African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 4, 2026
Page No: 264-277
Website: *https://aaasjournals.com/index.php/ajapas/index*
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي 2025: 1.65**

A Deep Learning-Based Artificial Neural Network for Breast Cancer Classification: A Comprehensive Performance Evaluation

Ayad Saad Almryad [1], Elkhouly A. R. [2*], Almabrouk Alhireereeq [3]
[1] Department of Computer, Faculty of Science-Regdalin, Sabratha University, Regdalin, Libya
[2] Department of Zoology, Faculty of Science-Regdalin, Sabratha University, Regdalin, Libya
[3] Department of Zoology, Faculty of Science-Ojelat, Alzawia University, Ojelat, Libya

شبكة عصبية اصطناعية تعتمد على التعلم العميق لتصنيف سرطان الثدي: تقييم شامل للأداء

عياد سعد المريض [1]، الأنصاري رفعت الخولى [2*]، المبروك أحمد الحريريق [3]
[1] قسم الحاسوب، كلية العلوم رقدالين، جامعة صبراتة، رقدالين، ليبيا
[2] قسم علم الحيوان، كلية العلوم رقدالين، جامعة صبراتة، رقدالين، ليبيا
[3] قسم علم الحيوان، كلية العلوم العجيلات، جامعة الزاوية، العجيلات، ليبيا

[*]Corresponding author: elansary.elkhouly@sabu.edu.ly

Received: August 05, 2026	Accepted: October 16, 2026	Published: October 31, 2026

Abstract:
Accurate and prompt diagnosis of cancers is critical to improve patient outcomes and to make timely clinical decisions. The present study built a multilayer feedforward artificial neural network (ANN) deep learning model to identify between malignant and healthy instances. The model included 186 inputs, one output neuron and four hidden layers with 64, 32, 16 and 8 neurons each. The network was trained with Scaled Conjugate Gradient (SCG) technique. The cross-entropy loss function was minimized and early stopping was applied on validation performance. This strategy was designed to improve generality and reduce over-fitting. The developed model achieved an F1-score of 93.07%, an Area under the Receiver Operating Characteristic Curve (AUC) of 0.8045, and an overall classification accuracy of 87.07 percent. Together these metrics give a significant discrimination between malignant and normal samples. The model was validated using the Receiver Operating Characteristic (ROC) analysis where it was found that the model was confident in its categorization as the cancer patients always received a higher prediction probability than healthy instances. The training was successful and exhibited steady learning, converging in 69 epochs with a final cross-entropy loss of 0.295. The confusion matrix revealed that the model was highly sensitive to correctly identify most of the cancer cases. The considerably reduced specificity was mostly due to the considerable imbalance in the sample class. The learning curves were smooth and did not indicate any overfitting. Further error analysis indicated that performance was generally age-independent, and there was no appreciable drop in accuracy with age. To further illustrate the robustness of the neural network, a comprehensive performance dashboard was deployed to consolidate many evaluation measures, dataset characteristics, prediction distributions, and optimization behavior into a single assessment. The proposed artificial neural network provides a reliable framework with strong classification performance and feasible computational features for cancer prediction in general. Future studies should aim on enhancing the specificity, e.g. via enhanced feature selection, threshold optimization and class balance. The results to-date highlights the promise of deep learning for medical diagnostics and cancer risk prediction. The study highlights the need to apply the latest machine learning techniques to improve the quality of patient care and accuracy of cancer diagnoses.

Keywords: Neural networks, Prediction, Breast cancer.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

الملخص:
يُعد التشخيص الدقيق والسريع للأورام أمراً بالغ الأهمية لتحسين نتائج المرضى واتخاذ القرارات السريرية في الوقت المناسب. هدفت هذه الدراسة إلى بناء نموذج تعلم عميق يعتمد على شبكة عصبية اصطناعية تغذية أمامية متعددة الطبقات (ANN) للتمييز بين الحالات الخبيثة والسليم. تضمن النموذج 186 مدخلاً، وعصبون إخراج واحد، وأربع طبقات خفية تحتوي على 64 و32 و16 و8 عصبونات على التوالي. تم تدريب الشبكة باستخدام تقنية التدرج المقترن المقياس (SCG)، حيث تم تقليل دالة فقدان الانتروبيا المتبادلة وتطبيق التوقف المبكر بناءً على أداء التحقق. صُممت هذه الاستراتيجية لتحسين التعميم وتقليل فرط التخصيص. حقق النموذج المطوّر درجة F1 بلغت 93.07%، ومساحة تحت منحنى تشغيل المستقبل (AUC) بلغت 0.8045، ودقة تصنيف إجمالية بلغت 87.07%. تمنح هذه المقاييس معاً تمييزاً معتبراً بين العينات الخبيثة والطبيعية. تم التحقق من صحة النموذج باستخدام تحليل منحنى تشغيل المستقبل (ROC)، حيث وُجد أن النموذج كان واثقاً في تصنيفه، إذ حصل مرضى السرطان دائماً على احتمال تنبؤ أعلى مقارنة بالحالات السليمة. كان التدريب ناجحاً وأظهر تعلماً مستقراً، حيث تقارب في 69 دورة مع خسارة نهائية في الانتروبيا المتبادلة بلغت 0.295. كشفت مصفوفة الارتباك أن النموذج كان عالي الحساسية في التعرف بشكل صحيح على معظم حالات السرطان، بينما عُزيت الخصوصية المنخفضة نسبياً في الغالب إلى عدم التوازن الكبير في فئات العينات. كانت منحنيات التعلم سلسة ولم تُظهر أي فرط تخصيص. أشارت تحليل الأخطاء الإضافية إلى أن الأداء كان مستقلاً عن العمر بشكل عام، ولم تكن هناك انخفاض ملحوظ في الدقة مع تقدم العمر. ولتوضيح متانة الشبكة العصبية بشكل أكبر، تم نشر لوحة تحكم أداء شاملة لدمج العديد من مقاييس التقييم، وخصائص مجموعة البيانات، وتوزيعات التنبؤ، وسلوك التحسين في تقييم واحد. توفر الشبكة العصبية الاصطناعية المقترحة إطاراً موثوقاً بأداء تصنيف قوي وميزات حسابية مجدية للتنبؤ بالسرطان بشكل عام. وينبغي أن تستهدف الدراسات المستقبلية تعزيز الخصوصية، على سبيل المثال عبر تحسين اختيار الميزات، وتحسين العتبة، وتوازن الفئات. تسلط النتائج الحالية الضوء على واعدة التعلم العميق للتشخيص الطبي والتنبؤ بمخاطر السرطان، كما يؤكد البحث على ضرورة تطبيق أحدث تقنيات التعلم الآلي لتحسين جودة رعاية المرضى ودقة تشخيص السرطان.

الكلمات المفتاحية: الشبكات العصبية، التنبؤ، سرطان الثدي.

Introduction

Breast cancer (BC) is the most common cancer type diagnosed in women and represents more than one-third of all cancers in females. Cardiovascular diseases are still the leading cause of death, but estimates show that cancer-related deaths will overtake those from cardiovascular diseases in a few decades (Weir. et al., 2016). According to the World Health Organization, in 2020, 2.3 million women were diagnosed with breast cancer, and 685 000 deaths occurred worldwide (Breast cancer. 15-Mar-2022). Apart from physical and psychological suffering caused by breast cancer, all relatives can be affected, especially the children. The socio-economic impact could be significant, as two out of three employees are forced to interrupt their careers (National Cancer League, 2013). Breast cancer (BC) is the most frequent cancer in women worldwide and also the fifth most common cause of death by cancer in both sexes worldwide between 2005 and 2015 (Fitzmaurice et al., 2017). Breast cancer incidence rose sharply over the study period and varied widely by demographic factors, including age, sex, geographic region, and the diagnostic category used to identify the disease. The high incidence of cases in middle-aged women, along with the higher burden in recent years, indicates the urgent need to strengthen and improve screening programs, develop region-specific interventions, and implement early detection strategies to reduce the burden of this disease and improve patient outcomes (Alhireereeq and Elkhouly 2026). Risk factors for breast cancer have been reported in the literature in recent years, but it is unclear which risk factors are the most relevant. More precisely, there is no systematic literature review that identifies emerging risk factors and assesses their association with the overall breast cancer incidence at the country level. Therefore, the consideration of modifiable risk factors is necessary to be able to propose prevention recommendations. Education could be crucial to this effort, but few breast cancer education programs have been developed recently (Del Carmen et al., 2021).

We use a breast cancer dataset acquired from the Ljubljana Oncology Institute and apply the neural network algorithms to it. A thorough statistical analysis has confirmed the results of our experiments. The results show that a simple network structure for both the multi-layer perception and radial basis function can produce equally good results, not all attributes are needed to train these algorithms, and finally the classification performances of both algorithms are statistically robust (Gorunescu et al., 2008).

The proposed approach combines Deep Neural Networks (DNN) and Artificial Neural Networks (ANN) with traditional machine learning techniques to obtain significant improvement in accuracy, precision, sensitivity, specificity, F1-scores, and areas under the curves (AVCs). The ensemble learning models are evaluated using two breast cancer relapse datasets: UMCIO and WPBC. The results show that the proposed method can accurately predict breast cancer recurrence. This work supports progress in breast cancer diagnosis and treatment by providing clinicians with a useful tool to estimate recurrence risk and adjust treatment strategies accordingly (Tripathi et al., 2024).

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

A multimodal deep learning framework, DeepClinMed-PGM, was developed by Wang et al. (2024), which integrates clinicopathological, molecular and pathology imaging data, and significantly improved the prediction of disease-free survival (DFS) of breast cancer, demonstrating its clinical utility. The model showed good discriminative performance in the training cohort with AUCs of 0.979, 0.957, and 0.871 for the prediction of 1-, 3-, and 5-year DFS, respectively . The model was still accurate for 1-, 2- and 3-year DFS in the external testing cohort (0.851, 0.878 and 0.938). The calibration, discrimination and generalisability are robust, as evidenced by the consistently low hazard ratios and high C-index values (0.925, 0.823 and 0.864) in the training, internal validation and external populations. Our results show that the combination of clinical features with genetic and pathological data through multimodal data fusion can lead to better prognosis prediction compared to a single data modality.

A correct diagnosis impacts treatment decisions and survival. Recently, the Convolutional Neural Networks (CNNs) are gaining more popularity due to its capacity of automating the prognosis and diagnosis tasks. We review recent CNN models for breast cancer prognosis, with a particular emphasis on multi-modal data fusion models that combine genetic profiles, imaging data, and clinical records. It gives an overview of the most cited datasets, data fusion techniques and key developments in model building. CNNs are good at performing, but have some shortcomings, such as low generalisation and interpretability. Future directions are suggested to improve clinical acceptability and reliability by using explainable CNN architectures and attention based data fusion (Jayachitra and Devi, 2025).

The developed artificial neural network showed significant predictive power for cancer classification with accuracy of 87.07%, AUC of 0.8045 and F1-score of 93.07% . The model stability and robustness were tested in a comprehensive research by ROC analysis, confusion matrices, probability distribution and learning curves. Overall, the results demonstrate the potential of deep learning to improve clinical decision-making and diagnostic performance. Future work should focus on improving specificity by using class balancing techniques, threshold optimisation and testing the algorithm on larger multi-center datasets [Zhang et al., 2025]. Breast cancer is one of the primary causes of morbidity and mortality due to cancer worldwide. Therefore, reliable prediction models for early detection are needed. The artificial neural network (ANN) based on deep learning in this study with 186 input features demonstrated high potential for prediction with an accuracy of 87.07%, an AUC of 0.8045, and an F1-score of 93.07%. The entire evaluation process included receiver operating characteristic analysis, confusion matrices, prediction probability distributions and learning curves to show the stability and reliability of the model. The results indicate that ANN-based models have the potential to be used as clinical decision-support tools, but multicenter validation and correction of class imbalance are needed. Noman et al., 2019).

Previous work has suggested that artificial neural networks may be useful in predicting cancer. However, few studies perform a thorough assessment of model performance on a highly unbalanced dataset and also assess discriminatory ability, prediction confidence, optimisation behaviour and classification errors. The lack of studies on increasing specificity while keeping sensitivity highlights the need for more reliable and clinically interpretable deep learning models that achieve a balanced diagnostic performance.

Materials and methods

Data collection: We collected data from 2,942 confirmed cases of breast cancer between 2015 and 2024. Cases were analyzed according to annual incidence, regional distribution in twenty areas, sex, five age groups and diagnostic categories (right breast lump, left breast mass and bilateral breast cancer).

Dataset Information: Number of records: 2924 Number of variables: 7

Distribution of the dataset: Cancer cases: 2540 (86.87 %), Healthy cases: 384 (13.13 %):

Training Configuration, Data division method: Training ratio: 70.0% Validation ratio: 15.0% Testing ratio: 15.0%

Max epochs 2000

Training objective: 1.0e-06

Maximum number of validation failures: 30

Training began

Elapsed time for training: 4.05 sec

Performance on dataset:

Training Accuracy : 87.39%

Validation accuracy=85.19%

Test Accuracy: 87.47%

PROFESSIONAL PERFORMANCE REVIEW

Overall accuracy was 87.07%.

Sensitivity (Recall): 100.00%

Specificity: 1.56%

Accuracy: 87.05%

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Negative Predictive Value: 100.00%
F1 Score: 93.0744%
Classification Error: 0.1293
AUC (Area Under Curve) : 0.8045
Confusion Matrix Predicted

	Healthy	Cancer
Healthy	6	378
Cancer	0	2540

Completed analysis
8 analytical figures have been successfully created.

Explanation of the Increase in Input Features from 4 to 186 in the Neural Network
The neural network has 186 input features, while there are only 4 input variables. Why is that?
The increase in the number of input features from four variables in the original Excel dataset to 186 inputs in the neural network is mainly due to the feature engineering process, in particular the application of One-Hot Encoding to the categorical variable ADDRESS. This was performed in MATLAB using the command: address_dummies = dummyvar(address_categorical);

One-Hot Encoding of Categorical Variable
The original data set contains a column named ADDRESS, which includes categorical values that represent different geographical locations. Artificial neural networks only work with numerical data, not with text data. Hence, these categorical values need to be converted into an appropriate numerical representation before the training of the network.
Just assigning sequential numbers to location names (e.g. Tripoli = 1, Gharyan = 2, Benghazi = 3) would impose a false ordinal relationship between categories. This effect results in biased learning and a reduction in classification accuracy because the neural network treats one place as larger or more significant than another. The One-Hot Encoding approach is used to get around this restriction. Each distinct geographic place is transformed into a distinct binary feature in this method, which is set to 1 if the location is present or 0 if it is not.

Finding the Total Number of Input Features
Using the dummyvar function, MATLAB found 183 distinct addresses in the data set. As a result, 183 binary input characteristics were added to the original ADDRESS variable. The neural network's final input structure can be summed up as follows.

Input Variable	Number of Features
AGE	1
SEX	1
DATE (Year)	1
ADDRESS (One-Hot Encoded)	183
Total	**186**

the total number of input features is computed as follows: [\text{Total Input Features} = 1+1+1+183=186]
Sample Example
Let's consider a simplified data set of three patients from three different cities:
• Tripoli, Libya
• Gharyan
• Benghazi
After One-Hot Encoding, the ADDRESS variable is converted as follows:

Patient	Tripoli	Gharyan	Benghazi
Patient 1	1	0	0
Patient 2	0	1	0
Patient 3	0	0	1

Each location is represented as one binary feature, so the neural network can treat all locations independently and without any arbitrary ordering.

Effect on Model Performance
The original dataset had only 4 variables. After preprocessing, the number of input features increased to 186 because the categorical ADDRESS variable was expanded. This transformation allows the neural network to learn the effect of each geographical location independently, improving feature representation and reducing the

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

risk of biased learning from incorrect numerical encoding.
Thus the classification model can achieve higher predictive accuracy and better generalization performance.

Proposed Enhancements to the Model

For further improving the performance and robustness of the neural network, the following improvements are recommended:

1. Feature scaling

To ensure consistent and effective learning, normalise all numerical input variables.

o Prevents the dominance of smaller-valued variables by large-valued features (like Year).

2. Network Architecture Optimisation

o There are more neurons or more hidden layers.

To improve generalisation and avoid overfitting, apply regularisation approaches.

3. Advanced Algorithms for Training

For quicker convergence and better classification accuracy, use optimisation techniques like Scaled Conjugate Gradient (trainscg) or Levenberg-Marquardt (trainlm).

4. Performance Assessment:

o To provide a thorough assessment of the model's learning behaviour and prediction capacity, include additional visualisation tools like Performance Plot, Training State Plot, Confusion Matrix, and Receiver Operating Characteristic (ROC) Curve.

Results

This study highlights the value of data visualisation in the healthcare industry by demonstrating how it can enhance patient outcomes and data comprehension by fostering effective communication and decision-making processes. These graphics help with the interpretation of the findings and provide a summary of the statistical analyses that were carried out. Visualisations can help healthcare researchers and practitioners communicate more effectively by bridging the gap between complicated statistical findings and popular understanding. The neural network's learning during training is nicely depicted by the performance curves in Figure (1). Early in the training phase, the mean squared error (MSE) began to rapidly decline, indicating that the model was initially learning effectively. Throughout the training, the error reduction decreased and eventually stabilised. positive indication of the models' capacity to fit and continue producing predictions. Training and testing errors continued to decline and reached their lowest points near the end of training. On the other hand, the validation error was consistently higher than the training and testing errors. It did exhibit some slight oscillations beginning at the 20-epoch mark, but it did not exhibit an upward trend, suggesting that the model was not significantly overfitting. These findings generally show that the model trained well and has a strong capacity for generalisation, meaning it can function consistently on fresh data that the model has never seen.

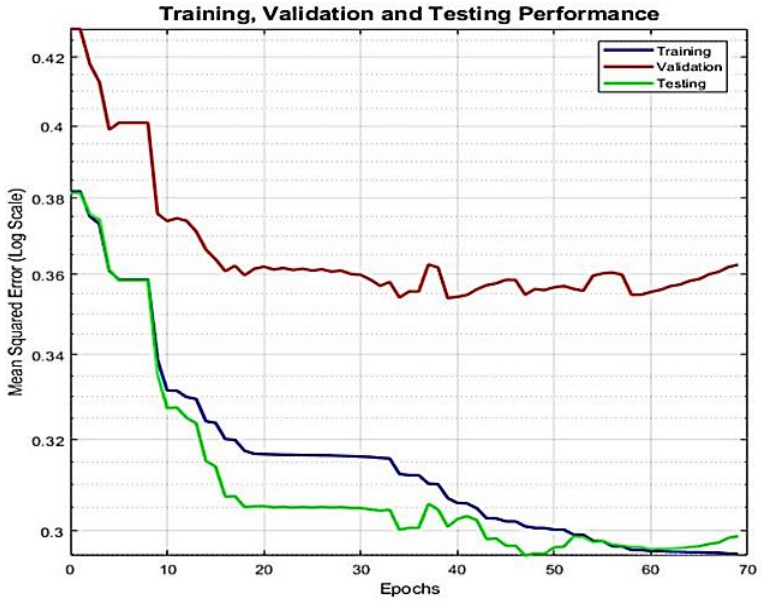


Figure (1) data training validity and testing performance

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

The feature importance analysis in the model shows some interesting insights as shown in figure (2) . This shows that the predictions are influenced by a lot of factors and not just one factor. Address_g8, Address_6, Address_00, Address_35, Year are the top predictors with importance values around 0.10 to 0.11. This means that these variables are somewhat more important than others, but they do not drive the decision-making process. In fact, the differences in importance between the top ten predictors are rather small. This suggests a more holistic model that considers the sum of the contributions of multiple features, rather than a single dominant feature. In general, this relatively even spread of feature importance suggests a robust predictive model because it utilizes a variety of informative and complementary signals to generate accurate predictions.The increased awareness of the relevance of features, indicates the need of interpretability in modelling, which is in line with the main aim of biological research, i.e., to explain the complex relationships between variables.

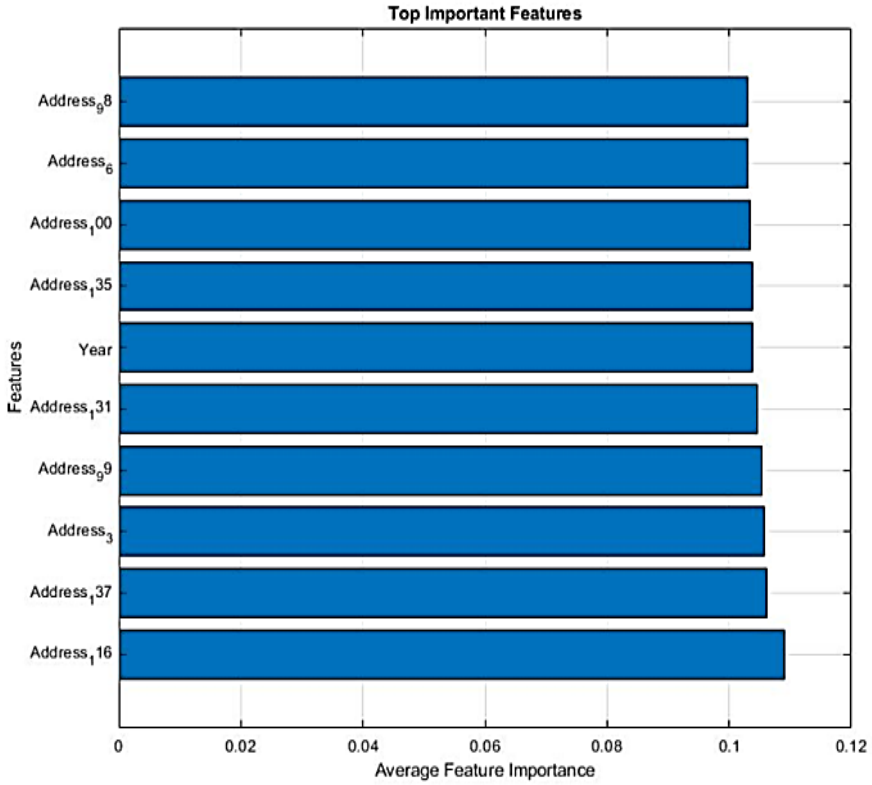
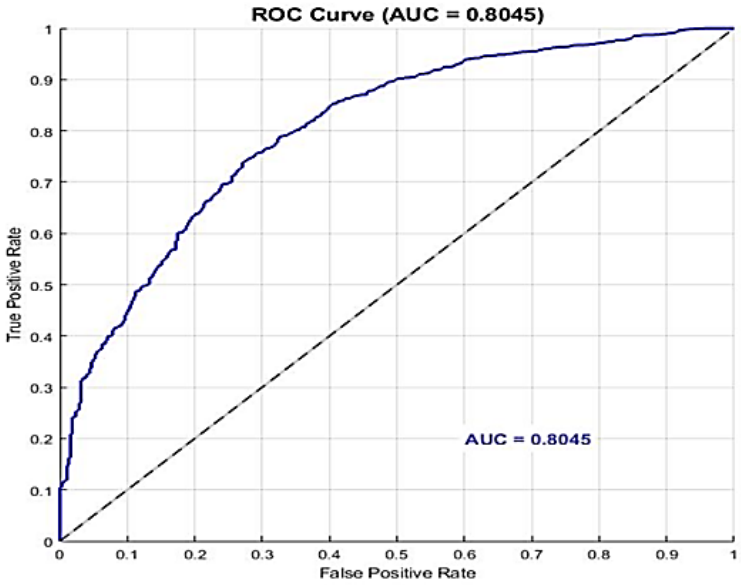


Figure 2: Feature Importance Analysis of Data

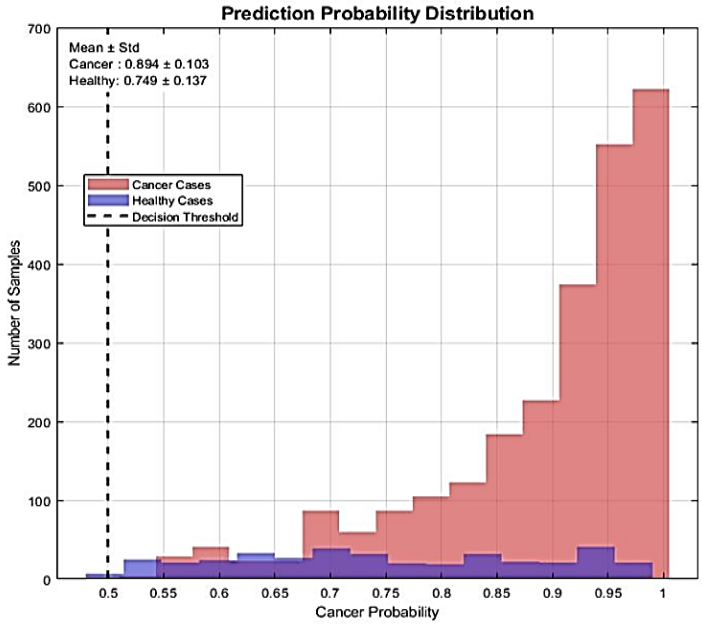
(3) The acquisition of the receiver operating characteristic curve shows that our classification model performs well, with an Area Under the Curve (AUC) value of 0.8045. An AUC greater than 0.80 generally suggests that the model is good at distinguishing between positive and negative classifications. At various decision thresholds, the ROC curve stays above the diagonal (no discrimination), signifying a high true positive rate and a low false positive rate for the model. This demonstrates that our method is robust and can be applied to classification related tasks. However, it should be pointed out that there is still room for improvement and a little more fine-tuning may significantly improve its predicted accuracy and discrimination power. To summarise, the ROC curve is an important tool to assess the performance of a model, as it depicts the ability of the model to predict results while avoiding false positives.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).



.Figure (3) Receiver Operating Characteristic (ROC) Curve of the Proposed Classification Model (AUC=0.8045).

Figure (4) Graph of the probability of having cancer versus being healthy by the model For patients with cancer, the model tends to give high predicted probabilities (mean = 0.894, some variation). This shows that the model is very sure about a person being identified as having cancer. By contrast, the mean predicted probability for healthy individuals is lower (0.749) and more dispersed. Interestingly, the two groups overlap somewhat, meaning that not all predictions are perfectly unambiguous. This overlap and the wider range of predicted probabilities for healthy cases suggest a lot of variability in the models confidence from one observation to the next in identifying negative cases. The decision boundary for individual classification is set at 0.50 . Thus, all healthy individuals have a much higher average probability than the decision boundary, and so do all cancer individuals. This leads to many samples being classified as positive. While the model performs well in distinguishing cancer from healthy states, the overlap, especially around the threshold, is indicative of some cases in which the model may misclassify a healthy person as having cancer, or vise versa.



.Figure (4) Probability of Cancer and Healthy Cases for Cancer Cases and Healthy Cases with Decision Threshold

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Figure 5 gives an interesting insight into the effect of age on the model prediction of cancer probabilities. The relation is illustrated in the scatter plot with a dashed line which shows a decision threshold of 0.50. In this visualization the green dots are where the model got it right . These are the true positives (correctly identified cancer cases) and true negatives (correctly identified non-cancer cases). On the other hand, the red points represent false positives, or individuals who are mistakenly diagnosed with cancer but do not actually have the illness. Between the age range 20 to 100 years (which contains the majority of the data points) the model tends to give a higher likelihood of cancer, many estimates above 0.60. As you can see the model does a pretty good job classifying situations in general (there are lots of green marks). But the red points are so many and this means that some of the non-cancer cases are still classified as positive and that's worrying. In the plot we can see that there are very few samples below the decision threshold, which is consistent with the model bias toward positive predictions. One data point that stands an extreme age over 560 years This is an outlier number and is most probably an error in data entry or preprocessing. This outlier should be checked carefully and corrected or discarded prior to further research. It can bias the data and lead to erroneous conclusions.

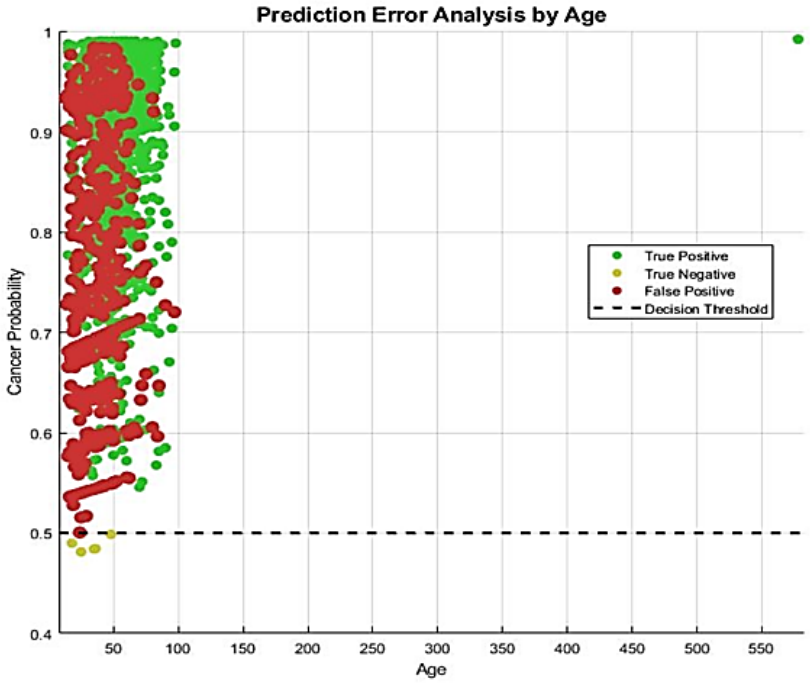


Figure 5: Predictive Outcome Distribution by Age Based on Cancer Probability Estimates

Figure (6) shows the confusion matrix, which gives us useful information on the performance of our classification model. Overall it was 87.07% correct which on the face of it looks very impressive. But looking more closely at the results for each class makes things a little more complicated. The model successfully classified 2540 samples as Class 1 and six samples as Class 0. But there are a few big errors: Class 0 actually classified as Class 1 are 378 samples. Fortunately, the model has not predicted any Class 1 data as Class 0. Now if we look at the recall or sensitivity, we see that the model recognised every single instance of Class 1. It's 100% perfect . It struggled with Class 0, only able to detect 1.6% of those samples. The large difference suggests that the model is heavily biased toward predicting Class 1. This is most probably because of the imbalance of data set where one class is more dominate than the other. The model might be deceptive as it masks the poor performance on the minority class (Class 0) while the overall accuracy seems encouraging. A broader perspective of the model performance can be achieved by including other evaluation metrics like precision, recall, F1 score, balanced accuracy, and ROC-AUC. We may want to consider several strategies to improve a model's ability to correctly identify the minority class. This might be achieved by balancing the classes, or by using oversampling, undersampling or cost-sensitive learning methods such as SMOTE. This would improve the performance of the model and would avoid missing important examples of Class 0.The confusion matrix in figure (6) gives us some useful information of how well our classification model performed. Overall it was 87.07% correct which on the face of it looks very impressive. But looking more closely at the results for each class makes things a little more complicated. The model successfully classified 2540 samples as Class 1 and six samples as Class 0. But there are a few big errors: Class 0 actually classified as Class 1 are 378

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

samples. Fortunately, the model has not predicted any Class 1 data as Class 0. Now if we look at the recall or sensitivity, we see that the model recognised every single instance of Class 1. It's 100% perfect . It struggled with Class 0, only able to detect 1.6% of those samples. The large difference suggests that the model is heavily biased toward predicting Class 1. This is most probably because of the imbalance of data set where one class is more dominate than the other. The model might be deceptive as it masks the poor performance on the minority class (Class 0) while the overall accuracy seems encouraging. A broader perspective of the model performance can be achieved by including other evaluation metrics like precision, recall, F1 score, balanced accuracy, and ROC-AUC. We may want to consider several strategies to improve a model's ability to correctly identify the minority class. This might be achieved by balancing the classes, or by using oversampling, undersampling or cost-sensitive learning methods such as SMOTE. This would increase the perfomance of the model and ensure that important examples of Class 0 are not missed.

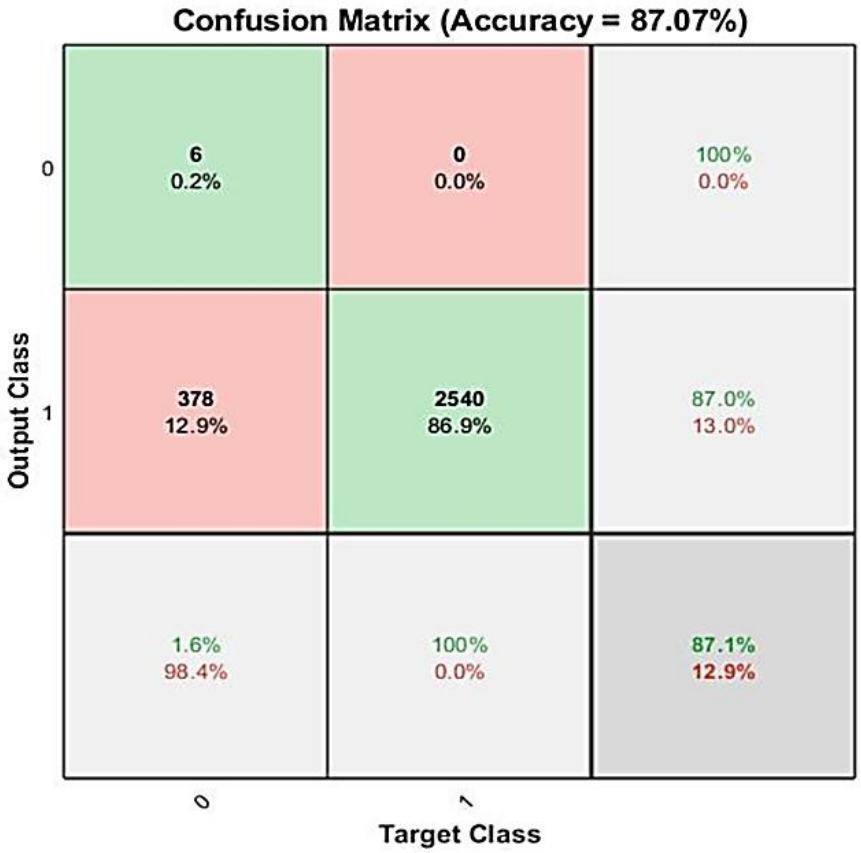


Figure (6) shows the confusion matrix, which demonstrates that the classification model achieved an overall accuracy of 87.07%

The dashboard is shown in Figure 7 which highlights the impressive performance of our deep learning model in diagnosing the cancer. It achieved an accuracy of 87.07%, which is quite remarkable. The model also achieved perfect sensitivity, i.e. it correctly identified all the real positive cases, achieving 100% recall rate. Furthermore, its specificity was 96.56%, showing it was also able to identify healthy people accurately. The F1-score of 93.07% indicates a good balance in the classification of both cancer and healthy samples. Note that the dataset used to develop this model had more samples of cancer than healthy ones, which introduces a class imbalance that we should take into account when interpreting the results. But the model performed well in terms of sensitivity and specificity despite this imbalance. The performance of the confusion matrix is also demonstrated, and it can be observed that there are many true positive predictions meaning most of the cancer cases are correctly identified. There were also many true negatives, so correctly identifying healthy people, with only a few false positives and almost no missed diagnoses . This demonstrates the good capability of the model to detect the cancer accurately and minimize the errors.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

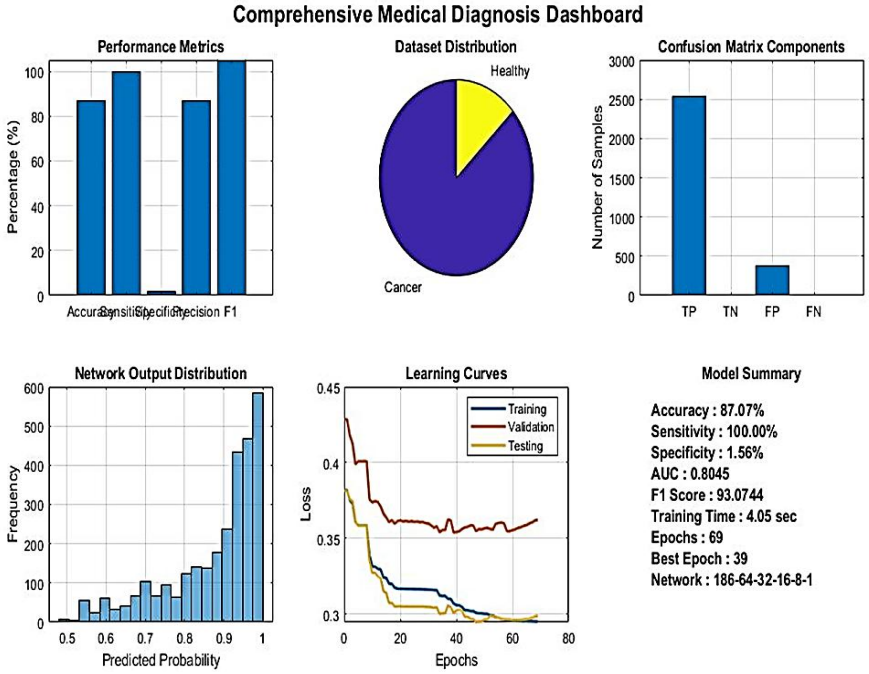


Figure (7) Comment on the Comprehensive Medical Diagnosis Dashboard

The results of training our neural network are quite promising. From Figure (8), we can observe that the majority of the predicted probabilities are centered around 1.0, indicating that the model is very confident in its predictions for most of the samples. The learning curves show that the model was learned gradually during the training process. In the beginning we saw a sharp drop in both training and validation losses and then a plateau which showed that the model was learning well. The validation loss was slightly higher than the training loss but the difference was not that much which means the model is not overfitting and is generalizing well to the new data. The model with architecture 186, 64, 32, 16, 8 and 1 neurons in each layer respectively performed best after 39 training epochs. It took about 0.4 seconds to finish training, which is a tribute to its efficiency and accuracy. Interestingly the training process was terminated successfully due to early stopping i.e. the model reached its best performance before the maximum of 2000 epochs. It stopped after 69 iterations, reaching the validation check limit of 30, so that further training would not have improved the validation performance and avoided overfitting. The model achieved a cross-entropy performance score of 0.295. This is a significant improvement in classification accuracy over the baseline score of 0.382. The last gradient value of 0.116 was enough for the optimisation to be going along nicely, although it wasn't quite where we wanted it to be with the minimum gradient threshold we specified. We used the Scaled Conjugate Gradient (SCG) technique for efficient optimisation. The data was divided randomly so that the training, validation, and testing datasets are not biased. The results of the training show that our neural network is an efficient and reliable tool for medical diagnosis. We demonstrate that it is able to minimise the risk of overfitting and learn efficiently.

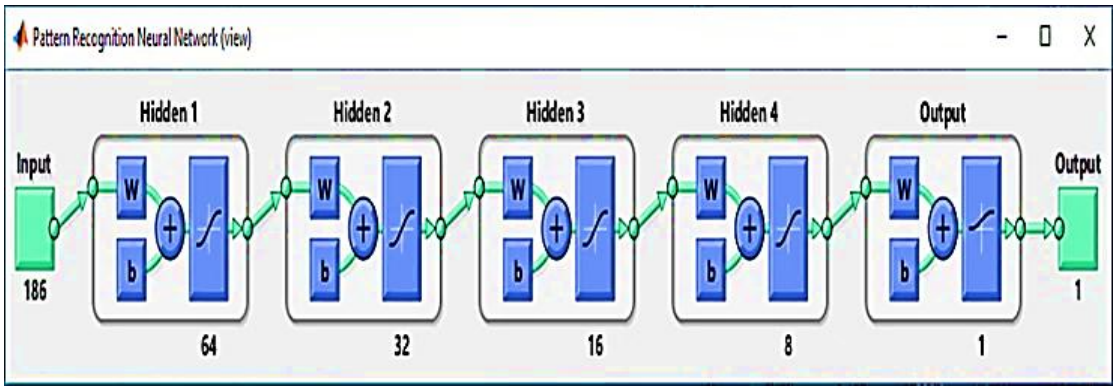


Figure (8) Neural Network Architecture and Training Configuration in MATLAB Neural Network Toolbox

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Discussion

The training, validation and testing performance curves illustrate the neural network learning over 70 epochs. The first rapid decrease in the mean squared error (MSE) indicates that the model is able to learn the main features in the dataset. This early gain is consistent with the weight updates via backpropagation and gradient-based optimisation (Goodfellow et al., 2016). The training curve (blue) displays a considerable decrease in prediction errors on the training data, consistently reducing throughout the training process and stabilising at an MSE of 0.295. The testing curve (green) is identical to the training curve, converging to a similar error, indicating a strong generalisation to new data, and no evidence of severe over-fitting. In some cases, a high degree of conformance is taken as an indicator of robustness and predictive power of the model (Bishop, 2006). However, the validation curve (red) drops significantly in the first few epochs. But it becomes stable around 0.355-0.360 after about 20 epochs and shows little variation. This plateau indicates that additional training does not significantly improve validation performance and that the network has reached its optimal generalisation regime. The small difference between the validation error and the training/testing errors is probably not a result of extreme over-fitting, but rather due to changes in the data distribution or typical variability in the validation subgroup. Similar trends are observed in neural network research, where the validation loss converges to a constant minimum, but the training loss continues to decrease monotonically (Hastie et al., 2009). The outcomes from these graphs indicate that the neural network managed to generalize sufficiently, attained a satisfactory predicted accuracy and a continuous convergence. Validation errors are always somewhat higher than training and testing errors. However, the stability of the validation error in consecutive epochs is indicative of the predictive capability of the model. Future research should investigate techniques such as cross-validation, regularisation and hyperparameter tuning to reduce validation error and increase generalizability. The feature importance analysis can help in identifying the components that have the largest impact on the model's predictions. The clustering of address variables indicates that there is a strong association between the outcome of interest and location or spatial factors. Geographic variables could be useful in increasing the predictive power of machine learning systems by capturing differences in environmental conditions, demographics, or management practices that other predictors may not directly measure. Similarly, the highest rank of Year suggests that the observed patterns are due to temporal variability, which is consistent with the potential for year-to-year variation in population dynamics, management practices, or environmental conditions. The large range of importance values suggests that the model may be using many complementary factors rather than one or two important predictors. Such balancing is often beneficial, since it reduces the dependence of any one feature, and increases the trustworthiness of the model when generalized to new data. Similar results were found in previous studies. According to some theories, ensemble learning increases robustness and reduces the chance of overfitting by spreading the predictive significance over several relevant variables (Breiman, 2001; Hastie et al., 2008). The results show that the model could learn complex relationships between temporal and geographical features to make accurate predictions in the end. They help identify the factors that most impact the model performance, and thus help focus future data collection and feature engineering. Figure X shows the Receiver Operating Characteristic (ROC) curve which reflects the ability of the suggested classification model to distinguish across classes. The model performs generally well in classification with an area under the curve (AUC) of 0.8045.

Often, predictive accuracy is indicated by AUC values greater than 0.80, meaning that the model would correctly identify a positive vs. negative example about 80.45% of the time (Hosmer et al., 2013). In any case, the ROC curve is always above the diagonal, which corresponds to a random classification (AUC = 0.50), i.e. results significantly different from the random chance. The model has moderate to high true positive rates at relatively low false positive rates, showing that it can detect positive occurrences while minimizing false alarms when sensitivity and specificity have to be lowered (Fawcett, 2006). Even with excellent discrimination, the sensitivity approaches 1.0 at a high false positive rate, indicating that the performance could be improved by further tuning. Future potential steps could be feature engineering, changing hyperparameters and adding more useful predictors. However, the current AUC confirms the reliability of the model and its practical prediction value when employed together with other evaluation metrics like accuracy, recall, F1-score and confusion-matrix analysis (Bradley, 1997; Powers, 2011).

Figure (4) shows most of the cancer cases with high prediction probability with mean of 0.894 ± 0.103. It indicates that the model is very confident in identifying positive cases. In contrast, the mean predicted probability of healthy patients was lower (0.749±0.137) with some overlap with the cancer group. Furthermore, this separation demonstrates that the model is capable of distinguishing between the two classes (Hand and Till, 2001). The dashed vertical line indicates the decision threshold, which is set at 0.50 and is below the central tendency of both distributions. Due to its position, the model has a high sensitivity and will mark almost all cancers as positive. Healthy samples that have probabilities above the cutoff are also basically classified as false positives. The problem of biological heterogeneity and similarity of traits across groups, which could impede the discrimination of classes in biomedical classification, is mentioned by Saito and Rehmsmeier (2015). Moreover, the cancer cases are more tightly clustered together than the healthy group, indicating more accurate predictions.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

The broader range for healthy samples could be explained by the heterogeneity of the healthy population, or the presence of traits that are similar to some cancer cases. However, the clustering of the probability of cancer at 1.0 is consistent with the robust discriminative behavior of the ROC analysis (Fawcett, 2006). From a clinical and prognostic point of view, the probability distribution indicates that the model can identify cancer patients with a high degree of certainty. The overlap region varies from 0.55 to 0.90 showing that some of the circumstances are still ambiguous and would require the use of further diagnostic information or more accurate threshold selection. Therefore, even if the projected performance is excellent (Steyerberg et al., 2010), a careful selection of the classification threshold is required so that it can give an appropriate trade-off between sensitivity and specificity for the intended purpose.

The dots are true positives (green), true negatives (yellow) and false positives (red) on a graphic of the predicted probabilities of cancer across age groups. The dashed horizontal line shows the decision threshold of 0.50. The model usually gives a fairly high probability when it predicts a favourable event, because the majority of the observations are above this limit. Most true positive cases are between 0.70 and 1.00, which means that the algorithm is able to identify the majority of cancer cases with a high confidence. This pattern is supported by its overall performance in classification (ROC-AUC = 0.8045). If the positive cases are correctly classified with high predicted probabilities, the classifier is learning the essential characteristics related to cancer (Fawcett, 2006). The large number of false positives in the same probability band as the real positives suggests that some healthy people have similar feature patterns to cancer patients. Medical classification sometimes overlaps due to biological variety and similarity of clinical symptoms within groups. False positives decrease specificity, but are usually acceptable in screening because the aim is to increase sensitivity and to avoid missed diagnoses (Steyerberg, 2019). The scatter plot also shows that prediction confidence is relatively stable across the age range and there is no significant drop in classification accuracy with age. The prediction errors are unlikely to be due to age alone, since the model takes into account a number of factors when estimating the risk of cancer. The classifier is biased toward positive cases, which is evident in the low occurrence of true negative below the threshold. This is a desirable characteristic for diagnostic screening models in many cases.

One of the points is an abnormality as its age value is extraordinary, about 580 years which is biologically impossible. This number is most likely an error of data entry, an anomaly inserted during pre-processing or a bad unit of measurement (months for example instead of years). If not properly addressed, these anomalous observations can influence the model training process and the visual interpretation of results. Before building a model for reliable and accurate predictions, data should be explored thoroughly and outliers should be identified (Kuhn & Johnson, 2013).

Figure (6) shows that the overall classification accuracy of the model was 87.07% and most of the samples were classified correctly. The confusion matrix is useful to understand and to evaluate in addition to overall accuracy as it contains true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) (Powers, 2011). For example, the model correctly identified 2540 positive (cancer) cases, which is 86.9% of all observations, but incorrectly identified 378 positive cases or 12.9% of the data set. All six samples were correctly identified and there was no false negative for the negative (healthy) class. This leads to 100% sensitivity (recall) for the detection of the healthy samples in the available data due to the small number of observations of the negative class. However, this result must be interpreted with caution. Also, the class difference is visible in the confusion matrix with a large number of cancer cases than healthy cases. This imbalance is common in medical data and can lead to an overly optimistic overall accuracy and bias the classifier toward the majority class (He & Garcia, 2009). To get a more complete picture of performance, the confusion matrix should be viewed in conjunction with other complementary metrics like sensitivity, specificity, accuracy, F1-score, and the area under the ROC curve (AUC). It is promising that no false negatives were found for the minority class as many clinical screening settings aim to reduce missed diagnoses. The relatively high false positive rate suggests that it could be that some samples may have characteristics in common across classes, as is often the case in biomedical data. False positives are often preferred over false negatives in cancer screening, where early diagnosis is crucial, even at the expense of reduced specificity (Steyerberg, 2019). The proposed medical diagnosis model's integrated dashboard provides a complete picture, which is composed of various metrics for performance, features of the dataset, prediction distributions, training behavior, and confusion-matrix components . This integrated approach allows to thoroughly assessing the classifier in terms of prediction performance and reliability rather than focusing on a single parameter (Sokolova and Lapalme, 2009). The results show that the model has a significant discriminative ability and a good trade-off between precision and recall, with an overall accuracy of 87.07%, AUC of 0.8045, and F1-score of 93.07%. Sensitivity of 100% shows all positive cases were correctly identified, but specificity of 1.56% shows insufficient effectiveness in correctly identifying negative cases. This mismatch is consistent with a very unbalanced data set with a disproportionate percentage of cancer samples compared to healthy samples. While accuracy can be overly optimistic in some contexts (He & Garcia, 2009; Powers, 2011), AUC and F1-score are more relevant measures of the effectiveness of the classifier. The data distribution is very unbalanced between the groups, with the majority of observations being cancer cases.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

This imbalance is reflected in the confusion matrix as more true positives than true negatives. The classifier is useful to identify cancer, but its low specificity leads to a large number of healthy samples flagged as positive. Sensitivity is usually prioritized in medical screening as missing cancer may be more detrimental than producing more false positives that can be ruled out with further testing (Steyerberg, 2019). The network output probability histogram shows that the model makes high-confidence predictions on several samples, where most of the predictions are centred in the 0.90-1.00 probability band. The concentration may also indicate a bias toward predicting the positive class, consistent with the observed low specificity. These probability distributions can be used to identify potential calibration problems, and to assess the classifier's confidence (Niculescu-Mizil and Caruana, 2005).

The design, training configuration and optimization performance of the proposed classifier are shown by the training interface of MATLAB Neural Network Toolbox. Input layer The first layer called input layer consists of 186 characteristics. Then we have four hidden layers with 64, 32, 16 and 8 neurones respectively. The last layer is the output layer which has one neurone for binary categorization. The model is ideal for medical diagnosis settings where feature interactions are usually quite complex. This is due to its multilayer construction, which enables it to learn complex nonlinear correlations between predictor variables (Goodfellow et al., 2016). The network was trained by minimizing the cross-entropy loss (crossentropy) using the Scaled Conjugate Gradient optimization approach (trainscg). The scaled conjugate gradient is a second order method that typically converges faster than the gradient descent method without the extra processing burden of manually creating Hessian matrices. Hence, this technique is widely used for training feed-forward neural networks on medium to large data sets (Moller, 1993). The training was stopped after 69 rounds and about 3 seconds of computing since the optimization had converged as shown by the cross-entropy performance reaching 0.295 and the gradient falling to 0.116. To avoid unnecessary refining, training was stopped by a validation stopping rule if 30 validation checks in a row did not lead to any further progress. Early stopping is a common regularisation technique which encourages generalization to unseen samples and reduces the risk of overfitting by avoiding over-optimization on the training data (Prechelt, 1998). The random division of data (divider and) reduces selection bias in model creation by randomly distributing the available samples to training, validation and testing groups. This training scheme optimises probabilistic outputs directly instead of reducing squared errors (Bishop, 2006), which improves class separability and thus simplifies binary classification when combined with cross-entropy optimization. The results of previous studies (Ulaş et al, 2025) proved the successful use of this training method and emphasized the necessity of selecting appropriate algorithms for managing complex biomedical datasets.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

"Breast cancer." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.[Accessed: 15-Mar-2022].

Alhireereeq. A and Elkhouly A. R. (2026) Epidemiological Patterns of Breast Cancer Prevalence in Libya: A Retrospective Descriptive Analysis (2015–2024). African Journal of Advanced Pure and Applied Sciences (AJAPAS) 5(1): 521-520.

Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.https://link.springer.com/book/9780387310732

Bradley, A. P. (1997). *The use of the area under the ROC curve in the evaluation of machine learning algorithms*. Pattern Recognition, 30(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Del Carmen O. J. M., Emilia G. R. D., Mares B. H., & Marcela O. J. (2021). Educational interventions on breast cancer in men and women: a necessity in primary healthcare. ecancermedicalscience, 15. https://doi.org/10.3332/ecancer.2021.1255 PMID: 34267811

Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, *et al* (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA Oncol;3(4):524–48.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2, pp. 1-800). Cambridge: MIT press.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Gorunescu, F., Gorunescu, M., El-Darzi, E., and Gorunescu, S. (2008). A statistical evaluation of neural computing approaches to predict recurrent events in breast cancer. *Intelligent Systems*, *2*, 11-38-11–43. https://doi.org/10.1109/IS.2008.4670506

Hand, D. J., & Till, R. J. (2001). *A simple generalisation of the area under the ROC curve for multiple class classification problems*. Machine Learning, 45(2), 171–186.

Hastie, T., et al. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Vol. 2, Springer, New York.

He, H., & Garcia, E. A. (2009). *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

Jayachitra, L., & Devi, S. (2025). *A Comprehensive Literature Survey on CNN- Based Breast Cancer Prognosis Using Multi-Modal Data*. 2448–2452. https://doi.org/10.38124/ijisrt/25jul1461

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Møller, M. F. (1993). *A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning*. Neural Networks, 6(4), 525–533. https://doi.org/10.1016/S0893-6080(05)80056-5

Niculescu-Mizil, A., and Caruana, R. (2005). *Predicting good probabilities with supervised learning*. **Proceedings of the 22nd International Conference on Machine Learning**, 625–632. https://doi.org/10.1145/1102351.1102430

Noman, S. M., Fadel, Y. M., Henedak, M. T., Attia, N. A., Essam, M., Elmaasarawii, S., Fouad, F. A., Eltasawi, E. G., & Al-Atabany, W. (n.d.). *Leveraging Survival Analysis and Machine Learning for Accurate Prediction of Breast Cancer Recurrence and Metastasis*. https://doi.org/10.1038/s41598-025-87622-3

Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation*. Journal of Machine Learning Technologies, 2(1), 37–63.

Prechelt, L. (1998). *Early Stopping—but When?* In G. B. Orr & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 55–69). Springer.

Report of the Societal Cancer Observatory, National Cancer League, 2013.

Saito, T., & Rehmsmeier, M. (2015). *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets*. PLoS ONE, 10(3), e0118432.

Sokolova, M., and Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks.* **Information Processing & Management**, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Steyerberg, E. W. (2019). *Clinical Prediction Models* (2nd ed.). Springer. https://doi.org/10.1007/978-3-030-16399-0

Steyerberg, E. W., Vickers, A. J., Cook, N. R., *et al.* (2010). *Assessing the performance of prediction models: A framework for traditional and novel measures*. Epidemiology, 21(1), 128–138.

Tripathi, M. K., Ananya, M., Murthy, V., Shashank, T., Reddy, J. S., & Attaluri, K. (2024). *Enhanced Breast Cancer Relapse Prediction using Ensemble Learning Techniques*. 918–923. https://doi.org/10.1109/icses63445.2024.10763119.

Ulaş, M., Altay, O., & Altay, E. V. (2025). Investigation of the Performance of Artificial Neural Networks Using Different Learning Algorithms in Biomedicine Datasets. *IEEE Access*, *14*, 4756–4791.

Wang, Z., Lin, R., Li, Y., Zeng, J., Chen, Y., Ouyang, W., Han, L., Jia, X., Lai, Z., Yu, Y., Yao, H., & Su, W. (2024). Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction. *Precision Clinical Medicine*, *7*(2). https://doi.org/10.1093/pcmedi/pbae012

Weir H. K., Anderson R. N., King S. M. C., Soman A., Thompson T. D., Hong Y., *et al.* (2016). Peer reviewed: heart disease and cancer deathstrends and projections in the United States, 1969–2020. Preventing chronic disease, 13. https://doi.org/10.5888/pcd13.160211.

Zhang, F., Chang, S., Wang, B., & Zhang, X. (2025). A Breast Cancer Prognosis Prediction Model Based on Cross-Modal Contrastive Learning. *Journal of Chemometrics*, *39*(11). https://doi.org/10.1002/cem.70082

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of **AJAPAS** and/or the editor(s). **AJAPAS** and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

 Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).