INDONESIAN JOURNAL OF LIFE SCIENCES

RESEARCH ARTICLE

Check for updates

OPEN ACCESS

Comparative Analysis for CNN and MLP Models in Breast Cancer Diagnosis

Priscilla Natalie Nurtanio[1], Darren Nathaniel[1], Temmy Sugiarto[1], Theresa Angelina[1], Raymond Tjandra[1], Yohana Joevanca Kurniawan[1], Maria Zefanya Sampe[1]*

[1]*School of Applied STEM, Universitas Prasetiya Mulya, BSD City, Tangerang, Banten, 15339, Indonesia*

**Corresponding author: maria.zefanya@prasetiyamulya.ac.id*

ARTICLE INFO

Article history:

Submitted June 5, 2025
Revised August 25, 2025
Accepted February 23, 2026

DOI: 10.54250/ijls.v8i01.254

KEYWORDS:

Breast cancer, Convolutional Neural Network (CNN), Diagnostic accuracy, Hyperparameter tuning, Multilayer Perceptron (MLP)

HIGHLIGHTS

- Optimized 1D-CNN architecture outperforms MLP models for breast cancer classification

ABSTRACT

Breast cancer remains a leading cause of global female mortality, necessitating accurate early diagnosis to improve survival rates. Traditional screening methods, such as mammography, often exhibit sensitivity limitations, particularly in dense breast tissue. This study evaluates the efficacy of Convolutional Neural Networks (CNN) and Multilayer Perceptron (MLP) models on the structured Wisconsin Diagnostic Breast Cancer (WDBC) dataset. A systematic comparison was conducted between baseline and hyperparameter-optimized architectures using a rigorous 5-fold stratified cross-validation framework to ensure statistical reliability. The results demonstrate that the Optimized CNN outperforms all other configurations, achieving a mean test accuracy of 98.86% and a critical recall rate of 99.09%. By effectively minimizing false negatives without compromising precision, the tuned 1D-CNN architecture offers a robust, automated diagnostic safeguard that significantly enhances the reliability of breast cancer detection workflows compared to traditional MLP approaches.

INTRODUCTION

Breast cancer remains one of the leading causes of death among women worldwide, especially for those aged 40 to 70 years (Sung et al., 2021). Each year, about 2.3 million new cases are diagnosed, and this number is expected to rise by 12–15% by 2025. Detecting breast cancer at an early stage, particularly stage 0 or 1, is crucial, as it can lead to survival rates of over 99% (American Cancer Society, 2024). Currently, mammography is the standard tool for early screening and clinical breast cancer detection (Reeves & Kaufman, 2023). However, it has notable limitations, especially in women with dense breast tissue, where it may fail to detect 15–20% of cancer cases (Lehman et al., 2015). These limitations highlight the urgent need for more accurate and supportive diagnostic tools to reduce missed diagnoses and assist radiologists in making better decisions. Breast cancer data is particularly chosen for this study as it addresses an important health concern for women. By focusing on early detection of breast cancer, we aim to contribute to improving healthcare outcomes for women globally and offer a solution to help combat this pressing issue.

Artificial intelligence (AI) has emerged as a promising solution to these limitations, offering enhanced accuracy and consistency in medical diagnostics. In particular, AI systems utilizing machine learning (ML) and deep learning (DL) have demonstrated capabilities that rival and sometimes surpass expert radiologists in breast cancer detection (Rodríguez-Ruiz et al., 2019; Araujo et al., 2017; Mannarsamy et al.,2025). Among these approaches, Convolutional Neural Networks (CNNs) are widely recognized for their ability to learn hierarchical feature representations and capture complex relationships within data, while Multilayer Perceptron (MLPs) remain effective for structured datasets and classification tasks.

Traditional machine learning models, such as Bayesian Networks and Multilayer Perceptrons (MLPs), have achieved strong results using structured data, with some reaching up to 97% accuracy (Kooi et al., 2017; McKinney et al., 2020). However, these models often rely on manual feature engineering and may be limited in capturing complex nonlinear relationships among variables (Yu et al., 2023). Convolutional Neural Networks (CNNs), while commonly associated with imaging applications, provide an alternative approach by enabling automated feature interaction learning and hierarchical representation extraction (Desai & Shah, 2021). Although CNNs have shown strong results in large-scale imaging diagnostics, their potential in structured diagnostic datasets remains an area of interest (McKinney et al., 2020). In this study, MLPs and CNNs are evaluated as complementary models, where MLPs serve as strong baselines for structured data and CNNs are adapted to capture feature interactions within numerical inputs. Nevertheless, both architectures are highly sensitive to hyperparameter configuration and training strategy, and improper tuning can lead to overfitting or unstable performance (Sahu et al., 2023).

To address this, our study investigates the optimization of CNN and MLP performance through systematic hyperparameter tuning and compares their diagnostic accuracy to that of both standard CNNs and MLPs. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which contains structured features derived from cytological images, we assess baseline and optimized models to examine improvements in predictive accuracy and generalization. We hypothesize that optimized CNN architectures can better capture complex feature relationships and improve diagnostic sensitivity, particularly in minimizing false negatives, compared with traditional MLP models (Alanazi et al., 2021).

MATERIALS AND METHODS

Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset is a well-established medical dataset widely used in machine learning research, particularly for breast cancer classification tasks. It consists of 569 samples obtained from digitized images of fine needle aspirates (FNA) of breast masses, with each

sample labelled as either benign or malignant. Feature extraction was performed on the cell nuclei in the images, calculating ten characteristics such as radius, texture, perimeter, area, smoothness, and fractal dimension, each summarized by their mean, standard error, and the largest value observed among the three most extreme cases, resulting in 30 real-valued features per sample. The dataset contains no missing values and is known to be linearly separable, making it suitable for evaluating the performance of various classification algorithms. It has been used extensively in both medical and machine learning literature due to its reliable structure and strong predictive potential, achieving up to 97.5% accuracy with linear models.

Although Convolutional Neural Networks (CNNs) are commonly associated with image-based analysis, this study employs the WDBC dataset, which consists of structured numerical features derived from cytological images rather than raw medical imaging data. Consequently, the models evaluated in this study learn patterns from tabular diagnostic features rather than pixel-level representations. While this allows controlled comparison between CNN and MLP architectures, the findings should be interpreted within the context of structured diagnostic modeling. Further validation using large-scale imaging datasets, such as mammography or histopathology images, is necessary to assess clinical generalizability.

Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are traditionally recognized for their superior performance in image classification due to their ability to learn spatial hierarchies from raw pixel data. However, in this study, the CNN architecture is specifically adapted to process the WDBC dataset, which consists of structured numerical features rather than raw medical images. To accommodate this data format, a 1D-CNN (One-Dimensional CNN) is utilized, treating the 30 diagnostic features as a sequential vector.

Unlike Multilayer Perceptrons (MLPs), which process data through global connections and often rely on manual feature engineering, the 1D-CNN applies local filters to identify complex correlations between adjacent diagnostic variables. While MLPs are effective for structured data, they are more susceptible to overfitting when handling high-dimensional diagnostic features without extensive regularization. The typical structure of a CNN includes convolutional layers, activation functions, pooling layers, and fully connected layers. The CNN mechanism implemented in this study follows these sequential steps:

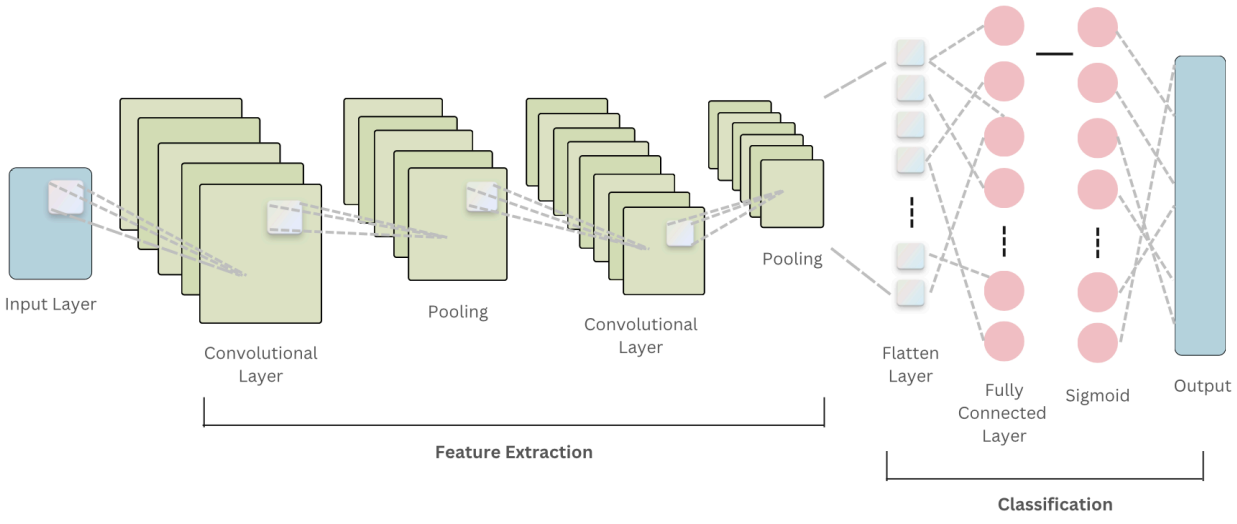

Figure 1. Convolutional Neural Network (CNN) algorithm

1. **Input data**
   The process starts with data input, which could be image data or structured numerical data. This data is fed into the CNN model for feature extraction and classification.
2. **Convolution layer**
   The convolution layer applies filters to the data to detect important features or relationships. These filters scan local regions of the input data, and an activation function is applied afterward to add non-linearity, enabling the network to learn more complex patterns.
3. **Pooling layer**
   The pooling layer reduces the dimensionality of the data, preserving the key features while reducing computational complexity. It helps make the model more efficient by focusing on the most important aspects of the data. These convolution and pooling layers can be multiplied several times depending on the model architecture.
4. **Flatten layer**
   The flatten layer transforms the multi-dimensional data into a one-dimensional vector. In the case of numerical data, this simply means preparing the data to pass into the fully connected layers for classification or prediction.
5. **Fully connected layer**
   The fully connected layer connects all the neurons from the previous layers and helps combine the learned features to make the final prediction. This step allows the model to make sense of the features learned and classify the data into different categories.
6. **Sigmoid layer**
   The sigmoid activation function outputs a probability between 0 and 1 for binary classification tasks. This helps the model determine the likelihood of the data belonging to a particular class
7. **Output**
   The final output represents the model's prediction, indicating the classification result based on the probability generated by the sigmoid function. If the output is closer to 1, the model classifies the data as malignant (for example), and if closer to 0, it classifies it as benign.

Multi-Layer Perceptron (MLP)

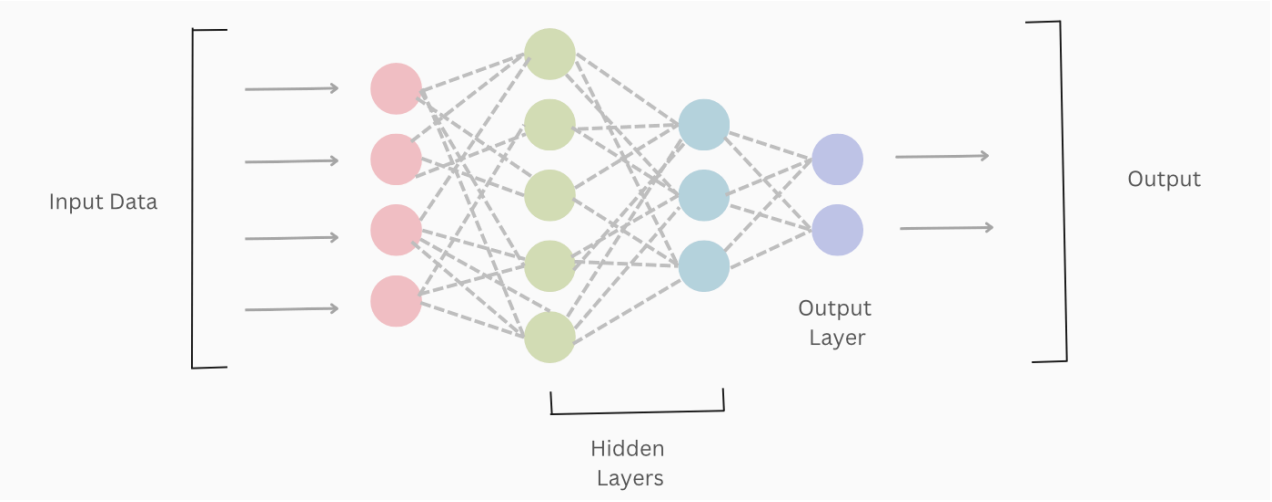

Figure 2. Multilayer Perceptron (MLP) algorithm

A Multilayer Perceptron (MLP) is a type of artificial neural network composed of an input layer, one or more hidden layers, and an output layer (Desai et al., 2021). It is used for tasks like classification and regression on structured data, where each neuron in one layer is connected to every neuron in the next layer. The input layer acts as the receiver, taking in the data, while the hidden layers process and compute the data through several iterations using activation functions (Lin et al., 2023). The output layer then predicts the results after processing the data. MLPs are ideal for problems where data lacks complex spatial relationships, and they rely on manual feature engineering, meaning features must be predefined before training. MLPs are commonly used for structured datasets, such as numerical values in breast cancer detection, where backpropagation is used to improve accuracy.

MLPs stand out from simpler models due to their ability to learn non-linear relationships through activation functions such as ReLU, sigmoid, or tanh. These functions enable the network to capture complex data patterns that linear models fail to represent. MLPs follow a feedforward architecture, where data flows in one direction from input to output. The learning process is guided by backpropagation, which calculates the error between predicted and actual outcomes and iteratively updates the model's weights using optimization algorithms like stochastic gradient descent (SGD). This mechanism allows MLPs to adjust internal parameters and gradually improve performance across training epochs.

The performance of an MLP is significantly influenced by the design of its architecture, especially the number of hidden layers and neurons. A shallow network may lead to underfitting, while a deep and overly complex model may result in overfitting if not properly regularized. To address these issues, techniques such as dropout, L2 regularization, and batch normalization are often implemented. Moreover, hyperparameter tuning plays a critical role in maximizing model generalization. Despite the emergence of more advanced neural architectures, MLPs remain a powerful and versatile choice for a wide range of structured data problems when configured effectively.

MLPs serve as a strong baseline for structured diagnostic modeling due to their ability to learn from numerical feature representations (Silas et al., 2025). In this study, their performance is compared with CNN-based architectures adapted for structured input to evaluate differences in learning feature interactions and diagnostic prediction capability.

Flowchart

Figure 3 shows the flow chart for this research. This flowchart outlines the workflow of developing a neural network model, starting from data input, preprocessing, feature engineering, and data splitting. Based on the selected model (MLP or CNN), its corresponding steps are applied. After parameter initialization, hyperparameter tuning may occur. The model is then trained iteratively and tested. Finally, predictions are made, and error metrics are evaluated.

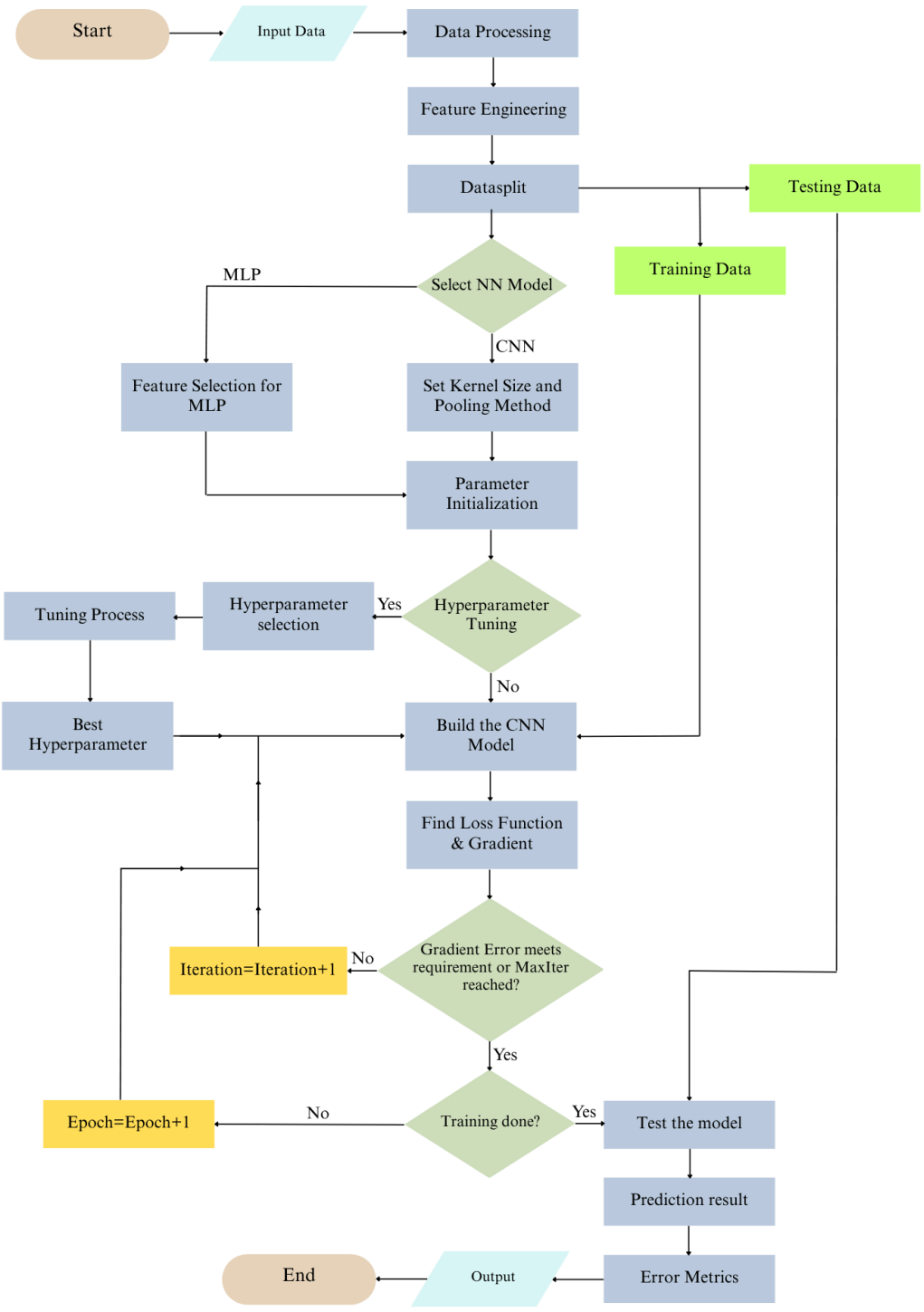

Figure 3. Flowchart

Cross-validation

As illustrated in the flowchart (**Figure 3**), the original dataset is partitioned during the datasplit stage into training and testing subsets through a 5-fold Stratified Cross-Validation process. This iterative partitioning maintains the original class distribution of benign and malignant cases within each subset to

prevent sampling bias and ensure that the performance metrics represent the model's true diagnostic capability across the entire dataset.

Evaluation metric

In breast cancer detection, minimizing false negatives is very important because missing a cancerous tumor can delay treatment and worsen the situation. To check how well the model works accurately, five key measurements are used:

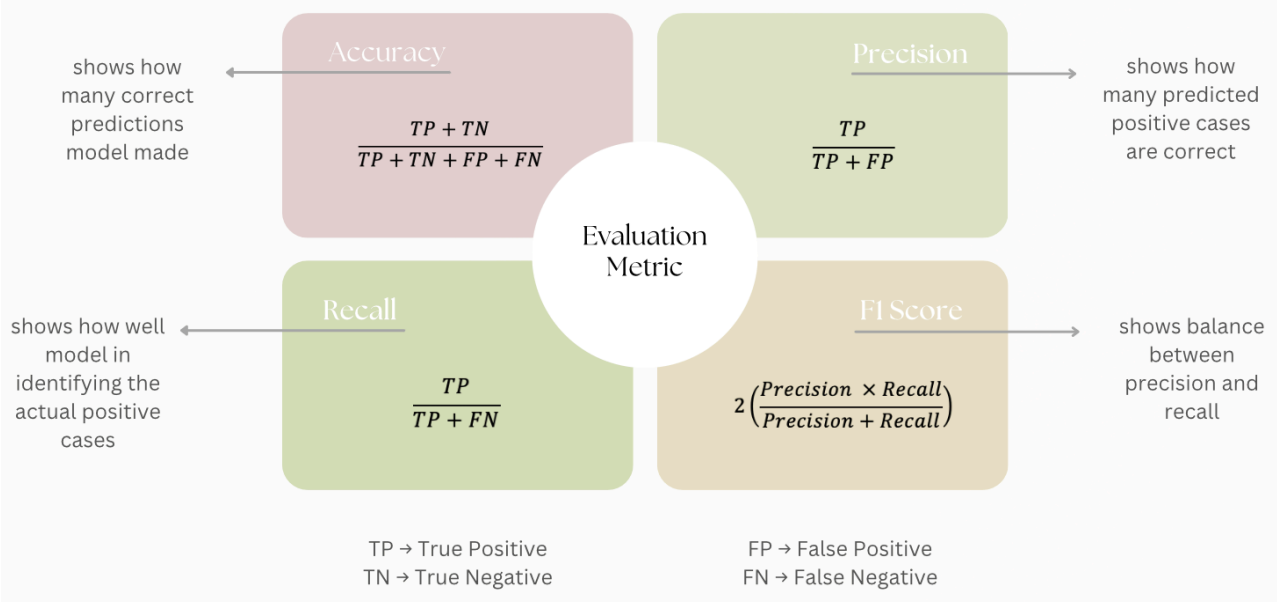

Figure 4. Evaluation metric

While the four metrics (Accuracy, Precision, Recall, F1-Score) provide essential insights into model performance, AUC (Area Under the Curve) serves as an additional critical measure. AUC (Area Under the Curve) is intrinsically linked to the ROC (Receiver Operating Characteristic) curve, as AUC quantifies the performance of a classifier by measuring the area under the ROC curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), giving us a visual representation of how well the model distinguishes between positive and negative cases. A higher AUC value indicates better model performance, as it reflects the model's ability to consistently rank positive instances higher than negative ones across various thresholds.

$$AUC = \int TPR\, d(FPR) \#(1)$$

Equation 1. Area Under the Curve (AUC) equation

The ROC curve is a fundamental tool for evaluating the performance of classification models, particularly when dealing with imbalanced datasets. It provides a comprehensive view of the model's ability to distinguish between positive and negative classes by plotting the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). Each point on the ROC curve corresponds to a different threshold value used to classify the data. A model that classifies instances well will produce a curve that bends towards the top-left corner of the graph, indicating a high true positive rate and low false positive rate. In contrast, a random classifier would produce a diagonal line from the bottom left to the top right, suggesting no ability to discriminate between classes.

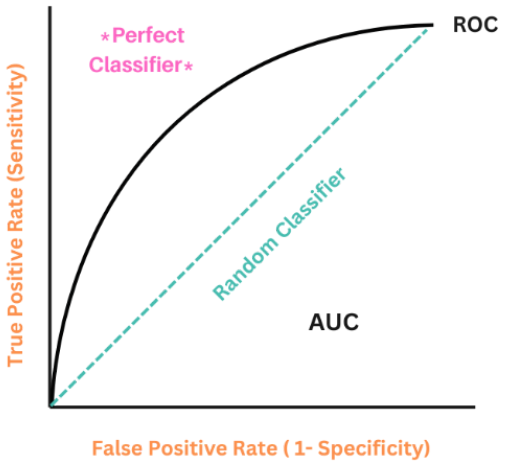

Figure 5. ROC curve visualization

The ROC curve helps to understand the classifier's performance across all possible thresholds, rather than relying on a single threshold. It is especially useful in scenarios where the costs of false positives and false negatives vary, allowing selection of a threshold that best suits the application. The AUC value derived from the ROC curve is a single scalar metric that summarizes the model's performance. An AUC of 1 represents a perfect classifier, while an AUC of 0.5 indicates a model that performs no better than random guessing. As AUC considers all thresholds, it offers a more complete evaluation of model performance compared to metrics like accuracy, which can be heavily influenced by class imbalance.

RESULTS

Data description and explanatory

The dataset used is the Breast Cancer Wisconsin (Diagnostic) Dataset, which consists of 569 samples with 30 numerical features extracted from digital images of breast tumor biopsies. Each sample is classified as either benign (non-cancerous) or malignant (cancerous), with 357 benign and 212 malignant samples. **Table 1** shows the variables used in this dataset.

Table 1. Variables of dataset

Variable	Description	Common Units
id	Unique identification number for each sample.	None (identifier)
diagnosis	Tumor diagnosis (M = Malignant, B = Benign).	Categorical (M = Malignant, B = Benign)
radius_mean	Mean distance from the center to points on the cell perimeter.	Millimeters (mm)
texture_mean	Mean standard deviation of gray-scale values in the image.	Unitless
perimeter_mean	Mean perimeter length of the cell.	Millimeters (mm)
area_mean	Mean area of the cell.	Square Millimeters (mm^2)
smoothness_mean	Mean of local variation in radius lengths.	Unitless

Variable	Description	Common Units
compactness_mean	Mean of (perimeter^2 / area - 1.0).	Unitless
concavity_mean	Mean severity of concave portions of the cell contour.	Unitless
concave_points_mean	Mean number of concave portions of the cell contour.	Unitless
symmetry_mean	Mean symmetry of the cell.	Unitless
fractal_dimension_mean	Mean approximation of the cell's fractal dimension (roughness).	Unitless
radius_se	Standard error for the radius.	Millimeters (mm)
texture_se	Standard error for the texture.	Unitless
perimeter_se	Standard error for the perimeter.	Millimeters (mm)
area_se	Standard error for the area.	Square Millimeters (mm^2)
smoothness_se	Standard error for the smoothness.	Unitless
compactness_se	Standard error for the compactness.	Unitless
concavity_se	Standard error for the concavity.	Unitless
concave_points_se	Standard error for the number of concave points.	Unitless
symmetry_se	Standard error for the symmetry.	Unitless
fractal_dimension_se	Standard error for the fractal dimension.	Unitless
radius_worst	Worst (largest) value for radius.	Millimeters (mm)
texture_worst	Worst value for texture.	Unitless
perimeter_worst	Worst value for perimeter.	Millimeters (mm)
area_worst	Worst value for area.	Square Millimeters (mm^2)
smoothness_worst	Worst value for smoothness.	Unitless
compactness_worst	Worst value for compactness.	Unitless

Figure 6 visualizes the distribution of diagnoses in the Breast Cancer dataset. Based on the available data, 357 individuals were diagnosed as "Benign" and 212 as "Malignant."

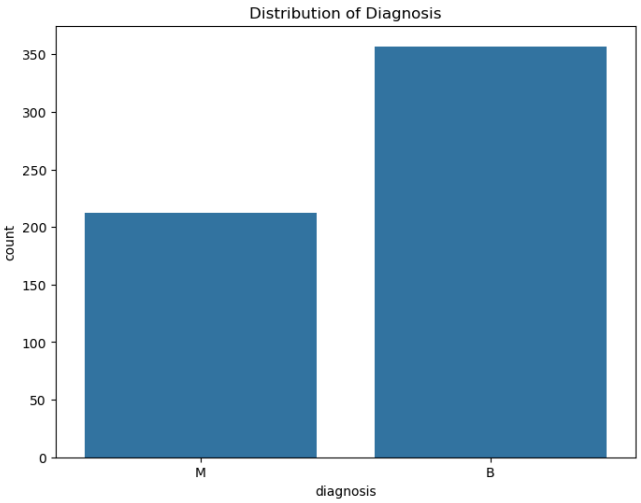

Figure 6. Distribution of diagnosis plot. *Abbreviations: B, Benign; M, Malignant*

To explore and visualize the data, an examination of the feature distributions for 'radius_mean', 'texture_mean', 'perimeter_mean', and 'area_mean' was conducted. **Figure 7** illustrates the overall distribution of each feature, without considering the diagnosis.

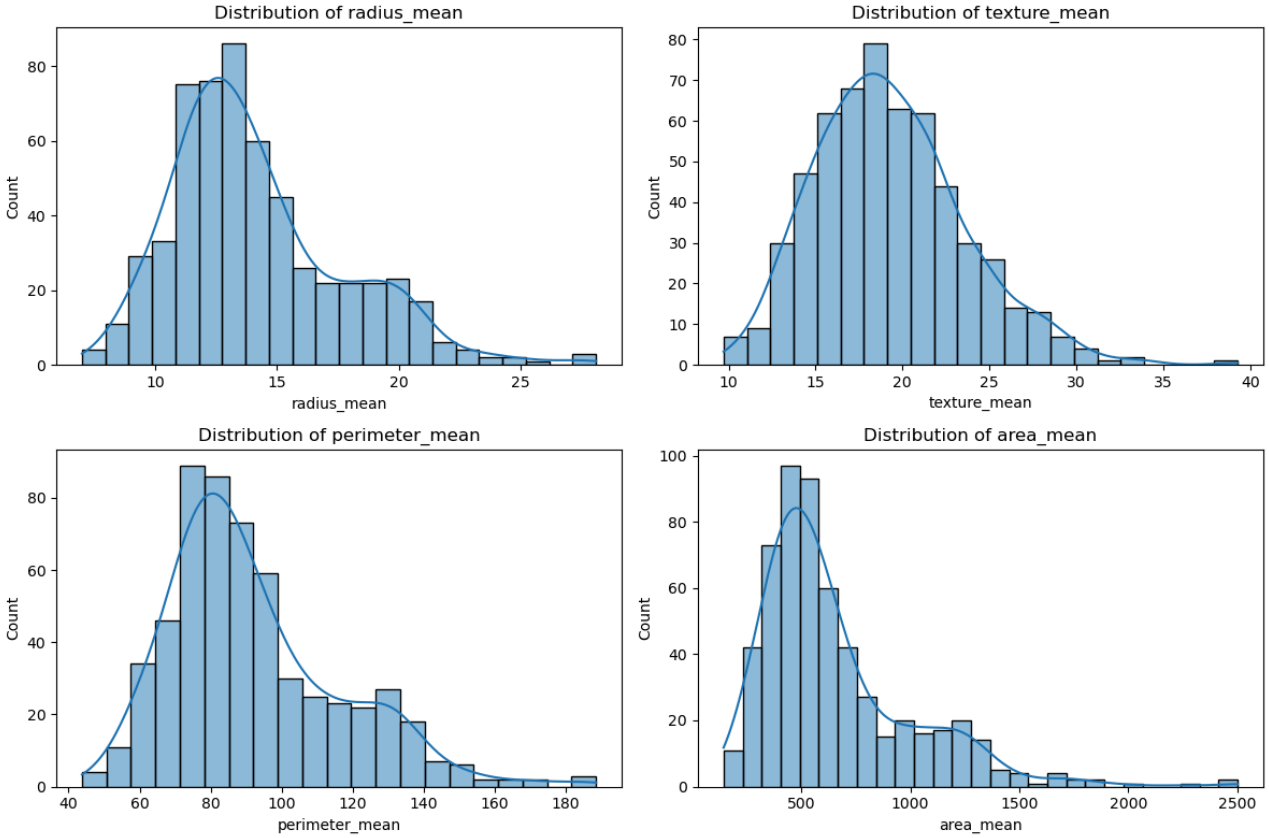

Figure 7. Distribution of mean variables plot

Based on **Figure 7**, radius_mean has a distribution that is close to normal but not perfectly symmetrical and is positively skewed. The texture_mean feature has a distribution that is almost normal and slightly right-skewed. Perimeter_mean shows a non-symmetrical distribution with more data

concentrated on the left, making it positively skewed. As for area_mean, its distribution is clearly non-normal and highly positively skewed, indicating the presence of many potential outliers. To highlight the differences in feature distributions between the two diagnosis classes, 'M' (Malignant) and 'B' (Benign), a box plot is used. This is useful for determining whether a feature effectively distinguishes between the two classes. **Figure 8** presents box plots illustrating the distribution of features categorized by diagnosis.

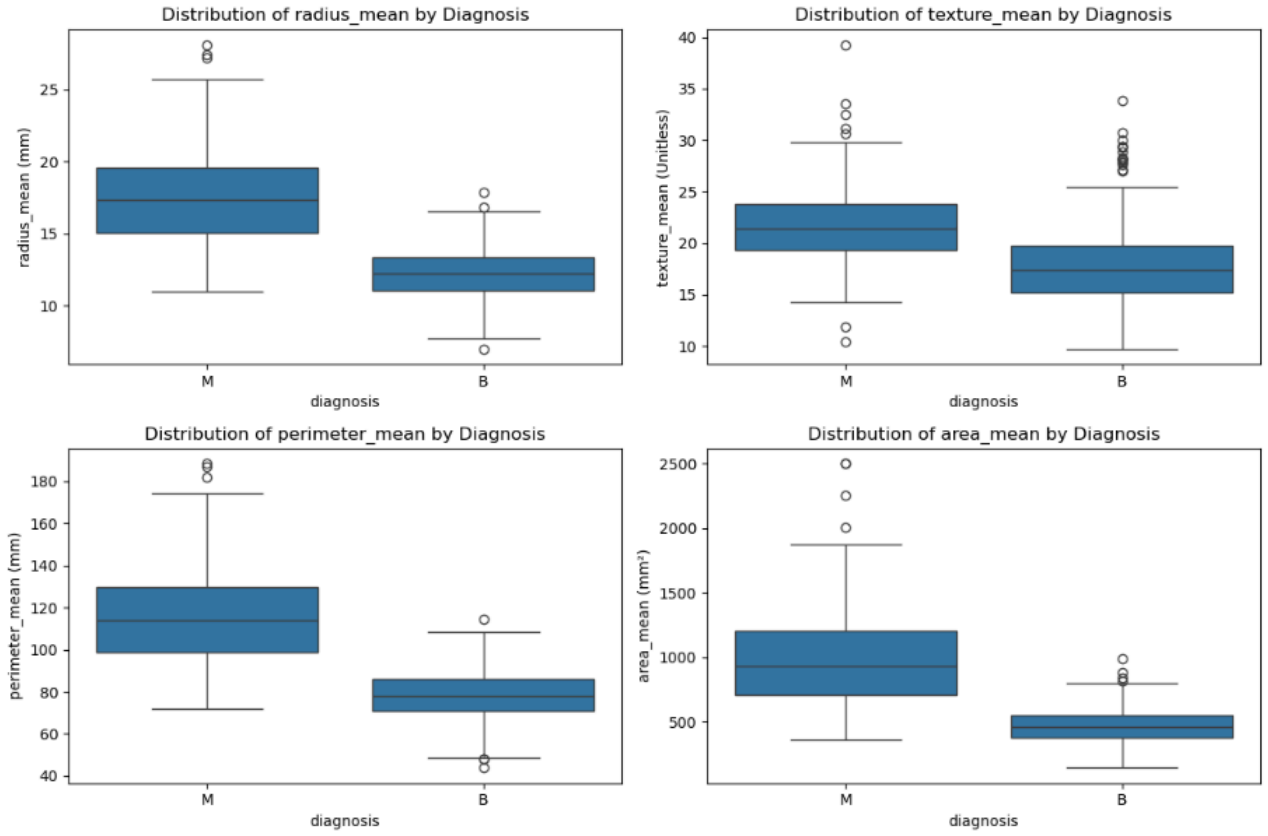

Figure 8. Distribution of mean variables by diagnosis box plots. *Abbreviations: B, Benign; M, Malignant*

Based on **Figure 8**, the radius_mean values are generally higher for malignant cases than for benign ones. This indicates that the feature may serve as a good indicator for diagnosis classification. In contrast, texture_mean shows less distinction between malignant (M) and benign (B), making it a weaker feature compared to radius_mean or area_mean. The perimeter_mean also tends to be higher in malignant cases, suggesting it is a fairly useful feature for distinguishing diagnoses. Lastly, area_mean has a significantly higher median and range for malignant than benign cases, making it the most effective feature among the four for distinguishing malignant and benign tumors. With this insight, feature selection will be carried out to improve feature quality for better accuracy, reduce skewness, and prepare the data for machine learning algorithms.

Feature engineering

Furthermore, feature engineering is performed as a crucial step for generating more informative and representative features that capture non-linear relationships and mitigate skewness, thereby enhancing the accuracy and performance of subsequent machine learning models. **Figure 9** below shows the distribution for each feature engineering performed.

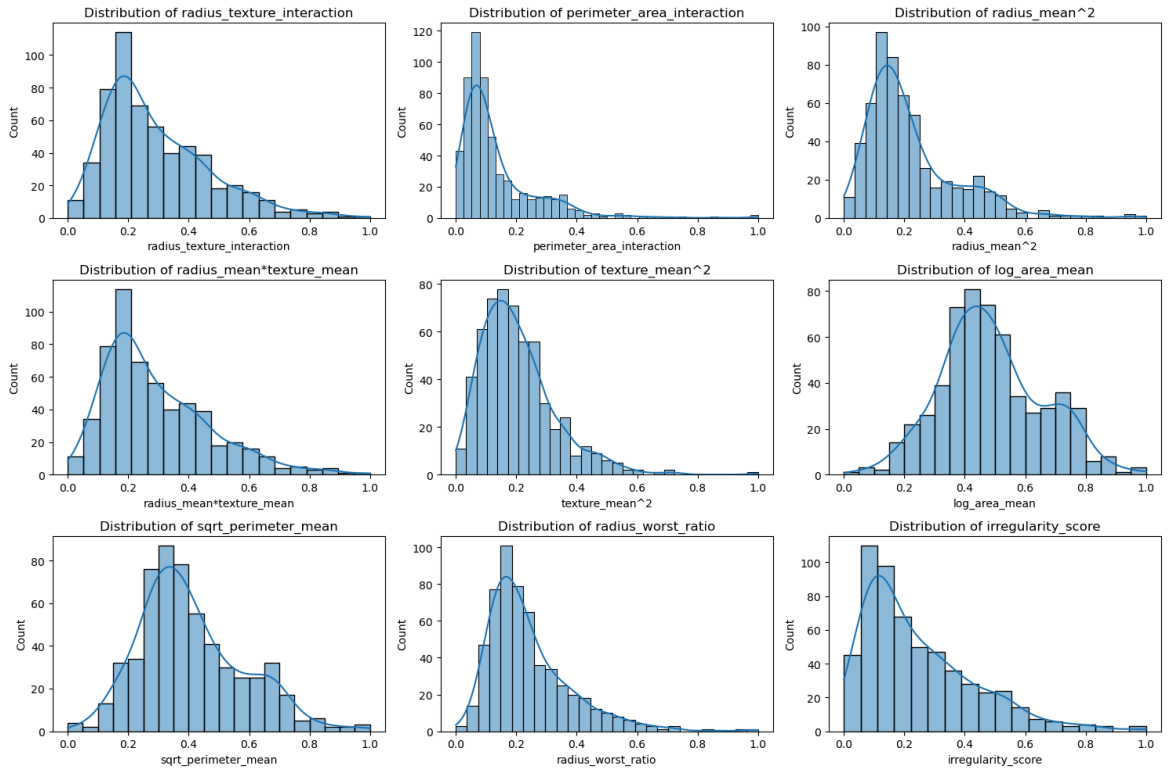

Figure 9. Distribution for feature engineering plot

In detail, the results of feature engineering include the creation of interaction features such as radius_texture_interaction and perimeter_area_interaction, which aim to capture the relationship between tumor size and texture. Additionally, polynomial features like radius_mean^2, texture_mean^2, and radius_mean multiplied by texture_mean were generated to enhance non-linear representations. A log transformation was applied to area_mean, and a square root transformation to perimeter_mean, both of which significantly reduced skewness and yielded more symmetrical distributions. After applying MinMaxScaler for normalization, the features log_area_mean, sqrt_perimeter_mean, and either radius_texture_interaction or radius_mean multiplied by texture_mean were considered the most promising for model inclusion due to their stable distributions and statistically significant relevance to cancer diagnosis.

Moreover, the newly engineered features radius_worst_ratio and irregularity_score emerged as strong candidates for improving model accuracy. The radius_worst_ratio displayed a skewed distribution, highlighting that extreme values of radius tend to be less common, which could provide valuable information in distinguishing more severe malignancies. The irregularity_score, with its varied distribution, suggests that tumors with irregular shapes may have distinct patterns that could further refine tumor classification, offering a robust feature for enhancing model performance.

Model configuration

Four machine learning models were configured, including a baseline model and an optimized model. Here is the configuration for each model:

Baseline Convolutional Neural Network (B-CNN) model

The B-CNN model incorporates two one-dimensional convolutional layers with 32 and 64 filters, followed by max-pooling layers to reduce dimensionality. Model optimizer for the baseline model is

adaptive moment estimation ('adam'). A dropout layer with a rate of 0.5 is included after the first dense layer to mitigate overfitting. The feature maps are then flattened into a vector before passing to a dense layer with 128 units and ReLU activation. Unlike the convolutional layers, the final output layer uses a sigmoid activation function, which is suitable for binary classification. This model has been trained with 5 epochs, and the batch size is 32.

Table 2. B-CNN model architecture

Layer (type)	Output Shape	Parameters
conv1d (Conv1D)	(None, 37, 32)	128
max_pooling1d (MaxPooling1D)	(None, 18, 32)	0
conv1d_1 (Conv1D)	(None, 16, 64)	6,208
max_pooling1d_1 (MaxPooling1D)	(None, 8, 64)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 128)	65,664
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129
Total parameters :	77,131	
Trainable parameters :	72,129	
Non-trainable parameters :	0	
Optimizer parameters :	2	

Optimized Convolutional Neural Network (O-CNN) model

The O-CNN model was optimized through a randomized hyperparameter search. This method randomly samples and evaluates various combinations of hyperparameters to find the configuration that yields the best performance (based on accuracy). The search space included the number of convolutional layers (1-3), filter sizes ([32] to [32, 64, 128]), kernel size (3 or 5), dense layer units (64-256), dropout rate (0.3-0.7), learning rate (0.0001-0.01), optimizer ('adam' or 'rmsprop'), batch size (16-64), and epochs (10-30). The best hyperparameters found through this process were 'adam' as optimizer, 1 convolutional layer, a learning rate of 0.01, a kernel size of 5, filter sizes of [32, 64, 128], a dropout rate of 0.7, 128 dense units, 10 epochs, and a batch size of 64.

Table 3. O-CNN model architecture

Layer (type)	Output Shape	Parameters
conv1d_3 (Conv1D)	(None, 35, 32)	192
max_pooling1d_3 (MaxPooling1D)	(None, 17, 32)	0

Layer (type)	Output Shape	Parameters
flatten_2 (Flatten)	(None, 544)	0
dense_4 (Dense)	(None, 128)	69,760
dropout_2 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 1)	129
Total parameters :	70,083	
Trainable parameters :	70,081	
Non-trainable parameters :	0	
Optimizer parameters :	2	

Baseline Multilayer Perceptron (B-MLP) model

Both Keras and scikit-learn MLP models were used. The Keras MLP features a sequential architecture with two hidden dense layers (64 and 32 units, both with ReLU activation), dropout regularization (0.5 after each hidden layer), and a sigmoid output layer. The model was optimized using 'adam' and trained for 10 epochs with a batch size of 32. In contrast, the scikit-learn MLP incorporates L2 regularization and trains for a maximum of 30 iterations with a batch size of 32 and an initial learning rate of 0.01. Unlike the CNN, which leverages convolutional layers to automatically learn spatial hierarchies from input features, MLPs operate on flattened input data, processing it through fully connected layers. Keras model offers more direct control over the network architecture and training loop, while the scikit-learn implementation provides a higher-level, more streamlined interface with built-in regularization and iteration control.

Table 4. B-MLP (Keras) model architecture

Layer (type)	Output Shape	Parameters
dense_6 (Dense)	(None, 64)	2,560
dropout _3 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 32)	2,080
dropout_4 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 1)	33
Total parameters :	4,675	
Trainable parameters :	4,673	
Non-trainable parameters :	0	
Optimizer parameters :	2	

Optimized Multilayer Perceptron (O-MLP) model

Similar to the O-CNN model, the O-MLP model was optimized by hyperparameter tuning for the Keras B-MLP model. The optimization was conducted only for the Keras B-MLP model since Keras offers direct control over the architecture, while scikit-learn has several built-in architectures. The search space included the number of hidden layers ([64], [128,64], [256,128,64]), dropout rate (0.3-0.7), learning rate (0.0001-0.01), optimizer ('adam' or 'rmsprop'), batch size (16-64), and epochs (10-30). The best hyperparameters for the O-MLP model were 'adam' as optimizer, a learning rate of 0.01, hidden layers of [128, 64], a dropout rate of 0.3, 10 epochs, and a batch size of 16.

Table 5. O-MLP model architecture

Layer (type)	Output Shape	Parameters
dense_12 (Dense)	(None, 128)	5,120
dropout _7 (Dropout)	(None, 128)	0
dense_13 (Dense)	(None, 64)	8,256
dropout _8 (Dropout)	(None, 64)	0
dense_14 (Dense)	(None, 1)	65
Total parameters :	13,443	
Trainable parameters :	13,441	
Non-trainable parameters :	0	
Optimizer parameters :	2	

Model evaluation

To ensure the reliability of these estimates and mitigate the risk of overfitting, the reported performance metrics represent the mean results aggregated across the five stratified cross-validation folds. In each iteration, one fold served as the independent validation set while the remaining four folds were used for model training. This approach ensures that every sample in the WDBC dataset is utilized for both training and evaluation, providing a comprehensive assessment of the models' generalization capability. **Table 6** presents a summary of these aggregated evaluation results across all models.

Table 6. Model metric evaluation

Metric	CNN				MLP (Keras)			
	Train Baseline	Test Baseline	Train Optimized	Test Optimized	Train Baseline	Test Baseline	Train Optimized	Test Optimized
Accuracy	0.9472	0.9398	0.9872	0.9886	0.9179	0.9123	0.9648	0.9649
Precision	0.9234	0.9762	0.9758	0.9742	0.9304	0.9211	0.9323	0.9318

Metric	CNN				MLP (Keras)			
	Train Baseline	Test Baseline	Train Optimized	Test Optimized	Train Baseline	Test Baseline	Train Optimized	Test Optimized
Recall	0.9528	0.9762	0.9883	0.9909	0.8425	0.8333	0.8764	0.8762
F1-Score	0.9250	0.9047	0.9642	0.9791	0.8843	0.8750	0.9538	0.9535
AUC	0.9876	0.9980	0.9946	0.9990	0.9610	0.9696	0.9971	0.9983

The experimental results indicate that both baseline and optimized CNN models exhibit high performance across training and testing sets, suggesting a robust and well-fitted architecture for breast cancer classification. As prioritized in this diagnostic analysis to minimize false negatives, the CNN models demonstrated superior and more consistent test recall (baseline: 0.976190; optimized: 0.990912) compared to the optimized MLP (0.876190). While the MLP models achieved high AUC values, the noticeable recall gap and fluctuations in test performance suggest a higher risk of missed malignancies compared to the tuned CNN. Consequently, an optimized CNN provides the most reliable balance between sensitivity and precision, which is critical for clinical safety.

The practical integration of this optimized CNN into clinical workflows offers a significant opportunity for "Second Read" diagnostic support. By embedding this model into existing Radiology Information Systems (RIS), it can act as an automated triaging tool that highlights high-probability malignant cases for immediate prioritized review by radiologists. This real-world application addresses the 15–20% miss rate associated with traditional mammography, especially in dense breast tissue, by providing a data-driven "safety net" to reduce diagnostic variability. Furthermore, the computational efficiency of the optimized CNN allows for near-instantaneous processing of Fine Needle Aspirate (FNA) digitized features, potentially accelerating the diagnostic timeline from biopsy to treatment planning. Integrating such AI-driven insights not only enhances the accuracy of early-stage detection but also provides radiologists with a consistent, evidence-based baseline to support complex decision-making in clinical oncology.

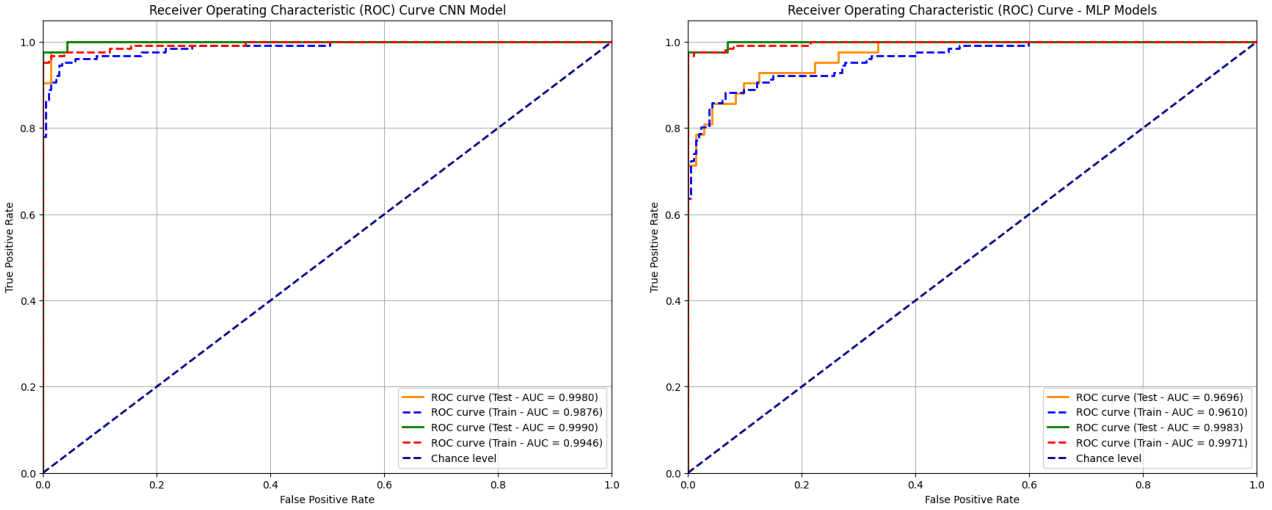

Figure 10. ROC curve for CNN model and MLP model

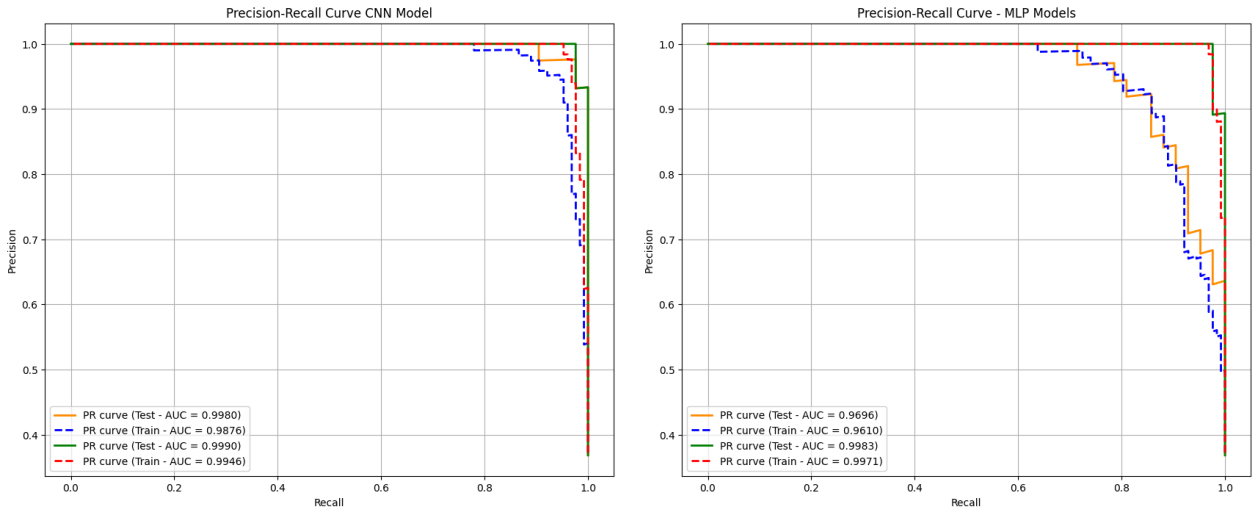

Figure 11. Precision-recall curve for the CNN and MLP model

The models' diagnostic capabilities are further substantiated by the ROC curves in **Figure 10** and PR curves in **Figure 11**. The CNN models exhibit exceptional and consistent discriminatory power, with both the Baseline and Optimized architectures achieving near-perfect test AUC values of 0.9980 and 0.9990, respectively. The curves rise sharply toward the top-left corner, indicating a high True Positive Rate with minimal False Positives across all thresholds. In contrast, the MLP models display greater performance variance between configurations. While the Optimized MLP achieves a competitive test AUC of 0.9983, the Baseline MLP lags noticeably with an AUC of 0.9696, highlighting the CNN's superior inherent robustness even prior to hyperparameter tuning. Similarly, the Precision-Recall analysis in **Figure 11** confirms the CNN's dominance: the Optimized CNN maintains perfect precision until very high recall levels, yielding an Average Precision (AP) of 0.9990. Conversely, the Baseline MLP shows a premature decline in precision as recall increases, resulting in a lower AP of 0.9696. This visual evidence reinforces that the CNN architecture offers a more stable and reliable diagnostic tool, minimizing false negatives without compromising precision.

Model simulation

For further evaluation of the Optimized CNN Model (O-CNN), which represents the most suitable model in this analysis, a simulation was conducted by testing it on ten random data points.

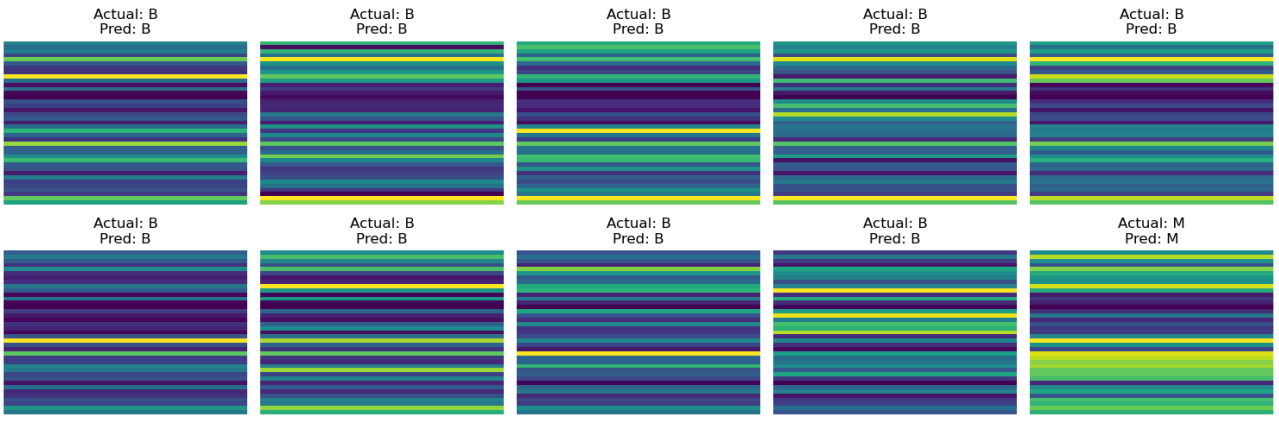

Figure 12. Feature map visualizations and classification outcomes (Actual vs. Predicted) of 10 random samples processed by the Optimized CNN (O-CNN) model. *Abbreviations: B, Benign; M, Malignant*

Figure 12 illustrates the simulation outcomes, where the model achieved 100% accuracy on this random subset, correctly classifying all nine benign cases and the single malignant case. Each panel in the figure represents the feature representation of an input sample processed by the CNN model. The color gradient in the visualization ranges from dark blue to yellow, where darker colors indicate lower activation values and brighter colors represent higher activation values within the feature map generated by the network. The abbreviations used in the figure denote the classification labels. While this simulation on a small sample provides encouraging evidence of the model's ability to generalize, indicating robustness and reliability, further validation on a larger dataset is necessary for a comprehensive assessment.

DISCUSSION

This study evaluated the comparative performance of CNN and MLP models for breast cancer diagnosis using structured features from the WDBC dataset. The findings indicate that CNN models consistently outperform MLP architectures across most evaluation metrics, particularly recall, which is critical in minimizing false negatives in clinical screening contexts (Alanazi et al., 2020; Sung et al., 2021).

The strong performance of CNN may be attributed to its ability to learn hierarchical interactions among numerical features, even when applied outside traditional imaging contexts (Yu et al., 2023). Unlike MLPs, which rely heavily on manually engineered representations, CNN architectures can capture localized feature dependencies through convolutional operations (Desai & Shah, 2021). However, several limitations must be acknowledged. First, the WDBC dataset is relatively small and structured, which may contribute to the high performance observed. Second, the dataset contains pre-extracted features rather than raw imaging data, meaning the CNN does not operate as a full imaging diagnostic system (Araujo et al., 2017; Kooi et al., 2017). Third, although stratified cross-validation was implemented to reduce overfitting risk, external validation using independent clinical datasets is required before real-world deployment (McKinney et al., 2020; Rodríguez-Ruiz et al., 2019).

From a methodology side, this study highlights the importance of preprocessing within cross-validation loops to prevent data leakage and ensure realistic performance estimates. Additionally, the results emphasize that hyperparameter tuning significantly improves model recognition ability while maintaining generalization. Future work should explore validation using mammography or histopathology image datasets, integration with explainable AI approaches, and testing in clinical decision-support environments.

CONCLUSION

This study explored the optimization of Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP) for breast cancer diagnosis using the Wisconsin Diagnostic Breast Cancer dataset. By implementing a systematic randomized hyperparameter search to refine critical parameters, such as kernel size, learning rate, and dropout rates, the study demonstrated that a tuned 1D-CNN architecture could effectively capture latent dependencies within digitized morphological features. Validated through a rigorous 5-fold stratified cross-validation framework, the Optimized CNN achieved a superior mean test accuracy of 98.86% and a critical recall rate of 99.09%. These results substantially outperform both the baseline configurations and the MLP models, substantiating that precise architectural tuning is a decisive factor in minimizing false negatives for clinical oncology. Future research should extend this work by validating the optimized architecture on large-scale raw mammography datasets (e.g., DDSM) and by investigating advanced deep learning models such as ResNet or EfficientNet to further enhance diagnostic generalizability across diverse patient populations.

AUTHOR CONTRIBUTIONS

PN: Conceptualization, Methodology, Writing - Original Draft. **DN**: Conceptualization, Software, Data Curation and Processing. **TS**: Software, Data Curation and Processing. **TA**: Introduction, Theory, Visualization, Writing - Original Draft. **RT**: Methodology, Knowledge Synthesis. **YJK**: Visualization, Formal Analysis. **MZS**: Supervision, Writing - Review & Editing.

ACKNOWLEDGEMENTS

None.

COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

None.

ADDITIONAL INFORMATION

None.

REFERENCES

Alanazi, S. A., Kamruzzaman, M. M., Sarker, M. N. I., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2020). Boosting breast cancer detection using convolutional neural network. *Journal of Healthcare Engineering*, 1-11. https://doi.org/10.1155/2020/1245179

American Cancer Society. (2024). Breast cancer facts & figures 2024-2025. American Cancer Society. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/2024/breast-cancer-facts-and-figures-2024.pdf

Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., & Polonia, A. (2017). Classification of breast cancer histology images using convolutional neural networks. *PLOS ONE, 12*(6). https://doi.org/10.1371/journal.pone.0177544

Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network and Convolutional neural network. *Clinical eHealth, 4*, 1-11. https://doi.org/10.1016/j.ceh.2020.11.002

Kooi, T., Litjens, G., Ginneken, B. V., Merida, A. G., Sanchez, C. I., Mann, R., Heeten, A. d., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis, 35*, 303-312. https://doi.org/10.1016/j.media.2016.07.007

Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine, 175*(11), 1828-1837. https://doi.org/10.1001/jamainternmed.2015.5231

Lin, G., Chen, M., Tan, M., Chen, L., & Chen, J. (2023). A dual-stage transformer and MLP-based network for breast ultrasound image segmentation. *Biocybernetics and Biomedical Engineering, 43*(4), 656-671. https://doi.org/10.1016/j.bbe.2023.09.001

Mannarsamy, V., Mahalingam, P., Kalivarathan, T., Amutha, K., Paulraj, R. K., & Ramasamy, S. (2025). Sift-BCD: SIFT-CNN integrated machine learning-based breast cancer detection. *Biomedical Signal Processing and Control, 106*. https://doi.org/10.1016/j.bspc.2025.107686

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., & Back, T. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*, 89-94. https://doi.org/10.1038/s41586-019-1799-6

Reeves, R. A., & Kaufman, T. (2023). *Mammography*. PubMed; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK559310/

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I., & Mann, R. M. (2019). Detection of breast cancer with mammography: Effect of an artificial intelligence support system. *Radiology*, *290*(2), 305–314. https://doi.org/10.1148/radiol.2018181371

Sahu, A., Das, P. K., & Meher, S. (2023). High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets. *Biomedical Signal Processing and Control, 80*(1). https://doi.org/10.1016/j.bspc.2022.104292

Silas, K. L. T., Alain, B. D.-T., Thierry, N., Pierre, L. T.J., & Ejuh, G. W. (2025). Breast cancer detection and classification: A study on the specification and implementation of multilayer perceptron analog artificial neural networks. *Computers in Biology and Medicine, 190*. https://doi.org/10.1016/j.compbiomed.2025.110060

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., DMV, A. J., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A Cancer Journal for Clinicians, 71*(3), 209-249. https://doi.org/10.3322/caac.21660

Yu, D., Lin, J., Cao, T., Chen, Y., Li, M., & Zhang, X. (2023). SECS: An effective CNN joint construction strategy for breast cancer histopathological image classification. *Computer and Information Sciences, 35*(2), 810-820. https://doi.org/10.1016/j.jksuci.2023.01.017