Invasive Ductal Carcinoma Prediction using Deep Convolutional Neural Networks

Adnane Deroui
Master in Data Science and Business Analytics
International University of Rabat
Rabat, Morocco

***Abstract*— This project focuses on automating the detection of Invasive Ductal Carcinoma (IDC), the most common and aggressive form of breast cancer, using deep learning techniques. IDC diagnosis traditionally relies on pathologists manually identifying cancerous regions in breast tissue, a time-consuming and expertise-dependent process. To address this, we developed a convolutional neural network (CNN) model capable of classifying and localizing cancerous tissue in histopathology images. The dataset consisted of over 277,000 patches from 279 patients, containing both IDC-positive and IDC-negative samples. To combat issues such as class imbalance and overfitting, we employed strategies like data augmentation by flipping cancerous images, as well as using dropout and max-pooling layers to optimize model performance. Our CNN model achieved a classification accuracy of 92%, effectively distinguishing IDC-positive from IDC-negative patches. The results show that our model provides a reliable, automated method for IDC detection, significantly streamlining the diagnostic process and improving accuracy. This work underscores the potential of deep learning in medical imaging, offering both efficiency and precision in clinical settings, ultimately aiding in timely and accurate breast cancer diagnosis.**

I. Introduction

Breast cancer is a major health concern worldwide, with early detection playing a key role in improving patient outcomes. Invasive ductal carcinoma (IDC) is the most common type of breast cancer, accounting for around 80% of cases. It is known for its aggressive nature and ability to spread to other parts of the body, making timely and accurate diagnosis crucial.

Currently, the diagnosis of IDC relies on manual evaluation of tissue samples by pathologists, which can be time-consuming and subjective. Advances in deep learning, particularly Convolutional Neural Networks (CNNs), offer the potential to automate this process. CNNs are specialized in image analysis and are highly effective in recognizing patterns within complex data like medical images. By leveraging large datasets of tissue images, CNNs can be trained to detect and locate cancerous cells, providing consistent and efficient analysis. This approach could reduce reliance on manual methods and offer improved diagnostic capabilities, especially in regions with limited access to medical experts. In this study, we explore the use of deep learning techniques, particularly CNNs, to automatically detect IDC, building on previous research while incorporating more modern methods to enhance accuracy and efficiency.

II. Literature Review

A. What is Deep Learning?

Deep learning is a subset of machine learning that uses artificial neural networks to mimic the way the human brain processes information. By stacking multiple layers of neurons, deep learning models can automatically extract complex features from raw data. This makes deep learning particularly effective for tasks like image and speech recognition, where traditional algorithms struggle with the high-dimensional nature of the data. [1]

In the medical field, deep learning has shown great potential, with use cases ranging from detecting tumors in medical imaging to predicting patient outcomes. For example, Convolutional Neural Networks (CNNs) are widely used in analyzing radiology scans, while Recurrent Neural Networks (RNNs) are applied in time-series data, such as electrocardiograms (ECG), to monitor heart conditions. [2]

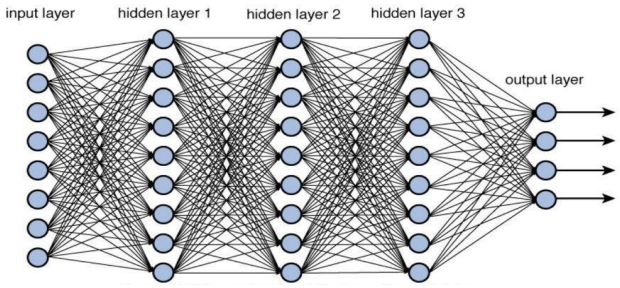

Fig. 1. Deep Neural Network with Multiple Layers

B. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks are a specialized type of deep learning model designed for processing structured grid data, such as images. Unlike traditional neural networks, CNNs use convolutional layers to automatically extract spatial features from input data. A convolution operation applies filters to the input, effectively scanning the image for patterns like edges, textures, and shapes. The output of these convolutions is passed through an activation function, typically ReLU (Rectified Linear Unit), to introduce non-linearity. [3]

CNNs also employ pooling layers, such as max pooling, to down sample feature maps, reducing dimensionality and computational complexity while retaining important information. This hierarchical approach allows CNNs to learn increasingly complex patterns in deeper layers. [4]

To optimize the model, CNNs utilize backpropagation along with gradient-based optimization techniques like stochastic gradient descent (SGD) or its variants, such as Adam, to minimize the loss function making CNNs both efficient and robust for tasks like image classification, object detection, and medical image analysis. [5]

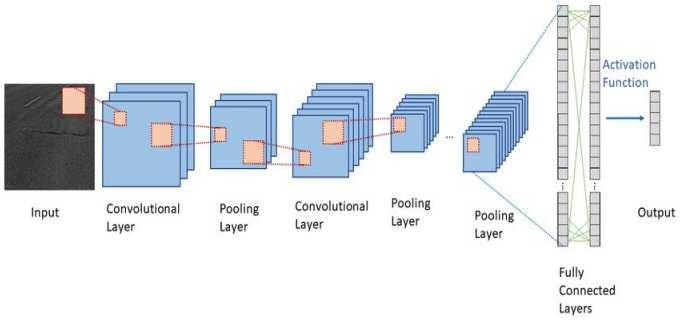

Fig. 2. Convolutional Neural Network Architecture

C. Convolutional and Pooling Layers

In CNNs, convolutional layers are the cornerstone of feature extraction. Each convolutional layer applies multiple filters (kernels) to the input image, performing convolution operations that slide the filters across the image to produce feature maps. These filters are designed to detect specific features, such as edges or textures, at various spatial locations within the image. The result is a set of feature maps that highlight different aspects of the input data. [6]

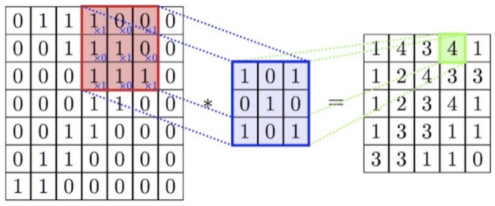

Fig. 3. Representation of a Convolutional Layer

Pooling layers, specifically max pooling, follow convolutional layers to reduce the spatial dimensions of the feature maps while preserving the most critical information. Max pooling works by dividing the feature map into non-overlapping regions and selecting the maximum value within each region, thus reducing the data and computational complexity. This process also helps in achieving translational invariance, allowing the network to recognize features regardless of their position in the input image. Together, convolutional and pooling layers enable CNNs to efficiently learn hierarchical features and improve performance on tasks like image classification and object detection. [7]

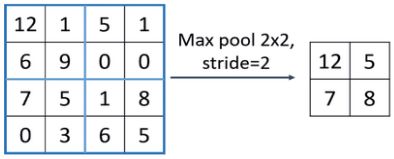

Fig. 4. Representation of a Max Pooling Layer

D. Activation Function and Loss Function

The hyperbolic tangent activation function is commonly used in neural networks to introduce non-linearity. It maps input values to a range between -1 and 1, allowing the network to handle both positive and negative activations, making it particularly useful for problems where the data may have a more balanced range of values. The output of the tanh function is centered around zero, which helps mitigate issues like vanishing gradients to some extent, compared to the sigmoid activation function. However, like sigmoid, tanh can still suffer from gradient saturation when inputs are in extreme ranges. [8]

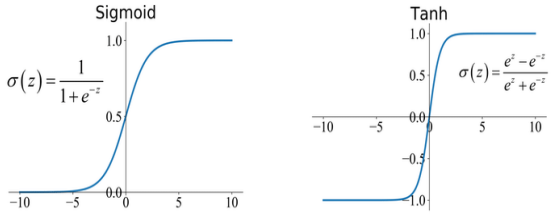

Fig. 5. Representation of the Tanh and Sigmoid Activation Functions

The binary cross-entropy, is used to measure the performance of a classification model where the output is a probability between 0 and 1. Log loss penalizes incorrect predictions more heavily as they deviate from the true label, making it particularly suitable for binary classification tasks. The function computes the negative log likelihood of the true labels, resulting in a loss value that the model seeks to minimize during training. By optimizing log loss, models are encouraged to output probabilities that are as close as possible to the true labels, leading to improved accuracy in classification tasks. [9]

$$J(W) = -\frac{1}{n}\sum_{i=1}^{n} y.\log(\hat{y}) + (1-y).\log(1-\hat{y})$$

E. Invasive Ductal Carcinoma (IDC)

Invasive ductal carcinoma (IDC) is the most common type of breast cancer, accounting for nearly 80% of all cases. IDC begins in the milk ducts of the breast but quickly invades surrounding tissues, allowing it to spread (metastasize) to other parts of the body if not detected early. The malignancy of IDC makes timely diagnosis crucial for patient outcomes. [10]

The diagnosis of IDC typically involves a biopsy, where a pathologist examines tissue samples to identify cancerous cells and assess the stage of the disease. This process, however, is labor-intensive and prone to human error, as the pathologist

must manually search for malignant cells. Recent advancements in deep learning, especially through CNNs, have paved the way for automated detection and localization of IDC in medical images, offering a faster and more objective means of diagnosis. By analyzing vast datasets of breast tissue images, CNNs can be trained to recognize IDC with high accuracy, improving diagnostic efficiency and aiding in treatment planning. [11]

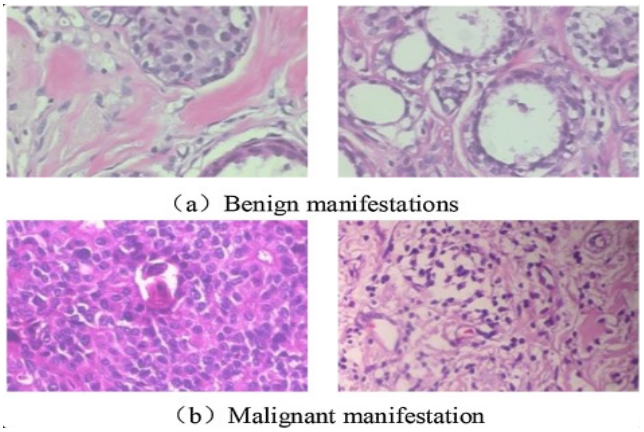

Fig. 6. Visualizing Benign and Malignant Breast Tissue

III. Methodology

A. Problem Statement

The primary problem this study aims to address is the accurate detection and localization of invasive ductal carcinoma (IDC) in breast tissue images to improve early diagnosis and treatment planning. Pathologists often rely on manual examination of tissue samples, a process that can be time-consuming, subjective, and prone to error. Given the aggressive nature of IDC and its potential to metastasize, timely and precise diagnosis is critical. Without efficient detection methods, delays in diagnosis may lead to suboptimal treatment outcomes, putting patients at higher risk.

B. Importing Libraries

Python offers a wide range of libraries that make data analysis, visualization, and machine learning more accessible and efficient. Below are the libraries used in this project

Pandas	*Python library for data manipulation and analysis*
Numpy	*Python library for linear algebra using arrays and matrices*
Matplotlib	*Python library for basic data visualizations*
Seaborn	*Python library for advanced data visualizations*
Scipy	*Python library for scientific and technical computing*
Scikit-learn	*Python library for machine learning algorithms*
TensorFlow	*Python library for deep learning algorithms*

Table 1. Libraries Used for this project

C. Overview of the Breast Histopathology Dataset

The dataset used in this study contains breast cancer image data from 279 patients, with each patient identified by a unique patient ID. From the whole mount slides of these patients, a total of 277,524 image patches were extracted, each sized 50x50 pixels. These patches are classified as either malignant (IDC-positive) or benign (IDC-negative), with 78,786 labeled as IDC-positive and 198,738 labeled as IDC-negative.

The primary goal of this dataset is to aid in the detection and localization of Invasive Ductal Carcinoma (IDC), the most common type of breast cancer. Pathologists typically focus on IDC regions when assigning aggressiveness grades to the cancer. This dataset provides a foundation for training machine learning models to automatically identify and delineate IDC regions, supporting faster and more accurate diagnostic processes.

	patient_id	path	target
0	9036	./Desktop/archive-6/IDC_regular_ps50_idx5/9036...	0
1	9036	./Desktop/archive-6/IDC_regular_ps50_idx5/9036...	0
2	9036	./Desktop/archive-6/IDC_regular_ps50_idx5/9036...	0
3	9036	./Desktop/archive-6/IDC_regular_ps50_idx5/9036...	0
4	9036	./Desktop/archive-6/IDC_regular_ps50_idx5/9036...	0
...	...	...	...
277519	8957	./Desktop/archive-6/IDC_regular_ps50_idx5/8957...	1
277520	8957	./Desktop/archive-6/IDC_regular_ps50_idx5/8957...	1
277521	8957	./Desktop/archive-6/IDC_regular_ps50_idx5/8957...	1
277522	8957	./Desktop/archive-6/IDC_regular_ps50_idx5/8957...	1
277523	8957	./Desktop/archive-6/IDC_regular_ps50_idx5/8957...	1

Fig. 7. Checking Data Types

D. Exploratory Data Analysis

In this section, we perform an Exploratory Data Analysis (EDA) to better understand the structure and key characteristics of the dataset. EDA is essential for identifying trends, spotting anomalies, and providing a general overview of the data before diving into any deep learning algorithms.

1) Distribution of Patches by Patient

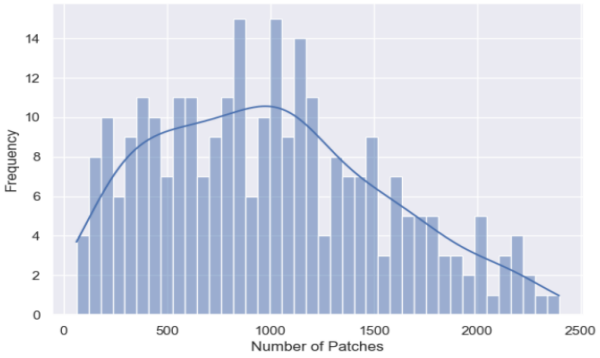

Fig. 8. Representation of the Number of Patches by Patient

The distribution of the number of image patches per patient in the dataset shows significant variation, which raises questions about potential differences in the resolution or size of tissue samples between patients. The dataset consists of 279 patients, with an average of approximately 995 patches per patient. The minimum number of patches for a patient is 63, while the maximum is 2,395, showing a substantial range. The interquartile range (IQR) also highlights this variation, with 25% of patients having fewer than 561 patches and 75% having fewer than 1,362 patches. This variability suggests that the size of the tissue samples, or the way they were processed, differs significantly between patients, which could impact the resolution or content of the image patches used for analysis.

2) *Distribution of the % of Patches containing IDC*

The distribution of the percentage of patches containing Invasive Ductal Carcinoma (IDC) across patients reveals significant variability in the concentration of cancerous regions. On average, 30.8% of patches per patient show IDC, but the standard deviation of 20.1% indicates wide fluctuations. Some patients have as little as 1% of their patches showing IDC, while others have up to 90%, suggesting that certain tissue slices are either densely cancerous or that only cancer-focused regions were sampled. The interquartile range shows that 25% of patients have fewer than 13.8% IDC-positive patches, while 75% have fewer than 44.6%.

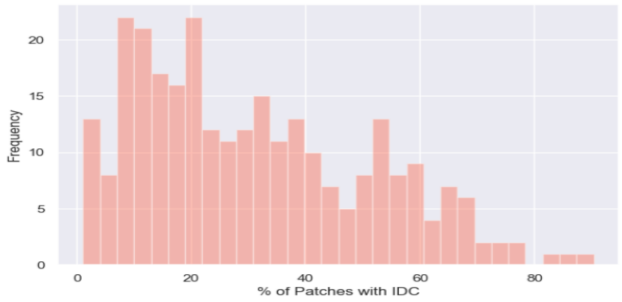

Fig. 9. Representation of the Distribution of Patches

This variability raises the question of whether the tissue slices per patient cover the entire region of interest or only specific sections where cancer is concentrated. In cases where over 80% of patches show IDC, it may indicate either a highly aggressive spread of the cancer or a selective sampling of cancerous regions, rather than a comprehensive scan of the breast tissue.

3) *Number of Benign and Malignant Patches*

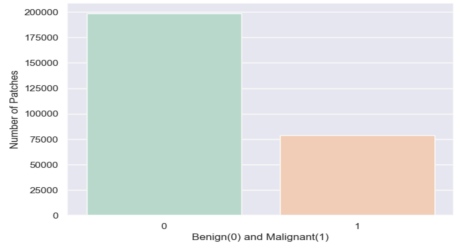

Fig. 11. Representation of the Benign and Malignant Patches

The dataset exhibits a noticeable class imbalance between IDC-positive (cancerous) and IDC-negative (healthy) patches. Out of a total of 277,524 image patches, 198,738 are IDC-negative (71.6%), while 78,786 are IDC-positive (28.4%). This imbalance can pose challenges for machine learning models, as they may become biased toward the majority class (healthy patches) and underperform in detecting cancerous regions.

To address this, it will be important to revisit the class distribution when setting up a validation strategy. Strategies such as adjusting class weights, oversampling IDC-positive patches, or undersampling IDC-negative patches can help ensure that the model accurately identifies cancerous patches, even with the imbalance present in the dataset.

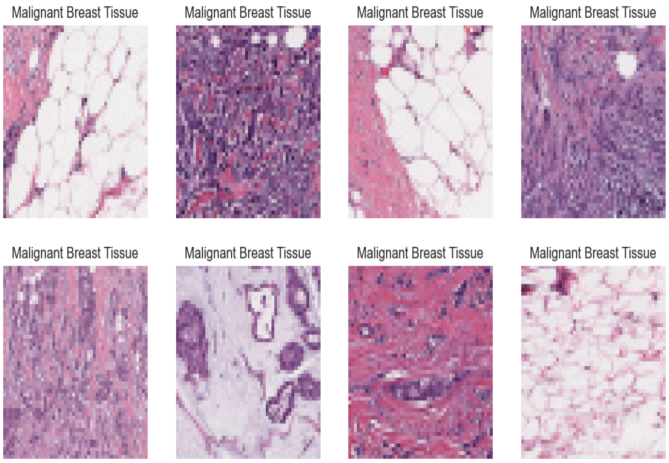

Fig. 12. Representation of Malignant Patches

When visually comparing normal and cancerous breast tissue, certain color differences become apparent. Cancerous tissue is often stained with a more intense red color, which makes it visually distinct from healthy tissue. In many cases, darker, more violet-colored areas tend to correspond to cancerous regions, while lighter, rose-colored tissue is often non-cancerous.

However, this is not always a reliable indicator, as some violet areas may not be cancerous. This raises a concern: if violet tissue is associated with mammary ducts rather than cancer, a model trained on these visual cues might mistakenly learn that the presence of mammary ducts is always linked to cancer. This potential bias highlights the need for caution when developing automated models, ensuring they differentiate between actual cancerous tissue and natural anatomical structures.

E. *Modeling*

1) *Data Augmentation*

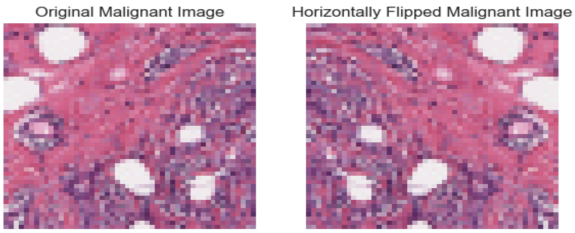

Fig. 13. Representation of the Data Augmentation Technique

Data augmentation has proven essential in addressing class imbalance and enhancing the robustness of our model. Specifically, the horizontal flipping of images has been a valuable technique in this context. Originally, the dataset was skewed with only 78,786 malignant images, leading to potential biases in model training. By horizontally flipping these malignant images, we effectively doubled their count, resulting in 157,572 malignant images. This augmentation has significantly improved the balance between malignant and benign images, which now stands at 45% and 55%, respectively.

The addition of these flipped images helps the model generalize better by exposing it to a greater variety of image orientations and reducing the risk of overfitting to the original, unaltered images. Overall, horizontal flipping has not only mitigated the initial imbalance but also enhanced the model's ability to learn more robust features from the augmented dataset.

2) *Data Partition*

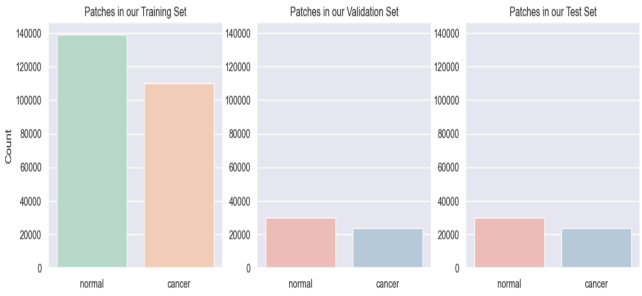

Fig. 14. Representation of the Different Sets

The data partitioning for this image classification task is designed to ensure balanced training, validation, and testing phases. The dataset has been divided as follows: *70%* for the training set, amounting to *249416* images; *15%* for the validation set, totaling *53447* images; and *15%* for the test set, also consisting of *53447* images.

Initially, the dataset was imbalanced, containing *78786* malignant images. However, after applying data augmentation techniques, the number of malignant images has increased to *157572*, balancing the dataset with *198738* benign images. This augmentation has effectively addressed the initial class imbalance, leading to a more equitable distribution of malignant (*45%)* and benign (*55%)* images across the entire dataset. This balanced distribution is crucial for training a robust model and ensuring fair evaluation across both classes.

3) *Data Preprocessing*

In this data preprocessing phase, we used the ImageDataGenerator class from Keras to load and preprocess images for a deep learning model.

```
BBATCH_SIZE = 32
IMG_HEIGHT = 224
IMG_WIDTH = 224

#create the ImageDataGenerator object and rescale the images
trainGenerator = ImageDataGenerator(rescale=1./255.)
valGenerator = ImageDataGenerator(rescale=1./255.)
testGenerator = ImageDataGenerator(rescale=1./255.)

#convert them into a dataset
train_dataset = trainGenerator.flow_from_dataframe(
    dataframe=train_df,
    class_mode="binary",
    x_col="image",
    y_col="label",
    batch_size=BATCH_SIZE,
    seed=42,
    shuffle=True,
    target_size=(IMG_HEIGHT,IMG_WIDTH))

val_dataset = valGenerator.flow_from_dataframe(
    dataframe=val_df,
    class_mode='binary',
    x_col="image",
    y_col="label",
    batch_size=BATCH_SIZE,
    seed=42,
    shuffle=True,
    target_size=(IMG_HEIGHT,IMG_WIDTH))

test_dataset = testGenerator.flow_from_dataframe(
    dataframe=test_df,
    class_mode='binary',
    x_col="image",
    y_col="label",
    batch_size=BATCH_SIZE,
    seed=42,
    shuffle=True,
    target_size=(IMG_HEIGHT,IMG_WIDTH)
)
```

Fig. 15. Data Preprocessing

First, the ImageDataGenerator is initialized with a "rescale" parameter that normalizes pixel values by dividing by 255, ensuring that all pixel values are scaled between 0 and 1. Next, the flow_from_dataframe method is used to convert the image paths and labels into datasets. For the training, validation, and test sets, the images are loaded from the file paths. Each image is resized to a fixed dimension of 50x50 pixels, which is specified by IMG_HEIGHT and IMG_WIDTH, and they are batched in groups of 32 using BATCH_SIZE. The "class_mode" is set to "binary" because this is a binary classification problem (Malignant vs. Benign). The datasets are shuffled to ensure random ordering of samples, and a seed value of 42 is provided to maintain reproducibility. This approach ensures that the images are correctly preprocessed and ready to be fed into the CNN model for training, validation, and testing.

4) *Model Architecture*

The chosen CNN architecture is designed to effectively handle the classification task with a focus on balancing performance and generalization.

Layer (type)	Output Shape	Param #
conv2d_20 (Conv2D)	(None, 224, 224, 32)	896
batch_normalization_24 (BatchNormalization)	(None, 224, 224, 32)	128
conv2d_21 (Conv2D)	(None, 224, 224, 32)	9,248
max_pooling2d_8 (MaxPooling2D)	(None, 112, 112, 32)	0
batch_normalization_25 (BatchNormalization)	(None, 112, 112, 32)	128
dropout_12 (Dropout)	(None, 112, 112, 32)	0
conv2d_22 (Conv2D)	(None, 112, 112, 64)	18,496
batch_normalization_26 (BatchNormalization)	(None, 112, 112, 64)	256
conv2d_23 (Conv2D)	(None, 112, 112, 64)	36,928
batch_normalization_27 (BatchNormalization)	(None, 112, 112, 64)	256
max_pooling2d_9 (MaxPooling2D)	(None, 56, 56, 64)	0
dropout_13 (Dropout)	(None, 56, 56, 64)	0
conv2d_24 (Conv2D)	(None, 56, 56, 128)	73,856
flatten_4 (Flatten)	(None, 401408)	0
dense_20 (Dense)	(None, 128)	51,380,352
batch_normalization_28 (BatchNormalization)	(None, 128)	512
dense_21 (Dense)	(None, 64)	8,256
batch_normalization_29 (BatchNormalization)	(None, 64)	256
dense_22 (Dense)	(None, 64)	4,160
dropout_14 (Dropout)	(None, 64)	0
dense_23 (Dense)	(None, 24)	1,560
dense_24 (Dense)	(None, 1)	25

Total params: 51,535,313 (196.59 MB)

Trainable params: 51,534,545 (196.59 MB)

Non-trainable params: 768 (3.00 KB)

Fig. 16. Summary of the proposed model

The model starts with two convolutional layers, each with 32 filters, a kernel size of 3x3, and ReLU activation. These layers use *'he_uniform'* initialization and 'same' padding to preserve the spatial dimensions of the input. Batch Normalization is applied after each convolutional layer to stabilize and accelerate training by normalizing the activations. MaxPooling is used after the first two convolutional blocks to downsample the feature maps, reducing their spatial dimensions and computational complexity. Dropout layers with a rate of 0.3 are included to prevent overfitting by randomly setting a fraction of the neurons to zero during training.

The model then progresses to two additional convolutional blocks with 64 filters each, followed by another pooling layer and dropout. This deepens the network and allows it to capture more complex features. The final convolutional block has 128 filters to extract high-level features before flattening the output for the dense layers.

The dense layers follow, starting with 128 units and decreasing to 64 units, all with ReLU activation and Batch Normalization. This dense architecture helps in learning complex patterns from the features extracted by the convolutional layers. The final dense layers, including an ultimate layer with 24 units and a final output layer with a single unit using a 'sigmoid' activation function, are designed for binary classification. The model is compiled with the *'Adam'* optimizer and *'binary_crossentropy'* loss function, optimizing performance for the classification task.

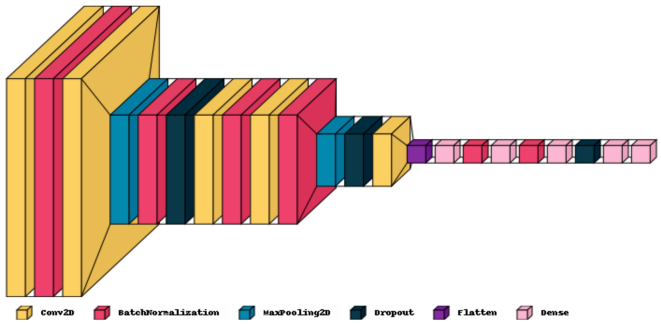

Fig. 17. Architecture of the proposed model

IV. Results and Experiments

The results of the training and validation accuracies, compared to the average pathologist accuracy of 85% in detecting invasive ductal carcinoma (IDC) show a promising performance of the model.

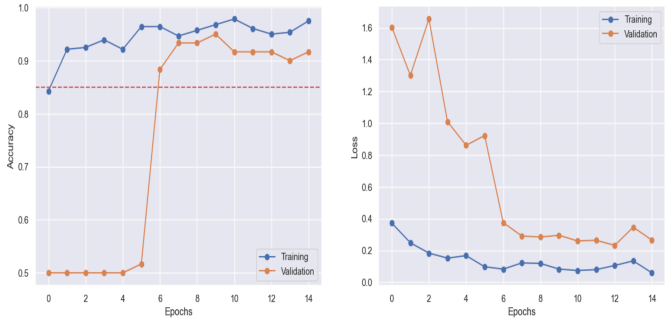

Fig. 18. Accuracy and Loss Evolution

During training, the accuracy steadily increases from around 84.3% in the first epoch to approximately 97.5% in the final epoch, indicating that the model is learning effectively. The validation accuracy starts at 50% for the initial epochs, which suggests the model struggled at the beginning to generalize. However, by the 7th epoch, the validation accuracy significantly improves, reaching up to 95%, which surpasses the pathologist's average accuracy of 85%.[12] This suggests that the model, after sufficient training, can potentially perform better than a human pathologist in detecting IDC.

The losses also reflect this trend. While the training loss decreases consistently, showing that the model fits the training data well, the validation loss begins high and gradually drops after the 6th epoch, further indicating improved generalization. The model's performance, particularly from epoch 7 onward, demonstrates that it has learned to generalize and provide high accuracy on unseen validation data, making it an effective tool for breast cancer detection. However, it's important to ensure that the model does not overfit and that further tests on different datasets are conducted.

V. Conclusion

In conclusion, this project demonstrates the application of deep learning, particularly convolutional neural networks (CNNs), in automating the detection of invasive ductal carcinoma (IDC) from breast tissue images. By leveraging image data preprocessing, model training, and evaluation strategies, we aimed to enhance the accuracy and efficiency of IDC classification. Although challenges such as class imbalance and potential overfitting were encountered, techniques like data augmentation and appropriate model regularization were employed to mitigate them. This automated approach not only helps in accelerating the diagnostic process but also offers potential improvements in precision, ultimately aiding in timely and accurate medical decision-making for breast cancer patients.

REFERENCES

[1] LeCun, Bengio, Hinton, “Deep Learning”, Nature, 2015

[2] Esteva A., Kuprel B., “Dermatologist-level Classification of skin cancer with deep neural networks”, Nature, 2017

[3] Kziehvsky A. et al., “ImageNet classification with deep convolutional neural networks”, Advances in Neural Information Processing Systems, 2012

[4] Goodfellow, I., Bengio, Y. & Courville, A. “Deep Learning” MIT Press, 2016

[5] Iandola et al., “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters”, 2016

[6] LeCun Y., Bengio Y., “Gradient Based Learning Applied to Document Recognition”, IEEE, 1998

[7] Goodfellow, I., Bengio, Y. & Courville, A. “Deep Learning” MIT Press, 2016

[8] Glorot X., Bordes A., “Deep Sparse Rectifier Neural Network”, International Conference on Artificial Intelligence and Statistics, 2016

[9] Goodfellow, I., Bengio, Y. & Courville, A. “Deep Learning” MIT Press, 2016

[10] American Cancer Society, “What is Invasive Ductal Carcinoma?”, 2021

[11] Cruz-Roa A. et al., “Automatic Detection of Invasive Ductal Carcinoma”, Medical Imaging, 2014

[12] Elmore et al., “Diagnostic concordance among pathologists interpreting breast biopsy specimens”, National Library of Medicine, 2015