Development and Validation of a Convolutional Neural Network Model for Early Detection of Invasive Ductal Carcinoma in Histopathological Images

Yawo Ezunkpe, Ankith Indra Kumar

Submitted to: JMIR Cancer
on: June 06, 2024

Table of Contents

Development and Validation of a Convolutional Neural Network Model for Early Detection of Invasive Ductal Carcinoma in Histopathological Images

Yawo Ezunkpe[1] PhD; Ankith Indra Kumar[1] MS

[1]San Jose State University San Jose US

Corresponding Author:
Yawo Ezunkpe PhD
San Jose State University
One Washington Square
San Jose
US

Abstract

Breast cancer is a major health concern for women worldwide and poses a significant challenge to healthcare systems. According to the World Health Organization(WHO), 2.3 million women were diagnosed with breast cancer in 2020, resulting in 685,000 deaths. Invasive Ductal Carcinoma (IDC) accounts for 80% of these cases, making it crucial to accurately diagnose it in a timely manner. Traditional breast cancer detection methods rely on diagnostic imaging techniques like mammography, ultrasound, and MRI, which are interpreted by trained radiologists. However, these methods' accuracy depends on the radiologist's experience and can be subjective, leading to variability in diagnosis. False positives and negatives are not uncommon, which can result in missed cancers or unnecessary biopsies. This paper proposes a Deep Convolutional Neural Network (DCNN) based model to detect IDC in histopathological images of breast tissue. The model uses a robust dataset and fine-tuned hyperparameters, highlighting the potential of deep learning in improving diagnostic accuracy in oncology. The model achieved an 87% accuracy on the test set.

(JMIR Preprints 06/06/2024:62996)
DOI: https://doi.org/10.2196/preprints.62996

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Development and Validation of a Convolutional Neural Network Model for Early Detection of Invasive Ductal Carcinoma in Histopathological Images

Ankith Indra Kumar, *Master's student in Computer Sciences, San Jose State University, CA,* and Yawo S. Ezunkpe, *Assistant Professor of Aerospace Engineering San Jose State University, CA*

Abstract—Breast cancer is a major health concern for women worldwide and poses a significant challenge to healthcare systems. According to the World Health Organization(WHO), 2.3 million women were diagnosed with breast cancer in 2020, resulting in 685,000 deaths. Invasive Ductal Carcinoma (IDC) accounts for 80% of these cases, making it crucial to accurately diagnose it in a timely manner. Traditional breast cancer detection methods rely on diagnostic imaging techniques like mammography, ultrasound, and MRI, which are interpreted by trained radiologists. However, these methods' accuracy depends on the radiologist's experience and can be subjective, leading to variability in diagnosis. False positives and negatives are not uncommon, which can result in missed cancers or unnecessary biopsies. This paper proposes a Deep Convolutional Neural Network (DCNN) based model to detect IDC in histopathological images of breast tissue. The model uses a robust dataset and fine-tuned hyperparameters, highlighting the potential of deep learning in improving diagnostic accuracy in oncology. The model achieved an 87% accuracy on the test set.

Index Terms—Convolutional Neural Networks, Invasive Ductal Carcinoma, Breast Cancer Detection, Machine Learning, Medical
Imaging

I. INTRODUCTION

Breast cancer continues to be the most common form of cancer among women all over the world, exerting a significant impact on healthcare systems and posing severe health risks. According to WHO, in 2020, 2.3 million women were diagnosed with breast cancer, which resulted in 685,000 fatalities. Invasive Ductal Carcinoma (IDC) accounted for 80% of these cases. Therefore, it is crucial to diagnose this condition accurately and promptly [1].

Breast cancer detection has traditionally relied on diagnostic imaging techniques such as mammography, ultrasound, and MRI. These techniques are interpreted by trained radiologists. However, the accuracy of these methods is highly dependent on the radiologist's experience and can be subjective, leading to variability in diagnosis [2]. False positives and negatives are not uncommon, which can result in unnecessary biopsies, missed cancers,.. etc.

Detecting breast cancer using these classical methods poses several difficulties. Key limitations include:

1) Inter-observer Variability: Different radiologists may interpret the same images differently, leading to inconsistencies in diagnosis [3].
2) Low Sensitivity in Dense Breast Tissue: Classical imaging techniques often have reduced sensitivity in dense breast tissue, which is common in younger women [4].
3) Time-Consuming: Manual review of mammograms is time-consuming and can lead to delays in diagnosis [3].
4) Limitation in Early Detection: These methods may not always detect tumors at an early stage, reducing the chances for successful treatment [4].

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that has the potential to transform breast cancer detection by using tools that can learn from data, recognize patterns, and assist humans in making decisions more efficiently and quickly. ML can help overcome many of the limitations of the classical methods and can be used to address them effectively:

1) Consistency: ML algorithms can provide consistent interpretations of imaging data, reducing the variability associated with human readers [5].
2) Enhanced Detection in Dense Tissue: ML algorithms can be trained to identify cancers in dense breast tissue more effectively than traditional methods [5].
3) Efficiency: ML can quickly analyze large volumes of imaging data, significantly speeding up the diagnostic process [5]
4) Early Detection: ML techniques, particularly deep learning, can detect subtle changes in tissues that may indicate early-

stage cancers [5].

Recent research in breast cancer detection has focused on developing advanced ML models, such as convolutional neural networks (CNNs), to improve the accuracy of breast cancer detection. Studies [6], and [7] have shown that CNNs can outperform classical methods by accurately classifying benign and malignant lesions. These models are trained using large datasets of annotated images, allowing them to learn complex features that are indicative of cancer.

Manjunathan et al.'s investigation into machine learning applications for breast cancer detection highlighted the performance of random forests, which achieved an accuracy of 96.5% on the Wisconsin Breast Cancer Dataset which contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. This research contributes to the field by underlining the potential of machine learning models to improve the accuracy of breast cancer detection, especially in the early stages that are typically challenging to identify with conventional methods like mammography [8].

Ahmed et al. focused on the potential of machine learning for breast cancer risk prediction using various algorithms on the Wisconsin Breast Cancer (Original) dataset. In their extensive comparative analysis, they found that SVMs achieved a marginally higher accuracy of 97.07%, with random forests and naive Bayes both yielding 97%. This finding is particularly relevant for the development of predictive modeling in breast cancer risk assessment and highlights the critical role of algorithm selection in enhancing predictive performance [9].

Deep learning techniques have the capability to autonomously identify distinctive features within extensive datasets of images, thereby eliminating the need for manually extracting features as seen in conventional machine learning approaches. This advancement has significantly enhanced the implementation of deep learning within the domain of medical image analysis [5]. In particular, Convolutional Neural Networks (CNNs) have demonstrated remarkable efficacy in analyzing medical imagery, as documented in several studies [7], [10]–[12].

A. Melek et al. provided an innovative perspective by employing deep learning for spatiotemporal breast cancer risk prediction. Their study introduced a Siamese neural network architecture that processed consecutive mammogram images, which addressed the limitations of classical risk models. The results indicated that this approach could significantly improve risk prediction, achieving an Area Under Curve (AUC) which is the measure of the ability of a binary classifier to distinguish between classes of 0.81 and outperforming single time-point CNN models. This advancement suggests a promising direction for personalized breast cancer screening and underscores the potential of deep learning applications in a clinical setting [13].

Our study advances the field by introducing a Convolutional Neural Network (CNN) architecture specifically designed to analyze histopathological image patches for the detection of Invasive Ductal Carcinoma (IDC). This targeted approach, focusing on localized image patches, marks a departure from traditional methods of examining entire histopathological slides and addresses key challenges in breast cancer detection, such as the high variability of tumor appearance and the vast size of histopathological images, which can make comprehensive analysis computationally intensive and potentially less sensitive:

1) Our model is trained on an extensive dataset comprising 277,524 image patches with equal representation of IDC-positive and negative cases. This balanced dataset composition enables the CNN to learn a broad spectrum of subtle discriminative features that are indicative of malignancy, enhancing both the sensitivity and specificity of cancer detection.
2) We innovate in CNN design by utilizing a consistent filter count across all convolutional layers, a shift from the common practice of increasing filter complexity in successive layers. Preliminary experiments indicated that typical filter augmentation resulted in significant information loss when analyzing small (50x50 pixel) image patches. Our approach aims to preserve the informational

content of these patches, which is crucial for detecting the nuanced patterns of IDC.

3) Additionally, we address a limitation inherent to the previously utilized Wisconsin Breast Cancer Dataset, where information loss from the abstraction of tumor characteristics into numerical features could potentially obscure crucial diagnostic details. In contrast, our model directly processes raw image data, maintaining the integrity of spatial relationships and morphological details, essential for accurate and robust cancer diagnosis.

Our methodology embodies a significant leap forward in computational pathology, emphasizing the necessity for models that align closely with the intricate and detailed nature of histopathological analysis. This approach not only raises the bar for diagnostic accuracy but also contributes to a deeper understanding of IDC characteristics, potentially opening new avenues for precision medicine.

II. DATA ACQUISITION AND PREPROCESSING

A. Data Collection and Labeling

The dataset employed in this study originates from a comprehensive collection of histopathological images designed to facilitate the detection of Invasive Ductal Carcinoma (IDC), the most prevalent subtype of breast cancer. This dataset, publicly available on Kaggle [14], consists of 277,524 patches of size 50x50 pixels extracted from 162 whole mount slide images of Breast Cancer (BCa) specimens. These specimens were scanned at a magnification of 40x, aiming to provide a detailed view for accurate analysis. The data encompass both IDC negative (198,738 patches) and IDC positive (78,786 patches) examples, offering a balanced perspective for training machine learning models.

Each image patch is uniquely identified by a filename following the format: *u xX yY classC.png*, where *u* denotes the patient ID, *X* and *Y* represent the coordinates from which the patch was cropped within the whole mount slide, and *C* indicates the class label, with 0 for non-IDC and 1 for IDC presence. This meticulous labeling process ensures the creation of a dataset that accurately reflects the variances and characteristics of IDC, thus providing a solid foundation for developing automated diagnostic tools.

The original images and dataset were curated with the intent of delineating the precise regions of IDC for aiding pathologists in assigning an aggressiveness grade to breast cancer samples. Such a focus is crucial, as IDC's identification and grading are paramount steps in determining the appropriate course of treatment for affected patients. The dataset is derived from a larger collection of breast cancer specimen images, originally hosted at the URL provided in the Kaggle dataset description [14]. This collection's significance is further underscored by its use in academic research, as cited in notable publications within the fields of medical imaging and diagnostics [3], [4].

B. Preprocessing

The preprocessing pipeline was meticulously crafted to engineer a dataset conducive to effective model training and evaluation. A stratified directory structure was established, segregating images into training, validation, and testing subsets. This hierarchical arrangement facilitated a streamlined flow for data processing and access. The dataset, composed of Invasive Ductal Carcinoma (IDC) histopathological image patches of 50x50 pixel resolution, underwent a comprehensive shuffling and allocation process, guided by a predefined train-test split ratio.

III. MODEL ARCHITECTURE

The model architecture constitute of an input layer that accepts images of size 50x50 with three channels. The data augmentation layer, integrated into the model, employs random horizontal flips, rotations, and zooms to enrich the training dataset and bolster the model's generalization ability. Subsequently, a series of four convolutional blocks—each constituting a convolutional layer with 256 filters, batch normalization, and ReLU activation, followed by a max-pooling layer—systematically extracts features while reducing spatial dimensions. A

dropout layer is introduced post-flattening to mitigate overfitting. The architecture culminates with a dense network, cascading from 512 to 32 neurons, and a final sigmoid activation function for binary classification.

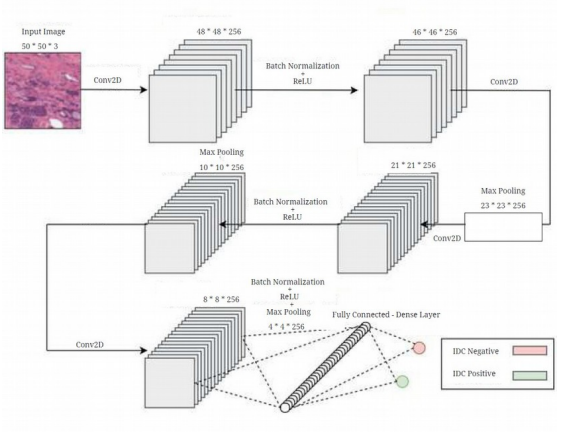


Fig. 1: model architecture.

IV. HYPERPARAMETERS AND FUNCTIONS

Hyperparameters are critical configurations that govern the training process and architecture of neural networks. In this study, several hyperparameters were meticulously chosen to optimize the performance of the custom convolutional neural network for breast cancer detection.

A. Optimizer

The chosen optimizer for the task is RMSprop, a gradientbased optimization technique. The RMSprop algorithm updates the weights by considering the moving average of the squared gradients to adjust the learning path. The mathematical formulation of RMSprop is shown below:

$$w_{t+1} = w_t + \left(\eta\sqrt{\nu_t + \varepsilon}\right) \cdot g_t \quad (1)$$

where w represents the weights, η is the learning rate, ν_t is the moving average of squared gradients, ε is a small scalar to prevent division by zero, and g_t is the gradient at time t.

B. Loss Function

The loss function utilized is binary cross-entropy, which is apt for binary classification tasks. It quantifies the difference between the true labels and the predicted probabilities. The binary cross-entropy loss for a set of predictions p and true outcomes y is depicted in the equation below:

$$L(y,p) = -\frac{1}{N}\sum_{i=1}[y_i \cdot \log\log(p_i) + (1-y_i) \cdot \log\log(1-p_i)] \quad (2)$$

In this formula, N is the number of observations, y_i represents the true label of the i^{th} observation, p_i is the predicted probability of the i^{th} observation, and the summation runs over all N observations. This loss function imposes a higher penalty for predictions that are confident and wrong—those far from the actual label, hence driving the model to make more accurate predictions over time.

C. Callback function

Callback functions serve as checkpoints and monitoring mechanisms during model training. In the current implementation, the ModelCheckpoint callback is employed to persist the model exhibiting the minimal validation loss, thus capturing the most optimal state of the model during training. An EarlyStopping callback is provisionally included but commented out, which, if activated, would cease training when the validation loss ceases to diminish, precluding unnecessary computations and potential overfitting.

D. Learning Rate

The learning rate which is an integral hyperparameter of the RMSprop optimizer, determines the step size at each iteration while moving toward a minimum of the loss function. A rate of η is typically set between 0.001 and 0.0001 in RMSprop to ensure that the model learns at an optimal pace, neither too slow (prolonging training) nor too fast (risking overshooting the minimum). The chosen rate balances efficiency and the risk of converging to local minima.

E. Batch Size

Batch size, another pivotal hyperparameter, influences the model's updating frequency and convergence stability. A size of 32 was selected, compromising the computational efficiency of larger batches and the fine-grained update of smaller batches. This size allows for a moderate estimation of the gradient while utilizing GPU resources effectively.

F. Dropout Rate

The dropout rate controls the proportion of neurons excluded from training at each update during a training phase and is set to 0.5, meaning half of the neurons in the layer will be randomly ignored. This prevents co-adaptation of neurons, effectively acting as a form of regularization to avoid overfitting and to promote the development of more robust neural representations.

G. Number of Filters

In the convolutional layers, 256 filters were employed, deviating from the common practice of incrementally increasing the filter count in deeper layers. This design was chosen due to the small input image size (50x50 pixels), which could lead to rapid information loss during dimensionality reduction. Starting with a high filter count helps to preserve critical features early in the network, essential for capturing subtle details in histopathological images.

H. Activation Function

The activation function plays a critical role in a neural network's ability to capture non-linear relationships. In this study, the Rectified Linear Unit (ReLU) function is employed following each convolutional layer to introduce non-linearity into the model, allowing it to learn complex patterns from the data. ReLU is defined mathematically as $f(x) = \max(0,x)$, which has been shown to speed up training without significantly reducing the network's capacity for data representation.

For the output layer, the sigmoid activation function is utilized due to its property of bounding the output between 0 and 1, making it ideal for binary classification tasks. The sigmoid function is given by

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

which maps any real-valued number into the range of 0 to 1, effectively modeling the probability that an input belongs to the positive class.

I. Image Size

The fixed input image size of 50x50 pixels is dictated by the dataset's native resolution. This dimensionality ensures that the model directly learns from the data without any additional scaling, preserving the original pathological features which are crucial for accurate diagnosis.

V. RESULTS

A. Training and Validation Performance

The learning dynamics of the proposed Convolutional Neural Network model are illustrated in Figure 2, where two graphs depict the training and validation accuracy and loss over epochs.

The graph titled "Training and validation accuracy" showcases a common phenomenon in machine learning models where the training accuracy surpasses the validation accuracy. Such a discrepancy typically implies that the model is overfitting to the training data, capturing patterns that may not generalize well to unseen data. The training accuracy is considerably high, peaking at around 88%, which is promising. However, the validation accuracy exhibits significant fluctuations, indicating that the model's performance on the validation set lacks stability. This instability could stem from several factors, such as inadequate regularization, a need for a more sophisticated data augmentation strategy, or possibly the requirement of a more intricate model to encapsulate the underlying data patterns.

On the other hand, the "Training and validation loss" graph displays volatility, with the training and validation loss demonstrating sharp ascents and descents. Ideally, a decrease in loss should be consistent as the model learns. The erratic nature

of the loss graph indicates a tumultuous learning process, potentially caused by an excessively high learning rate or the data being not entirely representative or sufficiently informative.

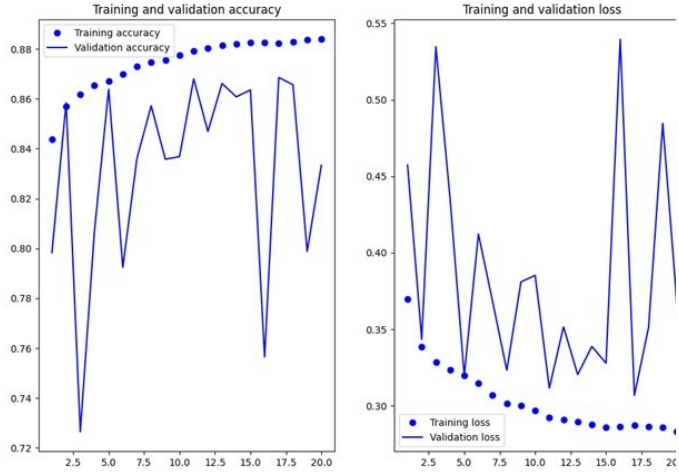


Fig. 2: Training and validation accuracy (left) and loss (right) over 20 epochs.

B. Confusion Matrix Analysis

The performance of the model can also be evaluated using a confusion matrix, which provides a visual representation of the model's predictive capabilities. The confusion matrix for our model is shown in Figure 3, and it delineates the counts of true positives, false positives, true negatives, and false negatives.

- True Negatives (Top-Left): The model correctly predicted the non-presence of breast cancer 38,267 times.
- False Positives (Top-Right): The model incorrectly predicted breast cancer 1,481 times when it was not present.
- False Negatives (Bottom-Left): The model failed to detect breast cancer 5,632 times when it was present, which could potentially lead to patients not receiving the necessary treatment.
- True Positives (Bottom-Right): The model correctly identified breast cancer 10,126 times.

The confusion matrix, presented in Figure 3, indicates that while the model has a commendable number of true positives, the relatively high number of false negatives is concerning. This aspect is particularly critical in the context of breast cancer detection, as minimizing missed diagnoses is paramount. Potential measures to address this issue include adjusting the decision threshold, enhancing the training data quality, or experimenting with different model architectures.

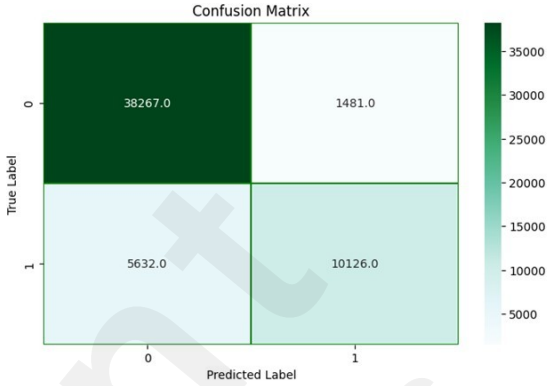


Fig. 3: Confusion matrix of the model's predictions.

VI. CONCLUSION

This investigation has substantiated the efficacy of a CNN with a uniform filter configuration in the precise detection of IDC in histopathological images. The model's 87 percent accuracy in test data evaluation is commendable, especially considering the diminutive image patch size and the challenge of training the model from scratch. Our findings reinforce the hypothesis that information preservation in the initial layers of a CNN is crucial when working with small input images. The consistent application of 256 filters across layers has proven to be a significant factor in enhancing model stability and accuracy. These insights could inform future research and model development strategies in digital pathology. In conclusion, the application of CNNs in medical image analysis, as demonstrated in this study, holds transformative potential for breast cancer diagnostics. The results underscore the viability of machine learning models to support pathologists, potentially leading to improved diagnostic outcomes and patient prognosis. As computational power and algorithmic innovations continue to advance, the integration of such models into clinical workflows is not merely aspirational but increasingly feasible,

marking a significant stride towards the confluence of artificial intelligence and healthcare.

REFERENCES

[1] World Health Organization, "Breast cancer," https://who.int/newsroom/fact-sheets/detail/breast-cancer, 2020, [Online; accessed 10February-2024].

[2] M. S. Reza and J. Ma, "Imbalanced histopathological breast cancer image classification with convolutional neural network," *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 619–624, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:67875563

[3] A. Cruz-Roa, A. Basavanhally, F. Gonzalez, H. Gilmore, M. Feldman,´ S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent," *Scientific Reports*, vol. 7, no. 46450, 2017.

[4] A. Cruz-Roa *et al.*, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *SPIE Medical Imaging*, pp. 904103–904103–8, 2014.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sanchez, "A survey on deep learning in medical image analysis,"´ *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[6] A. Cruz-Roa, A. Basavanhally, F. Gonzalez, H. Gilmore, M. Feldman,´ S. Ganesan, N. Shih, J. Tomaszewski, and A. Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 9041, 02 2014.

[7] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2560–2567.

[8] N. Manjunathan, N. Gomathi, and S. Muthulingam, "Early detection of breast cancer using machine learning," in *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, 2023, pp. 165– 169.

[9] M. Ahmed, M. A. Ali, J. Roy, S. Ahmed, and N. Ahmed, "Breast cancer risk prediction based on six machine learning algorithms," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2020, pp. 1–5.

[10] N. Bayramoglu, J. Kannala, and J. Heikkila, "Deep learning for magnification independent breast cancer histopathology image classification," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2440– 2445.

[11] Q. Li and W. Li, "Using deep learning for breast cancer diagnosis," Chinese Univ., Hong Kong, Tech. Rep., 2017.

[12] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li, "Breast cancer multiclassification from histopathological images with structured deep learning model," *Sci. Rep.*, vol. 7, no. 1, p. 4172, Dec. 2017.

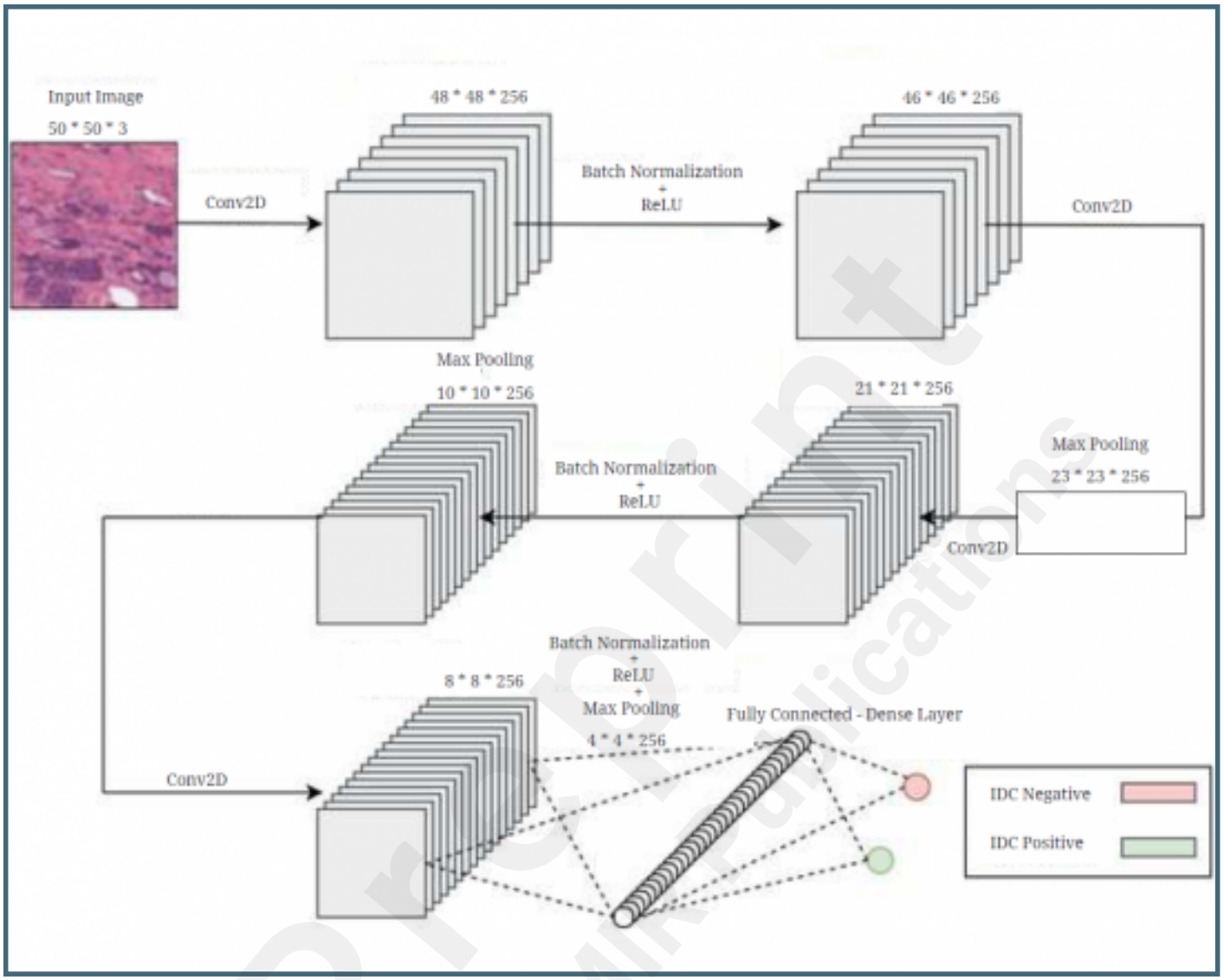
[13] A. Melek, S. Fakhry, and T. Basha, "Spatiotemporal mammographybased deep learning model for improved breast cancer risk prediction," in *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2023, pp. 1–4.

[14] P. Mooney, "Breast histopathology images," https://www.kaggle.com/datasets/paultimothymooney/breasthistopathology-images/data, 2019, accessed: 10- Feb- 2024.

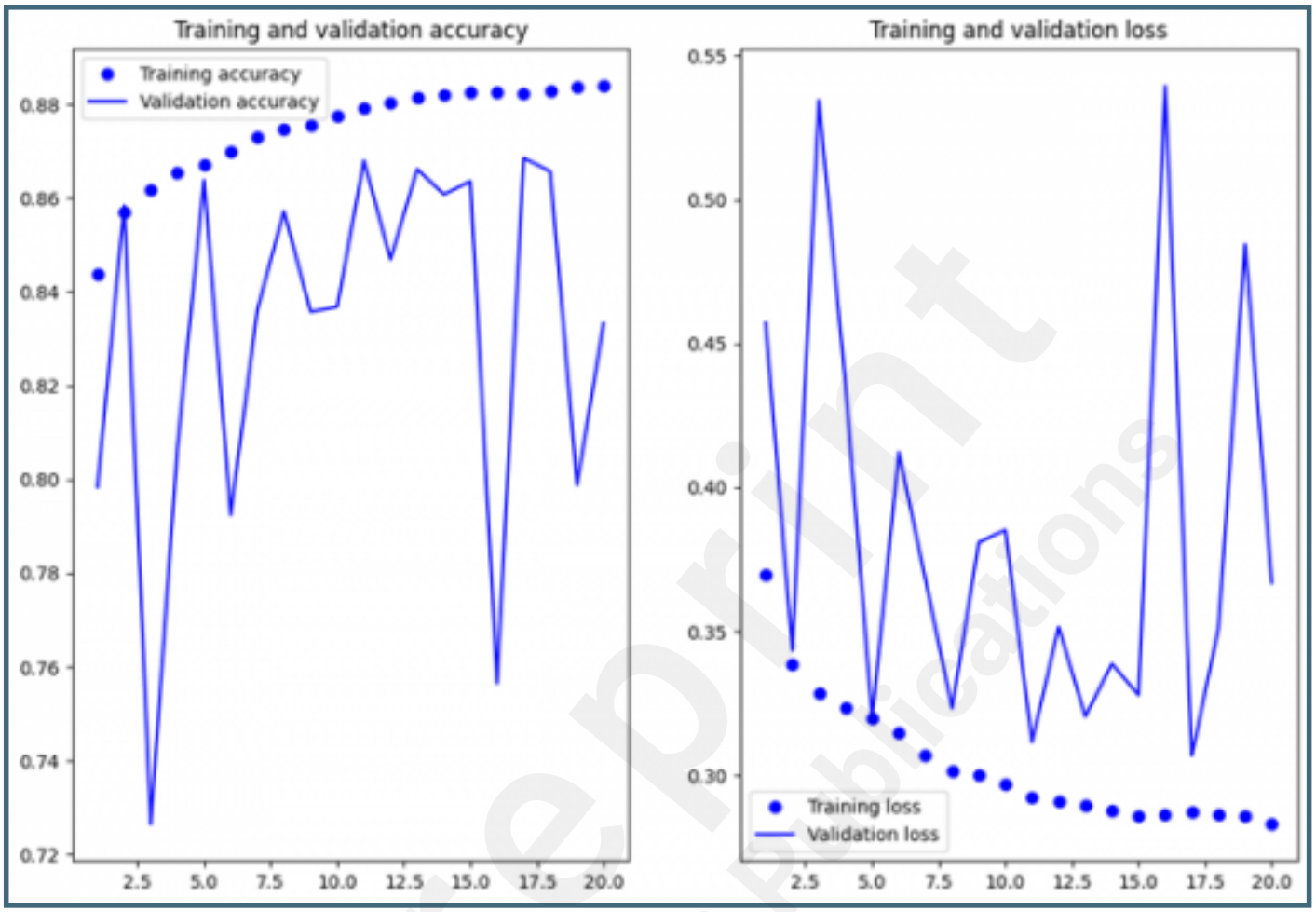
Supplementary Files

Figures

Model architecture.



Training and validation accuracy (left) and loss (right) over 20 epochs.



Confusion matrix of the model’s predictions.