ISSN (E): 2181-4570

ACCURATE BREAST CANCER CLASSIFICATION BY USING ARTIFICIAL INTELLIGENCE ALGORITHMS

Meliboyev Abdulaziz

Lecturer of Digital technology and mathematics department, Kokand University.

Abstract: Breast cancer is a significant cause of mortality for women worldwide, ranking as the second leading cause of death. In 2018, breast cancer accounted for the highest number of cancer-related deaths among women in 40 European countries. While it ranked as the second leading cause of cancer-related deaths in the EU-28, lung cancer held the top position. Detecting breast cancer at an early stage is vital for improving treatment outcomes and survival rates. Data mining has gained popularity as an effective tool for knowledge discovery in various fields, including medicine. Researchers have applied machine learning techniques, such as multiple classifier algorithms, to predict and analyze patient diagnoses using medical datasets. However, a challenge arises due to imbalanced training data, where the probability of not having the disease is higher than having it. This paper focuses on addressing this issue by comparing two distinct AI models: Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). The objective is to develop a suitable and reliable model capable of handling imbalanced datasets and missing values, thereby enhancing the overall performance of the breast cancer prediction model.

Keywords: Machine learning (ML), Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Breast cancer.

Introduction

Breast cancer ranks as the second leading cause of death among women worldwide. In 2018, breast cancer was responsible for the highest number of cancer-related deaths among women, accounting for 138,000 cases in 40 European countries [1]. However, in the EU-28, it was identified as the second leading cause of cancer-related deaths, with lung cancer taking the top spot. Over the course of a decade, from 2002 to 2012, breast cancer mortality rates in the EU showed a decline from 17.9 cases per 100,000 individuals to 15.2 cases per 100,000 individuals. Detecting the disease at an early stage is crucial for improving treatment outcomes and survival rates. In various fields such as marketing, social science, finance, and medicine, data mining has gained popularity as an effective tool for knowledge discovery. Particularly in medicine,

ISSN (E): 2181-4570

researchers have employed multiple classifier algorithms on medical datasets to predict and analyze patient diagnoses. For instance, machine learning techniques have been utilized to evaluate tumor behavior in breast cancer patients. However, a challenge arises due to an imbalance in the training data, as the probability of not having the disease is higher than that of having it. This paper addresses this issue by comparing two distinct AI models: Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). The objective is to develop a suitable and reliable model that can handle the imbalanced dataset and missing values, thereby enhancing the overall performance of the model.

Related work

There have been numerous research studies on the implementation of machine learning (ML) in breast cancer detection and diagnosis, with the goal of enhancing accuracy and improving patient outcomes. Common ML algorithms used in these studies include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision trees, and Random forests. To increase accuracy, some studies have employed a combination of these algorithms in an ensemble method. Additionally, ML has been utilized for breast cancer risk prediction, personalized treatment recommendations, and predicting patient outcomes. Ongoing research in this field indicates that ML has demonstrated the potential to improve breast cancer detection and diagnosis.

Yue, W. [2] et al describes various machine learning (ML) methods and their uses in diagnosing and predicting outcomes for breast cancer based on data from the WBCD benchmark database. The ML techniques have demonstrated impressive abilities to enhance accuracy and prediction in classification. They presents multiple methods, along with their references, algorithms, sampling strategies, and classification accuracies, creating a user-friendly and easy-to-understand resource.

S.Gcet et al. [3] conducted research on feature extraction for breast cancer detection, focusing on variance, range, and compactness. They used SVM classification to assess performance and found that the highest accuracy rates were 95% for variance, 94% for range, and 86% for compactness. Their results suggest that SVM is a suitable method for detecting breast cancer. In terms of cancer types, breast cancer had the highest percentage (41%) among the types studied. Arika. R et al. [4] proposed that using of classifier-based prediction models that incorporate data mining and machine learning has been helpful in reducing diagnostic errors and increasing the efficiency of

JOURNAL OF UNIVERSAL SCIENCE RESEARCH

ISSN (E): 2181-4570

cancer diagnosis. Their models have contributed to the prompt and thorough analysis of medical data while also minimising the occurrence of human errors. Mohammed S. A. et al. [5] focus is on addressing imbalanced data that contain missing values through the use of resampling techniques in order to improve the accuracy of breast cancer detection through classification. In this study, they employed three classifier algorithms (J48, NB, and SMO) on two distinct datasets related to breast cancer. The current paper provides an introduction to both breast cancer and machine learning and includes a comprehensive literature review of existing ML methods utilized in breast cancer detection. Maryam T. et al. [6] proposed research that the most commonly employed method for cancer detection applications is SVM. SVM was utilized either independently or in combination with other methods in order to enhance performance. The highest level of accuracy achieved through the use of SVM (either singularly or in hybrid form) was 99.8%, with the potential for further improvement up to 100%.

Data Description and Exploratory Data Analysis

We utilized the Wisconsin breast cancer database (WBCD) dataset, which is widely used and was created by the University of Wisconsin Hospitals[7]. This dataset includes health information from 212 breast cancer patients that are malignant and 357 healthy individuals that are benign, comprising 32 independent variables such as radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean and so on. The dataset also includes a target variable that is encoded as “benign” for healthy controls and “malignant” for patients with breast cancer. The data is stored in a CSV file with 31 columns without the target column. The first coulmn represented the id and the second one stored target label. The rest columns indicating whether the breast cancer features. The dataset has 569 rows and is split between the patient and healthy control groups. The count and distribution of the target variable can be seen in Fig 1.

ISSN (E): 2181-4570

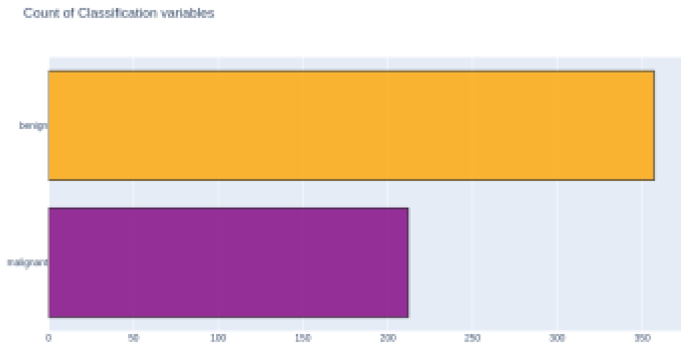


Figure 1

Experimental Results

This section focuses on the evaluation of multiple classification models using our dataset. A summary of the experiments conducted is presented in Table 1. The WBCD dataset was used to assess the performance of each model. The hyperparameters applied in the models include the "Adam" optimizer, the "binary cross entropy" loss function, and training for a total of 100 epochs. All models in this study were evaluated using two performance measures, namely Accuracy and Loss. These measures were utilized to assess the effectiveness and performance of the models in the classification task.

The results outlined in Table 1 highlight the dependable performance of both Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) in terms of accuracy, training loss, and validation accuracy.

In order to assess the performance of the two artificial intelligence (AI) models, distinct architectural configurations were employed, employing the well-known Wisconsin Breast Cancer Diagnosis (WBCD) dataset. The outcomes of this comparative analysis reveal that the Convolutional Neural Network (CNN) model yielded the highest accuracy rate, achieving an impressive score of 99.12%. Conversely, the Artificial Neural Network (ANN) model demonstrated a slightly lower accuracy rate, achieving 98.17%.

In addition to the accuracy metrics, the performance of both models was evaluated in terms of their validation accuracy and loss. The ANN model exhibited a commendable validation accuracy of 97.7%, indicating its ability to effectively generalize its predictions beyond the training dataset. Conversely, the CNN model

ISSN (E): 2181-4570

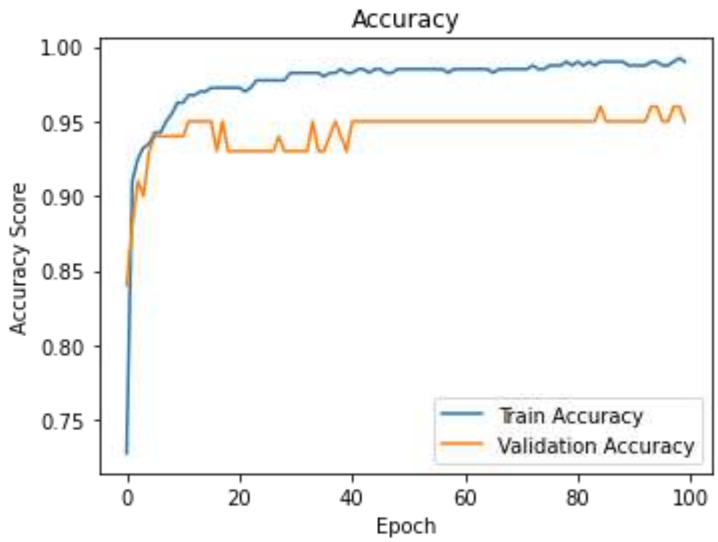
attained a validation accuracy of 95.1%, denoting a marginally lower but still substantial level of generalization capability.

Considering the loss metric, the ANN model incurred a loss of 25.1%. This value quantifies the discrepancy between the predicted and actual values, suggesting that the model experienced a relatively higher level of error during its training process. In contrast, the CNN model achieved a significantly lower loss of 4%, implying a superior ability to minimize prediction errors and align the predicted outputs with the ground truth labels.

To summarize, the experimental evaluation of the two AI models on the WBCD dataset revealed noteworthy performance disparities. The CNN model emerged as the superior choice, exhibiting a remarkable accuracy rate of 99.12%. Furthermore, the validation accuracy and loss metrics provided valuable insights into the models' generalization capabilities and predictive accuracy. Specifically, the ANN model demonstrated a validation accuracy of 97.7% and a loss of 25.1%, whereas the CNN model showcased a validation accuracy of 95.1% and a loss of 4%.

Model	Accuracy	Loss	Validation accuracy	Validation Loss
ANN	0.98	0.13	0.97	0.25
CNN	0.99	0.03	0.95	0.04

Table 1.



ISSN (E): 2181-4570

Figure 2. CNN learning curve

Our model's accuracy and validation curve is demonstrated in Figure 2. After 20 epochs, the training model reached approximately 97% and was almost stable. Moreover, validation accuracy

Figure 3 illustrates the Loss curve of the Convolutional Neural Network (CNN) model. The depicted graph demonstrates a gradual reduction in the loss value, converging to approximately 0.03. This diminishing trend signifies the efficacy and reliability of the CNN model in its ability to minimize the discrepancy between predicted and actual values.

The observed descent in the loss value, denoted by the decreasing trajectory of the curve, is indicative of the CNN's progressive improvement in capturing and representing the underlying patterns and features within the dataset. As the CNN model undergoes training iterations, it successfully optimizes its internal parameters to minimize the disparity between the predicted output and the ground truth labels, leading to the aforementioned reduction in loss.

A loss value of 0.03 signifies a relatively low level of dissimilarity between the model's predictions and the true values. Consequently, this outcome implies that the CNN model has attained a high degree of accuracy and precision in its predictive capabilities, effectively approximating the target values with a high level of fidelity.

In summary, the Loss curve depicted in Figure 3 showcases the CNN model's proficiency in reducing the loss value, ultimately converging to an impressively low value of 0.03. This outcome serves as a testament to the reliability and efficacy of the CNN model, affirming its capability to accurately capture and represent the underlying patterns within the given dataset.

ISSN (E): 2181-4570

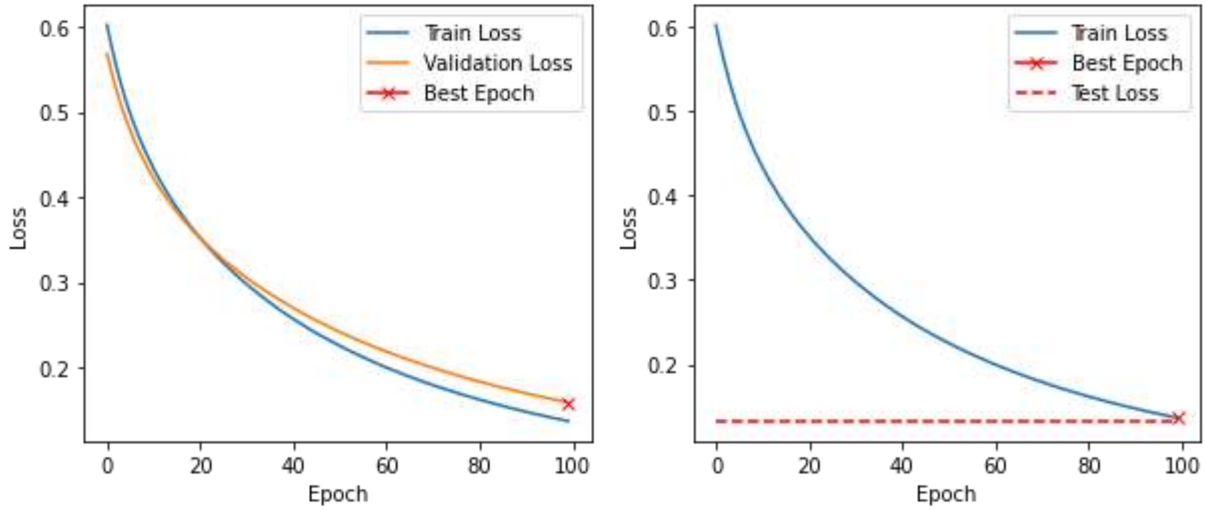


Figure 3.

Conclusion

Breast cancer is recognized as a prominent cause of mortality among women, emphasizing the importance of early detection for saving lives. Modern AI algorithms have proven valuable in breast cancer detection. The study use two AI algorithms, namely ANN and CNN, on breast cancer datasets. This research paper specifically addresses the challenge of CNN and ANN models for WBCD structured data for breast cancer classification. The results highlight the effectiveness of these classifiers in enhancing the detection accuracy of breast cancer. Furthermore, future work will expand these experiments to encompass different classifiers and diverse datasets.

References:

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. *Eur J Cancer.* 2018 Nov;103:356–87.

2. Yue, Wenbin, et al. "Machine learning with applications in breast cancer diagnosis and prognosis." *Designs* 2.2 (2018): 13.

3. Gc, Sailesh, et al. "Variability measurement for breast cancer classification of mammographic masses." *Proceedings of the 2015 Conference on research in adaptive and convergent systems*. 2015.

4. Arika, Rebecca Nyasuguta, and Agnes Mindila W. Cheruiyot. "Machine Learning Algorithms for Breast Cancer Diagnosis: Challenges, Prospects and Future Research Directions." (2023).

ISSN (E): 2181-4570

5. Mohammed, Siham A., et al. "Analysis of breast cancer detection using different machine learning techniques." *Data Mining and Big Data: 5th International Conference, DMBD 2020, Belgrade, Serbia, July 14–20, 2020, Proceedings 5*. Springer Singapore, 2020.

6. Tahmooresi, Maryam, et al. "Early detection of breast cancer using machine learning techniques." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10.3-2 (2018): 21-27.

7. Wolberg, William H., W. Nick Street, and Olvi L. Mangasarian. "Breast cancer Wisconsin (diagnostic) data set." *UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/]* (1992). DATASET. Online available: https://www.kaggle.com/datasets/ibrahimkaratas/breast-cancer-wisconsin-diagnostic