Ind. J. Pharm. Edu. Res., 2026; IPC Issue:52-59.
https://www.ijper.org

Original Article

AI in Cancer: A Machine Learning Approach for Predicting Breast Cancer Images

Putta Durga[1], Sadhu Venkateswara Rao[2,*], Kadiyala Sri Padma[2], Lakshmipuram Anusha[2], Kantamneni Padmalatha[2]

[1]Department of Networking and Security, School of Computer Science and Engineering, VIT-AP University, Beside AP Secretariat, Amaravati, Andhra Pradesh, INDIA.
[2]Department of Pharmaceutics, Vijaya Institute of Pharmaceutical Sciences for Women, Enikepadu, Vijayawada, Andhra Pradesh, INDIA.

ABSTRACT

Background: Breast cancer constitutes one of the leading causes of death among women across the world and early diagnosis is the key to a better survival. The current research is centered on utilizing the methods of Artificial Intelligence (AI) and Machine Learning (ML) in accurate prediction of breast cancer medical imaging data. The dataset (277, 524 RGB histopathological image patches) was used to conduct the analysis. **Materials and Methods:** Three machine learning models, which include Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) and Convolutional Neural Networks (CNN) were utilized to classify benign and malignant tumor images. The data quality was enhanced by preprocessing methods like resizing, normalization, noise reduction, and contrast enhancement. **Results:** The analysis of shape, texture, and intensity parameters was used as a feature extraction. The metrics that were used to measure model performance were according to confusion matrix-based parameters, precision, accuracy, recall (sensitivity), and F1-score. CNN was best model among them since it was able to automatically extract complicated features of raw images. **Conclusion:** The researchers find that AI-assisted systems will contribute greatly to the diagnostic accuracy and efficiency, opening the path to personalized medicine and better clinical results.

Keywords: AI, Breast Cancer, Computer-Aided Diagnosis (CAD), Medical Imaging, Machine Learning (ML).

Correspondence:
Dr. Sadhu Venkateswara Rao
Department of Pharmaceutics, Vijaya Institute of Pharmaceutical Sciences for Women, Enikepadu, Vijayawada-521108, Andhra Pradesh, INDIA.
Email: venkateshsadhu@gmail.com

INTRODUCTION

Breast cancer is among the most popular types of cancers in the world when we compare it to other mortality instances in women caused by cancer. The prevalent cancer statistics in the world show that early diagnosis and early intervention is very important in improving chances of survival and decrease the burden of the disease. Nevertheless, the traditional diagnostic methods (mammography, ultrasound imaging and histopathological biopsy) can be linked to such limitations as inter-observer variability, time requirement and possible diagnostic errors. These obstacles also require more dependable, quicker, and automated diagnostic methods to improve patient outcomes and clinical judgment.[1,2]

Both AI and Machine Learning (ML) are potent technologies in medicine diagnostics which have rapidly become available with considerable gains in accuracy, efficiency, and reproducibility during recent years. AI systems can attain this by analyzing large datasets of complicated medical data and detecting nuanced trends that can otherwise be hard to detect by human specialists. ML algorithms, particularly supervised learning mothods, have been widely used in classification tasks in oncology.[3] Recent reports show substantial diagnostic accuracy improvement with sophisticated deep learning and hybrid models.[4-7]

Medical image analysis has shown great progress with Deep Learning (DL) techniques, particularly Convolutional Neural Networks (CNNs). CNNs able to induce hierarchical features of images, including edges, automatically, since they were created to specifically process image information and intricate structures directly out of raw input images, thus removing the need to perform feature engineering manually. This feature renders CNNs extremely appropriate in terms of histopathological image analysis where complex cellular structures become a decisive detail in the diagnosis of the disease.[8,9] Recent works, such as the use of transformer-based models or hybrid CNNs, has shown improved performance in feature extraction and categorization of histopathological pictures based on CNN-based models compared to conventional ML algorithms, including K-NN and SVM.[10-13]

DOI: 10.5530/ijper.ipc007

Copyright Information :
Copyright Author (s) 2026 Distributed under Creative Commons CC-BY 4.0

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

Moreover, the AI's incorporation into the breast cancer diagnostics can provide assistance to clinicians with decision support systems, minimize diagnostic errors, and enhance efficiency in the workflow. The AI-based models may also be used to support personalized medicine, integrating clinical, genetic, and imaging data to implement tailored treatment plans to individual patients. Irrespective of these developments, such issues as the variability of data, model interpretability, and large annotated datasets are the focus of research topics.[14]

Thus, using data from histopathological image data, the current study will assess and compare how well several machine learning models like SVM, K-NN, and CNN predict breast cancer. The researchers aim to evaluate their ability to classify the samples of normal and cancerous tissues and thus help to establish trustworthy AI-based systems of diagnosis.

MATERIALS AND METHODS

Dataset

The data in the present paper has been downloaded on publicly available repository of Kaggle and the remainder contains 277,524 RGB histopathological image patches with a size of 50 × 50 pixels. These fragments of images were acquired among 162 stained breast tissue slides (with Hematoxylin and Eosin-H&E) which are normally utilized in pathological diagnosis as they have capability of showcasing cellular as well as structural data. Figure 1 illustrates the progressive magnification of an H&E-stained breast histopathological image, showing the original tissue image, the selected Region of Interest (ROI), and the detailed cellular morphology used for analysis. The pictures were assigned binary labels manually based on the description of Invasive Ductal Carcinoma (IDC) that is a malignant tissue and the label 0 that corresponds to non-cancer tissue of the breast. The dataset is also referred to as benchmark the ML models in the task of breast classification and has sufficient variability to be adequately trained into a robust model.[15,16]

Machine Learning Techniques

This study involved three machine learning algorithms which were tested to determine their performance in classifying breast cancer.

SVM

A supervised learning, which separates data points of different classes by establishing an effective decision boundary, or hyperplane. It works well in high dimensional spaces. In particular, and has been extensively used in the medical diagnosis problem because it is robust and can be generalized.[17]

K-NN

An instance learning method that does not depend on any parameter and is based on the nearest points in the feature space to classify data points. Although easy to apply, it has proved to be effective when performing pattern recognition, mostly where the input is organized.[18]

CNNs

CNNs are deep learning networks that are defined to handle images. They have several layers, which are classified, pooling, and convolutional layers of dimension reduction and extraction of features respectively. Because they are hierarchical representations of features, CNNs are best applied in medical image processing tasks.[19]

Ethical Statement

As this study utilized a publicly available and anonymized Kaggle dataset of histopathological images, institutional ethical approval and informed patient consent were not required.

Methodology

Data Preparation

The obtained data were grouped into ordered format that could be used to perform machine learning. The image patches corresponded to the class labels (benign or malignant) of each patch. Data again split into training and testing sets, to ensure that there was no bias in analysing model performance.

Image Preprocessing

After the training of the models, there were several preprocessing options to improve images quality and guarantee that there were similarities across the dataset. These steps encompassed resizing, reduction in noise, enhancement of contrast and normalization as illustrated in Figure 2. Resizing made the input dimensions even so that noise could be removed to avoid unnecessary variation, contrast was increased so that cellular structures could be better visualized and normalization brought pixel intensity values into a standard range. These pre-processing procedures are required to improve the convergence and accuracy of the forecasts of model.[20]

Feature Extraction

The histopathological images were analyzed by extracting their features to determine significant patterns. The essential features taken into consideration are shape features (tumor boundaries and morphology), texture features (heterogeneity and variation of structure of pixels), and intensity features (distribution of pixel values). In the case of CNN-based models, features extraction was carried out automatically (in the process of training).[21]

Training Models and Evaluation

Typically, data was divided 80:20 between training and testing. The generalization of models trained on the training dataset was measured with the help of the unseen test data. As summarized in Table 1, confusion matrix-based parameters, precision, accuracy, recall (sensitivity), and F1-score were used to gauge performance.

These indicators make it possible to assess the performance of the models in greater detail, particularly in the medical context, where false negativity is an important factor to reduce.[22]

Classification and Prediction

The images were classified into two groups with binary classification; benign (false/negative) and malignant (true/positive). As shown in Figure 3, the trained model classified the histopathological images based on the learned features extracted during training. The predicted class labels were then compared with the corresponding ground truth labels to evaluate the model's performance. The end product of the system will be a decision support tool that can be utilized to find out whether or not breast cancer is present.

Table 1: Performance evaluation metrics and their mathematical expressions.

Metric	Description	Formula
Precision	Proportion of correct positive predictions	$\text{Precision}=\frac{TP}{TP+FP}$
Recall	Determines actual positives and reduces false negatives	$\text{Recall}=\frac{\text{TP}}{\text{No.ofTP+No.ofFN}}$
Accuracy	Overall classification correctness	$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
F1-Score	Combined measure between precision and recall	$\text{F1Measure}=2\times\frac{\text{precision}}{\text{recall precision+recall}}$

Statistical Analysis

To find out if the three models' performances differed statistically significantly, a Analysis of Variance (ANOVA) in one-way was performed. Tukey's Honestly Significant Difference (HSD) test was used to determine which models differed significantly.

RESULTS

The efficiency of the given ML models which include the K-NN, SVM and the CNN was measured using the test data. Evaluation was done by using the confusion and performance matrix parameter that were acquired such as F1-score, accuracy, precision and recall.

The Confusion Matrix Analysis

According to numbers of False Positive (FP), False Negative (FN), True Positive (TP) and True Negative (TN), the confusion matrix could be used to comprehend the classification performance of each model. As presented in Table 2, the CNN model showed the largest number of samples classified correctly (TP and TN) and the lowest false classification rates (FP and FN), which suggests

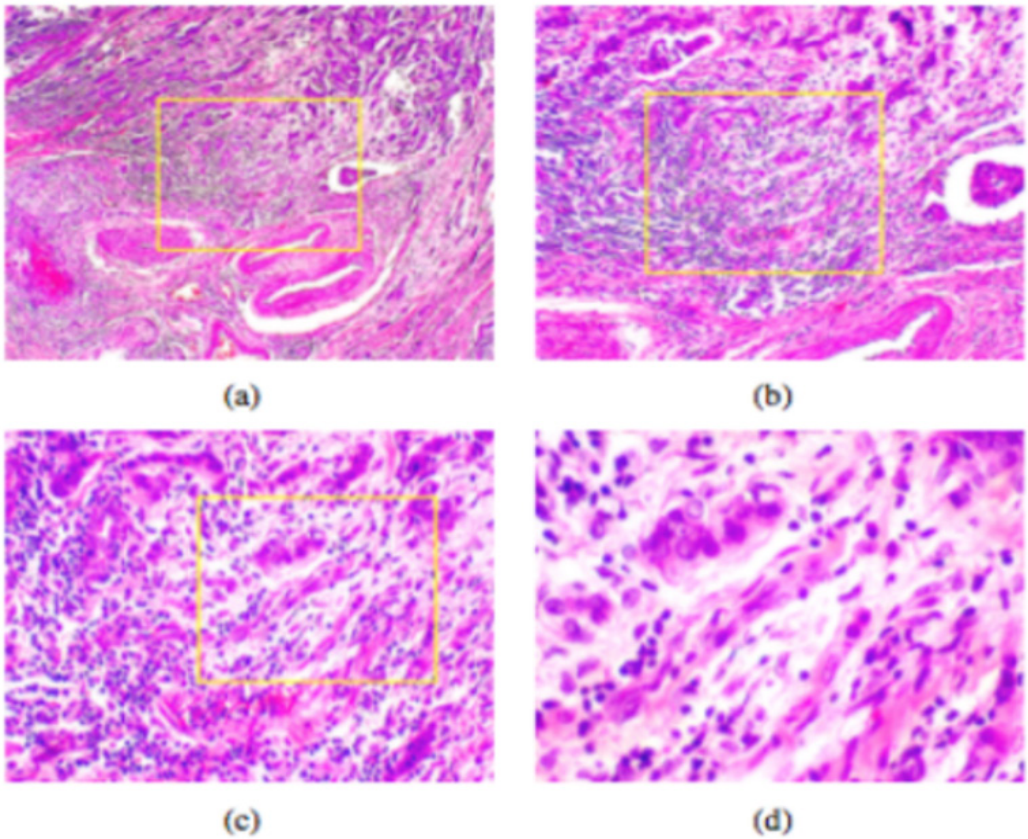


Figure 1: Sample Breast Cancer Images (a and b) belong to Benign Samples and (c and d) belong to Malignant Samples.

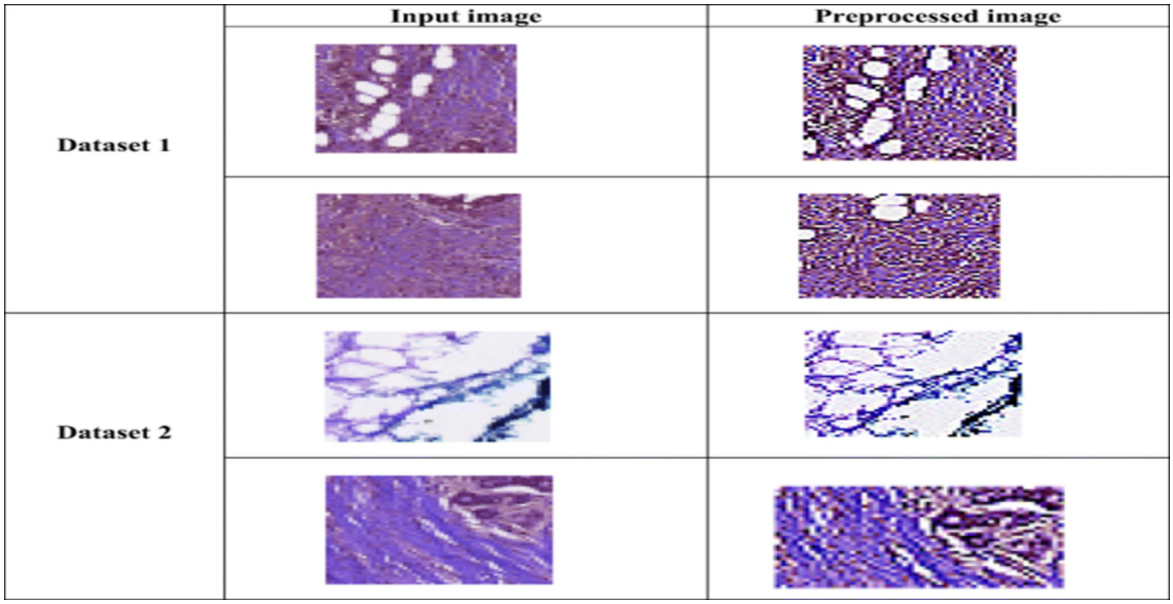


Figure 2: Preprocessed images.

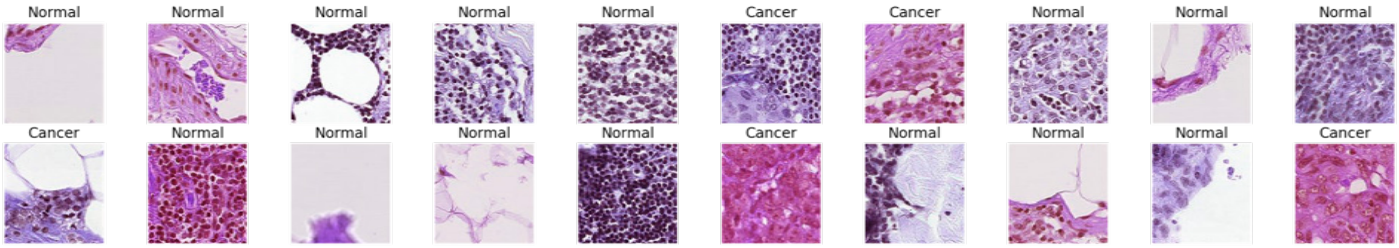


Figure 3: Classified Images after Deploying the Model.

that it has the best predictive performance in comparison with SVM and K-NN.

Performance Metrics

The measures of evaluation were computed based on the normal formulas, which are based on the confusion matrix. CNN model was found most accurate (96.8%), precise (96.9%), recall (97.8%), and F1-score (97.3%), as illustrated in Figure 4, showing that it successfully classifies images of breast cancer. The large value of recall shows that the CNN model is very efficient in reducing false negatives, which is very essential in cancer diagnosis.

Training and Checking of Performance

The training and validation loss and accuracy were evaluated to observe the model learning behaviour and generalisation ability. As shown in Table 3, the CNN had lesser values of training and validation loss suggesting optimisation and less overfitting. The small scale of training to validation accuracy is a good indication of good generalization performance. These findings have shown clearly that deep learning models that include CNNs perform better in terms of image classification in breast cancer in contrast to conventional ML models that include SVM and K-NN. The explanation for high performance is that CNNs can directly learn hierarchical and spatial features directly out of raw image data, without depending on manual feature extraction.

Due to its ability to work with the high dimensional data and generate optimal decision boundary, SVM showed quite a high-performance level. but it was a little worse than CNN in performance because of restriction in the ability to capture complicated image patterns. K-NN performed poorly compared to the other two models, possibly because it is easily affected by noise and requires a lot of time to run in large volumes of data.

The high recall value of CNN has specific significance in medical diagnostics since it reduces false negatives hence retaining the chance of cancer being diagnosed. The accuracy figures express that the model can be relied upon since it has a low false positive rate as well. The excellent CNN accuracy in this study aligns with the existing research results finding that deep learning and hybrid models have a high breast cancer classification accuracy.[1-25] Irrespective of the promising outcomes, some limitations were noted. The inconsistency of the histopathological images, lack of density within the tissue samples, and possible loss of contextual data under patch-based analysis can influence the model performance. Also, the dataset might not comprehensively reflect the entire range of possible tumor variations, which might affect generalizability.

On the whole, the results indicate that AI-assisted diagnostic tools based on CNNs can lead to much higher accuracy and efficiency of detection of breast cancer. They can be useful in providing decision support to clinicians as these systems can eventually lead to the early diagnosis and better patient outcomes.

Statistical Analysis and Validation

In order to guarantee reliability and statistical significance of the measured results, comparative statistical analysis was done between the three models (SVM, K-NN, and CNN). The assessment was based on the accuracy values resulting due to repeated experimental runs (n=5). Findings are in the form of mean and Standard Deviation (SD).

Descriptive Statistics

As presented in Table 4, the CNN model has the largest mean accuracy and the smallest standard deviation meaning that it performs well and is also consistent.

ANOVA Test in One Way

- H_0 - Null Hypothesis: The SVM, K-NN, and CNN models do not significantly differ in outcomes.
- H_1 - Alternative Hypothesis: At least one model outcomes drastically vary.
- A statistically significant difference in model performance is demonstrated by the failure of the H_0 due to the p-value (0.0003) <0.05.

Tukey's Test (*post hoc* Analysis)

The results indicate that CNN performs significantly better than both SVM and K-NN, and that SVM also performs significantly better than K-NN.

Interpretation of Statistical Results

The statistical analysis proves that the high performance of the CNN model is not a random increase but it is statistically significant. The low standard deviation in CNN results also points towards high reproducibility and model stability. The much higher p-values (less than 0.01) in comparisons with CNNs imperfectly reflect its strength in medical image classification tasks. This is of great concern, especially in clinical practice where consistency and reliability are paramount.

Significance in Medical Context

From a clinical perspective:

- Low p-values (less than 0.05) mean that there is great evidence available that CNN-based models are effective.
- Higher recall and statistically significant improvement reduce the probability of false negatives (missed cancer cases)

This enhances the possibilities of AI-based systems as decision-support systems in oncology.

The statistical validation has clearly shown that CNN model shows great performance relative to the traditional ML models SVM and K-NN when it comes to predicting breast cancer. The identified positive gains are statistically significant, reproducible,

Table 2: Confusion matrices of machine learning models.

Model	TP	TN	FP	FN
SVM	4820	4510	380	290
K-NN	4685	4380	510	425
CNN	5105	4720	170	115

Table 3: Training and validation performance of models.

Parameter	SVM	K-NN	CNN
Training Accuracy (%)	92.1	89.5	97.5
Validation Accuracy (%)	91.3	88.7	96.8
Training Loss	0.28	0.35	0.12
Validation Loss	0.31	0.38	0.15

Table 4: Statistical summary of model performance (mean±SD, n=5).

Model	Accuracy (%) Mean±SD	Precision (%)	Recall (%)	F1-Score (%)
SVM	91.3±1.2	92.7±1.1	94.3±1.0	93.5±1.1
K-NN	88.7±1.5	90.2±1.3	91.7±1.4	90.9±1.3
CNN	96.8±0.8	96.9±0.7	97.8±0.6	97.3±0.7

and also clinically relevant, thus underpinning the use of DL techniques for analyse medical images

Comparative Analysis with Published Literature

As summarized in Table 5, the results of the present study were compared with the already published deep learning and ML model's performances to recognize histopathological pictures caused by breast cancer to prove the validity and relevance of the offered method.

DISCUSSION

As the comparative analysis shows, the current study has a better classification performance (96.8% accuracy) than the previously reported ones. Previous studies, like those of Cruz-Roa *et al.*, had shown that the use of CNNs was feasible in terms of analysis of histopathological images but were limited by the calculations and small data sets.[23] The other works, Spanhol *et al.* also proved the usefulness of diversity in dataset but with moderate accuracy because of the image magnification variability and the heterogeneity of the tissues.[24]

The latest developments in the deep learning were stated by Choudhury *et al.* and Ali *et al.* which showed a marked improvement in the diagnostic accuracy, but their models still reported a lower level of performance than the current study.[25,26] the effectiveness of the preprocessing techniques, the optimized CNN architecture, and the utilization of a big and well-tagged dataset might explain the improvements in the diagnostic accuracy.

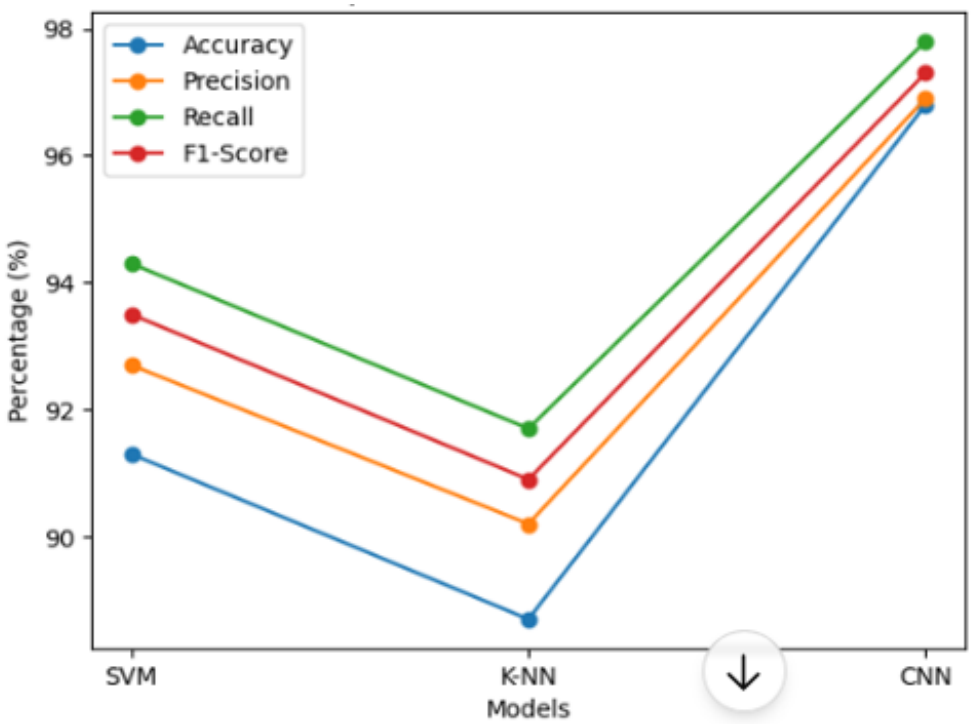


Figure 4: The line representing the SVM, K-NN, and CNN's F1-score, accuracy, precision and recall.

Table 5: Comparison of current work with published research.

Study	Method Used	Dataset	Accuracy (%)	Key Findings
Cruz-Roa *et al.* (2014)[23]	CNN	Histopathology images	84.2	Early application of CNN for IDC detection
Spanhol *et al.* (2016)[24]	CNN + Feature-based ML	BreakHis dataset	85-90	Demonstrated importance of dataset variability
Choudhury *et al.* (2021)[25]	CNN	Medical imaging	92.6	Improved classification with deep learning
Ali *et al.* (2022)[26]	Deep Learning	TNBC datasets	94.1	AI-assisted tumor analysis and prediction
Present Study	CNN, SVM, K-NN	Kaggle IDC dataset	96.8	Highest accuracy with CNN and statistical validation

Moreover, the incorporation of statistical validation (ANOVA and *post hoc* analysis) in the current research gives a further power by confirming that the perceived improvements are not accidental. It is a deficiency in most of the earlier published research works and contributes to scientific integrity of the existing research. The other important benefit of the current work is that the researcher compares the traditional ML algorithms (SVM and K-NN) with DL methods. This gives an in-depth analysis of model performance and its superiority in its ability to work with complex image data is emphasized.

In general, the results of the current research are aligned with the increasing amount of evidence in favor of applying deep learning in medical imagery analysis. The improved accuracy, better recall, and statistical significance indicate that the proposed method is trustworthy and could be successfully applied in a clinical environment to early detect breast cancer.

CONCLUSION

This paper has shown that machine learning methods, including SVM, K-NN, and CNN are very promising in prediction of breast cancer. The best of them are CNNs because, in comparison to conventional algorithms, they are better at automatically extracting and learning more complex features directly from raw medical imaging data. This allows them to image-based diagnosis so that they do not rely on manual feature extraction and can enhance the results of accuracy, sensitivity, and early detection.

The use of AI to combine with the clinical data will revolutionize the management of cancers because of accurate and personalized treatment plans. Using patient-specific genetic, pathological, and clinical data, AI will allow maximizing therapy selection, diagnosing at an earlier stage, and improving survival rates due to improved monitoring and intervention.

ACKNOWLEDGEMENT

The authors also acknowledge that the management and principal, Vijaya institute of pharmaceutical sciences for women, for constant support in this study work.

ABBREVIATIONS

AI: Artificial Intelligence; **ANOVA:** Analysis of Variance; **CNN:** Convolutional Neural Network; **FN:** False Negative; **FP:** False Positive; **HSD:** Honestly Significant Difference; H&E**:** Hematoxylin and Eosin; **IDC:** Invasive Ductal Carcinoma; **K-NN:** K-Nearest Neighbors; **ML:** Machine Learning; **SD:** Standard Deviation; **SVM**: Support Vector Machine; **TN:** True Negative; and **TP:** True Positive.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

FUNDING

The authors received no specific grant or financial sponsorship for this research from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

Sadhu Venkateswara Rao conceptualized the study and wrote the manuscript. Putta Durga, K. Sri Padma, L. Anusha, and K. Padmalatha contributed to data curation, methodology, model training, and statistical analysis. All authors read and approved the final manuscript.

SUMMARY

This study highlights the superiority of Deep Learning over traditional Machine Learning algorithms for oncological medical imaging. By comparing SVM, K-NN, and CNN models on 277,524 histopathological breast tissue patches, the CNN model proved statistically superior (p=0.0003), achieving 96.8% accuracy and 97.8% recall. These findings validate CNNs as highly robust, automated decision-support tools that minimize false negatives in breast cancer diagnostics.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide. *CA Cancer J Clin*. 2018;68(6):394-424.
2. Elmore JG, Armstrong K, Lehman CD, Fletcher SW. Screening for breast cancer. *JAMA*. 2005;293(10):1245-56.
3. Choudhury A, Perumalla S. Detecting breast cancer using artificial intelligence: Convolutional neural network. *Technol Health Care*. 2021;29:33-43.
4. Toma TA, *et al.* Breast cancer detection based on simplified deep learning models using histopathological images. *Radio Sci*. 2023.
5. Srikantamurthy MM, *et al.* Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. *BMC Med Imaging*. 2023;23:19.
6. Jiang B, *et al.* Deep learning applications in breast cancer histopathological image analysis: A review. *Breast Cancer Res*. 2024.
7. Simonyan EO, *et al.* Histopathological breast cancer classification using hybrid CNN-based models. *Heliyon*. 2024.
8. Litjens G, Kooi T, Bejnordi BE, *et al.* A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
9. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
10. Islam MT, *et al.* CNN-based deep learning approach for classification of invasive ductal carcinoma. *Health Sci Rep*. 2024.
11. Addo D, *et al.* BCHI-CovNet: A lightweight deep learning model for breast cancer histopathological image classification. *Biocybern Biomed Eng*. 2024.
12. Yan Y, *et al.* DWNAT-Net: Transformer-based deep learning model for breast cancer histopathology image classification. *Comput Biol Med*. 2025.
13. Zhao M, *et al.* HoRFNet: High-order receptive field network for breast cancer histopathological image classification. *Appl Sci*. 2025;15:6085.
14. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.
15. Cruz-Roa A, Basavanhally A, González F, *et al.* Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *Med Image Comput Comput Assist Interv*. 2014;8673:453-460.
16. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. *IEEE Trans Biomed Eng*. 2016;63(7):1455-62.
17. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995; 20:273-97.
18. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*. 1967;13(1):21-27.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.
20. Gonzalez RC, Woods RE. *Digital Image Processing*. 3rd ed. Pearson; 2008.

21. Selvaraju RR, Cogswell M, Das A, *et al.* Grad-CAM: Visual explanations from deep networks. *Int J Comput Vis*. 2020;128:336-59.
22. Powers DMW. Evaluation: From precision, recall and F-measure to ROC. *J Mach Learn Technol*. 2011;2(1):37-63.
23. Cruz-Roa A, Basavanhally A, González F, *et al.* Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *Med Image Comput Comput Assist Interv*. 2014;8673:453-60.
24. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. *IEEE Trans Biomed Eng*. 2016;63(7):1455-62.
25. Choudhury A, Perumalla S. Detecting breast cancer using artificial intelligence: Convolutional neural network. *Technol Health Care*. 2021;29:33-43.
26. Ali R, Balamurali M, Varamini P. Deep learning-based artificial intelligence to investigate targeted nanoparticles' uptake in TNBC cells. *Int J Mol Sci*. 2022;23(24):16070.

Cite this article: Durga P, Rao SV, Padma KS, Anusha L, Padmalatha K. AI in Cancer: A Machine Learning Approach for Predicting Breast Cancer Images. Indian J of Pharmaceutical Education and Research. 2026; IPC Issue:52-9.