Student: Alessio Akabuogu

ID Banner: 98191224

Module: Computer Vision and AI

BREAST CANCER

Introduction

Breast cancer is one of the most frequent and severe malignancies that afflict women worldwide. It remains a leading cause of cancer-related mortality, with millions of new cases recorded each. According to the World Health Organisation (WHO), breast cancer is the most often diagnosed cancer in the world and continues to be the top cause of death among women. Therefore, early detection is vitally crucial. Early detection of breast cancer improves treatment efficacy, increases survival rates, and reduces difficulties for patients.

Medical pictures, such as mammograms and histology slides, are important diagnostic tools for breast cancer. These pictures allow medical professionals to thoroughly examine the breast tissue and check for early warning indications. However, deciphering these visuals proves difficult. Even skilled specialists may differ at times, and weariness or stress might complicate the diagnosis. This necessitates the development of robust, dependable computer systems that can assist clinicians in better analysing these pictures.

What the Project Aims to Solve

Benign (not cancerous)

Malignant (cancerous)

This is referred to as a binary classification issue. The objective is to train a computer model to identify visual distinctions between healthy and malignant tissue. Improving this categorisation is critical since incorrect predictions might have severe consequences.

A false negative might cause a delay in treatment, lowering a patient's probability of survival. A false positive might cause undue worry, further testing, and increased expenditures. Because medical diagnosis is so sensitive, improving the accuracy of models like this may directly benefit actual patients and medical professionals alike.

Why Breast Cancer Detection Matters

The urgency of resolving this issue extends well beyond technology; it affects actual individuals and communities.

Early identification of breast cancer has been shown in studies to cut death rates by up to 40%.

Thousands of photos must be examined by pathologists. This wears you out and makes mistakes more likely. By providing a "second opinion," computer models can help them.

Many times, hospitals with little funding are unable to swiftly assess every patient. Physicians may concentrate on more important patients because to automated tools that lessen their burden.

Too few experts exist in emerging nations. By offering diagnostic support in areas where knowledge is scarce, AI-based solutions aid in closing this gap.

For these reasons, one of the most beneficial uses of AI in healthcare now is the categorisation of breast cancer.

Why Deep Learning Is Useful for This Problem

The processing of medical pictures has been altered by deep learning. CNNs can automatically identify patterns in photos without the requirement for human feature design. CNNs are extremely effective in detecting cancer because of this feature.

CNNs can learn characteristics like:

- The texture of cells
- Changes in colour
- Tissue structural alterations: asymmetrical forms or dense clusters of cells

These factors assist the model in determining whether a picture contains cancer. Earlier machine-learning systems were primarily reliant on hand-crafted features, which reduced accuracy. This constraint is removed by deep learning, which learns straight from raw pictures.

According to research, CNN-based models routinely outperform traditional approaches for detecting breast cancer.

This effective strategy is utilised in your project. The CNN algorithm learns from hundreds of pictures before predicting benign or malignant classes.

Understanding the Dataset Structure

The first step in dealing with the dataset is understanding how it is organised within the folder. After downloading the file from Google Drive, you printed the directory contents to determine how many photographs were contained within each folder. The photos in the dataset are divided into two categories: benign and malignant breast cancer. Both the training and testing folders contain these categories. When you tallied the files, you discovered that each training category had 4000 photographs and each testing category had 1000 images. This results in a total of 10,000 photos across the collection. The dataset is cleanly split into class folders, making photos easy to import, read, and label accurately.

Recognizing the Number of Classes and Labels

Once you've determined the folder layout, you'll need to determine how many classes or categories the model needs to learn. This dataset has only two categories: benign and malignant. This transforms the issue into a binary classification challenge, requiring the model to learn to discriminate between two potential outcomes. There is no need for a separate annotation file because the labels for these photographs are automatically generated from their folder names. There are no boundary boxes or segmentation masks since the dataset is only interested in identifying the general class, not individual forms or sections inside the image. The labels are straightforward and easy to use, which reduces training complexity.

Checking Image Formats and Dimensions

Before you can train a model, you must first comprehend the sort of pictures you will be dealing with. Although the dataset does not explicitly specify the picture types, the fact that OpenCV and Keras load them without problem implies that they are JPG images, which are widely accepted. All photos were downsized to 512 × 512 pixels, using three colour channels (RGB). When you construct the data generator, it automatically resizes. All pictures should be the same size so that the neural network can process them consistently and avoid shape-related mistakes. This phase guarantees that the whole dataset is uniform prior to being input into the model.

Identifying the Dataset Source and Quality

The dataset is extracted from a ZIP file saved in your Google Drive. This suggests it's either a well-known public dataset you obtained before or a bespoke dataset you created yourself. Many breast-cancer picture datasets utilised in projects like this are sourced from reputable websites like Kaggle or medical research archives. Because public repositories are often vetted before to publishing, your dataset is likely to be well-annotated and consistent. If it is customary, believability is determined by who produced the labels. Label quality is particularly crucial in medical datasets since inaccurate labels might lead to model misinterpretation. Regardless of the source, your dataset seems consistent since the quantity of photographs is balanced, the images load correctly, and the folder organisation is clean.

Initial Handling and Verification

Before training any model, it is recommended that the dataset be clean and accurate. You accomplished this by developing code to traverse through all of the folders and output the number of files in each one. This ensures that no directories are empty and that all required photos are present. You also showed random sample photos from benign and cancerous folders. This visual check confirms that the pictures are understandable, appropriately coloured, and free of corruption. It also allows you to determine whether the photographs look to be in the correct class. This type of rapid inspection is particularly useful since it keeps you from identifying problems after training has begun.

Creating Data Generators and Preprocessing Images

The following step was to configure the ImageDataGenerator. This utility is used to load photographs in tiny batches, scale the pixel values, and perform helpful changes. In your situation, you used rotation, zooming, shearing, and horizontal flipping. These adjustments produce somewhat different copies of the original photos. This is beneficial since it strengthens the model and reduces the likelihood that it will memorise the training data. You also used the validation_split option to split the dataset into training and validation sections. This guarantees that a subset of the training pictures is set aside for validation, allowing you to check whether the model is learning efficiently or overfitting.

Exploring Class Distribution and Balance

Your dataset is properly balanced, which is one of its assets. Both benign and malignant categories have an equal number of photos. This is useful since unbalanced datasets frequently drive models to favour the class with the most photos. When this happens, the model may provide high accuracy only by forecasting the majority class, which is deceptive. This is especially troublesome in medical applications, where false negatives can be hazardous. A balanced dataset, such as yours, allows the model to learn properly and decreases the possibility of bias.

Using Visual Tools to Understand the Data

Even if your snapshot does not include plots, you may generate various graphs to help you visualise the data. A bar chart indicating the number of photos per class would show two equal bars: one for benign and one for malignant. Because histopathology photos frequently contain numerous darker cell structures, a histogram of pixel intensities would most likely show more dark pixels. To see textural patterns, you may use heatmaps or correlation graphs. These visual tools, while not essential for training, provide a more in-depth understanding of the nature of the data.

Identifying Data Issues and Solutions

Finally, analysing the dataset entails looking for issues like missing files, outliers, and inconsistent samples. In your situation, no missing photographs were discovered, and all folders included the appropriate quantity of files. Minor difficulties, such as differences in illumination or microscope settings, may persist, but these can be addressed by data augmentation and normalisation. Resampling or class-weight changes are unnecessary since the dataset is balanced. The primary preprocessing procedures you used: scaling pixel values, enhancing pictures, and using train-validation splits; are successful for dealing with common difficulties in medical image datasets.

Conclusion

Breast cancer detection is a major worldwide health concern. Deep learning provides strong tools for early diagnosis, reducing human error, and improving healthcare access worldwide. By developing a CNN model to categorise breast tissue pictures, your effort makes a direct contribution to this essential topic. Your work is relevant and influential because it combines strong motivation, a scientific grounding, and defined objectives.

References

Araujo, T., et al. (2017). Classification of breast cancer histology images using convolutional neural networks. PLOS ONE.

Berry, D. A., et al. (2005). Effect of screening and adjuvant therapy on mortality from breast cancer. The New England Journal of Medicine.

Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based deep multiple instance learning. ICML.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. NIPS.

Spanhol, F. A., et al. (2016). Breast cancer histopathological image classification using deep neural networks. IJCNN.

World Health Organization (2021). Breast cancer — Key facts. WHO.