithms. *New Advances in Machine Learning*, 3, 19–48.

[49] Imandoust, S. B., & Bolandraftar, M. (2013). Utilization of the k-nearest neighbor (KNN) methodology for forecasting economic occurrences: Theoretical background. *International Journal of Engineering Research and Applications*, 3(5), 605–610.