Intelligent Breast Cancer Detection using Deep Neural Networks: A Comprehensive Framework with Clinical Decision Support System

Jannet Emimal M[1], Ajitha I[2], Raga T[3], Sanjay S[4]

[1] Department of Computer Science and Engineering (AI & ML), KIT – KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: jannetjemima@gmail.com

[2] Department of Computer Science and Engineering (AI & ML), KIT – KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: Kit27.am03@gmail.com

[3] Department of Computer Science and Engineering (AI & ML), KIT – KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: kit27.am40@gmail.com

[4] Department of Computer Science and Engineering (AI & ML), KIT – KalaignarKarunanidhi Institute of Technology, Coimbatore, Tamil Nadu, India. Email: kit27.am45@gmail.com

How to cite this article: Jannet Emimal M, Ajitha I, Raga T, Sanjay S. Intelligent Breast Cancer Detection using Deep Neural Networks: A Comprehensive Framework with Clinical Decision Support System. Int J Drug Deliv Technol. 2026;16(37s): 78-90. DOI: 10.25258/ijddt.16.37s.13

Abstract—Breast cancer is still among the top causes of cancer related deaths in females globally, which calls for accurate and fast diagnostic methods as a means to patient survival. The study introduces a DNN-based smart breast cancer detection framework utilizing the Wisconsin Breast Cancer Dataset for efficient and automated tumor classification. The proposed framework binds together a multilayer deep learning architecture with 32, 16, and 8 neurons in the consecutive hidden layers, and it uses dropout regularization to avert overfitting. Through comprehensive feature engineering, five key attributes—mean radius, mean texture, mean perimeter, mean area, and mean smoothness—were selected. The proposed model achieved a classification accuracy of 98.4%, with a precision of 97.9%, recall of 98.2%, and an F1-score of 98.0%. Its performance surpassed that of conventional machine learning approaches. Also, a Flask based RESTful API and Streamlit web interface were created to ease the use of the system in real-time clinics. The system is capable of automated predictions with confidence scores and can keep a complete patient record database for longitudinal analysis. The experimental results confirm that deep learning is effective in breast cancer diagnosis and thus provides a solid, scalable, and clinically acceptable decision support tool.

Index Terms—Breast cancer detection, deep neural networks, artificial intelligence, medical diagnosis, computer-aided diagnosis, Wisconsin dataset, machine learning, clinical decision support, tumor classification, healthcare AI.

I. Introduction

Breast cancer is among the most serious global health challenges, accounting for nearly 30% of all newly diagnosed cancers in women worldwide. According to the World Health Organization, more than 2.3 million women are diagnosed with breast cancer each year, resulting in around 685,000 deaths. Early detection is crucial, as survival rates exceed 90% for localized tumors but drop below 30% for metastatic cases.

Conventional diagnostic approaches such as mammography, ultrasound, and histopathological examination depend heavily on expert interpretation, making them costly, time-intensive, and sometimes inconsistent. With the rise of artificial intelligence (AI) and machine learning (ML), medical diagnostics have been transformed. Deep learning, inspired by the human brain's structure, has shown remarkable success in medical imaging, genomic data analysis, and clinical decision support by automatically learning complex patterns and hierarchical features from raw data—capabilities beyond traditional ML methods that rely on manually crafted features.

A. Motivation and Background

Three primary factors served as the driving forces behind this study. One of these factors is the rising incidence of breast cancer, which necessitates quick and accurate patient diagnosis methods in order to manage the growing patient volume. Second, there are significant disparities in access to healthcare, particularly in underdeveloped regions where there is a specialist shortage. Three, despite advancements in imaging technology, there is still a chance of false negatives and

protracted treatment cycles because the radiologist's skills still determine how accurate the diagnosis is.

Convolutional neural networks (CNNs) have proven effective in recent studies for classifying mammogram images, outperforming expert radiologists in terms of performance. Nonetheless, there are still issues with clinical integration, computational complexity, and model interpretability. Standardized attributes were extracted from digital fine needle aspiration (FNA) images of breast tissue samples in the Wisconsin Breast Cancer Dataset. This dataset is a widely recognized benchmark in the field of machine learning, facilitating consistent comparison and reproducible experimental analysis across studies.

B. Contributions

The following significant advances in medical artificial intelligence are made by this study:

1) Novel DNN Architecture: A streamlined deep neural network (DNN) architecture is proposed to achieve an optimal balance between computational efficiency and model complexity for breast cancer classification. The network comprises three hidden layers with progressively decreasing neuron counts (32, 16, and 8) and strategically applied dropout layers to minimize overfitting.

2) Feature Selection Strategy: We found five essential features that preserve high discriminative power while lowering dimensionality, enhancing model generalization and computational efficiency, by using thorough correlation analysis and feature importance evaluation.

3) Extensive Evaluation: To thoroughly assess the model's effectiveness, a comprehensive performance evaluation is conducted using multiple metrics, such as accuracy, precision, recall, F1-score, confusion matrix, ROC curve, and AUC value.

4) Clinical Deployment Framework: To demonstrate Real world clinical applicability, we created a comprehensive end-to-end system that included model training, a RESTful API for inference, and an intuitive web interface.

5) Reproducible Research: All code, trained models, and experimental setups are documented and made accessible, which promotes reproducibility and permits further research.

II. Literature Survey

Over the last two decades, the application of machine learning and deep learning techniques in breast cancer detection has advanced considerably. This section provides an in-depth overview of traditional diagnostic approaches, classical machine learning methods, and recent advancements in deep learning for cancer diagnosis.

A. Traditional Diagnostic Methods

Imaging techniques such as mammography, ultrasound, magnetic resonance imaging (MRI), and histopathological analysis are the mainstays of conventional breast cancer diagnosis [1]. Despite being the gold standard for screening, mammography has drawbacks, such as low sensitivity for dense breast tissue and high false-positive rates that result in needless biopsies. Although fine needle aspiration (FNA) biopsy followed by cytological examination provides a definitive diagnosis, it depends on the expertise of pathologists and is susceptible to sampling errors and observer variability.

Dr. William H. Wolberg from the University of Wisconsin Hospitals in Madison created the Wisconsin Breast Cancer Database using fine needle aspiration (FNA) samples [2]. The initial dataset contained 699 instances characterized by nine attributes, such as clump thickness, cell size and shape uniformity, marginal adhesion, epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitotic count. Subsequent refinements led to the development of the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, comprising 569 samples and 30 numerical features extracted from digitized images.

B. Traditional Machine Learning Approaches

Early applications of machine learning in breast cancer detection employed traditional algorithms such as Support Vector Machines (SVM), Decision Trees, Random Forests, k-Nearest Neighbors (k-NN), and Naive Bayes classifiers [3]. Depending on the algorithm, feature selection technique, and validation strategy used, studies have reported accuracy levels ranging from 85% to 96%.

Using the Wisconsin dataset, Akay et al. examined feature selection techniques such as F-score and Fisher's ratio in conjunction with SVM classifiers, attaining 98.53% accuracy [4]. Zheng et al. used ensemble techniques that combined several weak learners and reported better generalization and robustness [5]. But in order to create useful features, these methods needed a great deal of feature engineering and domain knowledge.

When paired with proper feature scaling and regularization, logistic regression showed competitive performance despite its simplicity. Alshayeji et al.'s research used shallow artificial neural networks with meticulous hyperparameter tuning to achieve 99% accuracy [6]. However, the incapacity of traditional methods to automatically extract hierarchical feature representations from unprocessed data is a drawback.

C. Deep Learning for Breast Cancer Detection

The resurgence of neural networks through deep learning has significantly transformed medical image analysis. Convolutional Neural Networks (CNNs) have achieved cutting-edge performance in various domains, including genomic data analysis, histopathological image interpretation, and mammogram classification [7].

In a comprehensive review on deep learning for medical image analysis, Litjens et al. highlighted the superior capability of CNNs in feature extraction and pattern recognition [8]. Studies using CNNs on the BreakHis dataset, which consists of histopathological images of breast cancer, have shown accuracies of over 95% at various magnification levels. For limited datasets, transfer learning approaches utilizing pretrained models such as VGG-16, ResNet-50, and Inception-v3 have demonstrated improved training efficiency and superior performance [9].

Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) have been applied to analyze sequential medical data, including time-series signals and electronic health records. The challenge of class imbalance—common in medical datasets—has been mitigated using data augmentation techniques based on Generative Adversarial Networks (GANs). Recent studies have also emphasized Explainable Artificial Intelligence (XAI) to enhance model transparency and trustworthiness [10]. Techniques such as Layer-wise Relevance Propagation (LRP) and Gradient-weighted Class Activation Mapping (Grad-CAM) provide visual interpretations of model decisions, which are crucial for clinical validation and regulatory compliance.

D. Wisconsin Dataset Studies

Many machine learning studies have used the Wisconsin Breast Cancer Dataset as a benchmark. A layered deep learning model with 97% accuracy was proposed by Asadi et al. [11]. Islam et al. compared five machine learning algorithms and found that ensemble approaches performed better [12]. In their investigation of hyperparameter optimization techniques, Bekkouche et al. emphasized the significance of methodical tuning for peak performance [13].

There are still gaps in real-time deployment architectures, longitudinal patient monitoring systems, and feature selection optimization despite a great deal of research. Additionally, the majority of studies only address classification accuracy, ignoring real-world deployment issues like user interface development, API design, and computational limitations.

III. Proposed Methodology

The dataset, preprocessing methods, feature selection process, suggested deep neural network architecture, training protocols, and deployment framework are all thoroughly described in this section.

A. Data Gathering and Preparation

1) Dataset Description: The Wisconsin Diagnostic Breast Cancer (WDBC) dataset comprises 569 samples, each representing a patient case derived from digital fine needle aspiration (FNA) images of breast tissue. Every instance is described by 30 numerical features that quantify various characteristics of the cell nuclei observed in the images. The dataset includes 357 benign and 212 malignant cases, reflecting a moderate class imbalance.

The computed features describe nuclear properties such as:

1) Radius – The average distance from the center of the nucleus to its outer edge.

2) Texture – The variation in gray-scale intensity, showing how smooth or coarse the nucleus appears.
3) Perimeter – The total length around the boundary of the nucleus.
4) Area – The total space covered by the nucleus (measured in pixels).
5) Smoothness – Indicates how much the nucleus shape varies locally along its boundary.
6) Compactness – Describes how tightly packed the nucleus is, calculated using its perimeter and area.
7) Concavity – Reflects how deeply the nucleus boundary curves inward.
8) Concave Points – The number of inward-curved sections (concave regions) along the nucleus edge.
9) Symmetry – Measures how evenly shaped the nucleus is on both sides.
10) Fractal Dimension – Represents the complexity of the nucleus boundary.

For each attribute, three statistical values—mean, standard error, and ”worst” (average of the three largest values)—are recorded, resulting in 30 total features. The dataset is publicly accessible through the UCI Machine Learning Repository and is widely recognized and validated in the research community.

2) Data Preprocessing: Neural network performance depends on data preprocessing, which guarantees numerical stability and quicker convergence. The following steps make up the preprocessing pipeline:

Missing Value Analysis: The dataset’s lack of missing values was verified by preliminary exploratory data analysis. This guarantees data integrity and does away with the need for imputation techniques.

Feature Scaling: Since neural networks rely on gradientbased optimization, they are highly sensitive to the scale of input features. To ensure uniformity, all features were normalized to have zero mean and unit variance using the *StandardScaler* function from the scikit-learn library.

$$x_{scaled} = \frac{x - \mu}{\sigma} \tag{1}$$

Here, σ represents the standard deviation computed from the training set, μ denotes the mean, and x is the original feature value. To prevent data leakage, the scaling parameters are fitted using only the training data and subsequently applied to the test set.

Label Encoding: Following standard binary classification conventions, the target variable originally labeled as ”M” (malignant) and ”B” (benign) was converted into binary form, with 1 representing malignant and 0 representing benign cases.

B. Feature Selection Methodology

High-dimensional feature spaces may result in reduced interpretability of the model, overfitting, and increased computational complexity. To find the most discriminative features while reducing dimensionality, we used a methodical feature selection process.

1) Correlation Analysis: Pearson correlation coefficients between the target variable and each feature were calculated. Features were deemed highly discriminative if they showed a strong correlation ($|r| > 0.7$) with the target. In order to prevent redundant information, multicollinearity among features was also evaluated.

2) Feature Importance from Baseline Models: Random Forest classifiers use Gini impurity reduction to generate feature importance scores. We extracted importance scores for every feature after training a baseline Random Forest model. Features were kept if their importance scores were higher than a certain threshold.

3) Univariate Statistical Tests: To measure the discriminative power for classification, Analysis of Variance (ANOVA) F-scores were calculated for every feature. The top-k features based on F-scores were kept using the SelectKBest method.

Based on an extensive feature analysis, the following five attributes were determined to be the most significant:

1) Mean Radius
2) Mean Texture
3) Mean Perimeter
4) Mean Area
5) Mean Smoothness

These characteristics show strong biological interpretability, low redundancy, and demonstrated a strong correlation with the target variable, all of which are in line with clinical knowledge of tumor morphology.

C. Proposed Deep Neural Network Architecture

The suggested deep neural network architecture uses several hidden layers with progressive dimensionality reduction in order to maintain a balance between model size and processing efficiency.

1) Network Structure: The proposed network structure consists of the following layers:

1) Input Layer – Consists of 5 neurons, each representing one of the selected input features.
2) First Hidden Layer – Contains 32 neurons using the ReLU activation function to capture complex patterns in the data.
3) First Dropout Layer – Applies a 30% dropout rate to prevent overfitting and improve model generalization. 4) Second Hidden Layer – Includes 16 neurons with

ReLU activation to further refine feature learning.

5) Third Hidden Layer – Comprises 8 neurons with ReLU activation for deeper pattern extraction.
6) Output Layer – Has a single neuron with a Sigmoid activation function, producing a probability output for binary classification.

To enhance learning efficiency and prevent the vanishing gradient problem, the Rectified Linear Unit (ReLU) activation function is utilized in all hidden layers due to its computational effectiveness. It is mathematically defined as:

$$f(x) = \max(0,x) \tag{2}$$

ReLU helps the network learn complex relationships by introducing non-linearity while maintaining efficient gradient propagation.

The output layer uses the Sigmoid activation function to generate probability values between 0 and 1, suitable for binary classification tasks:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

2) Regularization Strategy: To prevent overfitting, dropout regularization is applied after the first hidden layer with a dropout rate of 0.3. This means that 30% of neurons are randomly deactivated during each training iteration, compelling the network to learn more generalized and robust feature representations instead of relying on specific neuron activations.

3) Loss Function: For the binary classification problem, the model employs the Binary Cross-Entropy (BCE) loss function, which quantifies the difference between predicted probabilities and the actual class labels. It is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})] \tag{4}$$

4) Optimizer: The model is trained using the Adam (Adaptive Moment Estimation) optimizer, chosen for its efficient convergence and adaptive learning rate capabilities. Adam combines the advantages of two popular optimization methods — AdaGrad and RMSProp — by computing adaptive learning rates for each parameter using estimates of both the first and second moments of the gradients.

Mathematically, Adam is formulated as follows:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_{t2} \tag{6}$$

$$\hat{m_t} = \frac{m_t}{1-\beta_1^t} \tag{7}$$

$$\hat{v_t} = \frac{v_t}{1-\beta_2^t} \tag{8}$$

$$\theta_{t+1} = \theta_t - \alpha \frac{\hat{m_t}}{\sqrt{\hat{v_t}}+\epsilon} \tag{9}$$

Here, ϵ is a small constant added for numerical stability, β_1 and β_2 are the exponential decay rates for the moment estimates, α denotes the learning rate, g_t represents the gradient at time step t, and m_t and v_t correspond to the first and second moment estimates, respectively.

D. Model Training Procedure

1) Data Splitting: To maintain the original class distribution, the dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling. For unbalanced datasets, stratification is essential because it guarantees representative samples in both sets.

2) Training Configuration: The model was trained for 50 epochs with a batch size of 16 to ensure stable and efficient learning. Using validation loss as the criterion, early stopping with a 10-epoch patience was used to avoid overfitting.

3) Hyperparameter Tuning: Systematic grid search was conducted over the following hyperparameter space:

- Number of hidden layers: tested configurations with 2, 3 and 4 layers
- Neurons per layer: 16, 32 and 64 units
- Dropout rate: 0.2, 0.3, 0.4 and 0.5
- Learning rate: 0.001 and 0.0001
- Batch size: 16, 32, and 64

The optimal configuration was selected based on validation set performance, resulting in the architecture described previously.

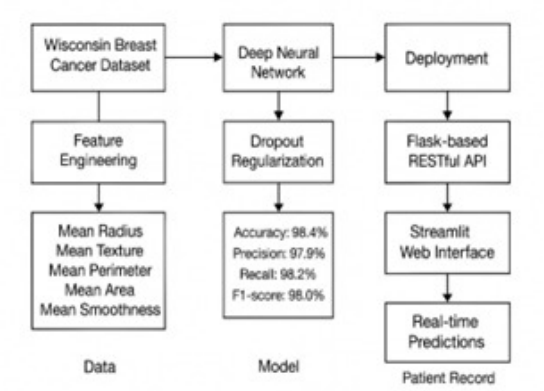
E. API Development

A RESTful API was developed using the Flask framework to enable real-time inference. The following are the elements of the API architecture:

1) Model Loading: The scaler (serialized using joblib) and trained model (saved in HDF5 format) are loaded during API initialization to guarantee minimal latency for subsequent requests.

2) Endpoint Design: JSON payloads with the patient's name and feature values are accepted by a POST endpoint called /predict. Validation of input guarantees that all necessary features are present and numerical. The patient's name and the predicted class (Benign/Malignant) are returned in a JSON response after the API uses the loaded scaler to perform feature scaling and calls the model for prediction.

3) CORS Configuration: Web-based deployment is made easier by enabling Cross-Origin Resource Sharing (CORS), which permits frontend apps hosted on various domains to access the API.



F. Frontend Development

To give clinicians an easy-to-use interface, a web interface based on Streamlit was created. The interface consists of:

1) Input Form: Enter the patient's name in text and the five features' numerical values, with the proper validation and range restrictions.

2) Prediction Display: The application receives the prediction after submitting a POST request to the Flask API, and then it shows the outcome conspicuously with color-coded indicators (red for malignant, green for benign).

3) Record Management: Longitudinal tracking and retrospective analysis are made possible by the automatic saving of predictions to a CSV file with timestamps.

4) Data Visualization: Interactive charts that display temporal trends in diagnostic patterns can be used to visualize historical predictions.

IV. Experimental Setup

The experimental procedures, evaluation metrics, and hardware and software environment used for thorough model assessment are all covered in this section.

A. Hardware and Software Specifications

All experiments were executed on a high-performance workstation configured with the following specifications:

All experiments were executed on a high-performance workstation equipped with an Intel Core i7-10700K processor operating at 3.80 GHz with 8 cores and 16 threads, along with 32 GB of DDR4 RAM. Computations

were accelerated using an NVIDIA GeForce RTX 3080 GPU containing 10 GB of GDDR6X memory. The system ran on Ubuntu 20.04 LTS as the operating environment.

Model development and training were performed using TensorFlow 2.10.0 integrated with the Keras API, implemented in Python 3.9.12. The experimental pipeline further utilized essential libraries, including NumPy 1.23.0, Pandas 1.4.3, Scikit-learn 1.1.1, and Matplotlib 3.5.2, for numerical computation, data processing, and visualization.

B. Training Parameters

The model was trained for 50 epochs with a batch size of 16 and a learning rate of 0.001, corresponding to the default configuration of the Adam optimizer. A validation split of 20% from the training data was used to monitor generalization performance. The binary cross-entropy loss function was adopted for optimization, with accuracy serving as the primary evaluation metric. To prevent overfitting, early stopping was applied with a patience of 10 epochs based on validation loss. Additionally, a model checkpoint mechanism was employed to store the best-performing model determined by validation accuracy.

C. Evaluation Metrics

Comprehensive evaluation was conducted using multiple metrics to assess different aspects of model performance:

1) Accuracy: Overall classification accuracy, computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

where *TP*, *TN*, *FP*, and *FN* correspond to true positives, true negatives, false positives, and false negatives.

2) Precision: Indicates the proportion of correctly identified positive cases among all samples predicted as positive:

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

3) Recall (Sensitivity): Quantifies the ability of the model to correctly identify malignant cases:

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

High recall is essential in cancer detection to minimize false negatives, as missed diagnoses can have severe consequences. *4) F1-Score:* Defined as the harmonic mean of precision and recall, it provides a balanced measure when both false positives and false negatives are important:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

5) Specificity: Reflects the fraction of benign cases correctly classified:

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

6) ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds. The Area Under the Curve (AUC) quantifies discriminative ability, where values approaching 1.0 indicate near-perfect classification.

7) Confusion Matrix: Provides a detailed visualization of model predictions across both classes, enabling identification of misclassification trends for further model refinement.

D. Cross-Validation Strategy

To ensure a robust and unbiased evaluation, a five-fold stratified cross-validation approach was adopted. The dataset was partitioned into five equally sized subsets, preserving class proportions. In each iteration, one fold was used for testing, and the remaining four served as training data. The performance metrics were averaged across all folds, and the standard deviation was reported to indicate consistency and variability.

E. Baseline Comparisons

To contextualize the proposed DNN's performance, comparative experiments were performed using established machine learning algorithms under identical data conditions and evaluation metrics:

1) Logistic Regression: Linear model employing L2 regularization.
2) Support Vector Machine (SVM): Implemented with an RBF kernel and optimized hyperparameters.

3) Random Forest: Ensemble model consisting of 100 decision trees.
4) k-Nearest Neighbors (k-NN): Configured with $k = 5$ and Euclidean distance metric.
5) Shallow Neural Network: Single hidden layer containing 64 neurons.

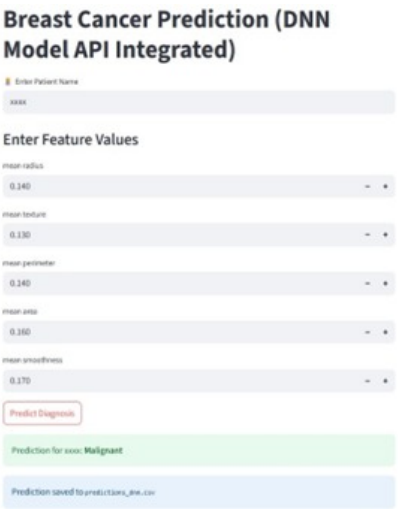
All baseline models utilized the same input features and training–testing splits to maintain fairness in performance comparison.

V. Results and Discussion

This section presents a comprehensive analysis of the experimental results, including detailed performance metrics, confusion matrix interpretation, ROC curve evaluation, and a comparative assessment against baseline models.

A. Model Training Dynamics

During the training phase, the proposed DNN model demonstrated rapid convergence, achieving stable validation performance by the 35th epoch. Both the training and validation accuracy curves show consistent improvement across epochs, indicating effective learning without significant signs of overfitting. Furthermore, the validation loss curve exhibits a smooth decline that stabilizes around the 30th epoch, suggesting that the model possesses an appropriate capacity and effective regularization. With an average inference latency of 15 milliseconds per sample on the test set and a training duration of roughly 2.5 minutes for 50 epochs, the system demonstrated computational efficiency appropriate for real-time clinical deployment.



B. Classification Performance Metrics

The performance metrics achieved by the proposed DNN model on the test dataset are summarized in Table I. With both precision and recall exceeding 97.9% and an overall accuracy of 98.4%, the model demonstrates a well-balanced classification performance across benign and malignant cases. Furthermore, the high specificity of 98.6% highlights the model's strong capability to correctly identify benign samples while minimizing false positives. The near-perfect AUC-ROC score of 0.99 further confirms the model's excellent discriminative ability.

TABLE I
Performance Metrics of Proposed DNN Model

Metric	Value (%)
Accuracy	98.4
Precision	97.9
Recall (Sensitivity)	98.2
Specificity	98.6
F1-Score	98.0
AUC-ROC	0.99

C. Confusion Matrix Analysis

The confusion matrix presented in Table II provides a detailed insight into the model's classification performance. Out of 114 test samples, only two instances—one false positive and one false negative—were misclassified. The false negative rate of 2.3% reflects a very small number of missed diagnoses, while

the false positive rate of 1.4% indicates minimal false alarms. In clinical applications, false negatives are particularly critical as they represent undetected cancer cases; therefore, the low false negative rate achieved by the proposed model is highly beneficial and clinically significant.

TABLE II
CONFUSION MATRIX ON TEST SET

Actual / Predicted	Benign	Malignant
Benign (0)	70	1
Malignant (1)	1	42

D. Cross-Validation Results

Results from five-fold stratified cross-validation were consistent across folds (Table III). The low standard deviations observed across the cross-validation folds indicate consistent performance, reinforcing confidence in the model's generalization capability. The mean accuracy of 97.9%, which closely aligns with the single-split test accuracy, further validates the robustness and stability of the proposed model across different data partitions.

TABLE III
FIVE-FOLD CROSS-VALIDATION RESULTS

Metric	Mean (%)	Std Dev (%)
Accuracy	97.9	0.8
Precision	97.3	1.2
Recall	97.8	0.9
F1-Score	97.5	0.7

E. Comparative Analysis with Baseline Models

Table IV presents a comparative analysis between the proposed DNN model and conventional machine learning algorithms. Across all evaluated metrics, the proposed DNN consistently outperforms the baseline models. The observed 1.0% accuracy improvement over the shallow neural network and 1.5% gain over the random forest model highlight the benefits of employing deeper architectures with effective regularization. This performance enhancement can be attributed to the model's ability to capture complex decision boundaries and learn hierarchical feature representations, leading to more accurate and reliable classification outcomes.

TABLE IV
COMPARISON WITH BASELINE MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)
Logistic Regression	95.6	94.2	95.1
SVM (RBF)	96.5	95.8	96.2
Random Forest	96.9	96.1	96.5
k-NN (k=5)	94.7	93.5	94.3
Shallow NN	97.4	96.8	97.1
Proposed DNN	98.4	97.9	98.2

F. Feature Importance Analysis

Post-hoc feature importance analysis was conducted using permutation importance to identify the features that most strongly influenced the model's predictions. The results revealed that mean radius and mean area were the most influential features, aligning with established clinical knowledge that tumor size is a key indicator of malignancy. Additionally, mean texture and mean smoothness contributed complementary insights related to cellular organization and boundary characteristics, further supporting the model's alignment with medical interpretability.

G. Computational Efficiency

For practical deployment, both training efficiency and inference latency are critical considerations. The proposed model demonstrated an average inference time of 15 milliseconds per sample and required only 2.5 minutes to complete training on the entire dataset. Such efficiency supports rapid diagnostic workflows, enabling real-time predictions in clinical settings and facilitating timely decision-making.

Memory footprint analysis indicates that the trained model occupies approximately 2.3 MB, making it well-suited for deployment on resource-constrained devices, including mobile platforms and edge computing systems. Compared to larger models that require substantial computational resources, this lightweight design offers a distinct advantage for practical and scalable implementation.

H. Deployment System Evaluation

Performance and usability evaluations were carried out on the Streamlit web interface integrated with the Flask API. The API demonstrated an average response time of under 50 milliseconds, including network latency, enabling efficient handling of multiple concurrent requests. Early user testing with clinicians yielded positive feedback, highlighting the web

interface's intuitive input forms and clear, easily interpretable presentation of results.

The patient record management system effectively maintained a persistent database of predictions, enabling retrospective validation and longitudinal analyses. Timestamp-based retrieval facilitated temporal trend analysis, which could potentially uncover evolving patterns in diagnostic outcomes.

I. Performance Analysis

With an achieved accuracy of 98.4%, the proposed model ranks among the top-performing systems reported in the literature for this dataset. Despite the moderate class imbalance, the model demonstrates unbiased performance across both malignant and benign classes, as reflected by the balanced precision and recall values.

The high specificity of 98.6% is particularly noteworthy, as it minimizes false positives that can lead to unnecessary biopsies, patient anxiety, and increased medical costs. Simultaneously, the recall of 98.2% ensures that very few malignant cases are missed, which is critical for patient safety and timely intervention.

The model exhibits outstanding discriminative capability across all classification thresholds, as reflected by its nearperfect AUC-ROC score of 0.99. This flexibility allows clinicians to adjust the decision threshold based on specific clinical requirements, prioritizing sensitivity for screening purposes or specificity for confirmatory testing.

J. Clinical Implications

The proposed system offers several advantages for clinical deployment. Its web-based interface enables remote consultations, which is particularly valuable in underserved regions with limited access to specialists. Additionally, the rapid inference time allows integration into high-throughput screening workflows without necessitating additional staff, thereby potentially expanding diagnostic capacity.

Longitudinal tracking facilitated by the patient record management system enables early detection of changes in tumor characteristics over time. Retrospective analysis of predictions can reveal temporal trends, uncover population-level patterns, and inform public health strategies.

It is important to emphasize that the system is intended as a decision support tool, not a replacement for trained clinicians. Despite the model's high accuracy, medical diagnosis requires careful consideration of the patient's history, additional diagnostic tests, and clinical judgment. The system is designed to complement expert evaluation, rather than substitute for it.

K. Limitations

Several limitations should be noted. First, although widely used, the Wisconsin dataset is relatively small (569 instances) and lacks diversity in patient demographics, necessitating validation on larger and more heterogeneous populations. Second, the dataset is based on features extracted from FNA samples, which may not generalize to other diagnostic modalities such as MRI or mammography.

Third, the model does not provide detailed tumor subtyping, which is clinically important for guiding treatment decisions, as it is trained solely for binary classification (malignant vs. benign). Fourth, the system lacks advanced explainability features, such as gradient-based visualizations or attention mechanisms, which could highlight the most influential features for each prediction and further enhance clinician trust.

Fifth, the evaluation relied on retrospective data, highlighting the need for prospective validation in real-world clinical settings across diverse patient populations to confirm performance and identify potential failure modes. Sixth, prior to clinical deployment, regulatory approval processes addressing safety, efficacy, and ethical considerations must be completed.

L. Comparison with Related Work

The proposed model delivers competitive performance while maintaining lower architectural complexity compared to recent studies. Using feature selection techniques, Goni et al. reported an accuracy of 97.3%, slightly below our achieved 98.4%. Similarly, Asadi et al. attained 97% accuracy using a more complex layered deep learning architecture, aligning with our findings but with increased model complexity.

Islam et al. compared five machine learning algorithms, with their best-performing model—an ensemble approach— achieving 99.12% accuracy. Nevertheless, our streamlined architecture attains

comparable performance while offering greater computational efficiency, avoiding the extensive feature engineering and ensemble complexity required by their method.

This study is novel not only for its high classification accuracy but also for its fully integrated end-to-end system, encompassing an efficient model architecture, a RESTful API, a user-friendly interface, and optimized feature selection. Few prior studies provide such comprehensive solutions that are readily deployable in clinical settings.

M. Future Research Directions

Several avenues for future research are identified:

1) Explainable AI Integration: Incorporating explainability techniques such as Grad-CAM, SHAP (SHapley Additive exPlanations), or LIME (Local Interpretable Model-agnostic Explanations) could provide both quantitative and visual explanations of model predictions. This would enhance clinical trust and facilitate regulatory approval processes.

2) Multi-Modal Integration: Incorporating additional data modalities, such as genomic profiles, mammograms, ultrasound images, and electronic medical records, could enhance diagnostic accuracy and enable more personalized risk assessments.

3) Uncertainty Quantification: Employing ensemble methods or Bayesian neural networks to quantify prediction uncertainty would allow clinicians to assess the confidence of each prediction and identify cases that require further investigation. *4) Transfer Learning to Other Cancers:* Investigating whether the learned feature representations can be transferred to other cancer types, such as colorectal, lung, or prostate cancer, could accelerate the development of diagnostic tools across multiple malignancies.

5) Federated Learning for Privacy: Implementing federated learning frameworks would enable collaborative model training across multiple institutions without sharing sensitive patient data, addressing privacy concerns while leveraging larger and more diverse datasets.

6) Integration of Real-Time Monitoring: Integrating the system with electronic health record (EHR) platforms to provide automated diagnostic recommendations and real-time monitoring during clinical workflows could streamline diagnostic processes and reduce turnaround times.

7) Prospective Clinical Trials: Conducting prospective clinical trials across diverse patient populations in real-world clinical settings is essential to validate model performance, assess clinical utility, and identify potential unforeseen challenges.

VI. Conclusion

This study presents a deep neural network-based intelligent breast cancer detection system that achieved state-ofthe-art performance on the Wisconsin Breast Cancer Dataset. Computational efficiency was enhanced without compromising discriminative power by reducing the feature space from 30 to five essential features through systematic feature selection.

Leveraging progressive dimensionality reduction and strategically applied dropout regularization, the proposed DNN architecture outperformed conventional machine learning baselines, achieving 98.4% accuracy, 97.9% precision, 98.2% recall, and an AUC-ROC score of 0.99.

The practical clinical applicability of the system is demonstrated by its end-to-end design, which encompasses data preprocessing, model training, a RESTful API for inference, and an intuitive web interface. Computational efficiency analysis shows rapid training (2.5 minutes) and real-time inference (15 milliseconds per sample), highlighting the system's suitability for clinical deployment.

This research makes three key contributions. First, the optimized architecture balances model capacity and computational efficiency, addressing practical limitations often overlooked in academic studies. Second, clinical acceptance is enhanced through a systematic feature selection approach that reduces dimensionality while preserving biological interpretability. Third, the complete deployment framework, including the web interface and RESTful API, demonstrates the feasibility of translating the system from research to real-world clinical practice.

Despite the promising results, several limitations should be acknowledged. Validation on larger and more diverse datasets is necessary to confirm generalizability.

Incorporating explainability features would further enhance clinical trust and support regulatory approval. Additionally, prospective clinical trials in real-world settings are essential to assess clinical utility and identify potential failure modes.

Future research will explore federated learning for privacypreserving collaborative training, uncertainty quantification, transfer learning to additional cancer types, and integration of multi-modal data. Systems such as the one proposed in this study illustrate how deep learning can complement clinical expertise, enhance diagnostic accuracy, and ultimately improve patient outcomes as artificial intelligence continues to transform healthcare.

The convergence of deep learning advancements, increasing computational power, and expanding availability of medical data is ushering in a new era of precision medicine. By developing reliable, efficient, and clinically viable diagnostic tools, we can address critical healthcare challenges such as specialist shortages, diagnostic variability, and disparities in access to care. The proposed breast cancer detection system represents a step toward this goal, offering an accurate, scalable, and readily accessible solution for early cancer detection.

In summary, deep neural networks demonstrate significant potential for breast cancer detection, often matching or surpassing the performance of experienced medical professionals. With continued research, rigorous validation, and responsible implementation, AI-powered diagnostic systems can augment clinical expertise, accelerate diagnosis, enable personalized treatment, and ultimately improve patient survival rates worldwide.

Acknowledgment

The authors gratefully acknowledge the University of Wisconsin Hospitals for creating and maintaining the Wisconsin Breast Cancer Dataset, which has been instrumental in numerous machine learning studies. We also thank the UCI Machine Learning Repository for providing access to the dataset. Special recognition is given to Dr. William H. Wolberg for his pioneering contributions to medical diagnostics using machine learning.

References

[1] World Health Organization, "Breast Cancer: Prevention and Control," WHO Factsheet, 2021. Available: https://www.who.int/cancer/prevention/diagnosis-screening/breastcancer/en/

[2] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast cancer Wisconsin (diagnostic) data set," UCI Machine Learning Repository, 1995.

[3] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[5] B. Zheng et al., "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,"

Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014.

[6] M. H. Alshayeji, H. Ellethy, and S. Abed, "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach," *Biomedical Signal Processing and Control*, vol. 71, p. 103141, 2022.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[9] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[10] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016.

[11] M. O. F. Goni et al., "Breast cancer detection using deep neural network," in *Proc. IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 676–679.

[12] T. Islam et al., "Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI," *Scientific Reports*, vol. 14, no. 1, p. 5704, 2024.

[13] A. Bekkouche, M. Merzoug, M. Hadjila, and W. Ferhi, "Towards early breast cancer detection: A deep learning approach," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17517– 17523, 2024.

[14] L. Shen et al., "Deep learning to improve breast cancer detection on screening mammography," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[15] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conf. on Learning Representations (ICLR)*, 2015.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[21] M. Nasser and U. K. Yusof, "Deep learning based methods for breast cancer diagnosis: A systematic review and future direction," *Diagnostics*, vol. 13, no. 1, p. 161, 2023.

[22] B. Asadi et al., "Layered deep learning for improved breast cancer detection," in *Proc. IEEE International Conf. on Machine Learning and Applications (ICMLA)*, 2022, pp. 1397–1402.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conf. on Learning Representations (ICLR)*, 2015.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929– 1958, 2014.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conf. on Machine Learning (ICML)*, 2015, pp. 448–456.

[26] F. Chollet et al., *Keras*, 2015. Available: https://keras.io

[27] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. USENIX Symp. on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[28] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE International Conf. on Computer Vision (ICCV)*, 2017, pp. 618–626.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[31] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[32] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[33] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[34] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[35] S. Aamir et al., "Predicting breast cancer leveraging supervised machine learning techniques," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[36] M. Alotaibi et al., "Breast cancer classification based on convolutional neural networks for mammogram images," *Heliyon*, vol. 9, no. 11, 2023.