HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Classification of Early Breast Cancer using Convolutional Neural Networks

Kaitlyn Leung

Havergal College, Canada

ABSTRACT

Breast cancer is a life-threatening disease that impacts millions of people worldwide, necessitating heightened awareness and effective early detection strategies. Current methods for determining the presence and type of breast cancer in a clinical breast exam conducted by healthcare providers may be inaccurate, imprecise, and prone to error. The goal of this research project is to offer a second opinion for physicians to predict whether or not a patient has early onset breast cancer through a machine learning model developed from an image dataset, to significantly reduce the error in breast cancer staging, making the process more efficient in healthcare. Early detection is crucial for successful treatments of breast cancer, as treatment options in late-stage diagnosis of breast cancer can have a much worse prognosis. In the US, 240,000 cases of breast cancer are diagnosed per year in both men and women with a mortality rate of approximately 42,000 per year. [1] To approach this problem, I have tested an MLP Classifier, Logistic Regression, Ridge Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, and trained a Convolutional Neural Network to compare various results and determine the best accuracy. The Convolutional Neural Network yielded impressive outcomes, reaching up to 79% testing accuracy and 98% training accuracy. The final model exhibits the capability to identify early signs of breast cancer and differentiate between malignant and benign tumours with relatively high accuracy showing the potential of AI in combination with medical imaging that can assist medical staff in diagnosis.

Introduction

Breast cancer is the second most common cause of death in women after lung cancer, where abnormal breast cells grow out of control in the ductal carcinoma, lobular carcinoma, or in other cells within the breast. [2] When left unchecked over time, the cancerous cells will spread to other organs including the brain, liver, lungs and bones becoming fatal especially as the first common detectable site is to the lymph nodes under the arm, however, it is possible to have cancer-bearing lymph nodes that cannot be felt. [5]

There are 5 categorizing breast cancer stages from stage 0 followed by stages 1 to 4. Early stages are where cells are fairly small (3cm) and have spread to a tiny area in the sentinel lymph node. [3] These are commonly treated with a lumpectomy or mastectomy, then radiation therapy to target any remaining cancer cells. [7] In metastatic breast cancer (Stage III), the tumor is large (5cm) and has spread to many nearby lymph nodes. This late-stage breast cancer is treated with systemic therapies, including chemotherapy, hormone therapy, and targeted therapy. [3] However, treatment options vary depending on the type of breast cancer, hormone receptor status, HER2 status, and the patient's overall health. [6]

Individuals with dense breast tissue pose a greater diagnostic challenge, as the detection of a lump during a physical examination becomes more challenging and a mammogram screening misses about 1 in 8 breast cancers. [3] A mammogram is an X-ray machine that takes a picture of the breast, where two plates will firmly press and flatten the breast for an X-ray picture to be taken, these steps will repeat until both breasts are screened thoroughly. [4]

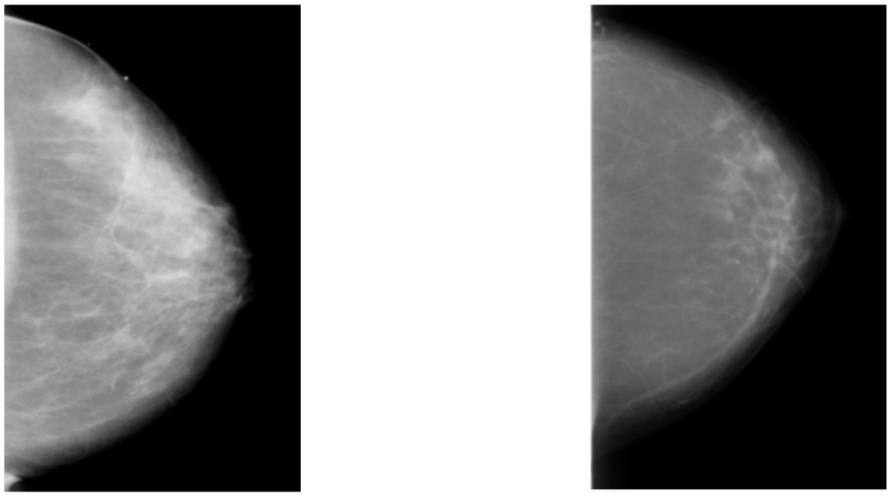
Image processing and classification through a machine learning model have been proven to be effective in autonomously detecting the presence of cancer by identifying patterns, increasing the speed of detection, and reducing human error. The machine learning algorithm built is a supervised Keras Sequential API classification early detection system that works with vision data outputting any presence of early signs of breast cancer, being an effective solution to decrease the number of deaths associated with breast cancer.

Dataset

The dataset used in this research project is a set of 2,620 images of scanned film mammography from CBIS-DDSM (Curated Breast Imaging Subset of DDSM) an updated and standardized version of the Digital Database for Screening Mammography (DDSM) curated by a trained mammographer.

The dataset consists of normal, benign, and malignant cases with verified pathology information.

The model used a total of 1,318 images: split into train and test with 50% of the data used for malignant training images, 43% benign testing images, and 7% benign without callback testing images.
Below is an example of each type of image:



Malignant Benign

Figure 1. Images of the two different classes in the dataset

In the figures screened by a mammogram on the left, there are microcalcifications (tiny deposits of calcium) that look like white patches or masses in the breast, indicating the presence of a malignant tumor. The image on the left also has a slightly spiculated outer edge, another indication of invasive tumor cells. In the dataset used by the model, there are many variations in stages, size, and quality of the images. Each image is converted to a CMAP array (255 x 255 x 3), as well as a turned-off axis for a cleaner display and flattened for training and testing purposes.

Methodology

Baseline Models

When first developing models, 4 out of 8 models showed the most promising data. 3 baseline models from scikit-learn; the Support Vector Classifier, Random Forest Classifier, and Ridge Classifier. Seeing results rang-

ing from 57% to 98.85% accuracy. The highest model accuracy created is the Keras Sequential API classification early detection system based on a Convolutional Neural Network model, which can be implemented to solve the goal of this research project.

Support Vector Classifier (SVC)

The Support Vector Classifier, also known as the Support Vector Machine, is a supervised algorithm that classifies the data points into categories. In this study, the model will be provided with training images and will label the images benign or malignant with a linear kernel. A decision boundary is used to separate data points belonging to two different classes by finding a hyperplane that maximizes the margin between the two classes.

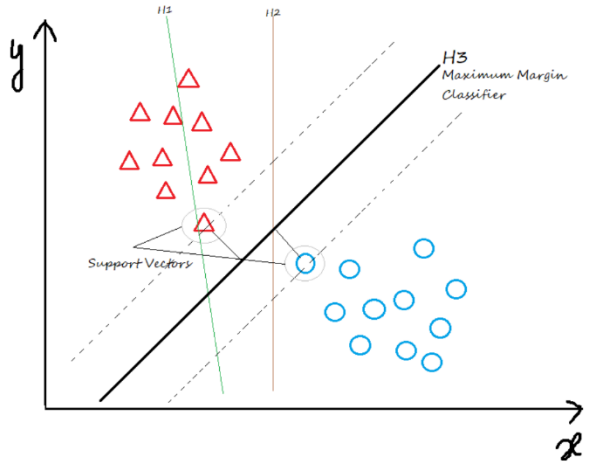


Figure 2. Image of Support Vector Classifier sorting data

Random Forest Classifier

The Random Forest Classifier is an extension of the Decision Tree Algorithm that combines predictions of multiple decision trees to improve accuracy and reduce overfitting. The Random Forest algorithm is made up of a collection of decision trees, each tree composed of data samples drawn from a training set. This classifier can handle a wide range of data types and complexities as it is known for its versatility and robustness.

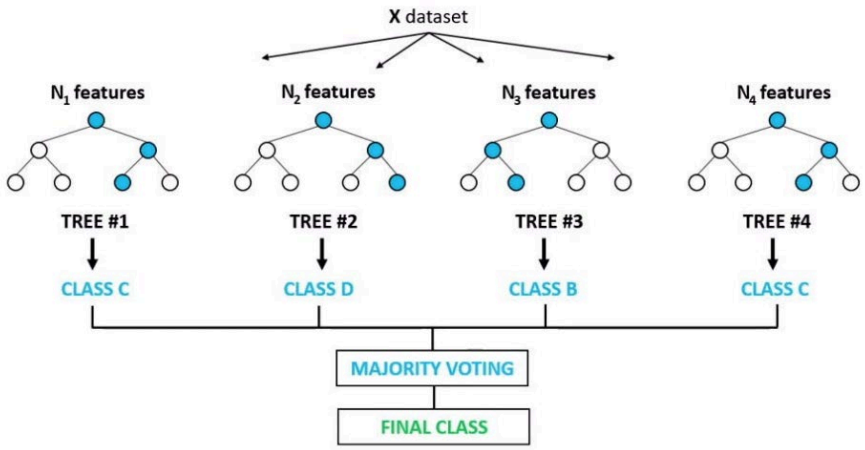


Figure 3. Image of Random Forest Classifier sorting data

Ridge Classifier

A Ridge Classifier is a linear classification algorithm that is a variation of Ridge Regression. The Ridge classifier converts target values into {-1, 1} then treating the problem as a regression task (multi-output regression in the multiclass case). Ridge classification introduces regularization, helping prevent overfitting often denoted as λ (lambda), encouraging the model to keep the coefficients and features small.

CNN (Convolutional Neural Network)

A Convolutional Neural Network is an Artificial Neural Network designed for processing images and recognition to process pixel data. [8] CNN's key feature is its ability to learn patterns and features from data, perfect for the recognition of benign vs malignant tumors. CNN receives input images, analyzes them, and produces output, categorizing the images into distinct classes. This is achieved through convolutional layers, neural network layers, pooling layers, dropout layers, and various other components to enhance the network's efficiency in image classification, collectively called the hidden layers.

The performance of the CNN model is evaluated on training samples, and validation accuracy is employed to evaluate its generalization to new data. Utilizing the train-test split function, I split the X train and Y train variables to replace the inputs with new cross-validation variables, testing it on the training samples.

Results and Discussion

Metrics

The evaluation of breast cancer classification from mammography images into the two primary classes was conducted using four performance metrics. A confusion matrix was used for further analysis for each model where the Precision, Recall, F1 and Accuracy can be determined. The number of epochs was also recorded to track the progression of the machine learning model's training over time and help determine the optimal point at which the model achieves the best performance on the given task.

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Baseline Models

Figure 4. Graph of Resulted Metrics in Baseline Models

Developing the initial models, I worked under the assumption that the baseline models would very likely underperform in contrast to the Convolutional Neural Network (CNN) models. The baseline models lack spatial awareness and overlook the 2D structure of the images, unlike CNNs which are purposely built for tasks involving images. Converting the images into a one-dimensional array, losing the ability to understand the spatial relationships and specific features in the visual data when comparing benign versus malignant tumors. The absence of mechanisms such as convolutional layers, feature hierarchy and parameter sharing in the baseline models contributed to their underperformance in tasks resulting in lower accuracies.

Convolutional Neural Network (CNN)

Model	Testing Accuracy	Accuracy	Number of Epochs
Model 10	77.01	96.05	30
Model 11	78.12	97.46	45
Model 13	78.64	97.76	54
Model 12	78.58	98.85	138

The highest classification testing and training accuracy was achieved when increasing the epoch number, however the model would be more likely to overfit. To combat this, I would increase the number of Neural Network (NN) layers and add dropout layers.

Out of all the models that were utilized and built, the best-performing model 13 yielded 78.64% testing accuracy and 97.76% training accuracy.

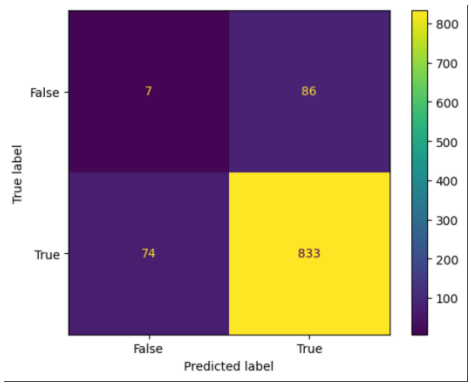


Figure 4. Image of Confusion Matrix a Model Produced. Confusion Matrix

The metrics predicted true positive, true negative, false positive, and false negative resulting in the models' ability to differentiate between each class. The goal is to reduce the number of false positives and false negatives. A false positive type 1 error occurs when a diagnosis incorrectly indicates the presence of the disease when it is not actually present. In healthcare, a type 1 error can led to unnecessary treatments, stress, and costs for patients. A false negative type 2 error occurs when the diagnosis fails to detect the disease when it is present. In healthcare, a type 2 error can be more serious, as it may result in missed treatment opportunities and delayed intervention, potentially allowing the disease to worsen. Displayed on the confusion matrix the model shown has made 74 false positives and 86 false negatives errors.

On average, the prevalent mistake observed across all models involved misclassifying a malignant tumor as benign. This tendency may stem from the processing of images, where I had to reduce the resolution, due to resource limitations, leading to an information loss. Consequently, this hindered the CNN model's ability to detect microcalcifications, contributing to the misclassification.

Conclusion

The objective of this research project is to develop a machine-learning model aimed at early breast cancer detection. The model is designed to provide a supplementary assessment for physicians, predicting whether a patient is likely to have early-onset breast cancer based on an image.

A Kaggle database sourced from CBIS-DDSM, consisting of 2,620 images, was pre-processed. The data was divided into 50% for malignant training, 43% for benign testing, and 7% for benign testing images without callbacks. Subsequently, this dataset was subjected to various models from scikit-learn, including the Ridge Classifier, Random Forest Classifier, and Support Vector Classifier, along with a curated trained Convolutional Neural Network model.

By training and refining components of a Keras Sequential API CNN model, I have developed a model that yields high-accuracy values. The model can serve as a breast cancer classification tool, the high accuracy values indicate that this can be trusted as a dependable indication of the presence of breast cancer. Nevertheless,

the model still exhibits weakness when classifying images. Therefore, it is advised to utilize this as a supplementary second reference as the effectiveness varies.

Lastly, this project would benefit from incorporating higher quality data sets as well as other forms of CNNs that could be tested on the dataset to see if the model's accuracy can be improved even more. Observing the application of this early detection system for breast cancer in the real-world for AI-based medical diagnosis for patients would be intriguing. Comparing the model's accuracy to traditional diagnostic procedures also adds an interesting dimension to its evaluation. It is important to note that these models are not designed to replace doctors; however, further research is crucial to determine the acceptable threshold at which they can effectively complement or enable nurses to streamline patient care. Achieving 100% accuracy with this model would be noteworthy, particularly in the context of real-world AI applications for medical diagnosis, enabling nurses and doctors to streamline patient care more efficiently.

Acknowledgments

I would like to thank Ronil Synghal for his guidance and advice during the course of this project.

References

1. U.S. Department of Health & Human Services. (2023, July 25). *How is breast cancer diagnosed?*. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm
2. American Cancer Society. (2023, September 14). *Breast cancer statistics: How common is breast cancer?*. Breast Cancer Statistics | How Common Is Breast Cancer? | American Cancer Society. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=Breast%20cancer%20is%20the%20second,kills%20more%20women%20each%20year.
3. *Treatment of breast cancer stages I-III*. American Cancer Society. (2022, April 12). https://www.cancer.org/cancer/types/breast-cancer/treatment/treatment-of-breast-cancer-by-stage/treatment-of-breast-cancer-stages-i-iii.html#:~:text=Most%20women%20with%20breast%20cancer,treatment%20you%20will%20likely%20need
4. Canadian Cancer Society / Société canadienne du cancer. (2023). Mammography. Canadian Cancer Society. https://cancer.ca/en/treatments/tests-and-procedures/mammography#:~:text=You%20will%20stand%20in%20front,the%20breast%20can%20be%20seen
5. World Health Organization. (2023, July 12). *Breast cancer*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
6. American Society of Clinical Oncology. (2023, August 18). *Breast cancer - types of treatment*. Cancer.Net. https://www.cancer.net/cancer-types/breast-cancer/types-treatment
7. DePolo, J. (2023, August 15). *Radiation Therapy*. Radiation therapy. https://www.breastcancer.org/treatment/radiation-therapy
8. Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018, June 22). *Convolutional Neural Networks: An overview and application in Radiology*. Insights into imaging. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108980/