**PROJECT REPORT**

*On*

***“*Breast Cancer Classification Using Deep Learning*”***

*Submitted in partial fulfillment of the requirements for the award of*

**Masters of Science in Technology**

**(Statistics And Data science)**

*In the department of*

**Mathematics**

*Submitted by*:

**Swarnamoy Ghosh**

**(PG/05/MSTSTDS/2023/002)**

*Under the guidance of*

**Dr. Kumari Pritee**

**(Assistant Professor)**

**Information Systems**

**Indian Institute of Management (IIM) Sambalpur**

**Certificate**

This is to certify that the project report entitled “**Breast Cancer Classification Using Deep Learning**,” submitted to the **Indian Institute of Management (IIM) Sambalpur**, is presented in partial fulfilment of the requirements for the summer internship between the 2nd and 3rd semesters of the **Master of Science in Technology (Statistics and Data Science)** in the **Department of Mathematics at Adamas University**. This report is a record of bonafide work carried out by **Swarnamoy Ghosh (PG/05/MSTSTDS/2023/002)** under your guidance.

All assistance and resources from various sources have been duly acknowledged.

No part of this report has been submitted elsewhere for the award of any other degree.

Dr. Kumari Pritee

Guide Name

Assistant Professor

Indian Institute of Management (Sambalpur)

Dr. Nav Kumar Mahato

Summer Internship Coordinator

Assistant Professor

Adamas University

**Acknowledgement**

The satisfaction and fulfilment that come with the successful completion of any project are incomplete without expressing gratitude to those whose guidance, support, and encouragement have been invaluable. It is with great pleasure that we present our project report, “**Breast Cancer Classification Using Deep Learning**,” which is the culmination of extensive research and knowledge application. We are deeply grateful to all who have contributed to the successful completion of this endeavour. Their support, guidance, and encouragement have made this work possible.

We extend our heartfelt gratitude to our guide, **Dr. Kumari Pritee, Assistant Professor at the Indian Institute of Management (Sambalpur)**, for her invaluable guidance, constant encouragement, and insightful suggestions throughout this project. Her expertise and dedication have been instrumental in shaping our understanding and approach.

Our sincere appreciation also goes to **Dr. Nav Kumar Mahato, Summer Internship Coordinator and Assistant Professor at Adamas University**, for his unwavering support, coordination, and guidance. His assistance and encouragement have been pivotal in steering our efforts in the right direction.

We would like to acknowledge the cooperation and support of our colleagues and fellow researchers, whose feedback and insights during discussions have greatly enriched our understanding and helped us overcome various challenges encountered throughout this project.

Furthermore, we are grateful to the faculty and staff of Adamas University for providing a conducive environment, knowledge resources, and assistance whenever needed. Their commitment to academic excellence has greatly supported our research efforts.

Finally, we would like to express our sincere gratitude to all participants involved in this research. Their willingness to contribute has been essential in gathering the necessary data and insights for our study.

**Declaration**

I, the undersigned, declare that the project entitled “**Breast Cancer Classification Using Deep Learning**,” submitted in partial fulfilment of the requirements for the summer internship between the 2nd and 3rd semesters of the **Master of Science in Technology (Statistics and Data Science) at Adamas University**, is our original work.

**Swarnamoy Ghosh**

**(PG/05/MSTSTDS/2023/002)**

**M.Sc. (Tech) Statistics and Data Science**

**Adamas University**

**Abstract**

Breast cancer remains one of the most prevalent and life-threatening cancers affecting women worldwide. Early and accurate diagnosis is essential for improving patient outcomes and guiding treatment decisions. This project utilizes deep learning, specifically Convolutional Neural Networks (CNNs), to classify breast cancer images into categories that indicate the presence of invasive ductal carcinoma (IDC) versus non-IDC. The image data for this study was sourced from Kaggle, a popular platform for open datasets, ensuring a diverse and high-quality dataset suitable for training deep learning models. Data preprocessing was conducted using the Pillow library, which facilitated image resizing, normalization, and augmentation, thereby improving model performance and generalization.

The CNN model achieved an impressive accuracy of 87% and an AUC (Area Under the Curve) score of 0.94, reflecting its strong ability to distinguish between IDC-positive and IDC-negative cases. These results demonstrate the potential of deep learning as a reliable tool in medical diagnostics.

This project has significant applications in the medical field, particularly in aiding pathologists by providing an automated second opinion, which could reduce diagnostic time and minimize human error. In research, the model and methodology can serve as a foundation for further studies in cancer detection and classification. This approach could be expanded to cover other types of cancer, making it a valuable tool for medical researchers. Ultimately, this work contributes to ongoing efforts to combat breast cancer and reduce its impact on women’s health worldwide by enhancing early detection capabilities with advanced technology.

**Table of Content**

|  |  |  |
| --- | --- | --- |
| Chapter | Title | Page |
|  | Title page | 1 |
|  | Certificate | 2 |
|  | Acknowledgement | 3 |
|  | Declaration | 4 |
|  | Abstract | 5 |
|  | Table of contents | 6 |
| 1 | Introduction | 7 |
| 2 | Technology and Limitation | 8 |
| 3 | Dataset and Preprocessing | 9 |
| 4 | Model Architecture and Performance | 11 |
| 5 | Conclusion | 15 |
| 6 | References | 16 |

**Chapter 1**

**Introduction**

**Purpose of the project**

Breast cancer is one of the most common and serious forms of cancer affecting women worldwide. It poses a significant threat to women's health, with a high incidence rate and, in advanced cases, a high mortality rate. Among various types of breast cancer, invasive ductal carcinoma (IDC) is the most frequently diagnosed, characterized by the invasion of malignant cells into breast tissue. Early detection and accurate diagnosis are crucial in managing breast cancer, as they can significantly improve treatment outcomes and patient survival rates.

Traditional diagnostic methods for IDC involve the examination of histopathology images, where pathologists analyse tissue samples under a microscope to identify malignancies. However, this manual process is time-consuming, subject to human error, and requires extensive expertise. The development of automated methods for diagnosing breast cancer using computational tools has become a priority in medical research, with deep learning emerging as a powerful technique in this domain. In particular, convolutional neural networks (CNNs) have shown great promise in analysing and classifying medical images, making them an ideal choice for identifying cancerous patterns in histopathology images.

**Objective**

The primary goal of this project is to design and implement a deep learning model based on CNN technology to classify breast cancer images, specifically identifying IDC-positive and IDC-negative cases. By leveraging CNNs' capacity to recognize complex patterns within images, this project aims to build an accurate and reliable classification model that can aid pathologists in diagnosing breast cancer, ultimately reducing diagnostic time and enhancing decision-making.

**Chapter 2**

**Technology and Limitations**

**Technology Used**

For this project, a Convolutional Neural Network (CNN) was implemented to classify IDC-positive and IDC-negative breast tissue samples. The model was built using TensorFlow and Keras, leveraging the power of deep learning to extract features from histopathology images and perform accurate classification. To handle the large dataset, the image preprocessing was done using the Pillow library, which allowed for efficient manipulation and augmentation of the data. The training was accelerated using GPU support through Kaggle's cloud GPU, which enabled faster model training and evaluation.

**Limitations**

While the model achieved promising results, several limitations were encountered during the project:

1. **GPU Limitations**: The training process initially faced challenges due to limited GPU resources. The local GPU setup was not efficient enough to handle the large dataset and required significant time to process each epoch. This led to the decision to utilize Kaggle's cloud GPU resources, which provided better computational power for faster training.
2. **Dataset Availability**: One of the major challenges of this project was the difficulty in finding a suitable dataset for the classification of IDC-positive and IDC-negative samples. Although the *Breast Histopathology Images* dataset on Kaggle provided a strong foundation, it was not fully suited for every aspect of the problem, particularly due to the imbalance in the classes and the need for high-quality, well-labeled data for both training and validation.
3. **Preprocessing Challenges**: Preprocessing the dataset proved to be another challenge, particularly in handling image normalization, augmentation, and balancing the dataset. The images had to be resized, normalized, and augmented to improve the generalization of the model. This process was time-consuming and required careful tuning to avoid overfitting, as the dataset contained a variety of image qualities and structures.
4. **Model Generalization**: While the model performed well on the training set, its ability to generalize to unseen data remains an area of concern. Although techniques such as cross-validation and data augmentation were used to mitigate this, the limited variety in the dataset could have impacted the model's performance on different unseen cases.

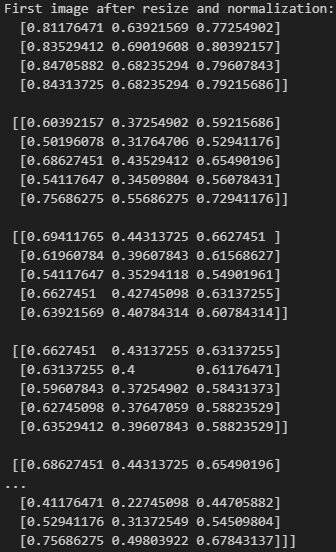
Despite these limitations, the project demonstrates the potential of using deep learning for medical image analysis and contributes to the growing body of work aimed at improving cancer detection methods. Future improvements could focus on obtaining more diverse datasets and optimizing GPU resources for better training efficiency.

**Chapter 3**

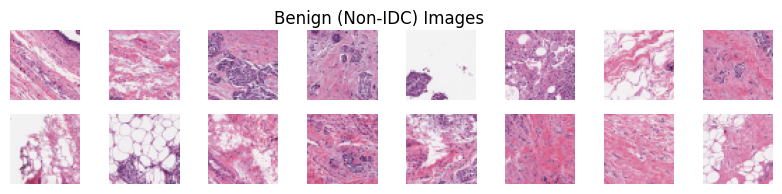
**Dataset And Preprocessing**

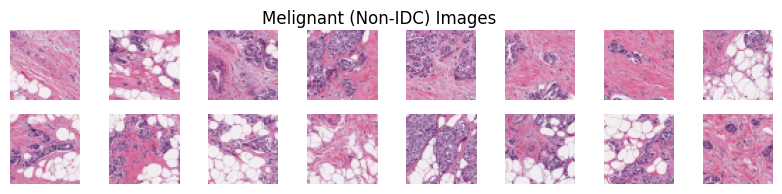
The dataset used in this project was obtained from Kaggle, containing high-resolution histopathology images labelled as IDC-positive or IDC-negative, depending on the presence of invasive ductal carcinoma. To prepare these images for input into a convolutional neural network (CNN), several preprocessing steps were applied, ensuring that the model could effectively learn from and generalize on the data.

* **Image Conversion and Normalization**: Each image was first converted into a numerical array format, allowing it to be processed by the CNN. This transformation facilitated pixel-wise manipulation, necessary for both normalization and resizing tasks. Each pixel value was scaled to a range between 0 and 1, normalizing the data and reducing input variability, which aids in stabilizing and accelerating model training.

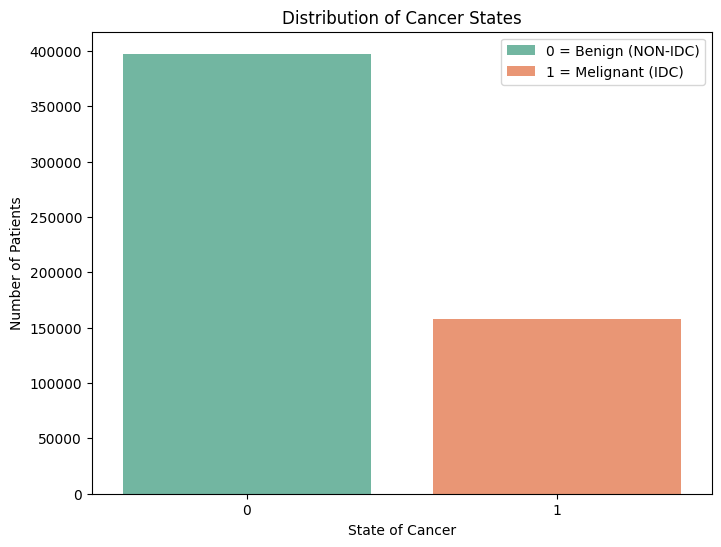


**Resizing**: Since CNNs require fixed-size inputs, each image was resized to a standard 50x50 pixel dimension. This resizing ensured consistency across all samples, reduced computational demands, and preserved essential features needed for classification.

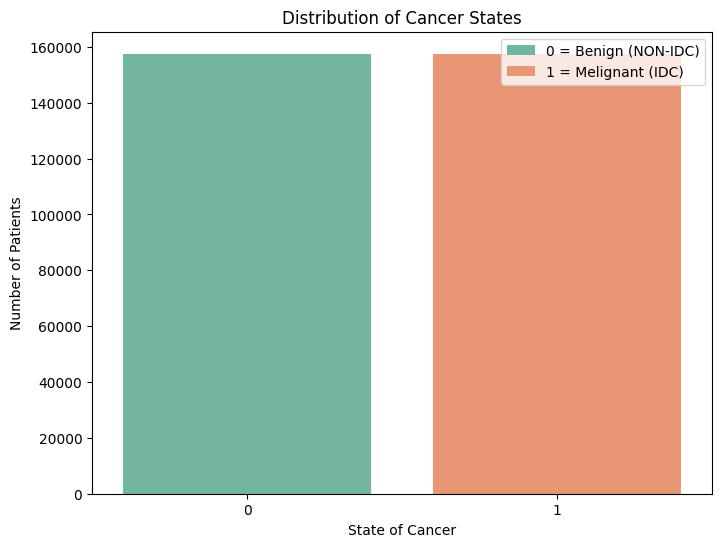




* **Categorization**: After preprocessing, images were categorized into two classes, IDC-positive and IDC-negative, based on labels provided in the dataset. This binary categorization served as the foundation for training the CNN model to identify cancerous and non-cancerous tissue regions accurately.
* **Class Imbalance and Under sampling**: The dataset was imbalanced, with a higher representation of IDC-negative samples compared to IDC-positive ones. Imbalanced datasets can lead to biased model predictions, as the model may favour the majority class. To address this, an under-sampling technique was applied to balance the dataset by reducing the number of IDC-negative samples, ensuring an equal representation of both classes. This adjustment improved the model's ability to distinguish between IDC-positive and IDC-negative cases.



After Under-Sampling:



These preprocessing steps, performed using the Pillow library in Python, enhanced the model’s performance by maintaining consistent image dimensions, improving feature extraction, and reducing overfitting tendencies.

**Chapter 4**

**Model Architecture and Performance**

**Model Architecture**

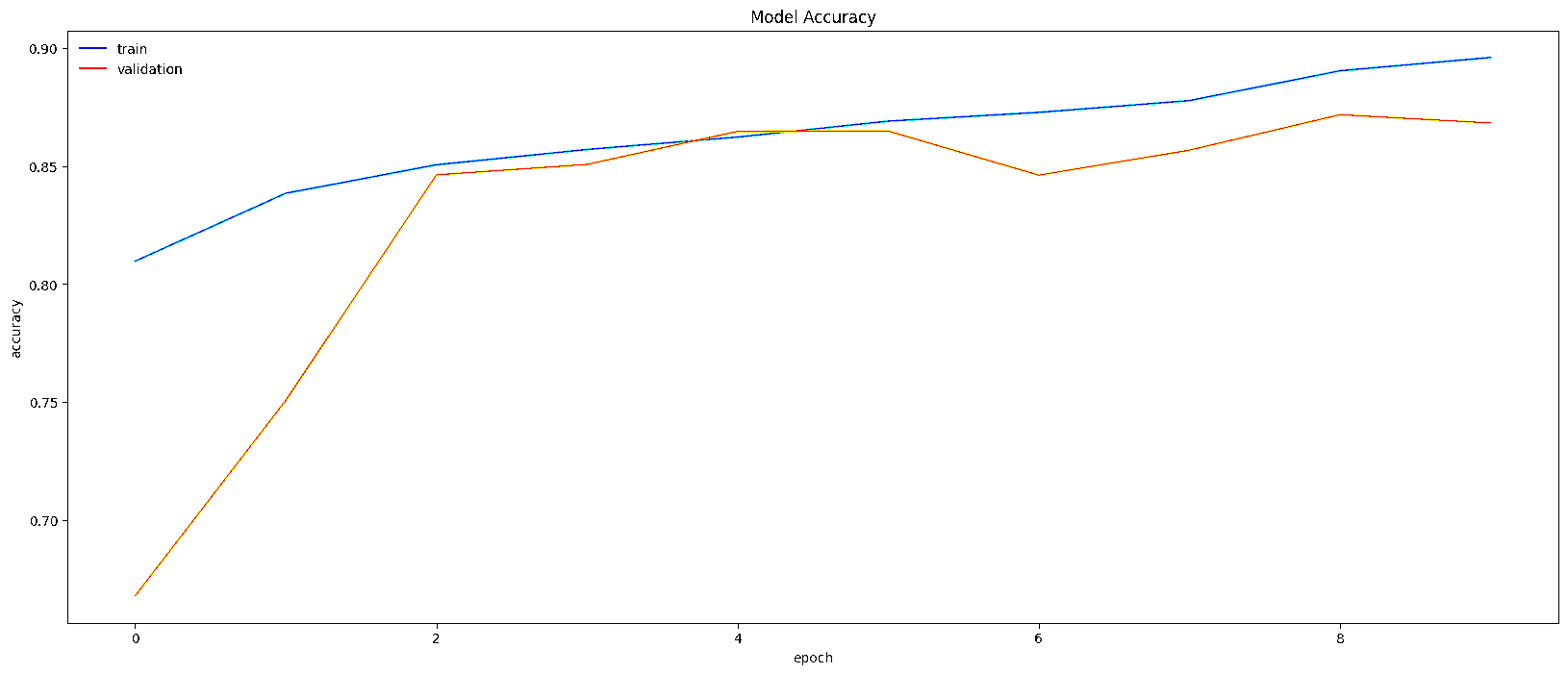
The convolutional neural network (CNN) architecture used for classifying breast cancer images includes a combination of convolutional layers, max-pooling layers, batch normalization, dense layers, and dropout for regularization.

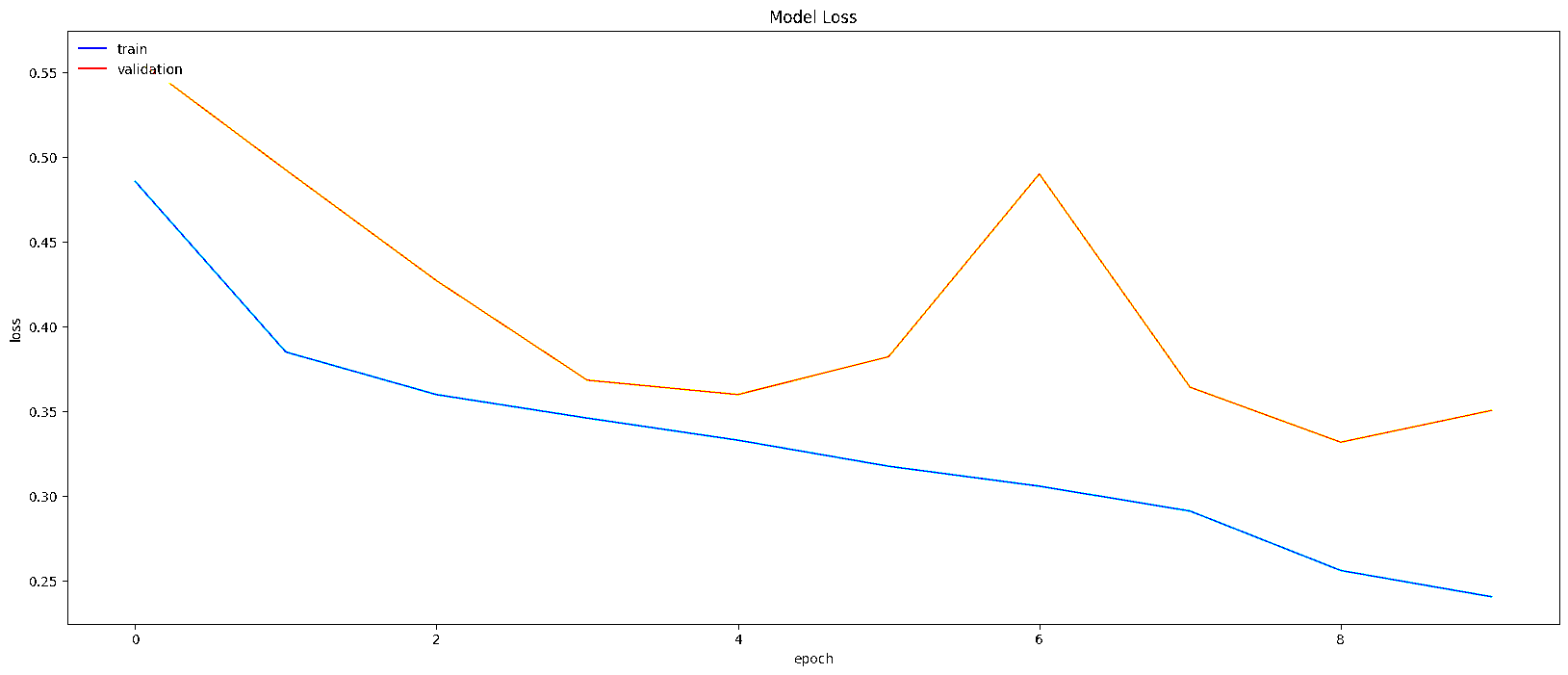
1. **Convolutional Layers:**
   * The model starts with a 2D convolutional layer with 32 filters and a 3x3 kernel size, followed by batch normalization to stabilize and accelerate training. This layer is followed by a max-pooling operation to reduce spatial dimensions.
   * A second convolutional layer is then applied with 128 filters and a 3x3 kernel, followed again by batch normalization and max pooling. These layers help the model capture essential patterns in the images that differentiate IDC-positive from IDC-negative samples.
2. **Fully Connected Layers:**
   * After flattening the pooled feature maps, a dense layer with 512 neurons is added, allowing the model to learn complex features. Dropout is applied here to prevent overfitting.
   * Another dense layer with 128 neurons follows, again with dropout, to further improve generalization.
3. **Output Layer:**
   * A final dense layer with a single neuron and a sigmoid activation function is used for binary classification (IDC-positive vs. IDC-negative).

**Model Training and Evaluation**

The model was trained for 10 epochs, during which both training and validation accuracy steadily improved, with final training accuracy reaching approximately 90% and validation accuracy stabilizing around 87%. The high performance on both datasets indicates a well-trained model with good generalization.

* **Training and Validation Accuracy**: The training and validation accuracy curves show that the model achieved an accuracy of 87% on the validation set.
* **Training and Validation Loss**: The training loss consistently decreased over epochs, showing convergence, while the validation loss reached stability, indicating a low level of overfitting.

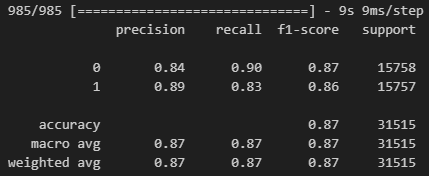




**Classification Report**

The classification report reveals strong performance across both classes:

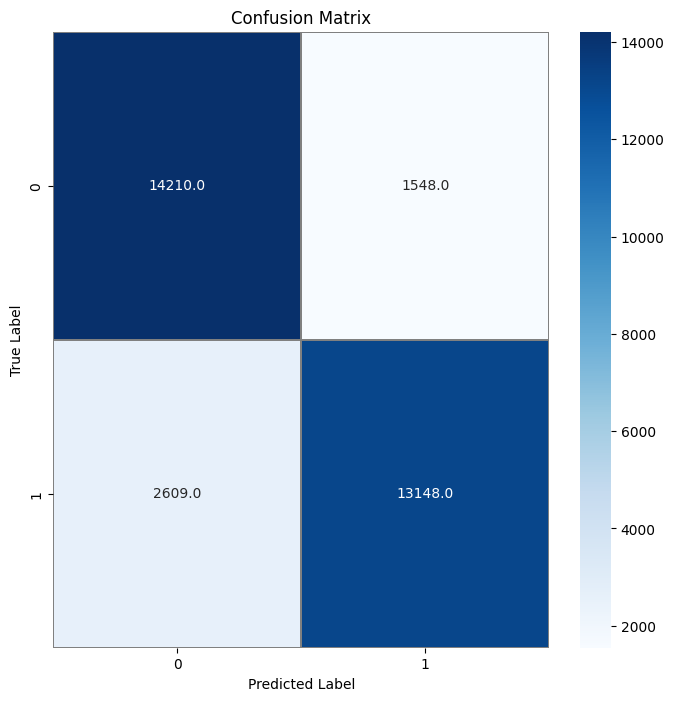
* **Precision**: The model achieved a precision of 0.84 for IDC-negative and 0.89 for IDC-positive, indicating that the model can accurately identify both classes.
* **Recall**: The recall values of 0.90 for IDC-negative and 0.83 for IDC-positive reflect a balanced performance, showing the model’s sensitivity to detecting IDC cases.
* **F1-Score**: The F1-score is balanced between both classes, with an average of 0.87, confirming consistent classification performance.



**Confusion Matrix**

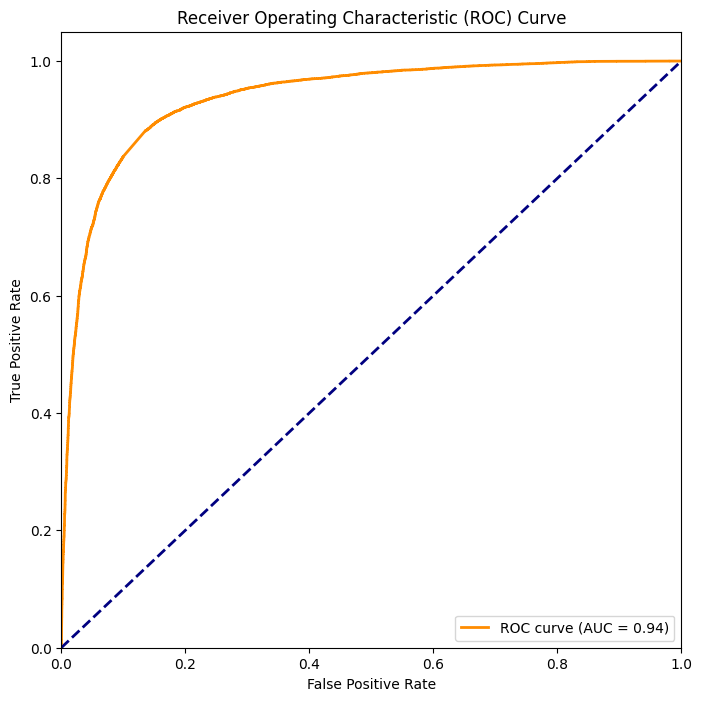
The confusion matrix shows:

* **True Negatives** (14210): IDC-negative samples correctly classified as negative.
* **True Positives** (13148): IDC-positive samples correctly classified as positive.
* **False Negatives** (2609): IDC-positive samples misclassified as negative.
* **False Positives** (1548): IDC-negative samples misclassified as positive.



**ROC-AUC Analysis**

The model achieved an impressive AUC (Area Under the Curve) score of 0.94, demonstrating its high discriminative ability between IDC-positive and IDC-negative samples. The ROC-AUC score is a strong indicator of the model’s effectiveness in correctly classifying the cancerous tissue regions.



**Chapter 5**

**Conclusion**

This CNN model shows promising performance in classifying IDC-positive and IDC-negative breast tissue samples. With an accuracy of 87% and an AUC of 0.94, the model highlights its potential for medical applications, particularly in early cancer detection and research. By assisting medical professionals in identifying malignant tissues, the model could improve diagnostic accuracy, minimize human error, and expedite the diagnostic process. This project underscores the effectiveness of deep learning in medical imaging analysis, contributing to better cancer diagnosis and ultimately improving patient outcomes. The success of this model emphasizes the growing role of artificial intelligence in healthcare, fostering more efficient and accurate diagnostic tools for clinicians.

**Chapter 6**

**References**

1. The dataset used in this study, *Breast Histopathology Images*, was provided by Paul Mooney on Kaggle.
2. Referred Breast Cancer Articles from National Cancer Institute.
3. Referred CNN model architecture of VGG16 from Tensorflow.