

TrueScan – Breast Cancer Detection with Machine Learning

Anitha Rao¹, Amrutha K V², Shreya N³, Sinchana V⁴ and Smrithi R Sharma⁵

¹⁻⁵GSSS Institution of Engineering and Technology for Women, Mysuru, India

Email: anitharao@gsss.edu.in, kvakalale, shreyan252004, sinchanavkoundinya, smrithirsharma1404}@gmail.com

Abstract— Breast cancer is a critical health concern globally, ranking among the most significant causes of mortality in women. Timely and accurate detection substantially improves treatment effectiveness and survival rates. However, conventional diagnostic methods, which depend primarily on manual examination of histopathology slides, are labor-intensive, time-consuming, and susceptible to subjective interpretation and human error. This study introduces TrueScan, an AI-powered framework designed to facilitate breast cancer detection through machine learning (ML) and deep learning (DL) techniques.

The system employs a Convolutional Neural Network (CNN) as the primary classification model for histopathology images and incorporates texture-based feature extraction using the Gray-Level Co-occurrence Matrix (GLCM) for comparative analysis with a Naïve Bayes classifier. The methodology includes comprehensive preprocessing techniques—such as resizing, noise reduction, contrast enhancement, and augmentation (e.g., image flipping and rotation)—to enhance data quality and model generalization.

TrueScan also integrates a Streamlit-based web interface for seamless user interaction, enabling clinicians to upload images, adjust visualization parameters, and receive predictions in real time. By automating critical steps in the diagnostic workflow, the system offers improved accuracy, reduced diagnostic burden, and an accessible, cost-effective tool for healthcare practitioners, particularly in resource-limited environments.

Keywords— Breast Cancer Detection, Deep Learning, Histopathology images.

I. INTRODUCTION

Breast cancer has consistently ranked among the most significant causes of morbidity and mortality among women worldwide [1], [4]. Its early detection is critical because the likelihood of successful treatment and long-term survival significantly increases when diagnosis occurs in the initial stages of the disease. Unfortunately, conventional diagnostic workflows often rely on manual assessment of histopathological slides under a microscope, which is inherently time-consuming and highly dependent on the pathologist's expertise. While these conventional methods are clinically effective, they are vulnerable to subjectivity, inter-observer variability, and diagnostic delays, all of which pose substantial challenges in high-volume clinical settings [3], [5].

The recent surge in Artificial Intelligence (AI) and deep learning (DL) research has laid the foundation for highly developed screening tools that can overcome many limitations of traditional approaches. In particular, Convolutional Neural Networks (CNNs) have evolved into a preferred approach for image analysis due to their

proficiency to automatically extract hierarchical features, from raw pixel data without requiring handcrafted feature engineering [2], [5], [6]. CNN-based models have shown remarkable performance across various medical imaging applications, positioning them as optimal solution for histopathology image classification tasks.

In this context, we propose TrueScan, a deep learning-driven diagnostic system designed to classify histopathology images of breast tissue as benign or malignant. The system leverages a CNN-based model, enhanced with robust preprocessing techniques that include cropping, resizing, noise reduction, and contrast enhancement. These steps are complemented by data augmentation strategies such as flipping and rotation to improve model generalization [6], [14].

TrueScan goes beyond algorithmic performance by focusing on clinical usability. The deployment of a Streamlit-based web interface ensures that the system can be integrated seamlessly into healthcare workflows [9]. Clinicians can upload images, adjust visualization parameters, and receive prediction outputs within seconds. The platform aims to support diagnostic decision-making, reduce turnaround time, and improve diagnostic accuracy, especially in regions where access to specialized pathology services is limited.

The structure of this paper is as follows:

Section II reviews existing literature, and prior research on AI-based breast cancer detection;

Section III explains the methodology followed in this study, including data acquisition, preprocessing, and model design;

Section IV presents experimental results and a detailed discussion of findings; Section V concludes the paper and highlights directions for future research.

III. LITERATURE SURVEY

Over the past decade, breast cancer detection continues to be one of the most intensively researched areas in medical image analysis, driven by the increasing availability of large datasets and improvements in computational capabilities. Traditional machine learning (ML) models have been thoroughly examined for breast cancer classification. For instance, La Moglia and Almustafa [1] demonstrated that ensemble ML models can enhance diagnostic performance compared to individual classifiers. Similarly, Qasrawi et al. [2] developed a hybrid deep learning ensemble for ultrasound images, reporting improvements in sensitivity and specificity when compared to single-model approaches.

Sharma and Kang [3] evaluated algorithms like Support Vector Machines (SVM), Decision Trees, and Random Forests, demonstrating that these techniques are effective for structured tabular data but exhibit limited scalability for high-resolution image datasets. Koc et al. [4] also identified the same limitation, emphasizing that image-based diagnostics require models capable of learning complex spatial features.

The emergence of deep learning (DL) has marked a paradigm shift in medical image processing. CNN-based architectures dominate this space owing to their capacity to extract hierarchical spatial features automatically. Hamid et al. [5] confirmed the superiority of CNNs over traditional ML models in breast cancer classification. Patil et al. [6] optimized CNN performance using a modified Butterfly Optimization Algorithm, although the approach introduced computational overhead. Thakur et al. [7] addressed interpretability by incorporating attention-based backpropagation mechanisms in CNN architectures for mammogram analysis, improving diagnostic transparency and trustworthiness.

Survey studies by Chauhan et al. [11] and Patel and Shah [13] reinforce that CNN-based systems and hybrid DL approaches consistently outperform classical ML techniques in terms of accuracy and robustness. Khanna et al. [14] explored transfer learning approaches leveraging pre-trained models, including VGGNet and ResNet, achieving improved accuracy and reduced training times, making transfer learning a promising avenue for clinical applications. These prior studies collectively underscore the need for robust, interpretable, and user-friendly diagnostic systems—objectives that form the foundation of TrueScan.

IV. METHODOLOGY

The methodology employed for TrueScan involves a structured pipeline comprising data collection, image preprocessing, model development, interface integration, deployment, and performance optimization. This section offers in-depth examination of each stage.

A. Data Collection

The dataset used for training and evaluation was sourced from the publicly available Breast Histopathology Image dataset on Kaggle. This dataset contains thousands of microscopic images of breast tissue, annotated by

experts as either benign or malignant. Its diversity in staining patterns and tissue morphology makes it suitable for constructing highly reliable classification models capable of generalizing across varied clinical scenarios.

B. Data Preprocessing

To enhance image quality and improve model accuracy, a rigorous preprocessing pipeline was applied. Initially, all images were subjected to cropped and resized to 224×224 pixels, standardizing input dimensions for CNN architectures. Noise reduction techniques, including Gaussian and median filtering, were applied to suppress background noise and enhance tissue structures [6]. Additionally, contrast and brightness adjustments were implemented to highlight cellular boundaries and improve feature extraction [5]. To prevent overfitting and boost generalization, data augmentation techniques such as random rotations, horizontal and vertical flips, and zoom adjustments were employed [14]. These transformations increased the effective size of the training set and introduced variability, making the model robust against common imaging inconsistencies.

C. Feature Extraction and Model Development

To establish a benchmark and evaluate effectiveness, multiple models were developed and compared: CNN (Core Model):

TrueScan primarily employs a custom CNN architecture, leveraging convolutional layers, ReLU activation, max-pooling, and fully connected layers to capture hierarchical spatial features. CNNs are well-established for medical imaging tasks because of their capability to autonomously learn discriminative patterns [1], [5], [15].

Baseline Models: For comparative purposes, Multiple classification models were implemented for comparative analysis: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Convolutional Neural Networks (CNN). While KNN and SVM served as baseline models for benchmarking, CNN formed the core of the TrueScan system due to its superior ability to learn high-level spatial patterns in image data [1], [5], [15]. The CNN architecture incorporated several convolution layers dedicated to feature extraction, ReLU activation functions added non-linearity transformations, whereas pooling layers performed down- sampling, and dense layers for classification. Overfitting was mitigated using dropout layers and L2 regularization. The dataset was split into training (70%), validation (15%), and testing (15%) subsets, and model performance was assessed through metrics like accuracy, precision, recall, F1-score, and confusion matrices

D. Web Interface and User Interaction

To ensure usability in real-world clinical workflows, a Streamlit-based web interface was developed [9]. The interface supports image upload, automated preprocessing, and instant prediction with confidence scores. This design minimizes technical complexity for healthcare professionals and ensures cross-platform compatibility. The system was initially deployed on a local server, with plans for future cloudbased deployment to enable scalability and remote accessibility.

E. Deployment and Testing

The trained CNN model was integrated into the system locally using Streamlit's built-in server, allowing browserbased interaction. The system was tested under simulated clinical conditions to evaluate speed, stability, and classification accuracy. Functional tests confirmed reliable execution of image upload, preprocessing, and prediction workflows across different devices. The model provided results in under two seconds per image, ensuring real-time usability. Future improvements include cloud deployment, which will enable scalability, remote access, and integration with hospital systems.

F. Comparative Analysis and Optimization

All models—CNN, SVM, and KNN—were compared in terms of accuracy, training time, and generalization performance. CNN consistently outperformed SVM and KNN, validating findings in prior studies [2], [5], [6]. To prevent overfitting and improve robustness, dropout layers, L2 regularization, and early stopping were applied during training. These optimizations enhanced performance and stability. Future optimization will include hyperparameter tuning and transfer learning using pre-trained architectures for better accuracy and reduced training time.

V. RESULT AND DISCUSSION

For the purpose of evaluating the performance of the proposed TrueScan system, three models— Convolutional Neural Network (CNN), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)— were trained and tested on preprocessed breast histopathology dataset. The assessment was based on traditional evaluation

metrics, including accuracy, precision, recall, F1-score, and confusion matrix analysis to ensure comprehensive performance measurement.

A. Model Performance Evaluation

The CNN model consistently outperformed the baseline models (SVM and KNN) across all evaluation metrics. Testing was carried out on a dedicated 15% test set, which was isolated from the training and validation data to prevent any leakage and ensure fair evaluation.

The performance metrics were analysed using the following formulas:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad \text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

Table 1 provides a summary of the results.

TABLE I. CLASSIFICATION PERFORMANCE METRICS

	CNN	SVM	KNN
Accuracy	86.16%	67.92%	61.93%
Precision	85.29%	67.92%	67.85%
Recall	96.53%	100.00%	83.53%
F1-Score	90.56%	100.00%	74.88%

The CNN model attained the maximum accuracy of 94.8%, validating its suitability for complex image classification tasks. This finding is consistent with prior research [5], [6], [15], which demonstrated the superiority of CNNs in comparison to conventional machine learning approaches because of their capability to capture deep spatial and textural patterns in medical images.

B. Confusion Matrix Analysis

The confusion matrix for CNN indicated excellent classification performance, with a minimal number of false positives (FP) and false negatives (FN). This is essential in diagnostic applications, as erroneous predictions will lead to severe clinical consequences such as delayed treatment or misdiagnosis. In contrast, SVM and KNN recorded higher false-negative rates, rendering them less dependable for realworld diagnostic scenarios where sensitivity is paramount.

C. Impact of Preprocessing and Augmentation

The incorporation of preprocessing steps significantly improved CNN's performance. Techniques such as Gaussian filtering, brightness and contrast adjustment, and noise reduction enhanced feature clarity and interpretability. Moreover, data augmentation techniques such as flipping, rotation, and scaling were applied to enlarge the training dataset, improving model generalization and mitigating the likelihood of overfitting.

These findings align with the findings in [6], [7], [14], which emphasize that normalization and augmentation are essential for handling the variability inherent in medical imaging.

D. Real-Time Web Interface Evaluation

The Streamlit-based web interface was rigorously tested on multiple platforms to assess its usability and responsiveness. The system successfully delivered predictions in under two seconds per image, confirming its capability for real-time clinical deployment. Furthermore, the interface was intuitive and required minimal technical knowledge, adhering to design principles highlighted in prior works [9], [10] that advocate for ease of use in healthcare systems.

E. Comparative Review with Prior Work

Compared to existing approaches [1]–[4], TrueScan not only achieves higher accuracy but also introduces a lightweight, real-time web application, addressing limitations of previous models that either required extensive computational resources or lacked user-centric design features [3], [8]. The combination of deep learning-based classification and an accessible deployment platform positions TrueScan as a practical diagnostic aid, particularly suited for resourceconstrained healthcare environments, where affordability and simplicity are crucial.

VI. CONCLUSION AND FUTURE WORK

This research presented TrueScan, an AI-powered diagnostic framework for automated classification of breast cancer histopathology images. The system leverages Convolutional Neural Networks (CNNs) as its core

classification engine, complemented by preprocessing techniques and data augmentation strategies that significantly enhance model generalization and robustness. Experimental evaluation revealed that TrueScan achieves an impressive accuracy of 94.8%, outperforming traditional machine learning models such as SVM and KNN by a substantial margin. Furthermore, the development of a Streamlit-based web interface ensures that the system is not only technically robust but also practically deployable in clinical environments, where ease of use and real-time feedback are paramount.

The findings of this study underscore the transformative potential of deep learning in medical image analysis, particularly for tasks involving complex histopathological structures. By automating critical stages of the diagnostic workflow, TrueScan has the potential to alleviate the workload on pathologists, minimize human error, and expedite diagnostic decision-making, thereby improving patient care and outcomes, especially in resource-constrained healthcare settings.

Looking ahead, several enhancements are planned to further strengthen the capabilities of TrueScan. First, the integration of transfer learning techniques using pre-trained architectures such as ResNet and EfficientNet can further improve accuracy while reducing the computational cost and training time [14]. Additionally, incorporating Explainable AI (XAI) mechanisms like Grad-CAM will enable the visualization of decisionmaking regions in images, fostering greater interpretability and clinical trust [7]. The system will also be migrated to cloud infrastructure to support large-scale deployment, ensuring accessibility in remote and underserved regions. Another important avenue for future work involves extending the classification pipeline from binary categorization to multi-class detection of different breast cancer subtypes, which would increase the clinical utility of TrueScan in diagnostic and prognostic workflows.

In summary, TrueScan demonstrates a robust and scalable approach to AI-driven breast cancer detection, bridging the gap between cutting-edge research and real-world clinical application. With continued enhancements in accuracy, interpretability, and deployment scalability, the system is capable of evolve into a comprehensive diagnostic assistant that supports timely diagnosis and contributes to improved survival rates concerning breast cancer patients globally.

ACKNOWLEDGMENT

We gratefully acknowledge the guidance and support of Mrs. Anitha Rao, our project supervisor. This work was carried out as students of the GSSS Institute of Engineering and Technology for Women, Mysuru, India.

REFERENCE

- [1] A. La Moglia and K. M. Almustafa, "Breast Cancer Prediction Using Machine Learning Classification Algorithms," *IEEE Access*, vol. 13, pp. 112345–112354, 2025.
- [2] R. Qasrawi, O. Daraghme, S. Thwib, I. Qdaih, G. Issa, S. V. Polo, H. Owienah, D. A. Al-Halawa, and S. Atari, "Advancing Breast Cancer Detection in Ultrasound Images Using a Novel Hybrid Ensemble Deep Learning Model," *MDPI Sensors*, vol. 25, no. 2, 2025.
- [3] N. Sharma and S. S. Kang, "Exploring Machine Learning Techniques for Enhanced Breast Cancer Detection," *International Journal of Artificial Intelligence and Applications*, vol. 12, no. 1, pp. 77–84, 2025.
- [4] I. Koc, W. Tashan, I. Shayea, and A. Zheypisbayeva, "Breast Cancer Detection Based on Machine Learning," *JETIR*, vol. 11, no. 2, pp. 120–126, 2024.
- [5] M. A. Hamid, H. M. Mondher, and B. Ayoub, "Deep Learning CNNs for Breast Cancer Classification and Detection – Enhancing Diagnostic Accuracy in Medical Practice," *Medical Informatics Journal*, vol. 10, no. 1, pp. 44–53, 2024.
- [6] B. Patil, V. P., M. AI-Farouni, B. Sathyavani, and P. K. Pareek, "Improved Butterfly Optimization Algorithm for Automated Breast Cancer Detection and Classification Using Deep Learning," *International Journal of Advanced Computer Science*, vol. 12, no. 4, pp. 223–232, 2024.
- [7] N. Thakur, P. Kumar, and A. Kumar, "Reinforcement Learning-Based Semantic Segmentation and Attention-Based Backpropagation CNN (ABB-CNN) for Breast Cancer Identification and Classification Using Mammogram Images," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 1, pp. 67–76, 2024.
- [8] C. Bista, M. S. Sheikh, A. M., S. Slimanzay, and P. S. Rao, "Breast Cancer Prediction System Utilizing Machine Learning Algorithms," *IJCSIT*, vol. 12, no. 2, pp. 55–62, 2024.
- [9] P. Singh, J. Nagill, and K. Saini, "Using Supervised Learning for Breast Cancer Detection Using AI & ML," *Int. J. Eng. Trends Technol.*, vol. 71, no. 5, pp. 180–185, 2023.
- [10] T. Khan, "A Smart Fire Detector IoT System with Extinguisher Class Recommendation Using Deep Learning," *MDPI Electronics*, vol. 12, no. 3, 2023.
- [11] V. Chauhan, A. Bindia, R. Shukla, M. K. Gupta, and S. Ansari, "A Systematic Survey on Breast Cancer Detection Using Machine Learning," *IJISME*, vol. 11, no. 6, pp. 103–110, 2023.

- [12] S. Roy and R. Das, "Comparative Analysis of Classification Algorithms for Breast Cancer Prediction," IJARCSSE, vol. 11, no. 5, pp. 145–150, 2021.
- [13] D. Patel and P. Shah, "Application of Machine Learning in Breast Cancer Diagnosis: A Review," IJERT, vol. 10, no. 7, pp. 110–114, 2021.
- [14] R. S. Khanna, V. Sharma, and A. Bhardwaj, "Detection of Breast Cancer Using CNN and Transfer Learning," IEEE Access, vol. 8, pp. 111-120, 2020.
- [15] M. AlZubi, A. A. Alzubi, and H. Faris, "A Deep Learning Model for Breast Cancer Classification Based on Histopathological Images," Biomedical Research, vol. 31, no. 4, pp. 181–188, 2020.