ESI Preprints **Not Peer-reviewed**

Neural Network Approaches for Early Breast Cancer Detection

Ala'a R. Al-Shamasneh
Nadia Alabdulkarim
Maya Ahmad
Dept. of Computer Science, Prince Sultan University, KSA

Doi: 10.19044/esipreprint.3.2024.p296

Approved: 09 March 2024
Posted: 11 March 2024

Cite As:
Al-Shamasneh A.R., Alabdulkarim N. & Ahmad M. (2024). *Neural Network Approaches for Early Breast Cancer Detection*. ESI Preprints.
https://doi.org/10.19044/esipreprint.3.2024.p296

Abstract

Research on breast cancer is crucial due to its significant impact on public health, with high mortality rates underscoring the urgency for improved diagnostic methods. Early detection plays a pivotal role in enhancing treatment outcomes and reducing mortality rates. This paper addresses the pressing need for more effective early disease detection methods, particularly focusing on breast cancer diagnosis. It proposes the utilization of neural networking techniques, known for their potential to enhance accuracy and efficiency in cancer diagnosis. The study aims to provide a comprehensive overview of breast cancer detection using neural networking, emphasizing its significance in improving patient outcomes. By showcasing the effectiveness of neural network approaches, the research contributes to advancing early cancer detection efforts, aligning with global health initiatives prioritizing early diagnosis and intervention.

Keywords: Breast Cancer, Neural Networks, Early Detection

Introduction

Detecting cancer at late stages remains a pervasive challenge in the medical field worldwide, often leading to critical health conditions and limited treatment options for patients. The urgency for timely diagnosis to prevent life-threatening situations underscores the importance of leveraging

technology to save lives. This paper explores the potential of neural networking, an artificial intelligence technique, in revolutionizing breast cancer detection and improving patient outcomes.

Neural networking, a key component of artificial intelligence, mimics the human brain's ability to process data through interconnected neurons or nodes arranged in layered structures. By utilizing deep learning principles, neural networks can continuously learn and enhance their ability to interpret data accurately over time. This approach holds promise for transforming cancer detection by providing faster, more accurate, and reliable results.

While mammogram screening has been a conventional method for breast cancer detection, recent advancements in technology have opened new avenues for leveraging artificial intelligence and neural networking. The integration of these modern techniques aims to augment existing diagnostic capabilities, enabling early detection and intervention for women worldwide. By harnessing the power of artificial intelligence, medical professionals can more effectively identify cancerous lesions and prioritize high-risk patients.

Although artificial intelligence models, such as those used in prostate cancer detection, offer promising results in automating cancer detection processes, they are not intended to replace human expertise. Instead, these technologies serve as valuable tools to assist oncologists and radiologists in identifying cancerous abnormalities and distinguishing them from benign masses. The inherent limitations of current detection methods, including false positives and missed diagnoses, highlight the need for more advanced and accurate approaches.

The application of neural networking in breast cancer detection holds immense potential for improving imaging speed, accuracy, and prognostic capabilities. This paper aims to present a comprehensive overview of the neural networking method, utilizing MATLAB as a powerful computational tool. MATLAB's specialized functionalities and toolboxes facilitate the analysis of large datasets, making it an ideal platform for integrating medical expertise with modern artificial neural network technology.

By combining insights from the medical field with cutting-edge artificial intelligence techniques, this research endeavors to contribute to the early detection of breast cancer. Ultimately, the goal is to empower healthcare professionals with innovative tools and methodologies to aid in the timely diagnosis and treatment of this life-threatening disease.

Literature Review

Breast cancer remains a significant global health concern, prompting ongoing research into more effective diagnostic methods. Recent studies have emphasized the importance of early detection in improving patient outcomes and reducing mortality rates (Duffy et al., 2020). Mammogram

screening has traditionally been the primary method for breast cancer detection, but its limitations, including false positives and missed diagnoses, have led to the exploration of alternative approaches (Yala et al., 2019).

Artificial intelligence (AI) and machine learning (ML) techniques have emerged as promising tools for enhancing cancer detection and diagnosis. Neural networking, a subset of AI, utilizes deep learning algorithms to analyze complex data patterns, mimicking the human brain's processing capabilities (Esteva et al., 2019). Neural networks have shown remarkable success in various medical imaging tasks, including the detection of breast cancer from mammograms (Rodriguez-Ruiz et al., 2019).

Recent advancements in AI-based breast cancer detection have demonstrated significant improvements in accuracy and efficiency. For example, deep learning models trained on large-scale datasets have achieved impressive results in identifying malignant lesions and reducing false positives (McKinney et al., 2020). These models leverage convolutional neural networks (CNNs) to extract relevant features from mammogram images and classify them with high precision (Wang et al., 2021).

Moreover, the integration of AI technologies into clinical practice has shown promise in streamlining diagnostic workflows and improving radiologists' productivity. AI-assisted interpretation tools, such as computer-aided detection (CAD) systems, provide radiologists with real-time feedback and decision support, enhancing their diagnostic accuracy and confidence (Rodriguez-Ruiz et al., 2020).

Despite the advancements, challenges remain in deploying AI-based breast cancer detection systems in real-world clinical settings. Issues such as data privacy, model interpretability, and regulatory compliance pose significant barriers to adoption (Liu et al., 2021). Additionally, ensuring the equitable distribution and accessibility of AI technologies across diverse patient populations is essential to mitigate healthcare disparities (Ward et al., 2020).

In conclusion, recent literature underscores the transformative potential of AI and neural networking in revolutionizing breast cancer detection. By leveraging advanced computational techniques and large-scale datasets, researchers and clinicians are poised to enhance early diagnosis and improve patient outcomes in the fight against breast cancer.

Methods

Below, we delve into the neural network approach facilitated by MATLAB, offering specialized toolboxes tailored for neural networks, leading to expedited, precise, and efficient outcomes.

A neural network comprises an input layer, one or more hidden layers, and an output layer, with numerous nodes or neurons in each layer.

These nodes receive inputs from every node in the preceding layer, establishing interconnectedness throughout the network. Modeled after the human brain, neural networks learn through this layered structure, enabling pattern recognition, data classification, and predictive analysis based on learned data.

Similar to human learning, neural networks can discern patterns in various data types, such as voice or images, through iterative training. This learning process involves adjusting connection weights between neurons based on a predefined learning rule until the network achieves desired performance. MATLAB simplifies neural network development, allowing users to create models and implement pattern recognition with minimal expertise.

MATLAB facilitates rapid neural network model creation, visualization, integration into existing applications, and deployment across diverse platforms. When developing artificial intelligence applications, particularly neural networks, common steps include data preparation, AI modeling, simulation, testing, and deployment.

A. Data Preparation:

In this stage, we gather a suitable dataset with labels and ensure its quality through human examination. The neural network then undergoes training using this dataset, with simulations generating the training process. Additional data may be incorporated into the training set to capture a broader range of patterns and variations.

B. AI Modeling:

When comparing the effectiveness of shallow neural networks with traditional machine learning methods like decision trees or SVMs, or when dealing with limited labeled training data, it's advisable to utilize the command-line functions available in the Classification and Regression Learner tool from the Statistics and Machine Learning Toolbox™. This toolbox is well-suited for neural networks, offering comprehensive support for specifying, training, and customizing neural network models. Additionally, it facilitates interoperability, allowing seamless sharing and utilization of information across different computer systems or applications.

C. Simulation and Testing:

This stage facilitates the integration of intricate systems, such as neural networks, and includes testing, simulation for larger systems, and validation and verification of the system's functionality.

D. Deployment:

In this report, pattern recognition will be employed for implementation instead of traditional coding methods.Overall, the methodology outlined in this report will encompass the following steps, which will be elaborated upon further:

A. Data preprocessing and classification.

B. Application of Artificial Neural Network (ANN) pattern recognition using MATLAB.

C. Results and discussion.

The tool utilized in this study to enable a neural network to identify signs and patterns of breast cancer is MATLAB. Within MATLAB's algorithm, there are several tools available, including the neural network pattern recognition (ANN classification pattern recognition) application. This application assists in detecting malignant tumors, utilizing data sets sourced from The UCI Machine Learning Repository.

According to MATLAB, the Artificial Neural Network (ANN) pattern recognition process involves training the neural network to classify data sets into specific categories, thus addressing classification challenges. Users can import data sets from various applications such as Excel or any other compatible data file recognized by MATLAB. Subsequently, the ANN is trained, validated, and tested, with results displaying the error margin and accuracy of the application.

The following subsections of the methodology detail the steps and procedures involved in utilizing pattern recognition for classifying two data sets: distinguishing between malignant and benign tumors, and identifying blood diagnosis data from breast cancer patients versus healthy individuals.

A. Data Pre-processing and Classification:

The initial phase involved gathering the necessary data for utilizing Artificial Neural Networks (ANN) in breast cancer detection. Data sets were acquired from the UCI Machine Learning Repository, with all preprocessing and organization documented and attached.

The primary dataset chosen from the repository is the "Breast Cancer Coimbra Data Set." This dataset comprises 10 attributes and a total of 116 instances of individuals' laboratory test results. Among these instances, 64 corresponded to breast cancer patients, while 52 were from individuals without breast cancer. The data was presented in an Excel table format, with column headers indicating attributes such as Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, and a label (1 for healthy controls, 2 for patients)

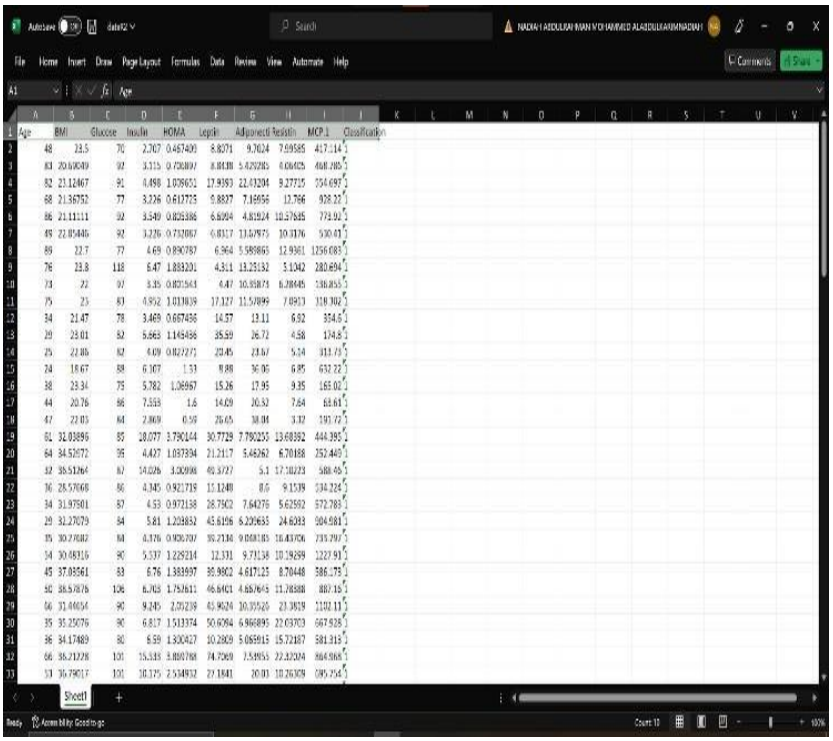


Figure 1. Breast Cancer Coimbra Data Set before preprocessing

To prepare the data for use in a neural network pattern recognition application, it needs to be structured into inputs and target outputs. This involves creating two new Excel sheets, which will be utilized in MATLAB.

The first sheet, named "Input-BreastCancer," includes the laboratory results of individuals, excluding age and label information. Subsequently, the rows of the dataset are shuffled to separate patients from healthy individuals, enabling further testing of the pattern recognition capabilities of the ANN in classification tasks.

Following this, the second Excel sheet, named "Output-TargetBreastcancer," is created to contain the labels. However, the labels are modified from 1 (indicating a healthy individual) and 2 (indicating a breast cancer patient) to 0 (representing a healthy individual) and 1 (representing a breast cancer patient). This adjustment makes the data readable to the pattern recognition program's target output, allowing it to recognize these numbers as a pattern. This completes the processing of the Coimbra dataset.

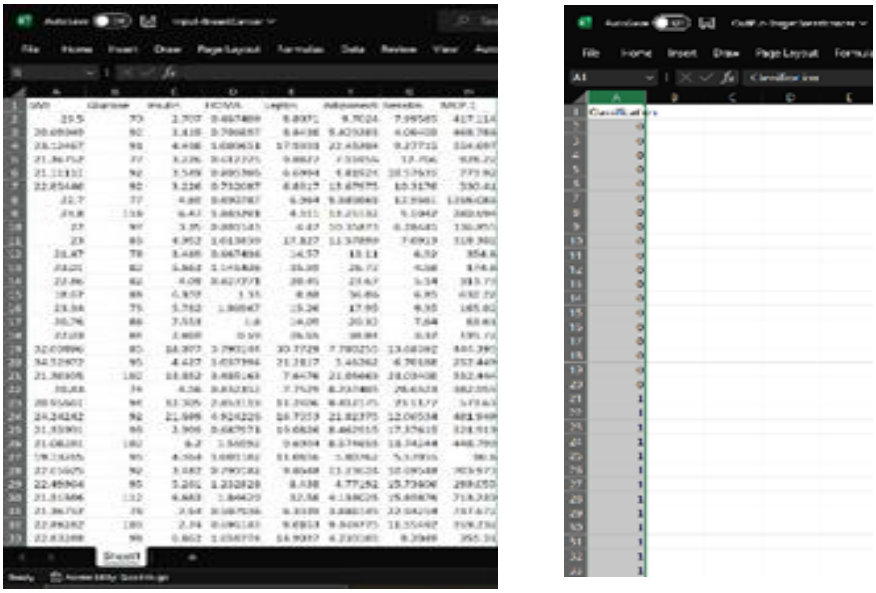


Figure 2. Breast Cancer Coimbra Data Set after processing, divided into two datasets of input and output

The second dataset chosen from the repository is the "Breast Cancer Wisconsin (Original) Data Set." This dataset comprises 11 attributes and a total of 699 instances of tumor test results. Among these tumors, 458 were classified as benign and 241 as malignant.The data is provided in the form of a .DATA file, which can be imported into Excel for further processing. Once imported, the data is formatted to default text with no formatted cells. Subsequently, attention is turned to the column headers, which include attributes such as Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class (2 for benign and 4 for malignant).

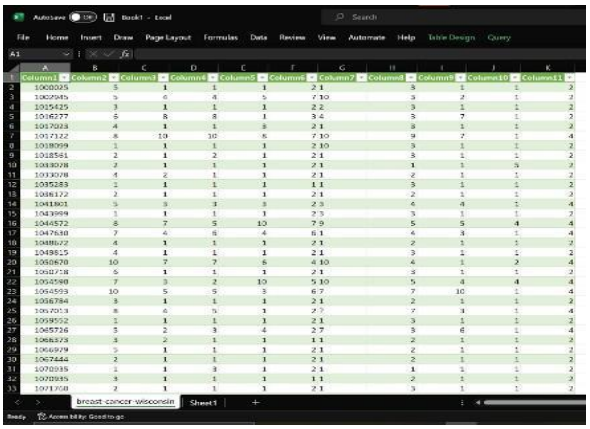


Figure 3. Breast Cancer Wisconsin Data Set before pre processing

To prepare this data for import into the neural network pattern recognition application, it needs to be structured into inputs and target outputs. This requires the creation of two new Excel sheets. The first sheet, named "WisconsinBreastCancerIN," includes all the tumor's specification attributes except for Sample code number and Class.

Following this, a second sheet called "WisconsinBreastCancerOUT" was generated. This sheet contains the class information, but it was transformed from 2 for benign and 4 for malignant to 0 for benign and 1 for malignant. This adjustment ensures compatibility with the pattern recognition program's target output. Thus, the processing and organization of the Wisconsin dataset are concluded.

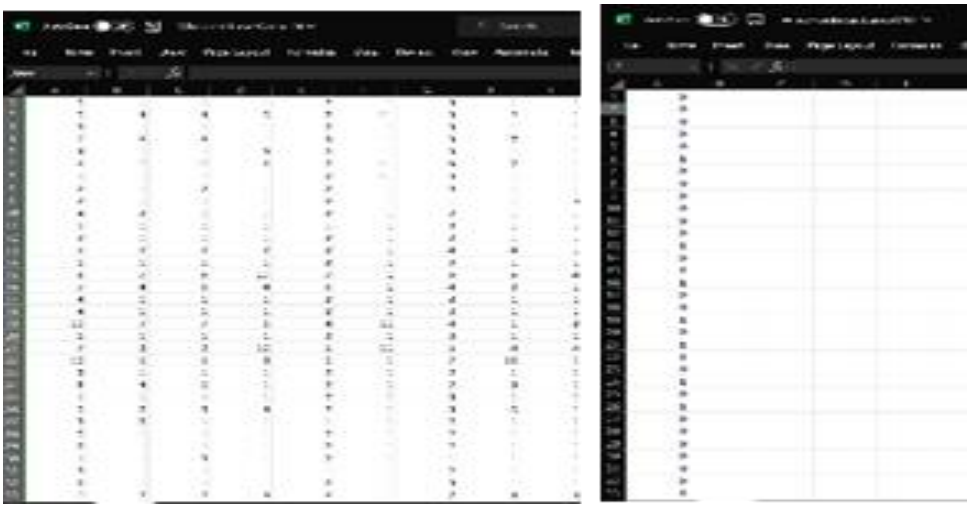


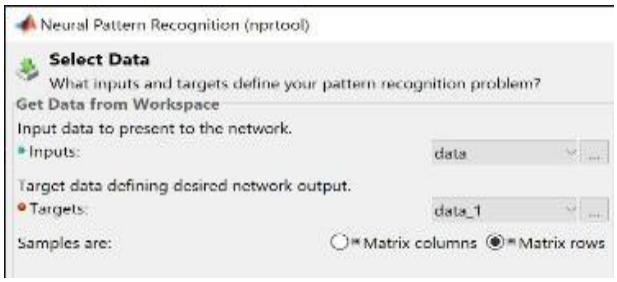
Figure 4. Breast Cancer Wisconsin Data Set after processing, divided into two datasets of input and output

Following the data preprocessing phase, the next step involves classifying the processed data using the neural network pattern recognition algorithm, as outlined in the forthcoming subsection.

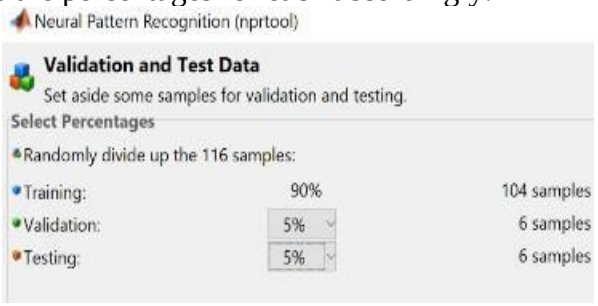
B. ANN Pattern Recognition Application in MATLAB

As previously explained in the methodology introduction, the ANN pattern recognition process entails categorizing datasets into distinct categories.

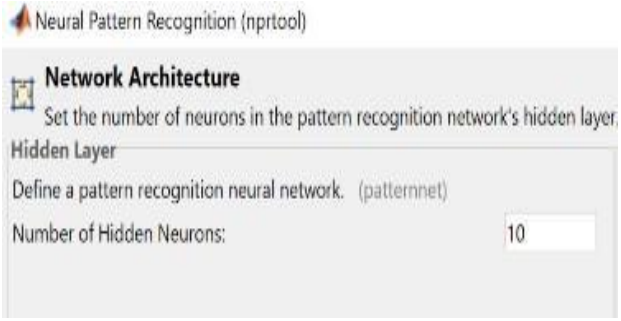
To elaborate further, the processed data, now labeled as input data and output data, are imported into MATLAB. The input data corresponds to the input, while the target output corresponds to the desired output. The illustration below demonstrates how the Coimbra breast cancer datasets are entered into MATLAB as an example:



For this research, it's important to note that all target outputs in the datasets denote 1 for breast cancer patients or malignant tumors, and 0 for healthy individuals or benign tumors.Subsequently, the datasets are randomized to train the ANN, as well as for validation and testing purposes. The user selects the percentages for each accordingly.



Afterwards, the hidden layers for the artificial neural network are entered as well:

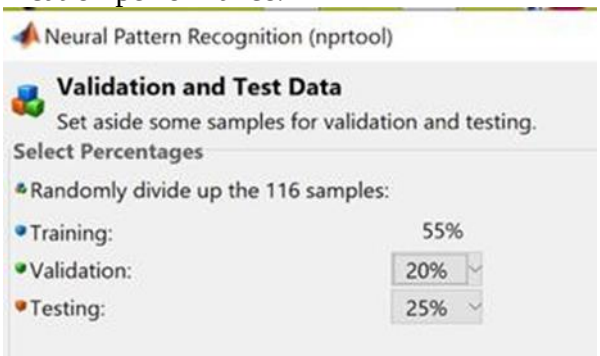


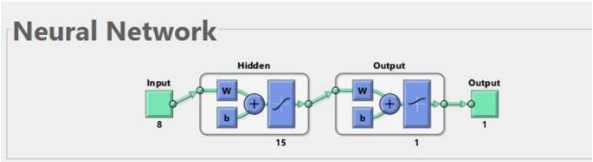
Lastly, the neural network training commences, yielding results including data, graphs, error percentages, and more upon completion. If the training results appear inadequate, adjusting the number of hidden layers can improve classification accuracy and reduce errors. Additionally, modifying the distribution of training, validation, and testing datasets can impact results positively. In the final segment of the methodology, the outcomes of the neural network training for classifying the Coimbra and Wisconsin datasets will be presented and scrutinized. The aim is to assess whether the ANN's pattern recognition successfully

identifies whether a tumor is benign or malignant in the case of the Wisconsin dataset, or discerns a breast cancer patient from a healthy individual in the Coimbra dataset.

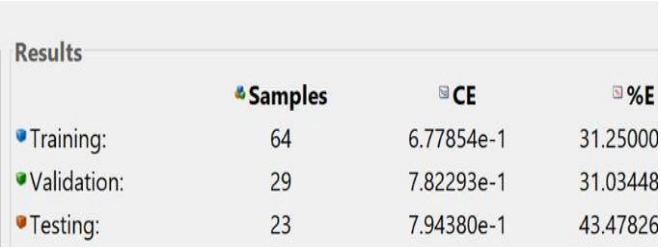
Results And Discussion

The Coimbra Breast cancer dataset was partitioned into 55% for training, 20% for validation, and the remaining 25% for testing purposes. Additionally, the number of hidden layers was adjusted to 15 to optimize the program's classification performance.





The result showed the following:



Before delving into the analysis, it's important to reiterate the ultimate objective of this pattern recognition endeavor: to distinguish between laboratory results belonging to breast cancer patients and those of healthy individuals within the Coimbra dataset. As depicted above, the dataset comprising 116 samples has been divided into 64 for training, 29 for validation, and 23 for testing. To gain deeper insights into the classification performance, it's crucial to examine the results of CE (cross-entropy) and

%E (percentage error). Below, we will discuss the analysis of training performance, training state, error histogram, confusion matrix, and ROC plot

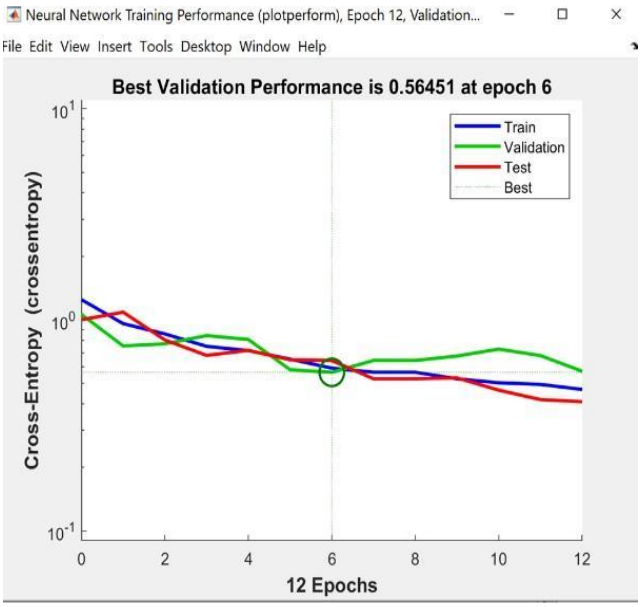


Figure 5. NN training performance plot

The plot above reveals that the validation achieved its peak performance at epoch 6, with a value of approximately 0.6. Additionally, a gradual decline in training errors suggests a reduction in errors over time. Moreover, the consistent decrease in the test line reaffirms the accuracy of the program's pattern recognition, indicating successful detection of patterns.

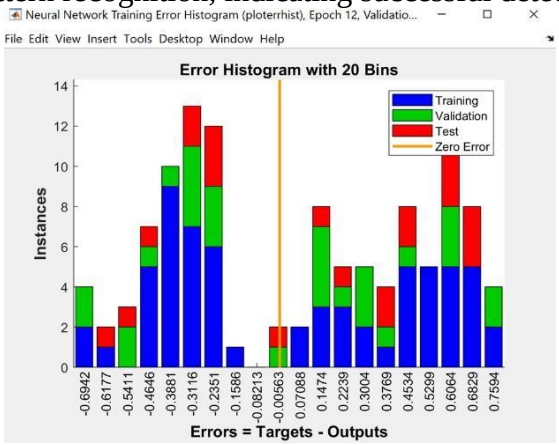


Figure 6. Error Histogram

With 20 bars or bins displayed in the histogram, it becomes evident that a zero error is observed around the -0.006 mark. These error histograms

provide insight into the discrepancies between our target and predicted digits post-training of our neural network. The remarkably low error percentage of 0.006 underscores the high accuracy of our pattern recognition in detection, indicating minimal errors in the process.

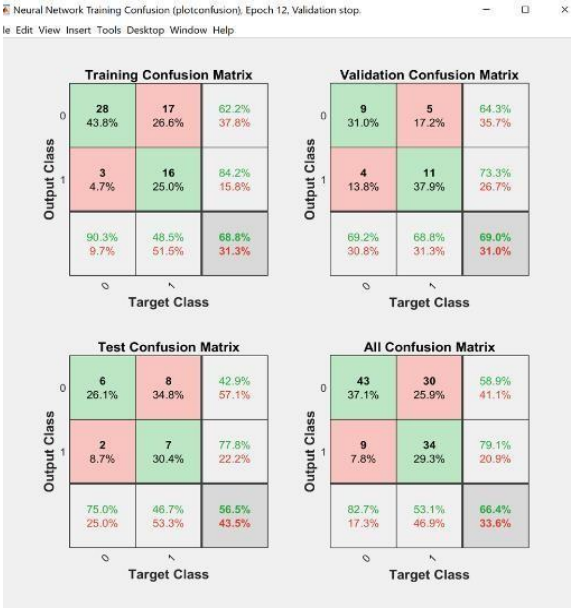


Figure 7. Confusion Matrix

The presented confusion matrix illustrates the breakdown of classifications for training, validation, test, and a combined matrix, respectively. Focusing on the training matrix, it reveals that the pattern recognition accurately identified 28 attributes as 0, signifying healthy individuals, and 16 attributes as breast cancer patients. The training's accuracy stands at approximately 70%, with the error rate around 30%.

When considering the combined confusion matrix encompassing training, validation, and testing, the neural network demonstrated a notable accuracy level. Specifically, it correctly identified 43 attributes representing healthy individuals while misclassifying 30 laboratory attributes belonging to breast cancer patients as healthy. Additionally, it accurately identified 34 laboratory attributes as indicative of breast cancer patients, yet erroneously categorized 9 healthy individuals as having breast cancer. Overall, the neural network achieved an accuracy rate of approximately 66%, with an error percentage of 33%. This suggests a reliable performance, as it surpasses the 50% threshold.

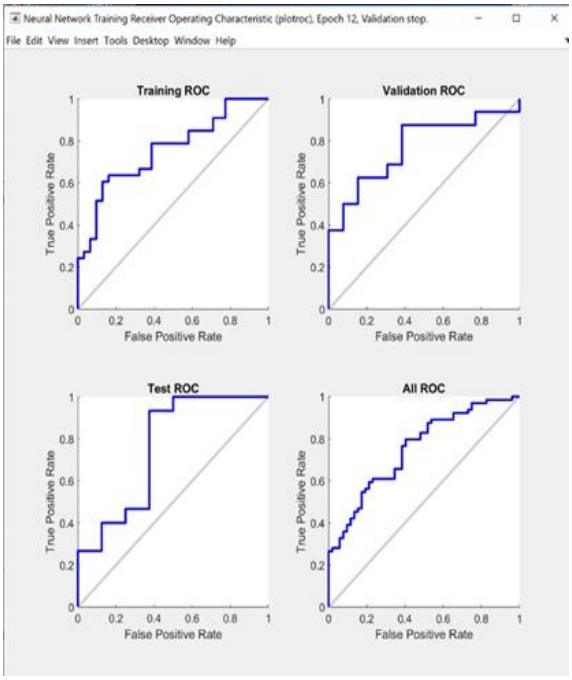


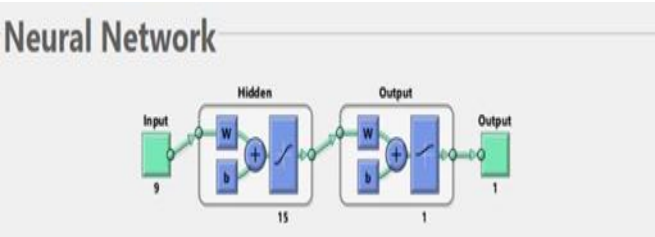
Figure 8. Receiver Operating Characteristic (ROC) Plot

From the ROC curves presented above, it's evident that the neural network's detection accuracy is steadily improving. The positive incline of the ROC curve, particularly noticeable when it surpasses the linear line in the center, indicates an enhancement in accuracy. Notably, when the ROC curve consistently lies above the linear line across all four graphs, especially in the combined ALL ROC, it underscores the neural network's proficiency in distinguishing between healthy individuals and breast cancer patients based on laboratory results.

In conclusion, the analysis of the Coimbra dataset results confirms the neural network's capability to accurately detect relevant patterns in the provided information.

Result and Discussion of Data and Graphs for Wisconsin:

Similar to the Coimbra dataset, the Wisconsin Breast Cancer dataset was divided into training, validation, and testing sets, with 55% allocated for training, 20% for validation, and 25% for testing. The neural network architecture included 15 hidden layers to optimize its performance in classification tasks, mirroring the setup used for the Coimbra dataset.



The result showed the following:

Results

	Samples	CE	%E
Training:	384	1.63032e-0	2.60416e-0
Validation:	175	2.74087e-0	3.42857e-0
Testing:	140	2.88176e-0	3.57142e-0

As done in the Coimbra Dataset, a quick reminder of the following before discussing the analysis, the end goal of this pattern recognition is to detect which of the tumor results in the Wisconsin dataset is a malignant tumor and which is a benign tumor. From the above results, the sample of 699 has been separated into 384 for training, 175 for validation, and 140 for testing.

Again, to further understand the show results of CE and %E, An analysis of the training performance, training state, error histogram, confusion matrix, and ROC plot will be discussed below.

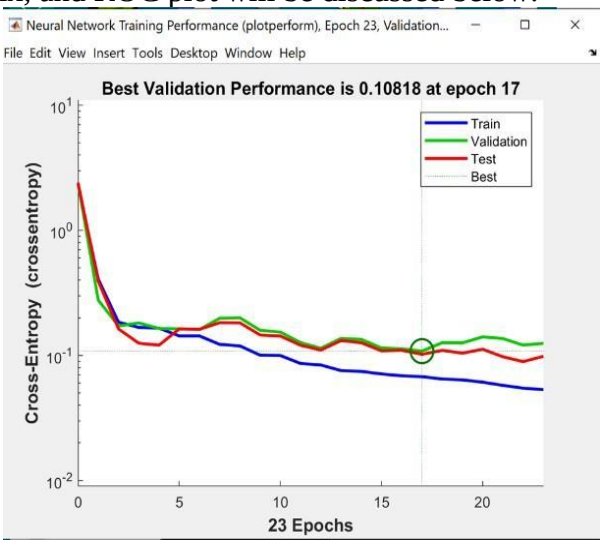


Figure 9. NN training performance plot

The graph above shows that the best performance of the validation is at epoch 17 with the value of about 0.11, as well as it can be shown that the

training is also at a decrease, meaning that less errors are occurring. Finally, to further harden that the detection of the program is majorly correct and accurate, we can see that the test line, just like in the Coimbra dataset result, has also been at a steady decrease line, which further indicates that yes, the pattern is being recognized correctly.

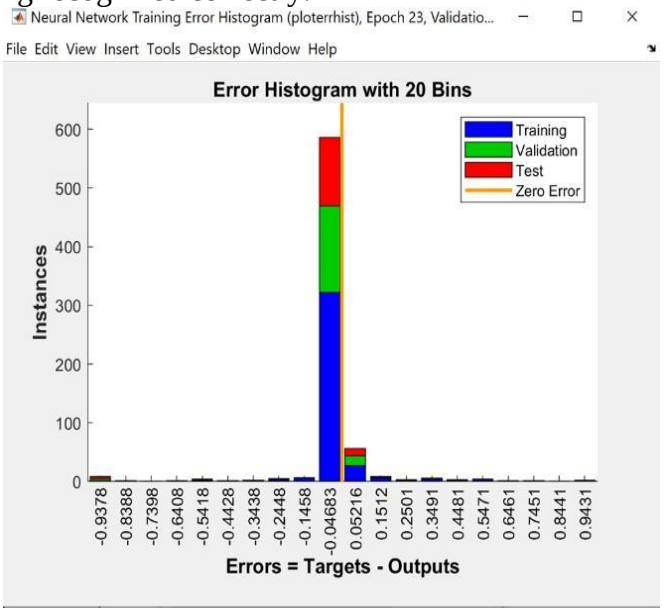


Figure 10. Error Histogram

The histogram contains 20 bars or bins, its results signify that there is a zero error laying at about - 0.01 since it's between -0.05 and 0.05 and it leans a bit to the left. As stated before , the error histograms aids in allowing the user to witness visually the errors between our target and our predicted digits after training the neural network, so with our error percentage being as low as 0.01, indicates that our pattern recognition in the detection has a high accuracy due to its low error rate.

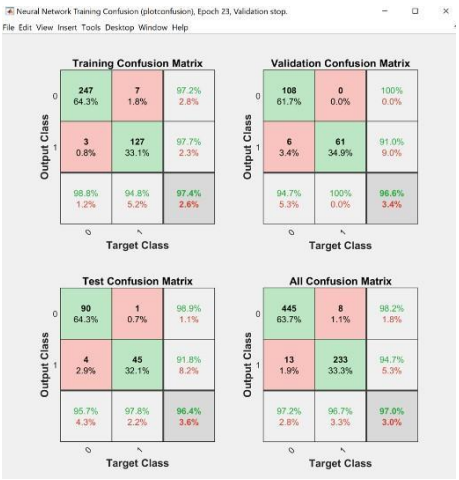


Figure 11. Confusion Matrix

Reading one of the matrices of the Wisconsin Breast cancer data set that is training, we can witness that the neural network recognition has correctly detected 257 attributes to be 0 , indicating benign tumors, and 127 attributes indicating to malignant tumors, with the training's accuracy is at an extremely high 97% with the error being as low as 3% , indicating high and accurate training that solidifies our ANN's training capabilities of detecting tumors. When we see the total confusion matrix, which again includes the validation test and training combined together, we can see that the neural network has correctly detected 445 tumor attributes that show to be benign tumors whilst mis detecting 8 tumor attributes that belong to malignant tumors as benign tumors, and has correctly detected 233 tumor attributes that show malignant tumors while mis detecting 13 of those to be malignant even though it is a benign tumor. It comes down to show that the total accuracy of this neural network is at a correct accuracy of 97% , and has an extremely low error percentage matrix of 3%, which shows that our ANN was more than capable of accurately distinguishing malignant tumor patterns from benign without having too many mistakes.

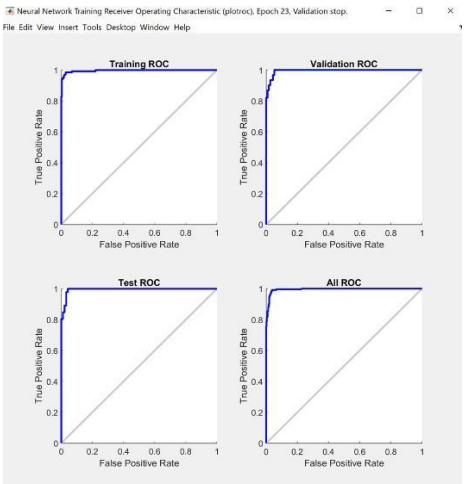


Figure 12. Receiver Operating Characteristic (ROC) Plot

Finally in terms of data analysis of Wisconsin, we can see from the ROC above, the curve is , again shown to be a positive incline above that of the linear line in the center, just like the Coimbra dataset , then it shows that the neural network is yet again beginning to increase in terms of detection accuracy. Since the curve is a positive incline above the line at the combined ALL ROC, this solidifies that the artificial neural network pattern recognition application is detecting the malignant tumors from the benign accurately, without much fault nor errors.

Thus, this concludes the analysis of the Wisconsin dataset results and its discussion. We finalize that yes, just like the Coimbra analysis, neural network pattern recognition can also detect quite accurately when it comes to the given information of the tumor attributes, thus concluding our results and analysis, as well as our methodology.

Conclusion

In summary, MATLAB has demonstrated its effectiveness in early breast cancer detection through neural network-based pattern recognition, effectively addressing classification challenges. The analysis conducted on the Coimbra breast cancer dataset showcased the program's ability to minimize errors while accurately recognizing patterns. Various metrics, including training performance, error histograms, confusion matrices, and ROC plots, confirmed its reliability in detecting breast cancer accurately.

Similarly, when applied to the Wisconsin Breast Cancer dataset, MATLAB exhibited high accuracy in pattern recognition and detection due to its low error rate. These findings underscore the significance of employing artificial neural networks and pattern recognition techniques in aiding healthcare professionals, such as doctors and radiologists, in diagnosing

breast cancer at early stages. MATLAB serves as a testament to how modern technology continues to support and enhance breast cancer detection efforts, benefiting the broader breast cancer community.

Conflict of Interest: The authors reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The authors did not obtain any funding for this research.

References:

1. Duffy, S. W., Parmar, D., Jatoi, I., & Miller, A. B. (2020). Fundamental Issues in Breast Cancer Screening. Journal of Clinical Oncology, 38(15), 1637–1643.
2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2019). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.
3. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., & Toma, M. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94.
4. Rodriguez-Ruiz, A., Krupinski, E., Mordang, J. J., Schilling, K., Heywang-Köbrunner, S. H., & Sechopoulos, I. (2019). Detection of breast cancer with mammography: Effect of an artificial intelligence support system. Radiology, 290(2), 305–314.
5. Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Broeders, M., Gennaro, G., Clauser, P., & Mann, R. M. (2020). Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. Journal of the National Cancer Institute, 112(9), 916–922.
6. Wang, T., Jing, C., Ma, Y., Lai, M., Zhang, Y., & He, Z. (2021). Convolutional neural networks in breast cancer screening: A comprehensive review. Frontiers in Medicine, 8, 664720.
7. Liu, Y., Chen, P.-H. C., Krause, J., Peng, L., LeCun, Y., & Deng, J. (2021). How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA, 325(6), 519–529.
8. Ward, R., Deaton, A., El-Sayed, A. M., & Gray, J. (2020). Data Science, Artificial Intelligence, and Machine Learning: Opportunities for Advancement in Breast Cancer Research. JAMA Oncology, 6(10), 1521–1522.