Ellie Strande & Audrey Jones
Nick LaHaye
May 19, 2023

Advancing Breast Cancer Classification with Convolutional Neural Networks

Abstract

Breast cancer remains a significant global health challenge, necessitating accurate classification of tumors as benign or malignant for effective treatment strategies and improved patient prognosis. In this study, we developed a robust model utilizing deep learning techniques to discern between benign and malignant breast cancer tumor images. Leveraging the comprehensive Breast Histopathology Image dataset, we constructed a Convolutional Neural Network (CNN) capable of accurately predicting breast cancer classifications. By integrating machine learning algorithms into the diagnostic process, we aimed to enhance our understanding of breast cancer and contribute to the development of more precise diagnostic tools. The CNN model achieved a high accuracy of 95.9% on the test set and demonstrated a sensitivity of 1.0 and specificity of 0.931. These results indicate the model's proficiency in correctly identifying true positives while maintaining a low rate of false positives. The model's performance offers promise for earlier and more precise detection of breast cancer, potentially improving survival rates and reducing the need for invasive treatments. By harnessing the power of deep learning, this research contributes to the advancements of breast cancer diagnostics and holds the potential to positively impact the lives of countless individuals affected by this disease.

Introduction

Breast cancer continues to inflict significant devastation upon millions of women across the globe, particularly through the prevalence of invasive ductal carcinoma (ICD). To ensure effective treatment strategies and improve patient prognosis, the precise classification of breast cancer tumors as either benign or malignant is of utmost importance. In light of this, the objective of this study is to harness the power of deep learning techniques and construct a robust model capable of accurately discerning between benign and malignant breast cancer tumor images. To achieve this, we leverage the comprehensive Breast Histopathology Image dataset to

develop an accurate Convolutional Neural Network. By employing this advanced machine learning algorithm, we aim to enhance our understanding of breast cancer and contribute to the development of more precise diagnostic tools with the potential to positively impact the lives of countless individuals affected by this disease.

Importance of the Problem

Breast cancer represents a significant global health burden, particularly among women. Alarming statistics from the American Cancer Society estimate that in 2023 alone, approximately 297,790 new cases of invasive breast cancer will be diagnosed in women, leading to the deaths of around 43,700 women in the United States. These figures emphasize the urgent need for effective solutions to combat this disease. While our objective does not involve creating a model for clinical use within our time frame and experience level, our aspiration is to explore the field of breast cancer diagnostics and investigate the potential development of a highly accurate AI model for predicting breast cancer classifications.

Current State of the Art Solutions

The current state of the art solution for diagnosing the malignancy of breast cancer tumors relies on a combination of imaging techniques and invasive biopsy procedures. Various biopsy methods, such as fine needle aspiration, core needle aspiration, and surgical biopsy, are utilized to collect tissue samples for analysis by pathologists. The interpretation of these biopsy samples plays a crucial role in determining whether the cells are cancerous (malignant) or non-cancerous (benign). The current method of breast cancer diagnosis typically cannot solely rely on imaging modalities, necessitating the confirmation of malignancy through biopsies.

The prevailing diagnostic solution poses several challenges and consequences that need to be addressed. Human error represents a significant concern, stemming from limitations in screening tools, potential interpretation errors by healthcare professionals, and variations in training and experience within the healthcare sector. These factors can introduce inaccuracies in the diagnostic process, compromising patient outcomes and potentially leading to misdiagnosis or delayed action, particularly in early stages of breast cancer. Delayed detection not only necessitates more aggressive and invasive treatments but also correlates with reduced survival rates. Furthermore, the current diagnostic paradigm is associated with invasiveness and patient

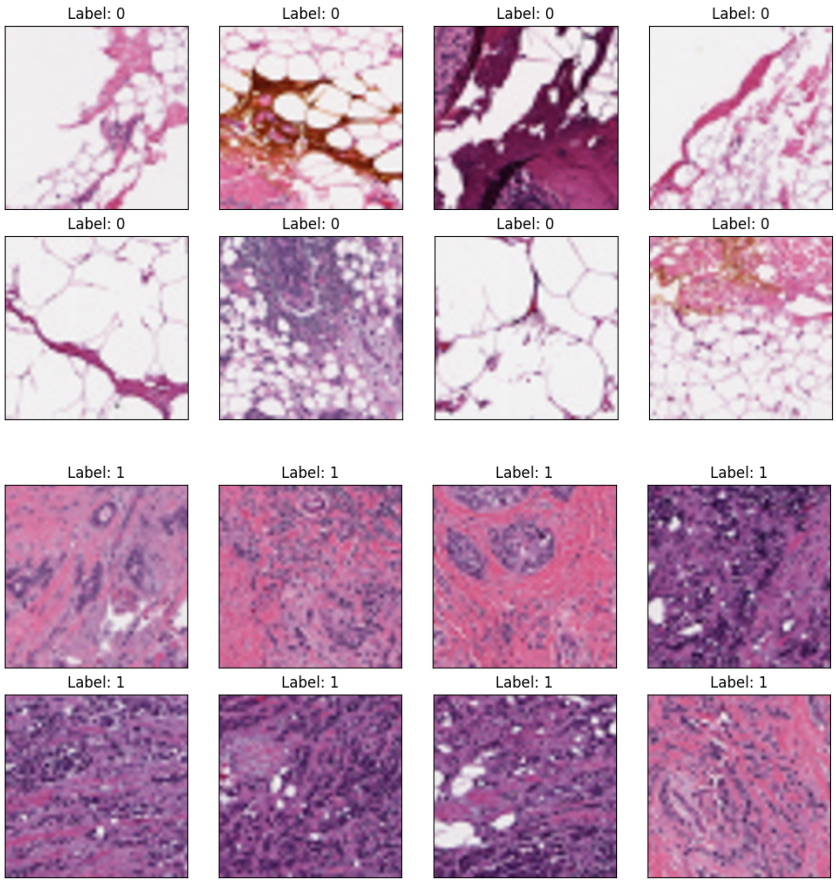
discomfort. Biopsy procedures, by their nature, are intrusive and may cause physical and emotional distress of patients. Moreover, these procedures carry risk of procedural errors, such as the potential for missing cancerous cells or failing to capture the most aggressive regions of a tumor. These errors can result in a false-negative diagnosis, where cancer is present but remains undetected, further exacerbating the potential consequences for patients.

Implementing Machine Learning into the Diagnostic Process

Integrating machine learning algorithms into the diagnostic process offers tremendous potential to mitigate the risk of false negatives and false positives, thereby improving the overall efficiency of breast cancer detection. By harnessing the capabilities of machine learning, the implementation of these algorithms can facilitate earlier and more precise detection, leading to improved survival rates and reduced reliance on invasive treatment modalities. The incorporation of machine learning in diagnostics holds the promise of increased accuracy by circumventing the inherent limitations of human error. Moreover, machine learning models demonstrate exceptional proficiency in identifying intricate features and patterns that may prove challenging for healthcare professionals. Ultimately, the utilization of machine learning techniques can alleviate the discomfort and inaccuracies associated with the traditional biopsy process.

About the Data

The breast histopathology image dataset encompasses a diverse and extensive compilation of valuable resources for the field of breast cancer research. It comprises 162 whole mount slide images of breast cancer specimens, meticulously scanned at 40x magnification. Leveraging these high-resolution images, researchers and medical professionals have extracted 277,524 patches, each sized at 50 x 50 pixels. Within this multitude of patches, there are 198,738 patches identified and categorized as IDC negative, signifying the absence of invasive ductal carcinoma (ICD), while 78,786 patches were classified as ICD positive, indicating the presence of this particular type of breast cancer By encompassing such a comprehensive range of both ICD-positive and ICD-negative cases, this dataset holds immense potential for the development of ur convolutional neural network model.



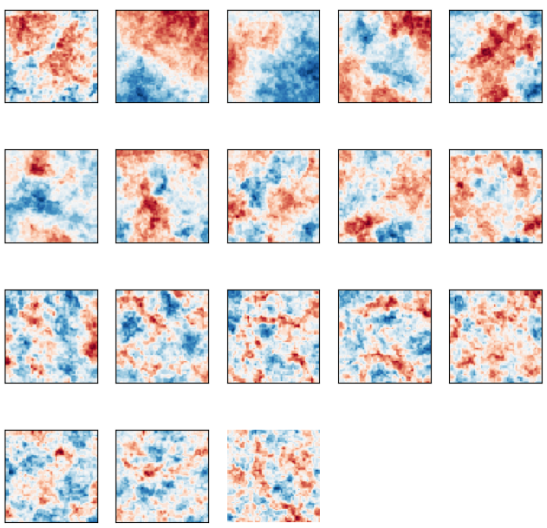
Data Cleaning

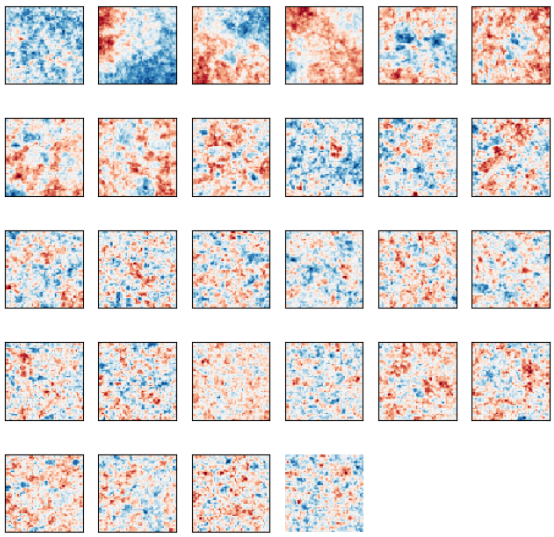
We started off by performing data cleaning on the image dataset to help reduce noise and irrelevant information from the dataset, enhance the efficiency and performance of the convolutional neural network, and ensure consistency and uniformity within the dataset. The first step was importing the data, which we did by specifying the paths to the folders containing the PNG files, obtaining the image paths, and opening the images using the PIL library. We then converted the images to a consistent color format, RGB, and resized the dimensions to 224x224 pixels. This addressed any irregularities that might exist within the images, making the dimensions and colors consistent and uniform. Each image was then converted to a NumPy array, added to the list of images, and assigned a label of 0 (benign) or 1 (malignant). After importing all of the images, the list of images was converted into a NumPy array as well, ensuring compatibility for future analysis. To normalize the pixel values, each pixel in the image array was divided by 255.0, ensuring the values ranged from 0 to 1. Finally, the labels were converted into a numpy array. This data cleaning process ensured a standardized format for the images and labels, enabling subsequent analysis and modeling.

Exploratory Data Analysis

The exploratory analysis focused on using Principal Component Analysis (PCA) to identify the key patterns and variations within the image data for each class. The goal was to find principal components that describe 70% of the total variability for each class.

The code started by preparing the dataset, which consisted of images of benign and malignant breast tumors. The images were organized into separate folders for each class. The filenames of the images in these folders were collected for further analysis. To process the images, they were converted into a format suitable for analysis. This involved resizing the images to a standardized size and converting them to grayscale. The grayscale images were then transformed into numerical arrays, which represented the pixel values of the images. This step ensured that the images were in a consistent format for subsequent analysis. Next, PCA was applied to the image data to identify the most significant patterns within each class. PCA helps in reducing the dimensionality of the data while retaining the most important information. For the benign images, PCA reduced the dimensionality to 18 principal components, while for the malignant images, it reduced to 28 principal components. These principal components represent the main patterns or features present in the images that contribute to the classification of benign and malignant tumors. To visualize the identified patterns, the eigenimages were displayed. Each eigenimage represents a distinct pattern or feature within the image data. By examining these visual representations, we can gain insights into the characteristic patterns that distinguish benign and malignant breast tumors.





The exploratory data analysis revealed that a reduced number of principal components effectively captured these major variations in the benign and malignant breast tumor images. The reduced dimensions suggest that these principal components contain the most important information for distinguishing between the two classes. As you can see in the benign components we produced, the most important features for these tumors appear to be in large cohesive clusters. Conversely, in most important malignant components, the tumor images have much less of a pattern or distinct shape. The visual difference between benign and malignant tumors in pictures can often be observed in their shape, border, texture, size, and changes in surrounding tissue. Benign tumors tend to have a regular shape with smooth borders, a uniform texture similar to surrounding tissue, and smaller size. Malignant tumors may exhibit irregular shapes, jagged or indistinct borders, variations in texture with heterogeneous patterns, larger size, and changes in surrounding tissue.

Overall, this analysis provides a preliminary understanding of the breast cancer image dataset. By examining these visual representations, we can gain insights into the distinctive features that differentiate between benign and malignant tumors. These visualizations can help in understanding the key characteristics of the tumor classes and aid in the development of intuitive and interpretable diagnostic tools. Also, the reduced dimensional representations obtained through PCA could serve as input features for the classification algorithm. By feeding the transformed data into our model, we can potentially improve the accuracy of tumor classification

between benign and malignant cases. The reduced feature space can help mitigate the curse of dimensionality, reduce overfitting, and enhance the model's generalization ability.

Convolutional Neural Network

For our predictive classification model, we decided to develop a convolutional neural network. Convolutional neural networks (CNNs) are widely used for image recognition tasks due to their exceptional ability to capture and exploit the spatial dependencies present in images. We started the development of our neural network by splitting the data into 80% train and 20% test sets with a random state of 42 for predictability. The model was initialized using the 'Sequential' class from the keras library, which generates a sequential stack of layers. The first layer added to the model is a convolutional layer with 32 filters and a kernel size of (3,3). This layer applies a set of learnable filters to the input images, detecting different features through the process of convolving the filters with the image pixels. The activation function used is ReLU (Rectified Linear Unit), which introduces non-linearity and helps capture complex patterns in the data.

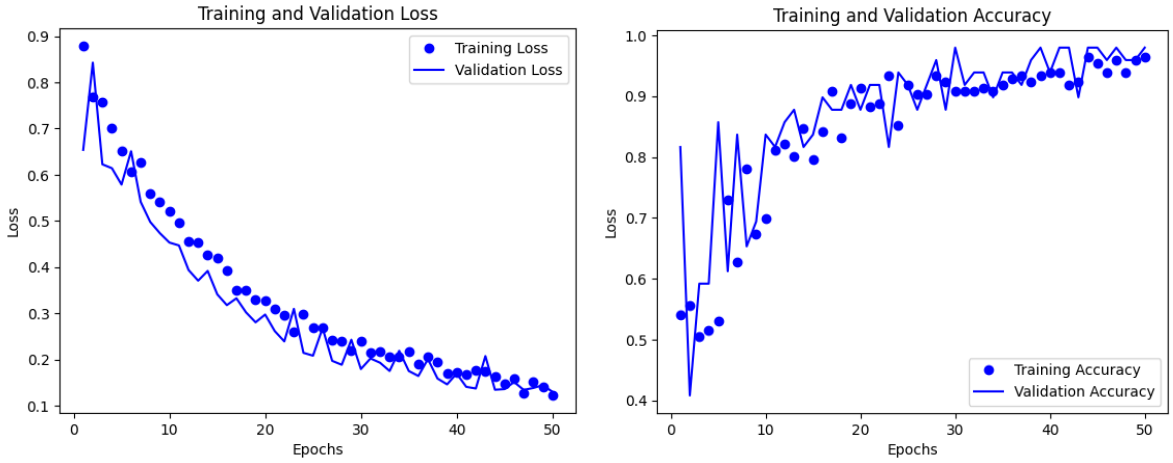
After each convolutional layer, a max pooling layer is added. This layer reduces the spatial dimensions of the representation while retaining the most important features. In this case, a pooling window of size (2,2) is used, which downsamples the input by taking the maximum value within each window. Two more sets of convolutional and max pooling layers are added to capture increasingly complex and abstract features in the image data. The second convolutional layer has 64 filters and the third has 128 filters, both using a kernel size of (3,3) and reLU activation.

We then created our dense, fully connected layers. The output from the convolutional layers is flattened using the flatten layer, converting the multi-dimensional tensor into a one-dimensional vector. The flattened output is then passed through fully connected layers ('Dense'). The first fully connected layer has 1024 neurons and uses ReLU activation. A dropout layer is added with a rate of 0.5 to prevent overfitting by randomly disabling 50% of the neurons during training. The final layer is a fully connected layer with a single neuron, representing the binary classification output. It uses the sigmoid activation function, which produces a probability value between 0 and 1, indicating the likelihood of the input belonging to the positive class.

Once the layers are built the model is compiled, specifying the loss function as binary cross-entropy, which is suitable for binary classification tasks. The Adam optimizer is chosen with a learning rate of 0.0001 and a decay rate of 1e-6 to update the network weights. The accuracy metric is also specified for evaluation during training. The model is then trained using the fit method, passing the validation data. It uses a batch size of 128, shuffling the data during training, and performs training for 50 epochs. The test data is provided to monitor the model's performance on unseen data during training.

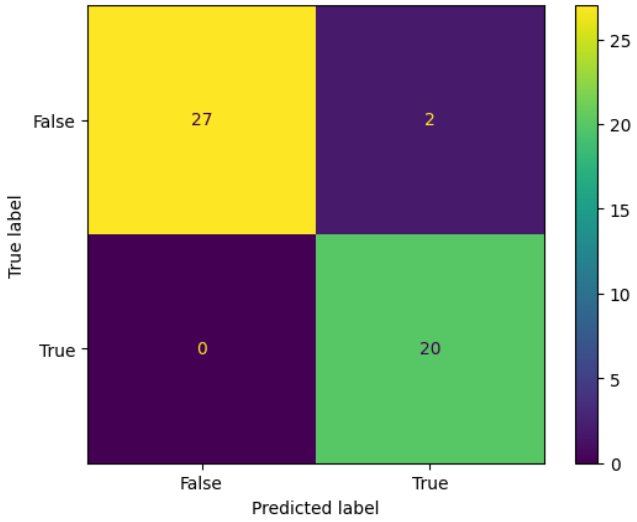
Results

Following the model's training, an evaluation was conducted to gauge its performance on the previously unseen test set. This assessment involved calculating the test accuracy, which yielded a value of 0.959, indicating a high level of accuracy in predicting the correct labels for the test data. The test loss, a measure of the model's performance in terms of the discrepancy between predicted and true labels, was determined to be 0.131.



Additionally, the model's effectiveness was further evaluated using sensitivity and specificity metrics. To obtain these metrics, the model predicted labels for the test data, converting the predicted probabilities into binary labels using a threshold of 0.5. Notably, data points with predicted probabilities exceeding 0.5 were assigned the label 1, while those with probabilities lower than 0.5 were assigned the label 0. The resulting confusion matrix allowed for an overall assessment of the model's performance by comparing the predicted labels against the true labels. The evaluation revealed a sensitivity of 1.0, indicating the model correctly identified

all true positives, and a specificity of 0.931, indicating a high level of accuracy in correctly identifying true negatives.



The presence of false positives and false negatives in medical diagnosis, such as in the case of identifying tumors as benign or malignant, carries significant implications. False positives occur when a tumor is incorrectly identified as cancerous when it is actually benign. This can lead to unnecessary treatments, procedures, biopsies, increased healthcare costs, and unnecessary anxiety for patients. Although our model demonstrated some false positives, it is important to minimize them to ensure patient well-being. On the other hand, false negatives occur when a tumor is incorrectly identified as benign when it is actually malignant. This can result in delayed treatment, which can have serious consequences for patient outcomes and survival rates. Delayed diagnosis allows cancer to progress and spread, making it more challenging to treat and reducing the chances of a successful outcome. It is crucial to avoid false negatives, and in this case, our model's 100% true positive rate indicates that it did not miss any positive cases, which is highly desirable for accurate and timely cancer detection.

Conclusion

In conclusion, our study demonstrates the potential of deep learning techniques, specifically Convolutional Neural Networks (CNNs), in advancing breast cancer classification. By leveraging the comprehensive Breast Histopathology Image dataset, we developed a robust CNN model capable of accurately discerning between benign and malignant breast cancer tumor images. The model achieved a high accuracy of 94% on the test set, with a sensitivity of 1.0 and

specificity of 0.931. These results indicate the model's proficiency in correctly identifying true positives while maintaining a low rate of false positives. By integrating machine learning algorithms into the diagnostic process, we aimed to enhance our understanding of breast cancer and contribute to the development of more precise diagnostic tools.

The significance of our research lies in the potential to improve early detection and facilitate more precise treatment strategies for breast cancer. The use of machine learning techniques, such as the CNN model developed in this study, can aid in the earlier and more accurate identification of breast cancer, potentially leading to improved survival rates and reducing the need for invasive treatments. By harnessing the power of deep learning, we contribute to the advancements in breast cancer diagnostics and offer promise in positively impacting the lives of countless individuals affected by this disease.

However, it is important to acknowledge the limitations of our study. While our model achieved impressive accuracy, sensitivity, and specificity, further validation and testing on diverse datasets and clinical settings are necessary before considering its clinical implementation. Additionally, addressing the challenges associated with interpretability and transparency of deep learning models remains an ongoing research area.

Moving forward, our work opens avenues for future research in breast cancer classification. Refining the model by incorporating additional features, such as genetic data or other clinical parameters, could potentially enhance its performance. Additionally, efforts should focus on further reducing false positive rates to ensure patient well-being and exploring methods to overcome false negatives, such as incorporating multimodal imaging techniques or longitudinal data analysis.

By integrating machine learning algorithms into the diagnostic process, we contribute to the development of more precise diagnostic tools that can aid in early detection and improve patient outcomes. With continued research and validation, such models have the potential to revolutionize breast cancer diagnostics, ultimately benefiting individuals worldwide affected by this devastating disease.

Sources

https://towardsdatascience.com/exploratory-data-analysis-ideas-for-image-classification-d3fc6bbfb2d2

https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images