IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 10, October 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

© 2024 IJIRCCE | Volume 12, Issue 10, October 2024| DOI: 10.15680/IJIRCCE.2024.1210010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Development of Deep Learning Model for Breast Cancer Detection using Histopathological Datasets

Zion Vas, Hithaishi K N

PG Student, Department of MCA, The National Institute of Engineering, Mysuru, Karnataka, India

PG Student, Department of MCA, The National Institute of Engineering, Mysuru, Karnataka, India

ABSTRACT:

Objectives: This study looks into the use of histopathological images and Convolutional Neural Networks (CNNs) for detection of Breast Cancer (BC). Conventional techniques for diagnosing BC frequently rely on labor-intensive, subjective manual analysis. But in order to eliminate bias and speed up the diagnosis process, this method integrates human expertise with Machine Learning (ML) and Deep Learning (DL) to improve detection accuracy as well as speed.

Methods: The CNN is especially well-suited for the interpretation of histopathological pictures since it is made to mimic the visual processing powers of the human brain and is particularly good at evaluating spatial information.

Findings: The results presented in this research demonstrate how well the CNN method works for BC detection. Convolutional layers are used by the CNN to extract and analyze spatial characteristics from histopathology images. The CNN provides a reliable technique for identifying cancers by improving the depiction of properties of malignant tissue through its capacity to recognize and decipher complicated patterns in these images. The model makes a substantial contribution to histopathological image analysis by increasing the precision and dependability of cancer identification through the use of learnt features and modifiable parameters.

Novelty: To evaluate the model's performance in identifying BC, ROC analysis and confusion matrices were employed as performance evaluation metrics. The outcomes show an 87% accuracy rate, indicating the effectiveness of the CNN-based method for increasing cancer identification using histopathological image analysis.

KEYWORDS: Deep Learning, Breast Cancer Identification, Medical Image Analysis, Histopathology Images, Convolutional Neural Network.

I. INTRODUCTION

BC is thought to be a fairly frequent form of cancer in women that starts in the breast cells. Following lung cancer, breast cancer poses a serious threat to a woman's life. BC is divided into different categories based on how the cell looks under a microscope. The Invasive Ductal Carcinoma (IDC) type is more harmful since it encompasses the entire breast tissue, whereas the DCIS type accounts for only 20% to 53% of instances. About 80% of BC patients fall into this group. [1].

BC can be successfully treated if it is discovered early. Therefore, it's critical to have access to appropriate screening techniques in order to identify BC's earliest symptom. A variety of imaging modalities are employed in the screening process to detect this illness; mammography, ultrasonography, and thermography are the most widely utilized methods. Mammography is one of the most important techniques for early BC detection. Since mammography is ineffective for breasts that are solid, diagnostic sonography or ultrasound techniques are frequently employed. In light of these concerns, thermography may be a more useful method than ultrasound for identifying smaller malignant masses since it can detect radiations that radiography cannot reach small tumors. [2].

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Many techniques have been developed to create and improve image processing since images have many challenges such as low contrast, noise, and lack of visual appreciation. These days, CNN, ML, and AI are the healthcare industry's fastest-growing subsectors. Artificial intelligence (AI) and ML are fields of study that address and enhance technological systems to handle complicated problems by lowering the need for human intelligence.

DL, a member of the Machine Learning family, relied on artificial neural networks. Common applications for deep learning architectures (DL architectures) include computer vision, natural language processing, medical image analysis, and histopathology diagnosis. DL architectures include CNN, CNN, RNN, and CNND. These new technologies can be used to increase the efficiency and diagnostic accuracy of cancer diagnosis, especially DL algorithms.[3].

The following is an outline of the primary goals: The first step is to preprocess the data from histopathological images so that tissue samples can be effectively compared to detect benign and malignant regions. Secondly, to use a CNN model with adjustable parameters to extract features and categorize these photos. Then, a confusion matrix and expanded metrics are used to evaluate performance. Lastly, a comparison between the suggested research and previous studies is done.

II. LITERATURE REVIEW

A number of novel strategies have emerged recently. This is a field that is still developing and has a lot of room for development. Therefore, the goal of our effort is to contribute significantly to the field. BC Diagnosis using DL Algorithm, Histopathological Image Analysis for Cubic SVM-Based BC Detection [4] published in the year 2020 uses the BreakHis dataset. The authors have explored six Support Vector Machines (SVMs) variations in conjunction with Random Forests and K-Nearest Neighbors (KNNs), for experimental purposes. They concluded that cubic SVM performed better than rest of the other methods. However, SVM has trouble predicting class labels in cases where the class size is huge. The BreakHis dataset was used in another method, called A Deep CNN Technique for Detection of BC Using Histopathology Images [5] in which they have used DenseNet networks.

A more recent method was published under the title BC Classification from Histopathological Images in 2021 using Patch-Based DL Modelling [6]. To categorize images, it first applies Logistic Regression using a Deep Belief Network (DBN). Their model's accuracy is only 86%. Nevertheless, it was a fresh technique, and not very reliable.

Also, another paper was published in 2022 titled BC Detection using CNN [7] for BC image analysis with created CNN model made predictions with a recall of 84% was achieved.

Further in 2023, Arslan Khalid, et al., conducted a study focusing on detection of BC in their paper titles BC Detection and Prevention Using ML [8] applied to three different feature selection modules: recursive feature elimination, univariate feature selection, and low-variance feature removal. A sizable EDA dataset of 3002 combined images collected from 1501 people was used to test it. Six distinct classification models were used: Support Vector Classifier (SVC), Logistic Regression (LR), Random Forest (RF), KNN, Decision Tree (DT). were the models utilized for the diagnosis of BC. The simulation results demonstrate how precise and low-power the CNN model is, demonstrating its excellent efficiency.

Our proposed model has employed CNN model to categorize images using the Breast Histopathological dataset with a minimal computation overhead.

III. DATA COLLECTION

The dataset is a valuable resource for research in BC detection using Histopathological Images. The data collection for the BC detection investigation consists of 162 complete slide images of BC cases that were scanned at a 40x magnification. A total of 277,524 images, measuring 50 X 50 pixels each, are taken out of these pictures. Two types of these patches are identified: 78,786 Malignant and 198,738 Benign.

The hierarchical naming approach in the dataset makes it easier to identify the classification of each patch and trace it back to its original location on the entire slide image. This strategy is essential for rigorous data processing and analysis

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

while evaluating and training deep learning models which in turn would help medical practitioners identify BC early on. Obviously, the result would ultimately lead to better patient outcomes and a stronger battle against this deadly illness. Histology dataset used satisfies the need to be a useful tool for practitioners and researchers working in the field of BC detection. Figure 1 and Figure. 2 shows the images of the sample datasets used.

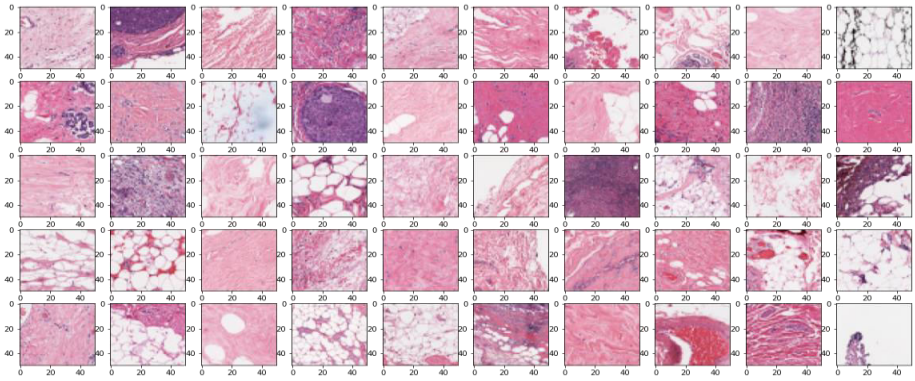

Figure 1: Benign Sample.

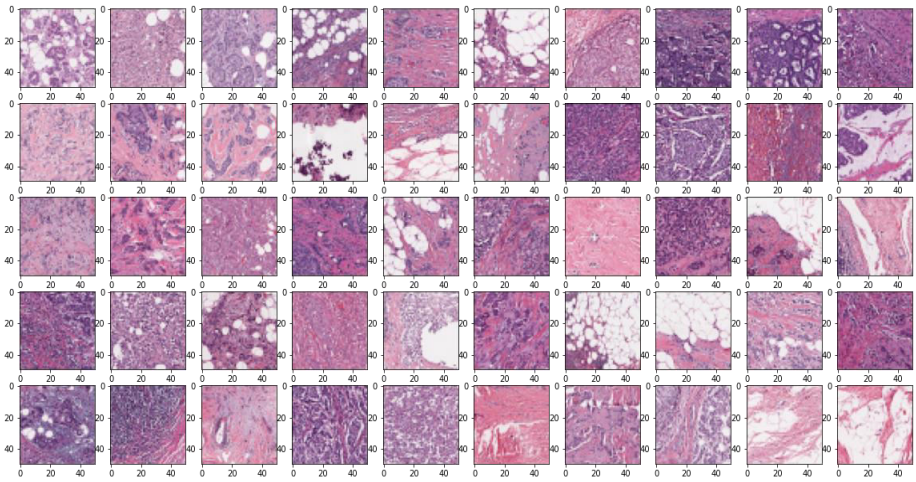

Figure 2: Malignant Sample.

IV. METHODOLOGY

The approach has made use of CNN to examine histopathological patches in order to find IDC. The patches are preprocessed to enhance features relevant for classification, including resizing, normalization, and augmentation to improve model robustness. The CNN architecture is designed to capture spatial hierarchies in the images through convolution layer followed by pooling layer and finally fully connected networks. A labeled dataset is used to train and validate the model. Recall and accuracy are performance indicators that are tracked to make sure the model reaches the intended threshold. Techniques such as Early Stopping, Learning Rate Reduction, and Model Checkpointing are employed to optimize training and prevent overfitting.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.1 Proposed CNN Algorithm

Steps Involved in CNN Algorithm

Step 1: Choose a Dataset

The dataset selected for this project is a group of histological pictures associated with IDC, a prevalent form of BC. The collection is made up of several tiny image patches that were taken from slides of breast tissue. Each image has an indicator of malignant or non-cancerous tissue, such as IDC positive or negative. IDC positive and IDC negative picture subfolders are found in each patient's folder in the dataset's first organization by patient ID.

Step 2: Prepare the Dataset for Training

In To ensure that the dataset is ready for training a DL model, preparation comprises multiple processes. In order to comprehend the distribution of photos among various patients and classifications, the project first examines the dataset's structure. Then, for simpler access and manipulation, it gathers all of the photos into a single directory. The pictures are then arranged into a pandas DataFrame with columns for the patient ID, target label (IDC positive or negative), image path, and image representation for each row. The dataset can be easily manipulated and analyzed with this DataFrame. Since there is an imbalance between the classes (IDC positive and IDC negative), balancing the dataset is a crucial part of preparation. To avoid the model being biased towards the dominant class during training, an equal amount of photos from each classes are sampled

Step 3: Create Training Data and Assign Labels

When the dataset is prepared, it is split into training and validation sets so that the model's performance can be evaluated during training. The directories containing the training and validation sets contain subdirectories for both IDC positive and IDC negative images. This directory structure is necessary for the image data generators of the TensorFlow Keras library to load and preprocess the images. The data generators create batches of images for training and validation, apply different augmentations like flips and rotations, and scale the source photos. The model can learn from the labelled data since it automatically assigns class labels based on the directory structure.

Step 4: Define and Train the CNN Model

The TensorFlow Keras package is used to define the CNN model. Convolutional layers, max-pooling layers, dropout layers, and fully connected dense layers are some of the layers that make up the model. The convolutional layers use convolutional filters to identify features in the images, while the max-pooling layers reduce the spatial dimensions of the feature maps to increase the computational efficiency of the model. In order to avoid overfitting, dropout layers randomly exclude particular neurons during training. The convolutional layers' information is used by the final dense layers to categorize the data. After training with the training data generator, the model is assembled using the binary cross-entropy loss function and Adam optimizer. The model is trained on the validation set, and its output is tracked. To alter the learning rate and maintain the optimal model, callback functions like ReduceLROnPlateau and ModelCheckpoint are utilized. Throughout the first several epochs, the model steadily increases its validation accuracy, with appreciable accuracy spikes. But occasionally, for a few epochs, the validation accuracy does not increase, suggesting that overfitting or learning plateaus may be occurring. Only when the validation accuracy increases does the model save its weights, as shown by notifications like saving model to "model.h5".

Step 5: Test the Model's Accuracy

The accuracy of the model is assessed on the validation set following training. The validation images are predicted by the trained model; the actual labels and the predicted probabilities are compared to determine metrics like accuracy, precision, recall, AUC (Area Under the Curve), and F1 score. These metrics provide a comprehensive evaluation of the functionality of the model. A confusion matrix is also generated in order to evaluate the model's ability to distinguish between photos that have IDC positives and negatives. In the end, a classification report is generated that includes an overview of the model's recall, F1 score, and precision for each class. This comprehensive evaluation helps understand the benefits and drawbacks of the model and makes recommendations for future developments.

4.2 CNN Architecture

The proposed CNN design is shown in Figure 3. This network is composed of several MultiLayer Perceptron (MLP) architectural elements, such as a single output layer, numerous hidden layers, and an input layer with numerous

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

neurons. A layer's neurons are connected to those of every other layer. The ensuing sections address the elements of the suggested architecture.

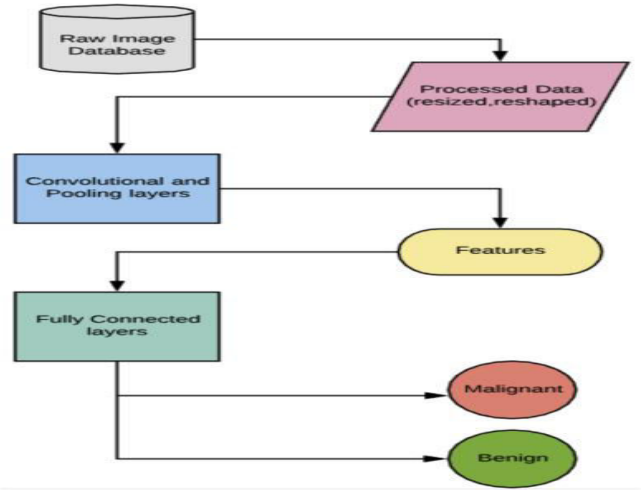

Figure 3: Proposed CNN Architecture.

4.2.1 Input Layer

The input layer is the first step where the model receives images with a size of 50x50 pixels and 3 channels representing the RGB colors. This layer is essential as it defines the shape of the data the model will work with. By setting the input shape as (50, 50, 3), the model knows the dimensionality of each image being fed into it for processing.

4.2.2 Convolution Layer

After convolution, the image is filtered using the convolution layer to improve the features. The convolution layers scan the image and identify patterns using filters or kernels. An activation function, Rectified Linear Unit (ReLU) is used to provide non-linearity which follows each layer in this structure. The data is separated into validation and training (90%) and Testing (10%).

Our proposed model is composed of three Convolutional Blocks. First Convolutional Block comprises of three convolutional layers, each with 32 filters and kernel size of 3x3 followed by a max pooling layer of size 2x2 and dropout with a rate of 0.3. Second Convolutional Block consists of three convolutional layers, each with 64 filters and kernel size of 3x3 followed by a max pooling layer of size 2x2 and dropout with a rate of 0.3. The last Convolutional Block consists of three convolutional layers, each with 128 filters and kernel size of 3x3 followed by a max pooling layer of size 2x2 and dropout with a rate of 0.3. Additionally, the use of padding='same' ensures that the dimensions of the output image remain the same as the input image after convolution, making it easier to handle the spatial data.

4.2.3 Pooling Layer

Pooling layers reduce the amount of processing and parameters in the network., which also minimize the spatial dimensions of the feature maps. They help in making the identification of features invariant to small translations of the input, thus enhancing the model's generalization. In the first Convolutional Block, MaxPooling2D with a pool size of (2, 2) is applied after the three convolutional layers with 32 filters. This layer reduces the spatial dimensions of the feature maps by a factor of 2. In the Second Convolutional Block, MaxPooling2D with a pool size of (2, 2) is applied after the three convolutional layers with 64 filters. Again, this reduces the spatial dimensions of the feature maps by a factor of 2. In the third Convolutional Block, MaxPooling2D with a pool size of (2, 2) is applied after the three convolutional layers with 128 filters. This reduces the spatial dimensions of the feature maps by a factor of 2. The pooling layers thus reduce the spatial dimensions by half, which helps in making the model more efficient and less prone to overfitting. By using max pooling, one can ensure that the most prominent features are preserved, contributing to the robustness and accuracy of the model.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.2.4 Flattening Layer and Fully Connected Layers

Flattening is a process of converting the 3D feature maps obtained from the convolutional and pooling layers into a 1D vector. This 1D vector is then fed into the fully connected (dense) layers. Purpose of the flatten layer is dimensionality reduction and compatibility. Every neuron in a dense layer is coupled to every other neuron in the layer before it. These layers are used for high-level reasoning.

4.2.5 Output Layer and Dropout Layers

The neural network model's output layer is the last layer. It generates the final predictions necessary to complete the task of binary classification for BC detection in our approach. The SoftMax function converts the logits into probabilities that add up to 1. Each output neuron's value is interpreted as the likelihood that the input is a member of the class.

Dropout is a regularization strategy that involves randomly removing neurons from neural networks during training in order to minimize overfitting. The purpose of dropout layer is to prevent overfitting and create redundancy.

4.2.6 Optimization Algorithm

The optimization algorithm Adam is employed to revamp the model's weights during training to bring down the loss function.

4.2.7 Evaluation Metrics and Training

The effectiveness of the model is assessed using metrics such as recall, precision, accuracy, ROC-AUC, and F1 score.

4.2.8 Hyperparameter Tuning

To maximize model performance, hyperparameters like learning rate, batch size, and network architecture are adjusted. Regular validation followed by testing are crucial to assess the model's efficiency and generalization on unseen data. Further, adjustments may be required based on results and feedback from medical experts.

4.3 Metrics to evaluate the Performance

Beyond accuracy, assessing a model's performance in binary classification with several measures provides a thorough understanding. Correctly predicted positive instances are referred to as True Positives (TP). Cases that were mistakenly predicted as positive are known as False Positives (FP) and True Negative (TN) is the case that were accurately anticipated to be negative, whereas False Negatives (FN) are the situations that were wrongly forecasted to be negative.

4.3.1 Accuracy

The ratio of positively and negatively anticipated instances to all instances is known as accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

4.3.2 Precision

The ratio of accurately anticipated positive cases to all predicted positives is called precision, or positive predictive value.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

4.3.3 Recall

It is the ratio of accurately predicted positive cases to all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

4.3.4 F1 – Score

The harmonic mean of recall and precision is known as the F1 Score. It offers a balance between recall and precision, particularly when there is a disagreement between the two measurements.

$$F1 - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (4)$$

4.3.5 Support

It is the total count of actual instances for every class in the dataset. All that is counted is the number of instances for each class.

V. RESULTS AND DISCUSSIONS

In this study, the Histopathological dataset's BC detection was achieved using the CNN model. Initially, kaggle.com was the source we gathered. For the simulation, we utilized Python Jupiter Notebook 3.8.10. Tensorflow, Pandas, Seaborn, Plotly, and other Python extensions are among the resources we have used. TensorFlow makes it easier to build, train, and implement CNNs. Pandas is a feature-rich data analysis package that provides DataFrames and Series. Plotly is a versatile framework for making interactive visualizations including dashboards and 3D plots, whereas Seaborn is a high-level interface for statistical visualization creation. 90% of the training data and 10% of the test data are included in each batch. The next subsections contain the Confusion Matrix and additional performance indicators for the proposed framework.

5.1. Graphical representation and Confusion Matrix

In order to understand the patterns that distinguish malignant tissues from non-cancerous ones, the CNN is trained on labelled histopathological patches. The training loss (hinted by dots) indicates how well the model performed using the training data or the images it has visualized. The loss indicates that as training progresses, the model gets more proficient at correctly classifying the photos.

The validation loss (shown by the blue line) assesses the model's performance to ensure that it learns from both training and unseen data. The validation loss likewise decreases with time, but it varies more than the training loss. This variability shows how difficult it is to anticipate using fresh data, which is essential for a strong model. The model appears to be learning to identify IDC with increasing accuracy and lowering errors, which is important for early and accurate BC detection. This is indicated by a continually decreasing validation loss along with training loss. Figure 4 shows the training and validation loss curves during the process of BC detection.

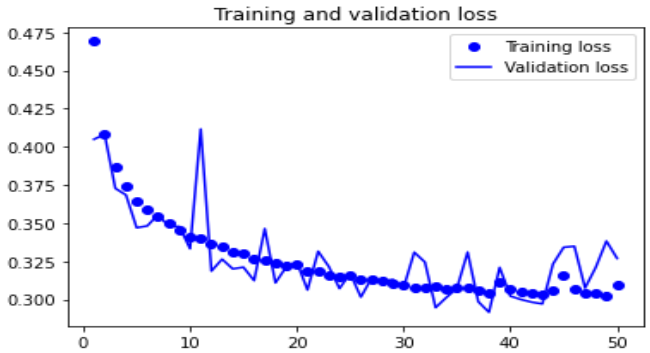

Figure 4: Graphical Representation of Cascaded Training with Validation Loss

The training and validation accuracy of a CNN model used to detect BC across several epochs is shown in Figure 5. Effective learning is demonstrated by the training accuracy, which is perceived as a continuous improvement as the model learns from the data and is depicted by the with markers. The validation accuracy, shown by the solid line, varies significantly, indicating that although the model performs better on training data, it performs less consistently on

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

unseen data. These fluctuations in validation accuracy could indicate overfitting, a circumstance where the model is too closely tailored to the training set and hence finds it difficult to generalize to new samples. This suggests that while the model may work well for BC recognition during training, its ability to identify cancer accurately in new histopathological pictures may be compromised. To ensure that the model remains reliable and efficient in practical applications, addressing this would necessitate the use of strategies like regularization or early halting.

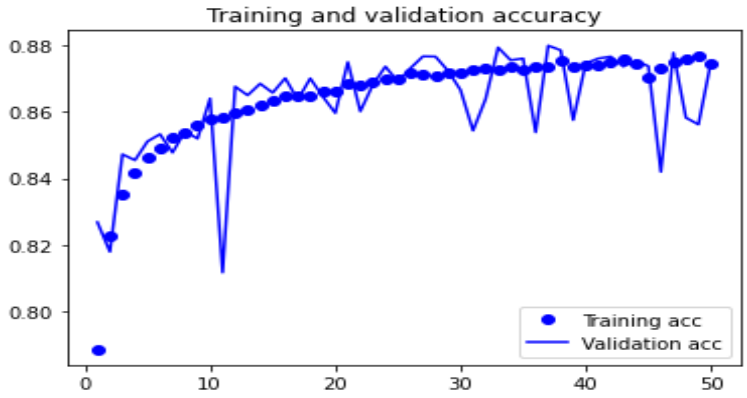

Figure 5: Graphical Representation of Cascaded Training with Validation of Accuracy.

One of the most important tools for assessing how well a CNN model detects BC is the Confusion Matrix that is shown in Figure 6. It provides an overview of the model's predictions for each of the four important metrics: FPs, FNs, TPs, and TNs. The situations where the model successfully recognized BC in histological images are called TP (7036), and the cases where the model correctly decided that no cancer was present are called TN (6746). The model appears to be able to forecast circumstances accurately in the case of absence of cancer based on these high values.

Additionally, areas of uncertainty in the model's performance are highlighted by the Confusion Matrix. The FPs (1,433) represent instances in which the model misdiagnosed BC, potentially causing patients to undergo needless stress and medical treatments. More importantly, there's a chance that delayed diagnosis and treatment could result from the FNs (843), which represent cases in which the algorithm was unable to identify BC. While the overall performance of the model is good, FNs in particular highlight that additional improvement is needed to ensure more reliable and accurate BC diagnosis while lowering the possibility of missing any instances.

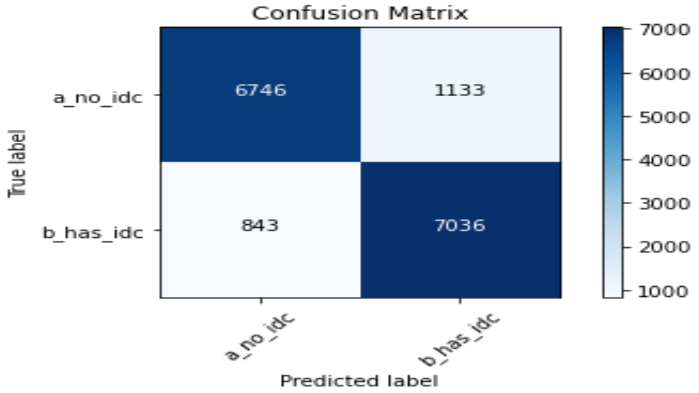

Figure 6: Confusion Matrix.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

A consolidated assessment of the classification model's performance represented in the confusion matrix in different quadrants is organized and presented in Table 1.

Table 1. Structure of prediction values for CNN

	Predicted Non match	**Predicted Match**
Actual Non match	TN	FP
Actual Match	FN	TP

The results findings are effectively presented by detailed numerical data in Table 2. The model's total accuracy, or the percentage of correctly categorized histopathological images, is 87%. The model achieved 89% precision for benign instances and 86% precision for malignant. The precision metrics are essential for comprehending the dependability of the model's classifications for every kind of tumour since they indicate the frequency with which the model's positive predictions come true.

The recall metrics, which are further divided into groups categorized as benign and malignant, demonstrate the model's capacity to distinguish true positive cases from all potential cases. The model demonstrated a higher 89% recall for malignant tumours and an 86% recall for benign tumours. This means that even if the model is good at spotting malignant tumours, it may still do a better job of spotting benign cases.

For benign tumours, the F1 Score, which combines precision and recall into a single score, is 87%; for malignant tumours, it is slightly higher at 88%. The CNN does a good job at differentiating between benign and malignant tumours, but there is still room for improvement, as seen by this score, which offers a fair assessment of the model's accuracy and dependability.

Table 2. Tabulation of Performance Indicators

Dataset Name	**Performance Measures**		**Results in %**
Histopathology Images	Accuracy		87%
	Precision	Benign	89%
		Malignant	86%
	Recall	Benign	86%
		Malignant	89%
	F1 Score	Benign	87%
		Malignant	88%

VI. COMPARATIVE ANALYSIS WITH EXISTING WORK

It is imperative that a comparison analysis be used to evaluate the performance. A comparison between the suggested method and the previous works is shown in the Table 3.

Several researches have demonstrated many breakthroughs in the field of BC diagnosis utilizing DL and ML techniques. Singh and Kumar [4] made one of the first contributions to this field when they investigated a number of machine learning classifiers, such as KNN, RF, and six different SVM variants. They found that cubic SVM performed better than other methods with an accuracy of 92.3% using the BreakHis dataset. SVM's scalability for practical applications was constrained by its inability to cope with huge class sizes.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Using a histology dataset, Wadhwa and Kaur [5] used a Deep CNN model in 2020 and achieved an amazing accuracy of 95.58%. This model is unique in that it can handle complicated image data better than other ML algorithms. In a similar vein, Hirra et al. [6] classified images using a Deep Belief Network (DBN). Although their method was unique, its accuracy was only 86%, which indicates that although DL approaches show promise, more tuning is required for improved reliability.

Ereken and Tarhan [7] used a CNN model for BC detection in another investigation, and while the recall metric suggested space for improvement, the model was able to identify real positive instances, as seen by the study's 84% recall rate.

By applying three feature selection strategies and evaluating six classification models on the EDA dataset, Khalid et al. [8] built on this work in 2023. The accuracy ranged from 87% to 96.49% according to their findings, with CNN-based methods being some of the most effective models examined.

The suggested CNN model in our study attained an accuracy of 87% when utilizing the breast histology dataset. These findings demonstrate the ongoing development of CNN-based methods, where outcomes are greatly influenced by model improvements and dataset quality.

Table 3. A comparative study of existing and proposed approach.

Citation	Dataset	Method	Results
S. Singh and R. Kumar [4]	Breakhis Dataset	KNN, RF and Six flavors of SVM	92.3%
G. Wadhwa and A. Kaur [5]	Histopathology Dataset	Deep CNN	95.58% Accuracy
I. Hirra et al. [6]	Histopathology Dataset	DBN	86% accuracy
Ö. F. Ereken and C. Tarhan [7]	Breast Cancer Dataset	CNN	84% Recall
Khalid A, Mehmood [8]	EDA Dataset	RF, DT, KNN, LR, SVC, Linear SVC	87% to 96.49%
Proposed CNN Model	Histopathology dataset	CNN model	87% Accuracy

VII. CONCLUSION

The acceptable recall values and encouraging results, the current BC detection model—which makes use of CNN and DL techniques demonstrates its efficacy in identifying true positives. But in order to reduce false negatives in real-world applications, a recall exceeding 0.85 is essential. Recall can be further improved by implementing advanced architectures, hyperparameter tuning, and data augmentation. By automating the first tissue slide screening process, this approach can significantly assist pathologists by decreasing the amount of time and manual labor needed. Because of its excellent recall and accuracy, cancer can be detected more quickly and consistently, freeing up pathologists to focus on more difficult cases and maximizing resource use. In the end, applying this model to workflows for diagnosis can help with early treatment, which improves patient outcomes.

In this work, publicly available histopathological imaging datasets are used to train CNN-based algorithms for BC prediction. 277,524 of size 50x50 pixel microscopic pictures total of 198,738 Benign and 78,786 Malignant—are included in the dataset. The dataset is divided into 90% and 10% portions for testing, validation, and training, respectively. By increasing the diversity of the training datasets, data augmentation approaches help the CNN model become more robust and more broadly applicable. Several criteria are used to assess the model's performance; the

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.625| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

findings indicate that it performed well, achieving 87% accuracy in binary classification. This has the potential to improve patient survival rates by greatly assisting in the early diagnosis of BC.

In the future, combining cutting-edge technologies with more predictive tools may improve the detection accuracy. Furthermore, for a more thorough identification, variables including cancer etiology, contaminated foods, and symptom analysis could be used.

REFERENCES

1.Masud M., Eldin Rashed A. E., and Hossain M. S., Convolutional neural network-based models for diagnosis of breast cancer, Neural Computing and Applications. (2020) 5, https://doi.org/10.1007/s00521-020-05394-5.
2.Muhammad G., Hossain M. S., and Kumar N., EEG-based pathology detection for home health monitoring, IEEE Journal on Selected Areas in Communications. (2021) 39, no. 2, 603–610, https://doi.org/10.1109/jsac.2020.3020654.
3.J. Zhang, X. Guo, B. Wang and W. Cui, "Automatic Detection of Invasive Ductal Carcinoma Based on the Fusion of Multi-Scale Residual Convolutional Neural Network and SVM," in IEEE Access, vol. 9, pp. 40308-40317, 2021, doi: 10.1109/ACCESS.2021.3063803.
4.S. Singh and R. Kumar, "Histopathological Image Analysis for Breast Cancer Detection Using Cubic SVM," 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2020, pp. 498-503, doi: 10.1109/SPIN48934.2020.9071218.
5.G. Wadhwa and A. Kaur, "A Deep CNN Technique for Detection of Breast Cancer Using Histopathology Images," 2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), Cochin, India, 2020, pp. 179-185, doi: 10.1109/ACCTHPA49271.2020.9213192.
6.I. Hirra et al., "Breast Cancer Classification From Histopathological Images Using Patch-Based Deep Learning Modeling," in IEEE Access, vol. 9, pp. 24273-24287, 2021, doi: 10.1109/ACCESS.2021.3056516.
7.Ö. F. Ereken and C. Tarhan, "Breast Cancer Detection using Convolutional Neural Networks," 2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2022, pp. 597-601, doi: 10.1109/ISMSIT56059.2022.9932694.
8.Khalid A, Mehmood A, Alabrah A, Alkhamees BF, Amin F, AlSalman H, Choi GS. Breast Cancer Detection and Prevention Using Machine Learning. Diagnostics (Basel). 2023 Oct 2;13(19):3113. doi: 10.3390/diagnostics13193113. PMID: 37835856; PMCID: PMC10572157.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details