JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 3, SEPTEMBER, 2026, Pp. 300 – 317
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Fine-Tuned Artificial Neural Network (ANN) with Explainable AI (XAI) for Classification and Interpretation of the Breast Cancer Wisconsin

Ikra Muhibban Ananta [(1)*], Hanif Amarudin [(2)], Restu Adi Akbar [(3)], M. Fajratul Ikhsan [(4)], Maria Ulfah Siregar [(5)]
Department of Informatics, Sunan Kalijaga State Islamic University, Yogyakarta, Indonesia
e-mail : {muhib.ikra,hanifamar17,restu6005,ikhsanm288}@gmail.com, maria.siregar@uin-suka.ac.id.
* Corresponding author.
This article was submitted on 4 January 2026, revised on 7 May 2026, accepted on 8 May 2026, and published on 25 September 2026.

Abstract

Breast cancer remains one of the leading causes of cancer-related mortality among women, making early detection and accurate diagnosis essential. Although machine learning and deep learning approaches have achieved high classification performance, their interpretability and evaluation robustness remain limited. This study proposes a Fine-Tuned Artificial Neural Network (ANN) integrated with Explainable Artificial Intelligence (XAI) for breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Model evaluation was conducted using Stratified 5-Fold Cross-Validation, while optimization was performed through learning rate adjustment and early stopping. Model interpretability was analyzed using SHapley Additive exPlanations (SHAP), including summary, decision, force, and waterfall plots. Experimental results show that the baseline ANN achieved an average accuracy of 97.54%, precision of 97.82%, recall of 98.32%, and F1-score of 98.04%. The Fine-Tuned ANN improved performance, achieving 97.89% accuracy, 98.08% precision, 98.60% recall, and 98.32% F1-score, with better stability across folds. SHAP analysis identified worst area, worst radius, worst texture, and mean concave points as the most influential features. Furthermore, the ablation study indicated that the Sigmoid activation function, Adam optimizer, and learning rate of 0.01 produced the optimal configuration.

Keywords: Breast Cancer, ANN, XAI, SHAP, Ablation Study

Abstrak

Kanker payudara merupakan salah satu penyebab utama kematian akibat kanker pada perempuan, sehingga deteksi dini dan diagnosis akurat sangat penting. Berbagai pendekatan machine learning dan deep learning telah menunjukkan performa tinggi, namun masih terbatas pada aspek interpretabilitas dan evaluasi yang robust. Penelitian ini mengusulkan model Fine-Tuned Artificial Neural Network (ANN) yang diintegrasikan dengan Explainable Artificial Intelligence (XAI) untuk klasifikasi kanker payudara menggunakan dataset Wisconsin Diagnostic Breast Cancer (WDBC). Evaluasi dilakukan menggunakan Stratified 5-Fold Cross-Validation, sedangkan optimasi model diterapkan melalui fine-tuning dengan penyesuaian learning rate dan early stopping. Interpretabilitas model dianalisis menggunakan SHapley Additive exPlanations (SHAP) melalui summary plot, decision plot, force plot, dan waterfall plot. Hasil eksperimen menunjukkan bahwa model ANN baseline memperoleh rata-rata akurasi 97,54%, presisi 97,82%, recall 98,32%, dan F1-score 98,04%, sedangkan Fine-Tuned ANN meningkat menjadi akurasi 97,89%, presisi 98,08%, recall 98,60%, dan F1-score 98,32%, dengan stabilitas performa yang lebih baik. Analisis SHAP mengidentifikasi worst area, worst radius, worst texture, dan mean concave points sebagai fitur dominan. Selain itu, studi ablasi menunjukkan bahwa aktivasi Sigmoid, optimizer Adam, dan learning rate 0,01 memberikan konfigurasi terbaik. Temuan ini menunjukkan bahwa Fine-Tuned ANN dengan XAI menghasilkan model yang akurat, stabil, dan interpretable untuk diagnosis kanker payudara berbasis komputer.

Kata Kunci: Kanker Payudara, ANN, XAI, SHAP, Studi Ablasi

1. INTRODUCTION

Breast cancer is a major concern because it is one of the most common types of cancer among women worldwide and can be fatal. According to the World Health Organization (World Health Organization, 2025), an estimated 2.3 million women will be diagnosed with breast cancer in 2022, resulting in 670,000 deaths globally. Breast cancer detection is a major factor in determining the level of risk. Early detection of breast cancer can effectively reduce the risk of malignancy. To achieve the goal of early and accurate detection, research needs to develop the best approach for diagnosing the level of malignancy of breast cancer. Another problem arises when the approach developed only diagnoses breast cancer and doesn't sufficiently explain the reasons behind the decisions made by AI. Explainable AI (XAI) can be a solution to this problem. XAI helps to reveal knowledge and explain how a model finds patterns in a dataset, so that we can understand the reasons behind a decision (Saeed & Omlin, 2023).

A study describes the comparison of several deep learning approaches to determine which approach has a high level of accuracy. This study compares Artificial Neural Network (ANN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) methods using a breast cancer dataset from the University of Wisconsin. The ANN model was built with six layers and used ReLU and sigmoid activation functions. The results show that the ANN method has the highest accuracy score, reaching 99.9% (Mohd Nasir et al., 2023). Other studies also compared several approaches, such as K-Nearest Neighbors (KNN) and ANN, on two datasets, namely Wisconsin Breast Cancer (WBC) and Wisconsin Diagnostic Breast Cancer (WDBC). The results for the WBC dataset showed that the model with the highest accuracy was obtained by KNN at 97.7%, while for the WDBC dataset, the model with the highest accuracy was obtained by ANN at 98.6%. This ANN model was built with two layers, while the dataset used was specific only to the "worst" feature. This study adds interpretation to AI decisions using Explainable AI (XAI) using several methods such as permutation importance, Partial Dependence Plots (PDP), and SHAP (Khater et al., 2025). The next study will use decision trees, ANN, and support vector machines (SVMs) to conduct experiments on public datasets from Kaggle. The models were optimized using correlation-based feature selection methods, resulting in an accuracy score that reached 97% and a precision score that reached 100%. The recall score was 92%, and the F1 score was 96% (Fahrurrozi et al., 2025).

Many studies have explored the best approach to achieve the highest level of accuracy. However, these studies still present several unresolved limitations. The comparison of ANNs, RNNs, and CNNs conducted by Mohd Nasir et al. (2023) did not incorporate model interpretability. The KNN and ANN approach integrated with XAI by Khater et al. was limited to a subset of the least relevant features and did not include fine-tuning optimization. Meanwhile, Fahrurrozi et al. (2025), who employed correlation-based feature selection, did not explain the contribution of features to individual classification decisions. Therefore, the simultaneous integration of fine-tuning-based model optimization and SHAP-based decision interpretation remains a research gap that has not been adequately addressed in previous studies. In addition, understanding the results provided by AI on a dataset can increase trust and transparency of the results, as well as understanding of the steps to be taken. Therefore, this study will conduct experiments to build an optimized model to obtain better results in detecting breast cancer. The model will explain and explore which features have the most influence on the results of the classification decisions.

This study's goal is to develop a model from one of the popular models, called ANN, for the Wisconsin Breast Cancer (diagnostic) case study. This model will be optimized with fine-tuning to obtain better decisions. The results of the experiment will be a comparison of the classification model before and after fine-tuning. The fine-tuned results will be analyzed to gain an understanding of the features affecting the decision results using XAI with the SHAP method. As a complement to the main experiment, this study will also conduct an ablation study to evaluate the influence of important components in the model architecture on breast cancer classification performance. This ablation study will focus on several key aspects, such as activation functions, optimizer types, and learning rate variations, to see how changes in the configuration of the

architecture affect the accuracy of a model. This ablation study approach can be used to select the optimal model configuration based on accuracy and computational complexity results (Montaha et al., 2022). For the record, the results of this ablation study will not be analyzed using XAI. Still, they will only be used to compare accuracy between configurations, providing an overview of the factors that contribute most to performance improvement. The research will provide insights into the development of optimized classification models. This will help people understand how well these models perform, particularly in health-related areas such as breast cancer. This will give people confidence in the decision results of ANN models.

2. METHODS

ANN is a discipline of artificial intelligence inspired by the structure and functioning of the human brain. ANN also has extraordinary capabilities for pattern recognition, clustering, information analysis, and function calculation (Dubaish & Jaber, 2024). Previous research has shown that ANN can demonstrate a high level of accuracy in classifying breast cancer in the Wisconsin dataset. This study relies on ANN as the primary method for classification to detect malignant and benign tumors. Figure 1 presents the details of the stages in this system.

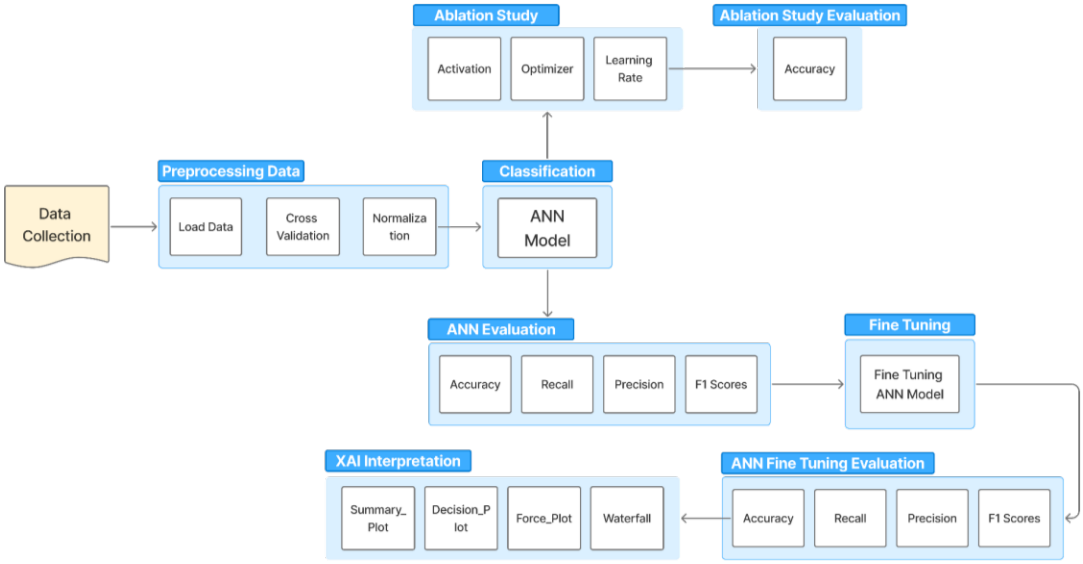

Figure 1 Flowchart

2.1 Data Collection

The dataset used in this study is Breast Cancer Wisconsin (Diagnostic), taken from the scikit-learn repository. This dataset is a public dataset used to classify malignant and benign tumors. The following is a summary of the dataset description, which can be seen in Table 1.

Table 1 Description of the Wisconsin Breast Cancer Dataset (Diagnostic)

Parameter	Values
Number of Instances	569
Number of Attributes	30 numeric
Number of Main Attributes	10
Missing Attribute Values	None
Class	2 (Malignant dan Benign)
Class Distribution	212 – Malignant, 357 – Benign

This dataset has a total of 569 samples. Each data sample has 30 numerical attributes used to predict whether it belongs to the malignant or benign category. These numerical attributes are

divided into three different statistical levels, including Mean, Standard Error, and Worst (the average of the 3 worst/largest values). Each of these three levels has 10 main features. This dataset has 10 main features, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

2.2 Data Pre-processing

The pre-processing stage is very important in machine learning because poor-quality information can directly affect prediction performance (Miftakhudin et al., 2025). The dataset consists of numerical features used to classify tumors into two categories: malignant and benign. Model evaluation was conducted using stratified 5-fold cross-validation. In this method, the dataset is divided into five folds while maintaining balanced class distributions across all folds. In each iteration, four folds are used as the training set, while one fold is used as the test set. This procedure is repeated until each fold has been used once as the test set. The use of stratified cross-validation was intended to produce a more robust and stable model evaluation while reducing bias caused by a single data partition.

Before model training, all numerical features were normalized using StandardScaler. The normalization was performed within each cross-validation fold by fitting the scaler only on the training set, thereby transforming the feature distribution into a standard scale with a mean close to zero and a standard deviation of one. This normalization process is crucial to avoid the dominance of features with much larger scale values during model training (Jlassi & Dixon, 2024). Next, the normalization parameters obtained from the training data are used to transform the testing data in the same fold. This step is important for maintaining feature scale consistency and preventing data leakage during the model evaluation process.

To ensure the consistency of experimental results and the reproducibility of the research, the data randomization process in cross-validation uses a random state parameter of 42. Additionally, model performance evaluation is conducted using several metrics, namely accuracy, precision, recall, and F1-score. The mean and standard deviation values from all folds are used to analyze the stability and generalization ability of the Artificial Neural Network (ANN) model.

2.3 Classification (ANN)

ANN is one of the supervised algorithms in machine learning, meaning that this algorithm requires a number of parameters to produce good results (Gunawan, 2020). ANN works well for both linear and non-linear regression and can be seen as a different statistical approach to solving the least squares problem. Both ANN terms refer to standard regression analysis, which has similarities (Saputra et al., 2020).

The classification model used in this study was built using a Sequential-based Artificial Neural Network (ANN) architecture from TensorFlow Keras. The model architecture consists of one input layer, two hidden layers, and one output layer. The input layer receives all numerical features from the Wisconsin Breast Cancer (Diagnostic) dataset, which has undergone normalization using StandardScaler. The first hidden layer consists of 64 neurons with a ReLU (Rectified Linear Unit) activation function, while the second hidden layer consists of 32 neurons with the same activation function. The use of the ReLU activation function aims to improve the model's ability to learn non-linear patterns in the data and reduce the vanishing gradient problem during the training process. The output layer uses a single neuron with a sigmoid activation function to generate binary classification probabilities between the malignant and benign classes.

To improve the model's generalization ability and reduce the risk of overfitting, the ANN architecture is equipped with two dropout layers. The first dropout, set to 0.3, is placed after the first hidden layer, while the second dropout, set to 0.2, is placed after the second hidden layer. The dropout strategy works by randomly deactivating some neurons during training so that the model does not rely too heavily on specific neurons and is able to produce more stable performance on the test data. During the model training phase, the Adam optimizer is used with

the binary cross-entropy loss function, as this study involves a binary classification task. Additionally, the evaluation metrics used include accuracy, precision, recall, and F1-score to evaluate the model's performance more comprehensively. The initial training process is conducted over 50 epochs with a batch size of 16.

To improve the reliability of model evaluation, this study employs stratified 5-fold cross-validation. This method divides the dataset into five folds with class proportions kept balanced in each fold. In each iteration, one fold is used as test data while the other four folds are used as training data. This approach allows the model to be evaluated across various data partitioning variations, making the evaluation results more robust and less dependent on a specific train-test split scenario.

After the initial training process is complete, a fine-tuning stage is performed by continuing to train the model using the weights from the previous training. The fine-tuning process is performed using the Adam optimizer with a smaller learning rate of 5×10^{-4} to improve learning stability and refine the model's weight adjustments. Additionally, an Early Stopping mechanism with a patience parameter of 10 is applied to automatically halt the training process when the validation loss no longer improves. This strategy is used to prevent overfitting while preserving the model's optimal weights during the fine-tuning process.

2.4 Evaluation

Model performance was evaluated using several classification metrics, namely accuracy, precision, recall, and F1-score. The evaluation was conducted using stratified 5-fold cross-validation to ensure that the model was tested across different data partitions while maintaining balanced class proportions in each fold. In each fold, the model was evaluated before and after fine-tuning to examine the contribution of fine-tuning to model performance improvement. Accuracy was used to measure the overall correctness of predictions, precision was used to assess the model's ability to minimize false positives, and recall was used to evaluate its ability to correctly identify positive cases. Meanwhile, the F1-score was used to measure the balance between precision and recall.

In addition to calculating the mean values across all cross-validation folds, this study also calculated the standard deviation of each evaluation metric. The standard deviation was used to assess the stability and consistency of model performance across different data partitions. A lower standard deviation indicates that the model has stronger generalization capability and more stable performance across testing scenarios. The purpose of this evaluation is to assess the extent to which the ANN model is able to classify the test data, while also providing an overview of the model's performance in terms of class imbalance (Ghanem et al., 2023).

Furthermore, the experimental evaluation was not limited to classification performance measurement but also included model interpretability analysis using SHAP-based Explainable Artificial Intelligence and architectural contribution assessment through ablation studies. SHAP-based XAI was used to analyze the influence of each feature on the decisions produced by the ANN model. Meanwhile, ablation studies were conducted to examine the impact of hyperparameter variations on ANN performance using stratified 5-fold cross-validation. The tested parameters included the number of hidden neurons, dropout rate, activation function, optimizer, and learning rate. Each configuration was evaluated using accuracy, precision, recall, and F1-score. The mean and standard deviation values across all folds were used to assess model performance and stability, and the best-performing configuration was subsequently selected for the fine-tuning stage.

2.5 Fine-tuning

After the basic structure of the ANN was completed, fine-tuning was performed to improve the model's performance. This adjustment aims to reduce overfitting and improve the demonstration's generalization. The adjustment process is an effort to set up a previously trained demonstration

to be more efficient for specific tasks or datasets. Previously trained demonstrations are generally built using very large and diverse datasets, so that the demonstration is able to understand the information in general. and one output layer with sigmoid activation for binary classification (Susilo et al., 2024).

Fine-tuning was performed using the same model architecture while leveraging the weights obtained from the previous training stage. This approach allowed the model to continue learning from the previously optimized parameters without restarting the training from random initialization. At this stage, the Adam optimizer was used with a learning rate of 5×10^{-4}. The lower learning rate was intended to stabilize the weight update process and enable more gradual parameter adjustments, thereby improving model performance. In addition, Early Stopping was applied by monitoring the validation loss. Training was automatically terminated when no improvement in validation loss was observed for 10 consecutive epochs. This mechanism was used to reduce the risk of overfitting while preserving the optimal model weights.

The fine-tuning stage was conducted for a maximum of 50 epochs with a batch size of 16, using the same training and validation partitions as in the initial training stage. After fine-tuning was completed, model performance was evaluated using accuracy, precision, recall, and F1-score. In addition, the classification report and confusion matrix were used as supplementary evaluation outputs to provide a more detailed assessment of the model's classification performance on breast cancer data.

2.6 Interpretation XAI (SHAP)

SHAP (SHapley Additive exPlanations) is one of the most widely used methods in Explainable Artificial Intelligence (XAI). As a model-agnostic approach, SHAP can be applied to various black-box models, including Artificial Neural Networks (ANN), after the training process has been completed (Purwono et al., 2025). A key advantage of SHAP is its ability to identify the features that contribute most significantly to model predictions. It provides greater transparency into neural network predictions, breaking down each prediction into the factors that pushed it one way or another. This enhances the transparency and interpretability of the model's decisions and helps people trust what it's doing.

SHAP also provides several visualization techniques for interpreting these explanations. There's the summary plot, for example, which provides a global overview of how all your features influence predictions. Each point shows a SHAP value for one observation. Larger absolute SHAP values indicate a greater influence on the prediction, whether positive or negative. The colors let you spot which feature values are at play (El Jihaoui et al., 2025). This enables rapid identification of the most influential features and which features are the main players across your whole dataset.

For analyzing individual predictions, the decision plot and force plot are particularly useful. These help you see how a particular mix of features pushed the prediction one way or another, either for a single case or a few at a time. The force plot lays out exactly how each feature nudged the prediction up or down from the starting point. The decision plot takes you step by step through the order in which features shaped the outcome, thereby enabling a step-by-step understanding of the model's decision process (Rahim & Basheer, 2025).

Then there's the waterfall plot. It lines up the feature contributions for one prediction, starting from the baseline and showing how each feature adds or subtracts from the final answer. You end up with a clear picture of which features contributed most substantially on that specific prediction (El Jihaoui et al., 2025). Altogether, these four types of plots make it much easier to understand what's happening inside ANN-based models. Researchers and decision-makers can see both the big picture and the fine details, which helps build real confidence in the AI system and keeps its decisions from feeling like a black box (Rahim & Basheer, 2025).

2.7 Ablation Study

To ensure an optimal and robust ANN model, an ablation study process is applied to the hyperparameter configuration. Referring to the research by (Meyes et al., 2019), the ablation method in ANN not only serves to reduce complexity but also as an important instrument to analyze the contribution of each component to the overall classification performance of the model. In this study, the ablation method was conducted by modifying the main components to observe the effect of each change on the evaluation metrics. The hyperparameter variables tested were:

a) Activation Function: testing to compare the effectiveness of non-linearity functions in the hidden layer, which includes the use of ReLU, Tanh, and Sigmoid (Pantalé, 2023).
b) Optimizer: evaluation conducted on optimization algorithms to determine the best method, comparing Adam, SGD, and RMSprop (Khan et al., 2022).
c) Learning Rate: sensitivity analysis of the model conducted on the learning rate with variations ranging from 0.1 to 0.0001 (Jepkoech et al., 2021).

3. RESULTS AND DISCUSSION

3.1 ANN Baseline & Fine-Tuning Results

An Artificial Neural Network model with two hidden layers consisting of 64 and 32 neurons was trained using the Wisconsin Breast Cancer Diagnostic dataset. The training was conducted using stratified 5-fold cross-validation to ensure a more robust evaluation that was not dependent on a single data partition. In each fold, preprocessing was performed using StandardScaler before the model was trained for 50 epochs with a batch size of 16.

The evaluation was conducted in two stages: baseline evaluation before fine-tuning and evaluation after fine-tuning. In the baseline stage, the model was trained using the Adam optimizer with its default learning rate. Subsequently, fine-tuning was performed by continuing the training process using a lower learning rate of 5×10^{-4} and applying Early Stopping to prevent overfitting and preserve the optimal model weights. Table 2 presents the evaluation results of the baseline ANN and fine-tuned ANN models, which were obtained from five experimental runs with different random seeds to assess the stability and consistency of model performance.

Table 2 ANN Model Evaluation Before and After Fine-Tuning

Model	Training	Accuracy	Precision	Recall	F1-Score
Baseline (ANN)	1	96.49%	98.55%	95.77%	97.14%
	2	97.37%	95.95%	100%	97.93%
	3	97.37%	96.00%	100%	97.96%
	4	97.37%	100%	95.83%	97.87%
	5	99.12%	98.61%	100%	99.30%
	Mean	97.54%	97.82%	98.32%	98.04%
	Std	±1.02%	±1.52%	±1.89%	±0.77%
Fine-Tuned (ANN)	1	96.49%	98.55%	95.77%	97.14%
	2	97.37%	95.95%	100%	97.93%
	3	98.25%	97.30%	100%	98.63%
	4	98.25%	100%	97.22%	98.59%
	5	99.12%	98.81%	100%	99.30%
	Mean	97.89%	98.08%	98.60%	98.32%
	Std	±0.89%	±1.37%	±1.78%	±0.73%

As shown in Table 2, the fine-tuned ANN model demonstrated consistent performance improvements over the baseline model across all evaluation metrics. The average accuracy increased from 97.54% to 97.89%, precision from 97.82% to 98.08%, recall from 98.32% to 98.60%, and F1-score from 98.04% to 98.32%. An important finding is the reduction in standard deviation after fine-tuning across nearly all metrics. The standard deviation of accuracy decreased from ±1.02% to ±0.89%, while the standard deviation of the F1-score decreased from ±0.77% to

±0.73%. This reduction indicates that fine-tuning not only improved the average performance but also enhanced the stability and consistency of the model across variations in weight initialization.

Although the average improvement was relatively moderate, with an accuracy increase of approximately 0.35%, the consistency of the results across the five experimental runs suggests that the improvement was systematic rather than attributable solely to random initialization. This finding supports the claim that fine-tuning made a measurable contribution to improving ANN performance in breast cancer classification. The classification results are further visualized using a confusion matrix, as shown in Figures 6 and 7 in Appendix A. The confusion matrix shows that the ANN model achieved strong classification performance on the WDBC dataset across all cross-validation folds. Both before and after fine-tuning, the model produced a high number of true positives and true negatives, along with a low number of false positives and false negatives.

Fine-tuning improved performance in several folds, particularly by reducing the number of false positives. In Fold 1 and Fold 3, fine-tuning increased the number of correct predictions for the negative class without reducing the model's ability to detect the positive class. In the remaining folds, the model maintained stable performance with relatively consistent results. Overall, the confusion matrix results indicate that the model demonstrated stable generalization capability across different data partitions. The low false negative rate also suggests that the model had high sensitivity in detecting breast cancer cases, which is particularly important for classification tasks in the medical domain.

3.2 XAI Results

The XAI method used is SHAP on an ANN model that has been fine-tuned. A small portion of the training data is taken as background data to help calculate the contribution value of each feature. Then, KernelExplainer is used to calculate the SHAP value on 50 test data samples, which shows how much each feature influences the model's prediction. The next step is to ensure that the SHAP calculation results have the appropriate dimensions so that they can be visualized and analyzed correctly. These results are visualized through summary plots, decision plots, force plots, and waterfall plots, as shown in Figures 2 to 5.

Figure 2 presents the SHAP summary plot, which provides a global interpretation of feature importance in the fine-tuned ANN model. The results indicate that worst area, worst radius, worst texture, mean concave points, and worst smoothness were the most influential features in breast cancer classification. Higher values of these features tended to increase the model output toward malignant tumor prediction, whereas lower values were more strongly associated with benign tumor classification.

These findings are consistent with established medical knowledge, as malignant tumors are generally characterized by larger size, irregular borders, higher concavity, and greater structural heterogeneity. The consistency between the SHAP-based interpretation and these clinical indicators strengthens the validity of the model interpretation. From a clinical perspective, this interpretability improves model transparency and enables healthcare professionals to identify the key factors contributing to the classification results. Therefore, the proposed model has the potential to increase clinical trust and support decision-making in breast cancer diagnosis.

Figure 3 illustrates the SHAP visualization results, indicating that the ANN model predictions were primarily influenced by several key features, including worst area, worst radius, worst concave points, mean radius, and mean area. These features contributed more substantially to the model output than other features and can therefore be regarded as dominant factors in the breast cancer classification task.

From a medical perspective, these findings are consistent with the clinical characteristics represented in the WDBC dataset, in which tumor size, cell nucleus morphology, and tissue irregularity are important indicators in breast cancer diagnosis. Features such as worst radius and worst area are associated with tumor or nucleus size, whereas worst concave points and mean

concavity represent the degree of cellular shape irregularity, which is commonly linked to malignancy. The consistency between the key SHAP-identified features and medical domain knowledge indicates that the model not only achieved high classification performance but also relied on clinically relevant features.

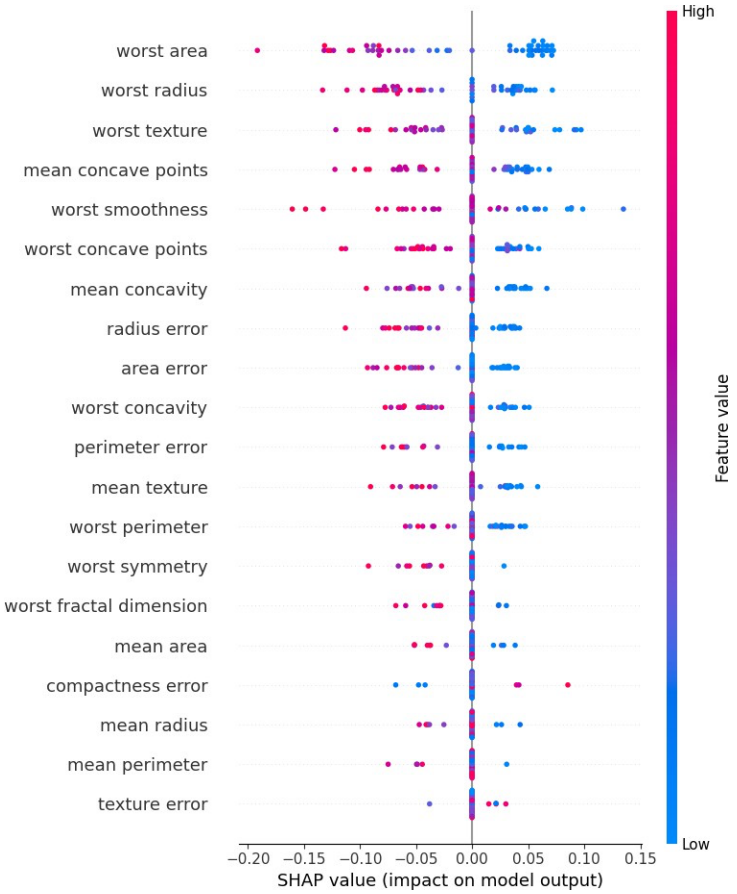

Figure 2 SHAP Summary Plot

In addition to improving model interpretability, SHAP also provides interpretative validation for ANN predictions, which are generally considered black-box models. By examining the contribution of each feature to the model output, the classification results become more transparent and more comprehensible to healthcare professionals and relevant stakeholders. From the perspective of clinical implications, SHAP-based interpretability can help increase trust in AI-based prediction systems for breast cancer diagnosis. Data on the most influential features can be used to support clinical decision-making, allowing the model to function not only as an automated prediction system but also as a more explainable and accountable analytical tool.

Figure 4 depicts the SHAP force plot, which shows the contribution of each feature to the individual predictions generated by the ANN model. In this visualization, features shown in red increase the model output, whereas features shown in blue decrease the model output. The results indicate that worst area, worst radius, worst concave points, and mean radius had dominant contributions to the model output for the analyzed samples.

From a medical perspective, these features are closely related to the morphological characteristics of breast cancer cells. Worst area and worst radius represent the size of abnormal cell masses or nuclei, whereas worst concave points indicate the degree of irregularity in cellular shape. In clinical practice, increased cell size and structural irregularity are important indicators for identifying malignant tumors. The consistency between the key features identified by SHAP

and established medical indicators suggests that the ANN model captured clinically relevant patterns rather than relying solely on random statistical associations in the data.

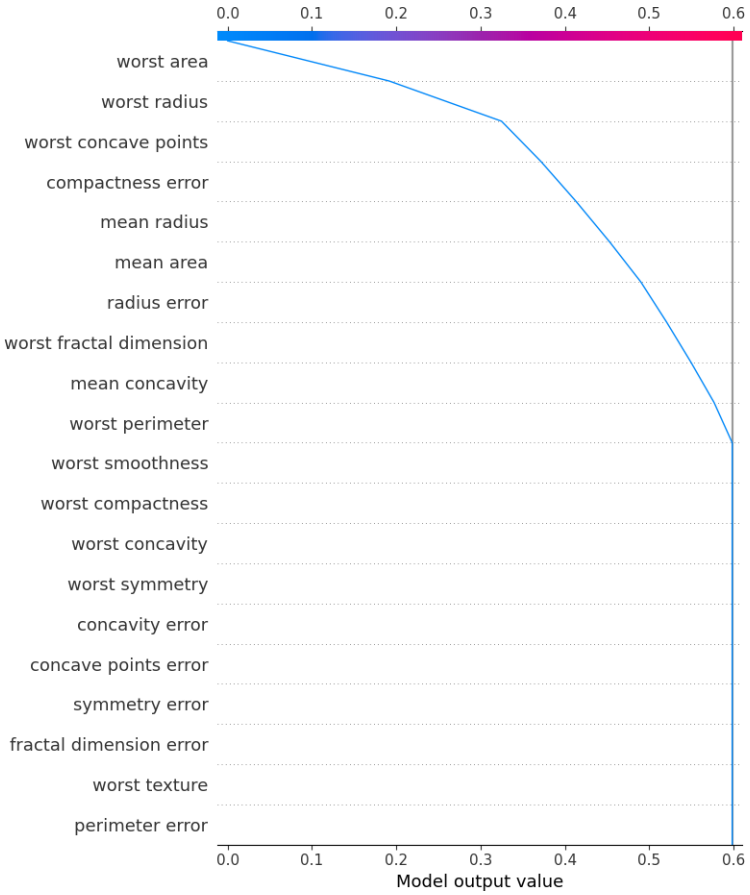

Figure 3 SHAP Decision Plot

Figure 4 SHAP Force Plot (Individual Explanation)

Furthermore, the SHAP force plot provides local interpretability by illustrating how each feature contributes to the prediction of a specific sample. This improves the transparency of the ANN model, which is typically considered a black-box model, and makes the classification output more understandable for healthcare professionals and relevant stakeholders. From the perspective of clinical implications, SHAP-based interpretability has the potential to support the use of AI as a decision support system in breast cancer diagnosis. Data on the features that most strongly influence predictions can help clinicians understand the factors contributing to the model's classification results, increase trust in the AI system, and support a more transparent and accountable clinical evaluation process.

Figure 5 presents the SHAP waterfall plot for a single prediction sample, illustrating the contribution of each feature to the final classification output. The features with the largest contributions were worst area, worst radius, and worst concave points. These features are clinically relevant indicators in breast cancer diagnosis, as malignant tumors are generally associated with larger size, irregular morphology, and higher concavity. The consistency between

the SHAP interpretation and medical domain knowledge further supports the validity of the model interpretation.

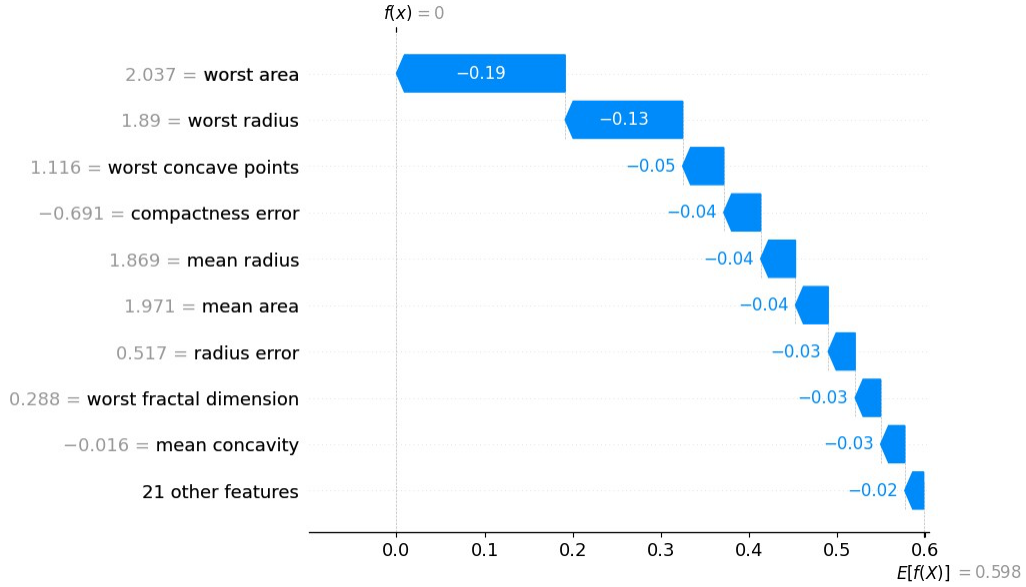

Figure 5 SHAP Waterfall Plot

From a clinical perspective, this visualization helps healthcare professionals identify the primary factors underlying the prediction for an individual case. Thus, model transparency can improve user confidence and support the potential implementation of the model as a decision support tool in breast cancer diagnosis.

3.3 Ablation Study

At this stage, an ablation study was conducted to evaluate the impact of several components of the Artificial Neural Network architecture on breast cancer classification performance. The ablation study was performed using a stratified 5-fold cross-validation scheme to ensure a more robust model evaluation that was not dependent on a single data partition. Each configuration was evaluated using accuracy, precision, recall, and F1-score, along with their standard deviations, to assess the stability of model performance.

The ablation experiments were conducted on several key components, including the number of hidden-layer neurons, dropout rate, activation function, optimizer, and learning rate. The activation functions evaluated in this study included ReLU, tanh, and sigmoid, whereas the optimizers examined were Adam, SGD, and RMSprop. In addition, learning rates of 0.1, 0.01, 0.001, and 0.0001 were evaluated. Each model was trained for a maximum of 50 epochs with a batch size of 16, and Early Stopping was applied to reduce the risk of overfitting. In each fold, the data were normalized using StandardScaler before model training. The scaler was fitted only on the training fold and subsequently applied to the corresponding test fold to prevent data leakage. The evaluation results from all folds were then averaged to obtain the final performance of each configuration. After training, the model generated class probabilities on the test data, which were subsequently converted into binary class labels. Classification accuracy was calculated by comparing the predicted labels with the ground truth labels. The evaluation results are summarized in Tables 3-5.

All three activation functions yielded relatively comparable accuracy values, with differences of less than 0.35%. The sigmoid activation function achieved the highest recall, with a value of 99.16% ± 1.12%, and the lowest standard deviation for recall, indicating a more consistent ability to detect positive, malignant cases. ReLU achieved the highest precision, with a value of 98.11%

± 1.82%, but produced the lowest recall, with a value of 97.77% ± 2.87%, and the highest standard deviation, indicating less stable performance across folds. Meanwhile, tanh showed an intermediate pattern with a relatively balanced performance profile. In the medical context, where high recall is essential to minimize false negatives, the sigmoid activation function can be considered the most suitable for this dataset.

Table 3 Ablation Study Evaluation: Activation Function

Metrik	ReLU	Tanh	Sigmoid
Accuracy	97.37%±1.11%	97.59%±1.16%	97.72%±1.19%
Precision	98.11%±1.82%	97.56%±1.97%	97.32%±2.20%
Recall	97.77%±2.87%	98.60%±1.25%	99.16%±1.12%
F1-Score	97.89%±0.92%	98.06%±0.91%	98.21%±0.92%

Table 4 Ablation Study Evaluation: Optimizer

Metrik	Adam	SGD	RMSProp
Accuracy	97.37%±0.96%	94.03%±1.87%	97.54%±1.02%
Precision	97.60%±2.42%	94.12%±2.72%	97.84%±2.17%
Recall	98.32%±2.06%	96.64%±2.26%	98.33%±1.63%
F1-Score	97.91%±0.74%	95.32%±1.45%	98.05%±0.79%

Performance differences among optimizers were more pronounced than those among activation functions. SGD showed a substantial decrease across all evaluation metrics, with an average accuracy of only 94.03% ± 1.87% and the highest standard deviation, indicating less stable convergence, which may be attributed to its non-adaptive learning rate mechanism. In contrast, Adam and RMSprop produced highly competitive performance. RMSprop slightly outperformed Adam in terms of F1-score, with values of 98.05% ± 0.79% and 97.91% ± 0.74%, respectively. However, Adam produced a lower standard deviation in accuracy, with ±0.96% compared with ±1.02% for RMSprop, indicating more stable performance. Therefore, Adam was used in the main model because it provided a strong balance between classification performance and stability.

Table 5 Ablation Study Evaluation: Learning Rate

Metrik	LR = 0.1	LR = 0.01	LR = 0.001	LR = 0.0001
Accuracy	97.54%±1.16%	98.60%±0.89%	97.54%±1.16%	96.49%±1.24%
Precision	96.49%±1.36%	98.37%±1.57%	97.86%±2.44%	97.28%±1.88%
Recall	99.72%±0.56%	99.44%±1.13%	98.32%±1.64%	97.21%±3.17%
F1-Score	98.07%±0.91%	98.89%±0.71%	98.05%±0.90%	97.19%±1.03%

The learning rate also had a substantial effect on model stability and performance. A learning rate of 0.01 produced the best overall performance, with an accuracy of 98.60% ± 0.89% and an F1-score of 98.89% ± 0.71%, as well as the lowest standard deviation for both metrics. Although a learning rate of 0.1 achieved the highest recall, with a value of 99.72% ± 0.56%, it produced the lowest precision, with a value of 96.49% ± 1.36%. This finding indicates that the model tended to overclassify samples as positive cases. Meanwhile, a learning rate of 0.0001 produced the weakest performance and the highest standard deviation for recall, with a value of ±3.17%, suggesting that an excessively small learning rate prevented the model from reaching optimal convergence within the specified number of epochs.

It should be noted that the learning rate used in the fine-tuning stage was lower than the optimal learning rate obtained in the ablation study. Although a learning rate of 0.01 produced the best result in the ablation experiment, fine-tuning required a smaller learning rate to enable more gradual and stable weight adjustments. Therefore, the lower learning rate used during fine-tuning was intended to refine the previously learned parameters rather than to retrain the model with large parameter updates.

The results of the ablation study were not treated as the final performance of the proposed model. Although several configurations in the ablation study produced accuracy values approaching 99%, these results should not be interpreted as the primary benchmark of the main model. The ablation study was designed to examine the effect of individual hyperparameter variations under a controlled experimental setting. In this stage, the ANN model was retrained using specific configurations to isolate the influence of each component, rather than to replicate the complete experimental pipeline of the proposed model.

In contrast, the main model was evaluated using the complete experimental pipeline, which included fine-tuning, Early Stopping, and comprehensive performance evaluation using multiple classification metrics. Therefore, the accuracy values obtained from the ablation study were used only for comparative analysis across configurations and not as the final performance indicator of the proposed model. The higher accuracy values observed in some ablation configurations may also have been influenced by the relatively small dataset size and the sensitivity of ANN performance to specific hyperparameter combinations.

No indication of data leakage was identified, as the separation between the training set and test set was consistently maintained within the cross-validation procedure across all experiments. Thus, the ablation study served as a complementary analysis to provide insight into the effects of hyperparameter variations, including activation functions, optimizers, and learning rates, on model performance, rather than as the sole basis for determining the final model configuration.

3.4 Comparison with Previous Studies

The results of this study were compared with three relevant previous studies in the domain of breast cancer classification using the Wisconsin Breast Cancer Diagnostic dataset, as presented in Table 6. The comparison indicates that the proposed model achieved an average accuracy of 97.89%, with the highest accuracy reaching 99.12%. These results place the proposed model within a competitive performance range compared with previous studies. Mohd Nasir et al. (2023) reported the highest accuracy of 99.9%; however, their evaluation was conducted using a single experimental run without cross-validation, variance-related stability analysis, fine-tuning, or XAI-based model interpretability. These limitations make it difficult to statistically verify the reliability and consistency of the reported model performance.

Table 6 Comparison with Previous Studies

Author	Best Model	Dataset	Optimization	Accuracy
Mohd Nasir et al. (2023)	ANN	WDBC	-	99.9%
Khater et al. (2025)	ANN	WDBC	-	98.6%
	KNN	WBC	-	97.7%
Fahrurrozi et al. (2025)	SVM	Kaggle	Hyperparameter tuning and feature selection	97%
This research	ANN Fine-Tuned	WDBC	Hyperparameter tuning + Fine tuning	99.12%

Khater et al. applied an XAI approach; however, their study was limited to the worst feature subset of the WDBC dataset, consisting of 10 out of 30 features. In contrast, the present study utilized all available features and applied a more comprehensive evaluation pipeline. Meanwhile, Fahrurrozi et al. (2025) employed hyperparameter tuning and feature selection, but their study did not incorporate model interpretability or standard deviation-based performance stability analysis.

Furthermore, most previous studies relied on a single train-test split or a single-run experimental design without cross-validation. Such an approach may lead to biased evaluation results because model performance can be highly dependent on a particular data partition. This study addresses

these limitations by employing stratified 5-fold cross-validation, resulting in a more robust, stable, and representative evaluation across different data partitions.

The main contribution of this study lies in the integration of several components that have not been jointly applied in previous studies, namely model optimization through fine-tuning, robust evaluation using stratified cross-validation with standard deviation-based stability analysis, and model interpretability using SHAP with four types of visualization. The inclusion of standard deviation values from the cross-validation results indicates that the proposed model not only achieved strong classification performance but also demonstrated consistent performance and stable generalization across different evaluation scenarios.

4. CONCLUSIONS

This study proposed a breast cancer classification model based on an Artificial Neural Network optimized through fine-tuning and enhanced with SHAP-based interpretability. The experimental results showed that the fine-tuned model produced consistent, although relatively moderate, performance improvements compared with the baseline model across several evaluation metrics. The fine-tuned model achieved an average accuracy of 97.89%, precision of 98.08%, recall of 98.60%, and F1-score of 98.32%. The application of stratified 5-fold cross-validation further indicated that the model demonstrated relatively stable performance across different data partitions.

The integration of Explainable Artificial Intelligence provided additional insight into the classification behavior of the model. The SHAP-based analysis consistently identified features such as worst area, worst radius, worst texture, and mean concave points as the most influential factors in the classification output. These findings are consistent with established clinical indicators, suggesting that the model was able to capture clinically relevant patterns in the dataset. The ablation study results indicated that model performance was sensitive to hyperparameter configurations, particularly the activation function, optimizer, and learning rate. Although several configurations produced higher accuracy values in specific experimental settings, these results were used only for comparative analysis and should not be interpreted as the final performance of the proposed main model.

However, this study has several limitations. First, the dataset used in this study was relatively small and well-structured; therefore, it may not fully represent the complexity of real-world clinical data. Second, the model evaluation was conducted solely through internal validation using cross-validation and did not involve external validation on an independent dataset. This limitation restricts the extent to which the model's generalizability can be confirmed. Third, the model has not yet been evaluated in a real clinical environment, where additional challenges such as data heterogeneity, noise, workflow integration, and operational constraints may arise.

Therefore, although the proposed approach demonstrated promising results in terms of classification performance and interpretability within the experimental scope of this study, further research is required before the model can be practically implemented in a clinical setting. Future studies are recommended to employ larger and more diverse datasets, conduct external validation using independent data, and evaluate model robustness under real-world clinical conditions.

REFERENCES

Dubaish, A. A., & Jaber, A. A. (2024). Comparative Analysis of SVM and ANN for Machine Condition Monitoring and Fault Diagnosis in Gearboxes. *Mathematical Modelling of Engineering Problems*, *11*(4), 976–986. https://doi.org/10.18280/mmep.110414

El Jihaoui, M., Abra, O. E. K., & Mansouri, K. (2025). Predicting and Interpreting Student Academic Performance: A Deep Learning and Shapley Additive Explanations Approach. *SHS Web of Conferences*, *214*, 01001. https://doi.org/10.1051/shsconf/202521401001

Fahrurrozi, F., Aminullah, M., Febriyanto, E., Adityo, H., & Setiawan, M. F. (2025). Optimasi Model Klasifikasi Kanker Payudara Melalui Pendekatan Decision Tree, Artificial Neural Network, dan Support Vector Machine. *Jurnal Ilmiah Kesehatan*, *14*(1), 90–101. https://doi.org/10.52657/jik.v14i1.2674

Ghanem, M., Ghaith, A. K., El-Hajj, V. G., Bhandarkar, A., de Giorgio, A., Elmi-Terander, A., & Bydon, M. (2023). Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain Sciences*, *13*(12), Article ID: 1723. https://doi.org/10.3390/brainsci13121723

Gunawan, I. (2020). Optimasi Model Artificial Neural Network untuk Klasifikasi Paket Jaringan. *SIMETRIS*, *14*(2), 1–5. https://doi.org/10.51901/simetris.v14i2.135

Jepkoech, J., Mugo, D. M., Kenduiywo, B. K., & Too, E. C. (2021). The Effect of Adaptive Learning Rate on the Accuracy of Neural Networks. *International Journal of Advanced Computer Science and Applications*, *12*(8), 736–751. https://doi.org/10.14569/IJACSA.2021.0120885

Jlassi, O., & Dixon, P. C. (2024). The Effect of Time Normalization and Biomechanical Signal Processing Techniques of Ground Reaction Force Curves on Deep-Learning Model Performance. *Journal of Biomechanics*, *168*, Article ID: 112116. https://doi.org/10.1016/j.jbiomech.2024.112116

Khan, I. U., Azam, S., Montaha, S., Mahmud, A. Al, Rafid, A. K. M. R. H., Hasan, Md. Z., & Jonkman, M. (2022). An Effective Approach to Address Processing Time and Computational Complexity Employing Modified CCT for Lung Disease Classification. *Intelligent Systems with Applications*, *16*, Article ID: 200147. https://doi.org/10.1016/j.iswa.2022.200147

Khater, T., Hussain, A., Bendardaf, R., Talaat, I. M., Tawfik, H., Ansari, S., & Mahmoud, S. (2025). An Explainable Artificial Intelligence Model for the Classification of Breast Cancer. *IEEE Access*, *13*, 5618–5633. https://doi.org/10.1109/ACCESS.2023.3308446

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks*. http://arxiv.org/abs/1901.08644

Miftakhudin, M., Murtopo, A. A., & Arif, Z. (2025). Integrasi Artificial Neural Network dan Algoritma Genetika untuk Prediksi Bencana Banjir Pesisir Kota Tegal. *RIGGS: Journal of Artificial Intelligence and Digital Business*, *4*(3), 840–848. https://doi.org/10.31004/riggs.v4i3.2068

Mohd Nasir, H., Brahin, N. M. A., Zainuddin, S., Mispan, M. S., Md Isa, I. S., & Sha'abani, M. N. A. H. (2023). The Comparative Study of Deep Learning Neural Network Approaches for Breast Cancer Diagnosis. *International Journal of Online and Biomedical Engineering (IJOE)*, *19*(06), 127–140. https://doi.org/10.3991/ijoe.v19i06.34905

Montaha, S., Azam, S., Rafid, A. K. M. R. H., Hasan, Md. Z., Karim, A., & Islam, A. (2022). TimeDistributed-CNN-LSTM: A Hybrid Approach Combining CNN and LSTM to Classify Brain Tumor on 3D MRI Scans Performing Ablation Study. *IEEE Access*, *10*, 60039–60059. https://doi.org/10.1109/ACCESS.2022.3179577

Pantalé, O. (2023). Comparing Activation Functions in Machine Learning for Finite Element Simulations in Thermomechanical Forming. *Algorithms*, *16*(12), Article ID: 537. https://doi.org/10.3390/a16120537

Purwono, P., Wulandari, A. N. E., & Nisa, K. (2025). Explainable Artificial Intelligence (XAI) in Medical Imaging: Techniques, Applications, Challenges, and Future Directions. *Advanced Mechanical and Mechatronic Systems*, *1*(1), 52–66. https://doi.org/10.53623/amms.v1i1.692

Rahim, M. N., & Basheer, K. P. M. (2025). A Hybrid Explainability Framework for Recommender Systems Using SHAP with Natural Language Interpretations. *International Journal of Innovative Research and Scientific Studies*, *8*(6), 687–703. https://doi.org/10.53894/ijirss.v8i6.9669

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *Knowledge-Based Systems*, *263*, Article ID: 110273. https://doi.org/10.1016/j.knosys.2023.110273

Saputra, E. P., Indriyanti, I., & Supriatiningsih, S. (2020). Classifications Using Artificial Neural Network Method In Protecting Credit Fitness. *Indonesian Journal of Artificial Intelligence and Data Mining*, *3*(1), 50–56. https://doi.org/10.24014/ijaidm.v3i1.9442

Susilo, A., Christanti, V., & Lauro, M. D. (2024). Fine-Tuning LLaMA-2-Chat untuk ChatBot Penerjemah Bahasa Gaul Menggunakan LoRA dan QLoRA. *MIND Journal*, *9*(2), 248–260. https://doi.org/10.26760/mindjournal.v9i2.248-260

World Health Organization. (2025). *Breast Cancer*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 3, SEPTEMBER, 2026, Pp. 300 – 317
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

APPENDIX A

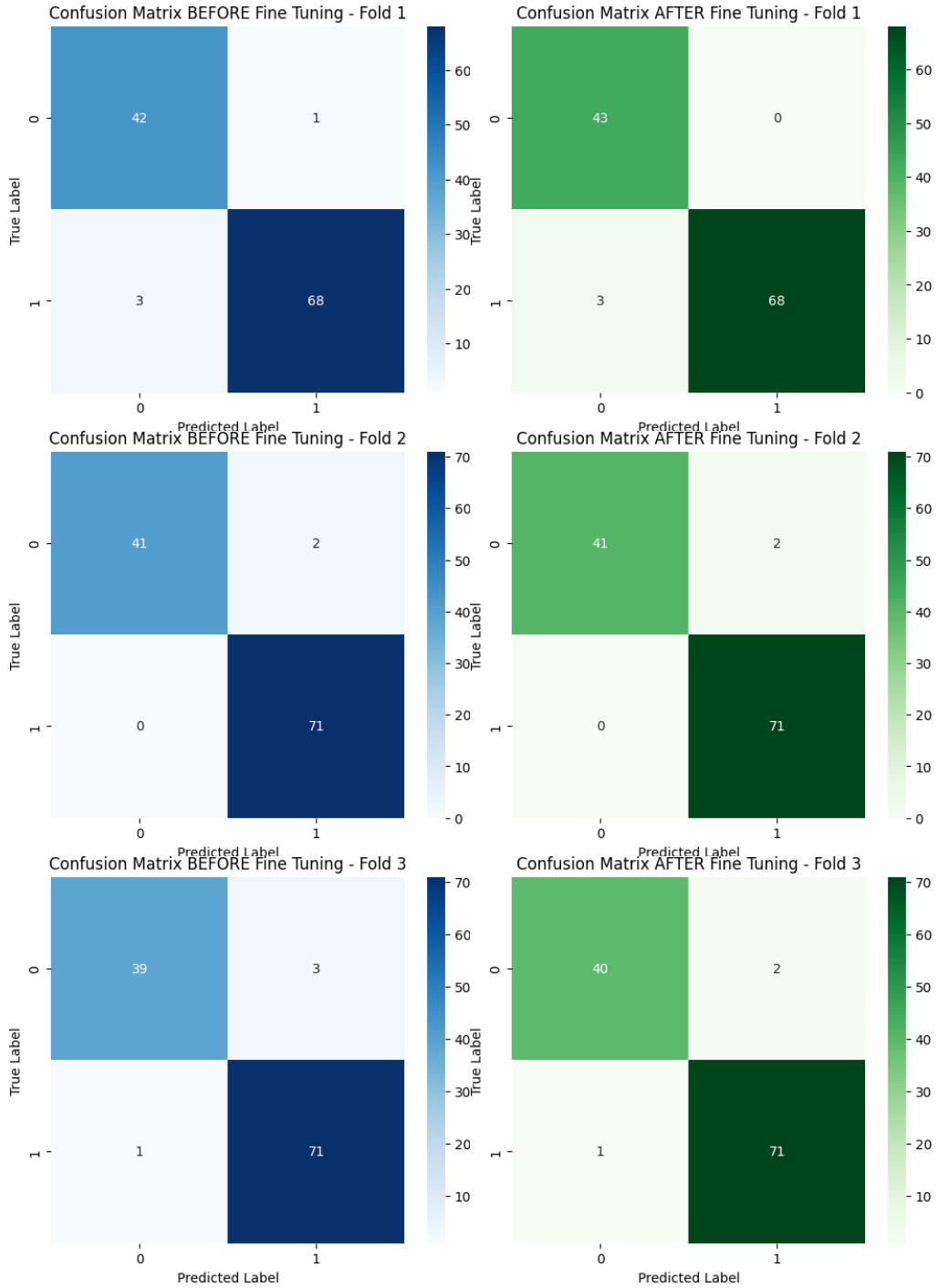

Figure 6 Confusion Matrix Before and After Fine-Tuning

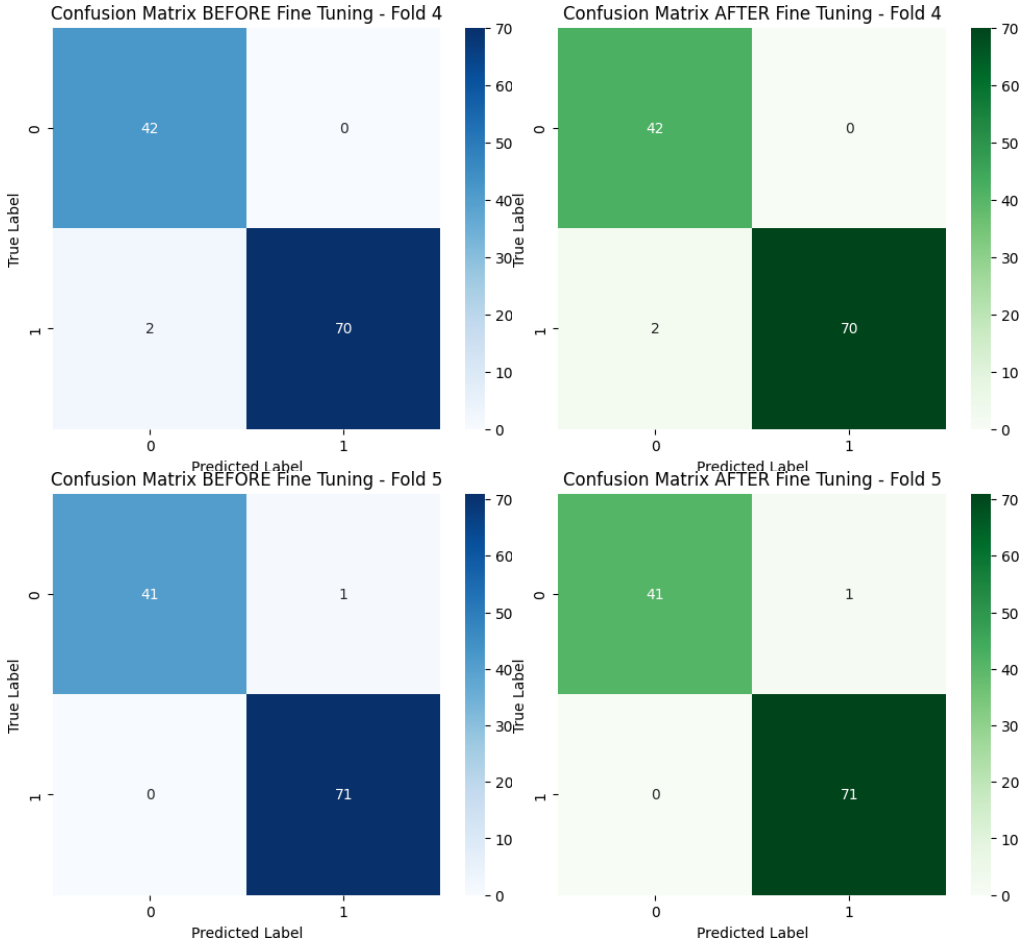

Figure 7 Confusion Matrix Before and After Fine-Tuning (Continued)