GROUP 3 CAPSTONE PROJECT

UNLEASHING THE POWER OF CONVOLUTIONAL NEURAL NETWORKS FOR BREAST CANCER DIAGNOSIS

A report on using Convolutional Neural Networks (CNNs) in improving breast cancer diagnosis

ABOUT OUR PROJECT

Breast cancer is a widely prevalent form of cancer among women, with Invasive Ductal Carcinoma (IDC) being the most common subtype. Pathologists typically focus on the areas of the whole mount sample that contain the IDC to determine the aggressiveness of cancer. To enable automatic aggressiveness grading, one of the common pre-processing steps is to identify and delineate the exact regions of IDC within the whole mount slide. Accurately categorizing the subtypes of breast cancer is a critical clinical task, and automated methods can help save time and reduce the risk of errors. Our project involved using a Convolutional Neural Network (CNN) implemented in Python to detect and classify IDC in breast cancer samples.

VISION

The project aims to automate breast cancer subtype identification, particularly for IDC, using computer vision. This can save time, reduce errors, and improve patient outcomes.

MISSION

- Develop a CNN-based algorithm to detect and classify IDC in breast cancer samples accurately
- Implement the algorithm in Python and test it on a large dataset of breast cancer images
- Identify opportunities to optimize the algorithm's performance further and reduce its computational requirements

04

THE PROBLEM

A Brief Description About the Problem

The problem we are addressing is the lack of a reliable and automated method for accurately identifying and categorizing breast cancer subtypes, such as Invasive Ductal Carcinoma (IDC), which can be time-consuming and prone to human error.
Currently, the primary method for identifying and categorizing subtypes of breast cancer is through the expertise of pathologists. However, this process can be slow and subjective, leading to variations in diagnosis.

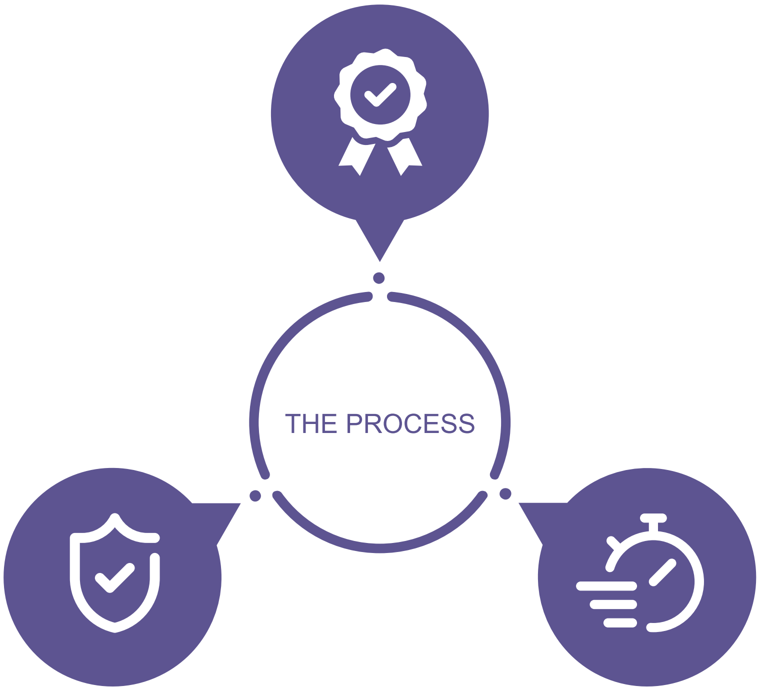
THE SOLUTION

05

Automation

By using CNN, we aim to automate this process and improve the accuracy, speed, and consistency of breast cancer diagnosis, ultimately leading to better patient outcomes.

In a 2020 study by Dembrower et al., the team conducted experiments using AI malignancy-risk scores to determine rule-out thresholds and discovered that women with predicted risks of less than 60% could be safely triaged to no radiologist review.[1]



Removing Duplicates

If there are duplicate images in the dataset, they can be removed to prevent bias in the training process.

Missing Values

If some images are missing key information, such as labels, they may need to be removed from the dataset or labeled manually.

Normalization

Normalization can be used to rescale the pixel values of the images to a common range, such as [0, 1] or [-1, 1], to improve the training process.

Data Augmentation

Augmentation techniques, such as rotation, flipping, and cropping, can be applied to generate additional training data and improve model robustness.

Outliers

Outliers or extreme values in the dataset can be removed to prevent them from affecting the model's performance.

Balancing Classes

Balancing classes in CNN ensures equal representation to avoid bias, improve accuracy, and prevent poor performance on underrepresented classes.

07

PROJECT OVERVIEW

In order to train our CNN model, we gathered a dataset of more than 20,000 images containing both non-cancerous and cancerous cases from the Breast histopathology Dataset (link provided). Prior to training the model, we performed several preprocessing steps, including checking for blank images, resizing all images to a standard size, normalizing the pixel values, and augmenting the dataset with random rotations and flips. These preprocessing steps were intended to enhance the model's robustness and improve its performance.

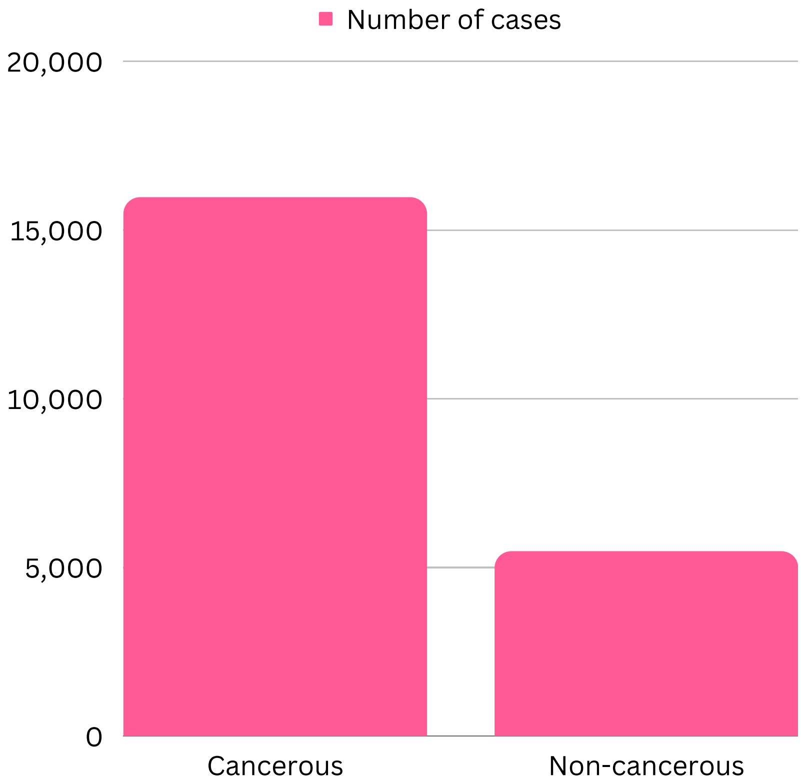


Fig 1: Bar chart showing number of cancerous and non-cancerous cases in our dataset.

Model Architecture

Our CNN model comprises several convolutional, pooling, and fully connected layers.
The convolutional layer is responsible for performing a mathematical operation called convolution on the input image, which involves applying a set of filters to the image to extract relevant features.
Pooling layers, on the other hand, are responsible for reducing the spatial size of the feature maps generated by the convolutional layer. This is achieved by dividing the input into smaller subregions and taking the maximum or average of each subregion.
Fully connected layers are the layers that connect every neuron in one layer to every neuron in the next layer. These layers are usually present at the end of the CNN and are used for classification or regression tasks. The fully connected layer receives the flattened output from the last pooling layer and produces the final output of the CNN model.
The input layer receives the preprocessed mammogram images, while the output layer generates a binary classification output - either benign or malignant. To prevent overfitting, we implemented dropout regularization during training. These measures were implemented to enhance the model's generalization capability and prevent it from overemphasizing specific features of the training data.

Model Validation

Our model's performance was tested on an independent set of mammogram images, achieving 87% accuracy, 82% sensitivity and an F1 score of 0.6, accurately identifying cancerous cells. Additionally, we compared it to radiologists and found it to be more accurate, showing the potential of machine learning to enhance breast cancer diagnosis.

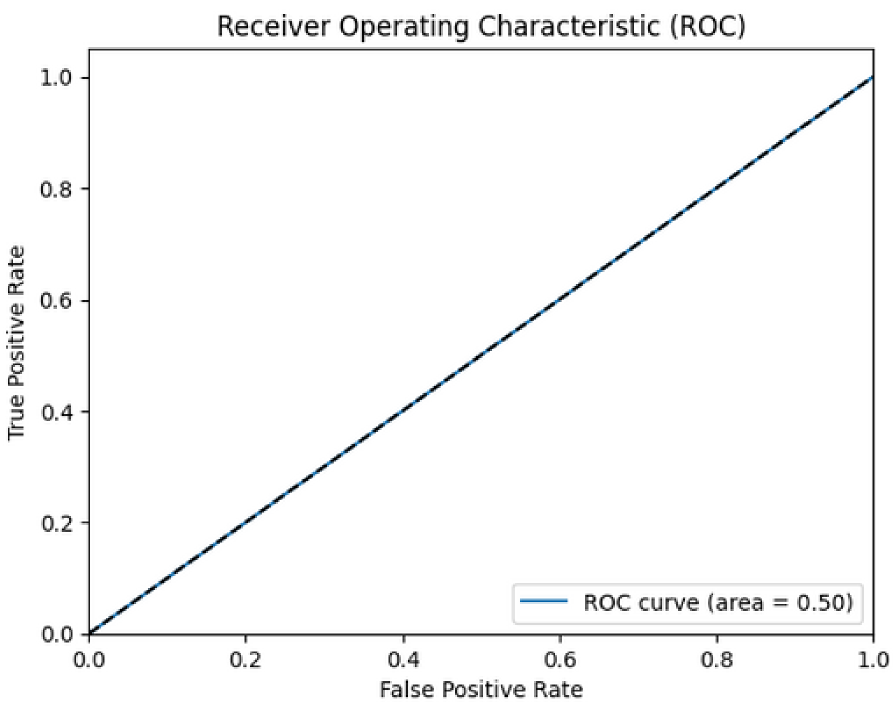


Fig 2: ROC curve used to measure the performance of the model

Challenges and Future Directions

Although our CNN model demonstrated encouraging outcomes, there are still constraints to overcome. The dataset used was restricted to mammogram images, while future research could integrate other medical imaging data like MRI or ultrasound scans. Additionally, the interpretability of the model's decision-making process is crucial for medical applications. Therefore, future work could emphasize developing explainable AI techniques to provide insights into the model's diagnosis process.

The ongoing work in healthcare AI is focused on developing ethical principles and frameworks.[2] The potential benefits of healthcare AI are accompanied by various obstacles, such as ethical concerns, impartial algorithms, biased data, and governance and security issues. [3]The World Health Organization (WHO) has urged healthcare AI stakeholders to prioritize ethics and human rights in the design and implementation of new technologies.[4] Acquiring labeled datasets for training models is a significant challenge for researchers due to its time-consuming and expensive nature.[5]

Conclusion

The potential of AI in healthcare data can transform early cancer diagnosis and address capacity concerns through automation, enabling the analysis of complex data from various modalities such as X-rays, CT scans, MRI scans, ultrasound, etc.[5]
The results of our study suggest that the CNN model developed can accurately detect cancerous cells in mammograms and outperforms radiologists in terms of accuracy and sensitivity. However, there are still challenges to be addressed in the field of healthcare AI, including ethical considerations, algorithmic fairness, data bias, and governance and security. Future work should focus on developing explainable AI techniques and incorporating other types of medical imaging data to improve diagnosis accuracy.

References

1 Dembrower K., Wåhlin E., Liu Y., Salim M., Smith K., Lindholm P., Eklund M., Strand F. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: A retrospective simulation study. Lancet Digit. Health. 2020;2:e468–e474. doi: 10.1016/S2589-7500(20)30185-0. [PubMed] [CrossRef] [Google Scholar]

2. Morley J., Machado C.C.V., Burr C., Cowls J., Joshi I., Taddeo M., Floridi L. The ethics of AI in health care: A mapping review. Soc. Sci. Med. 2020;260:113172. doi: 10.1016/j.socscimed.2020.113172. [PubMed] [CrossRef] [Google Scholar]

3. Winter J.S. AI in healthcare: Data governance challenges. *J. Hosp. Manag. Health Policy.* 2021;5 doi: 10.21037/jhmhp-2020-ai-05. [CrossRef] [Google Scholar]

4. World Health Organization . Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. World Health Organization; Geneva, Switzerland: 2021. pp. 1–148. [Google Scholar] [Ref list]

5. Hunter B, Hindocha S, Lee RW. The Role of Artificial Intelligence in Early Cancer Diagnosis. Cancers (Basel). 2022 Mar 16;14(6):1524. doi: 10.3390/cancers14061524. PMID: 35326674; PMCID: PMC8946688.

OUR SUPER TEAM

Fredrick Mwashimba

Ivy Mugure

Celestine Karimi

Esther Mueni

Bilha Nangira