A Machine Learning Model to Detect Early Breast Cancer

Classification of Early Breast Cancer using a Machine Learning Model

Abstract

Breast cancer is a prevalent and life-threatening condition 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 ineffective, vary from person to person, and are 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 reduce the error in classification significantly, 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 be very limited. In a single year over 240,000 cases of breast cancer are diagnosed in both men and women, having a mortality rate of 42,000 each year in the U.S. [1]

To approach this problem, I have tested a MLP Classifier, Logistic Regression, Ridge Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, and a trained Convolutional Neural Network to compare various results and determine the best accuracy. The final model is able to determine the presence of early signs of breast cancer as well as distinguish between malignant versus benign tumors with a relatively high accuracy showing the potential of medical imaging to assist medical staff.

After training the model on the different classes, the results were found to be impressive as the final accuracy was at 98.85%.

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]

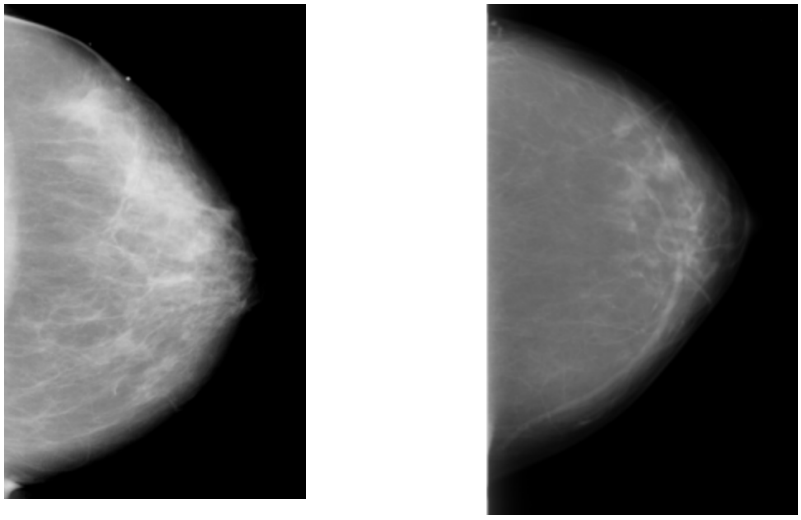
There are 5 categorizing Breast cancer stages from stage 0 followed by stages 1 to 4. Early stages are where cells are fairly small (3cm), have spread to only spread to a tiny area in the sentinel lymph node. [3] Commonly treated with a lumpectomy or mastectomy, then radiation therapy to target any remaining cancer cell. [7] 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 a 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 has been proven to be effective to autonomously detect the presence of a cancer from identifying patterns, increasing the speed of detection, reducing human error, is accessible, and can be a second opinion that can help physicians come to conclusions. 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. As well the image on the left has a slight spiculated outer edge, another indication of an invasion of tumor cells. In the dataset used by the model there were many variations of stages, size, and quality of image. Each image was 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/Models

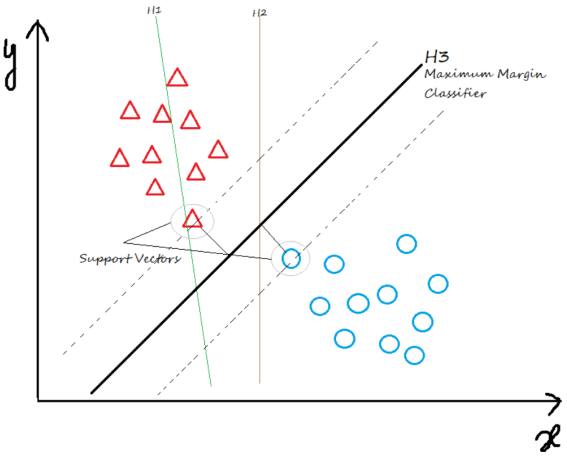
Baseline Models

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 ranging 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 is 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, where in this study they will be provided training images and label them benign or malignant with a linear kernel. With the decision boundary it separates data points belonging to two different classes by finding a hyperplane that maximizes the margin between the two classes.

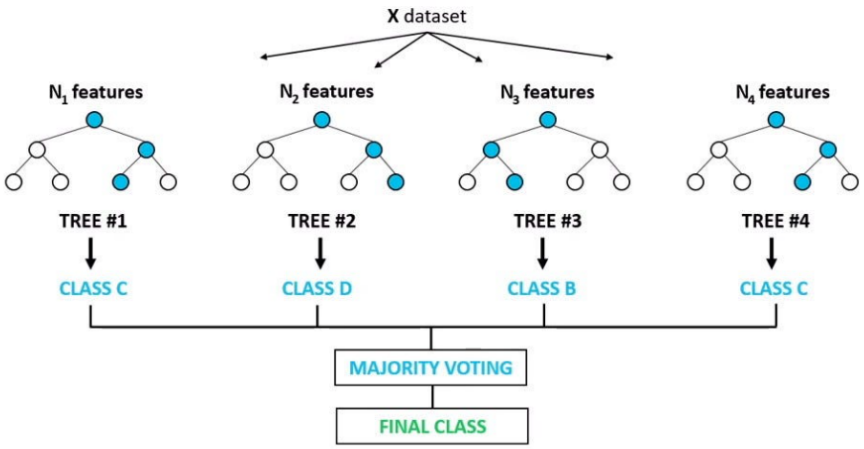
Figure 2



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 is able to handle a wide range of data types and complexities as it is known for its versatility and robustness.

Figure 3



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} and then treats 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] CNNs key feature is its ability to learn patterns and features from data which is perfect for recognition of benign vs malignant tumors. The CNNs take input images, process them, then output them, classifying them into different categories. This is achieved through the use of convolutional layers and pooling layers, to make the network more efficient and help classify the image, known as the hidden layers.

The CNN model's performance is evaluated on the training samples, then a validation accuracy is used to assess how well the model generalizes to new data. With the train test split function, I split the X train and Y train variable to replace the inputs with new cross validation variables, testing it on the training samples.

Results and Discussion

Metrics

The classification of breast cancer from mammography images into two main classes was evaluated on 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 were 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.

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

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

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

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

Figure 4

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 unlike CNNs which are purposely built for tasks involving imagines and the baseline models overlooked the 2D structure of the images. Converting the images into a one-dimensional

array, therefore 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)

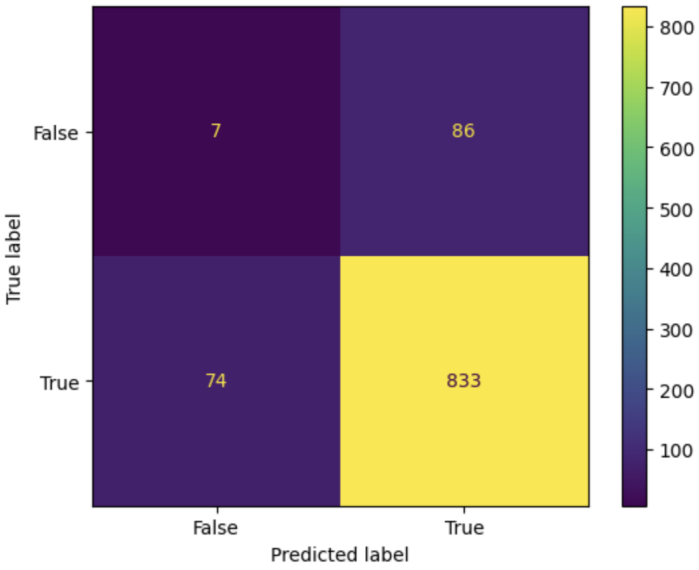
Model	Testing Accuracy	Accuracy	Number of Epochs
Model 10	77.01	96.05	30
Model 11	78.12	97.46	45
Model 13	78.64	97.76	54
Model 12	78.58	98.85	138

The highest classification testing and training accuracy was achieved when increasing the epoch number, however this 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 12 yielding an accuracy of 98.85%.

Figure 4

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 false positives and false negatives. A false positive type 1 error occurs when a diagnosis incorrectly indicates the presence of a condition or disease when it is not actually present. In healthcare, a Type 1 error can lead to unnecessary treatments, stress, and costs for patients. A false negative type 2 error occurs when the diagnosis fails to detect a condition or disease when it is actually present. In healthcare, a Type 2 error can be more serious, as it may result in missed treatment opportunities and delayed intervention, potentially allowing a condition 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 the reduced resolution, implemented due to resource limitations, led to an information loss. Consequently, this hindered the CNN model's ability to detect microcalcifications, contributing to the misclassification.

Conclusion

The objective of the research 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 dataset.

A database of Kaggle from CBIS-DDSM consisting of 2,620 images was pre processed, splitting the data into 50% malignant training 43% benign testing and 7 benign without callback testing images; This dataset was then run through multiple models from sklearn such as the Ridge Classifier Random Forest Classifier, and Support Vector Classifier as well as a curated trained CNN model.

By training and tweaking parts of a Keras Sequential API CNN model on the data set, I have created an accurate breast cancer classification tool that can be implemented into practice. The model's high accuracy values means that this can be trusted as a reliable 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 and certain weaknesses still exist.

Lastly, this project would benefit from incorporating higher quality data datasets as well as other forms of CNNs could be tested on the dataset to see if the model's accuracy can be improved even more. If this model reached 100% accuracy it would be cool to see it in use of real-world applications of AI models to medical diagnosis patients and comparing the models accuracy to the accuracy of a traditional diagnosis procedure. As AI will not replace doctors however there needs to be more research done to understand at what threshold would be acceptable to supplement or even enable nurses to streamline patient care. While this Early Detection System for breast cancer model is not intended to replace doctors, it would be interesting to compare the model's accuracy to traditional diagnostic procedures. Achieving 100% accuracy with this model would be fascinating, especially when applied to real-world AI

applications in medical diagnosis for patients, enabling nurses and doctors to streamline patient care more efficiently.

Acknowledgements

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

References

1. https://www.cdc.gov/cancer/breast/basic_info/diagnosis.htm
2. 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. 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. https://cancer.ca/en/treatments/tests-and-procedures/mammography#:~:text=You%20will%20stand%20in%20front,the%20breast%20can%20be%20seen.
5. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
6. https://www.cancer.net/cancer-types/breast-cancer/types-treatment
7. https://www.breastcancer.org/treatment/radiation-therapy
8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108980/

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html