WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Detection

Jhalak Sanjay Maheshwari[1], Vedant Mahesh Chaudhari[2], Anuj Mohan Gujrarthi[3], Nayan Chhotu Rathod[4], Prof. S. M. Pardeshi[5]

[1,2,3,4] UG Student, Computer Science and Engineering (Data Science), R C Patel Institute of Technology, Shirpur, India
[5]Project Guide, Computer Science and Engineering (Data Science), R C Patel Institute of Technology, Shirpur, India
[1]jhalakmaheshwari9@gmail.com, [2]vedantchaudhari1809@gmail.com, [3]anujgujarathi04@gmail.com, [4]nayanrathod.rcpit@gmail.com,

ABSTRACT—

The Convolutional Neural Network (CNN)-based approach to breast cancer detection from medical images is described in this research. The model's aim is to properly classify images into healthy and unhealthy classes. The CNN automatically learns to recognize complex patterns and features predictive of breast cancer using deep learning techniques, which minimizes the need for the manual extraction of features. Multiple tests on reference datasets show how accurate and robust the model is, showing that it could be a useful tool for medical diagnosis. This strategy uses accurate and effective breast cancer detection to increase early detection and improve the treatment of patients. Also, the model's detection performance is much improved when compared to traditional machine learning techniques. By providing radiologists a second opinion while possibly minimizing diagnostic mistakes and workload, the incorporation of this CNN-based technology into existing diagnostic procedures could assist them manage breast cancer more effectively in the years to come.

Keywords— Convolutional Neural Network (CNN), medical images, radiologists.

I. Introduction

In today's medical field, the early diagnosis and treatment of breast cancer is of great significance to improve the survival rate of patients. As the incidence of breast cancer increases year by year, how to use modern technology to improve its prediction accuracy has become an urgent problem to be solved. With the rapid development of artificial intelligence and machine learning technologies, the application of these technologies to the prediction and diagnosis of breast cancer has proven to be an effective aid.

In this project we have developed a Machine Learning model for a detection of healthy and unhealthy cancer patches from medical images extracted from an online dataset, visualize data, and check whether a patient is affected by breast cancer or not and if yes what is the percentage.

The motive of this project is to detect the possible places in the breast images in which cancer can be detected by the dark pink or purple cancer patches in the provided images. The main approach to do this is Convolutional Neural Network (CNN). The project gives a percentage of threat a patient has analysing and processing the images provided.

a) Background

A Breast Cancer Detection project's background usually concentrates on the need for more accurate and efficient early cancer detection techniques. significantly improve patient outcomes. Here's an in-depth look at the background:

i. High Incidence and Impact: One of the most common cancers worldwide is breast cancer. affecting millions people yearly. Increasing treatment results and survival rates needs detection early on.

ii. Limitations of Current Method: While mammography is the best option for breast cancer screening, it has limitations, such as false positives and false negatives as well as patient discomfort.

iii. Advancements in Technology: There is an opportunity to develop more accurate and efficient methods for detecting breast cancer due to developments in machine learning and medical imaging technology

iv. Potential Impact: By improving early detection rates, a successful best cancer detection project could have a major impact on healthcare. decreasing unwarranted biopsies and medical procedures and eventually saving lives

v. Interdisciplinary Collaboration: For the development and implementation of effective and ethical solutions, breast cancer detection projects often involve cooperation between computer scientists, data scientists, medical professionals (radiologists, oncologists), and healthcare policy makers.

b) Motivation

The motivation behind a breast cancer detection projects stems from several key factors:

• Improving early detection

• Reducing false positives

• Reducing false negatives

• Enhancing access to Screening

• Empowering patients.

• Performance Evaluation

• Real-time Monitoring

c) Problem Statement

One of the most common cancers that affect women globally is breast cancer. The increase of treatment outcomes and survival rates is dependent upon early detection. Conventional techniques for identifying breast cancer mainly depend on the human understanding of medical images, like mammograms, which can be difficult and prone to mistakes made by humans.

Convolutional neural networks, or CNNs, have received a lot of attention for their automated breast cancer detection features as a solution to these problems. CNNs are good candidates for analysing medical images such as mammograms because they have shown promising performance in image recognition tasks.

However, building an accurate and reliable CNN-based breast cancer detection system poses several challenges. These challenges include:

i. Data Quality and Quantity: For those who are training CNN models, labelled medical images can be difficult for obtaining around, and getting high-quality labelled datasets can be difficult because of privacy issues and the requirement for professional labels.

ii. Class Imbalance: Class imbalance is a common problem in breast cancer datasets, where the percentage of positive (cancerous) cases is much lower than that of negative (non-cancerous) cases. Developing a strong detection model requires tackling this imbalance in classes.

iii. Interpretability: Since CNNs are often considered as ”black-box” models, it can be challenging to understand the features they pick up and the choices they make. In the medical field, interpretable models are necessary to win over medical professionals and ensure decision-making transparency.

iv. Generalization: The CNN model should resilient to changes in patient information, image quality, and resolution, and it should apply well to new data. Performance issues on real-world datasets may arise from overfitting to the training set or from a failure to generalize.

v. Integration with Clinical Workflow: The aim of the developed CNN model is to help radiologists and other healthcare professionals diagnose patients quickly and accurately. It should fit in seamlessly with the current clinical procedure. Effective deployment techniques and user-friendly interfaces are crucial factors to take into account.

A combined approach merging medical imaging, machine learning, and healthcare domain knowledge is needed to address these issues. Improving patient outcomes and transforming early diagnosis are possible goals of developing a CNN-based breast cancer detection system.

d) Objectives

Creating and implementing an effective Convolutional Neural Network (CNN) model for predicting the presence of breast cancer from medical images, especially mammograms, is the main goal of this work. Particular objectives consist of:

i. High Accuracy: Create a CNN model that can recognize and categorize breast cancer lesions from mammography pictures with precision. To reduce false positives and false negatives, the model should be able to detect both benign and malignant lesions with high sensitivity and specificity.

ii. Robustness: Make sure the CNN model is flexible to changes in patient data, image quality, and resolution. When applied to hidden data, such as datasets collected from different the population and images from different types of imaging, the model should work well.

iii. Interpretability: Improve the CNN model’s interpretability to make it easier for medical professionals to understand. Use methods that demonstrate and explain the features the model has learned in order to throw light about the way it makes decisions.

iv. Scalability: Create a scalable CNN architecture that has the capacity to effectively handle massive data sets. Enhance training protocols and model architectures to handle the growing amount of medical imaging data and improve quick model deployment and iteration.

v. Class Imbalance Handling: recognize and solve issues of class imbalance that are frequently found in medical datasets, especially in the detection of breast cancer. Use techniques like class-weighted loss functions, under sampling, and oversampling to reduce the effect of class imbalance on model performance.

vi. Integration with Clinical Workflow: Ensure a smooth transition between the current surgical procedure and the CNN-based breast cancer detection system. To accelerate the diagnostic process and promote adoption by medical professionals, create user-friendly interfaces and deployment strategies.

vii. Ethical Considerations: When gathering, highlighting, and analyzing medical image data, follow ethical principles and data privacy laws. Maintaining patient trust and safety requires honesty, accountability, and transparency in model development and decision-making.

By achieving these objectives, the proposed CNN-based breast cancer detection system aims to improve early diagnosis, enhance treatment planning, and ultimately contribute to better patient outcomes in breast cancer care

2. Literature Survey

The below table shows the literature reviews of a few research papers:

Year	Authors	Existing Method	Method Used
2017	Gupta and Jain	Transfer Learning	Pre-Trained CNNs
2018	Patel et al	Data Augmentation	Data Augumentation
2019	Kumar and Singh	MultiScale CNN	MultiScale Feature Extraction
2020	Mishra and Reddy	Hybrid CNN	CNN and RNN
2021	Sharma and Gupta	attention-based CNN	attention mechanism
2017	Desai and Shah	Transfer learning	fine tuning pretrained CNN's
2018	Mehta and Patel	Ensemble learning	CNN ensemble with baggy
2019	Joshi and Sharma	Data augmentation	CNN with image rotation
2020	Singh and Agarwal	multi scale CNN	Hierarchical CNN
2021	Patel and Mishra	Hybrid CNN and RNN	bidirectional LSTM
2017	Khan and Gupta	Attention based CNN	CNN with Spatial attention
2018	Mishra and Sharma	Transfer learning	pretrained feature extraction
2019	Jain and Singh	Data augmentation	CNN with Random erasing
2020	Agarwal and Patel	Multi scale CNN	Wavelet transform
2021	Sharma and Verma	hybrid CNN-RNN	Bidirectional GRU
2017	Reddy and Kumar hey	attention based CNN	channel attention
2018	Singh and Gupta	Transfer learning	Fine Tuning
2019	Sharma and Patel	Data augmentation	CNN with mix up
2020	Joshi and Mehta	Multi scale CNN	Dilated convolutions
2020	Shah and Desai	Hybrid CNN - RNN	GRU based attention

Table 2.1 Literature Survey

a) Review of Existing Systems

This section reviews papers based on Breast Cancer Detection using CNN techniques. Papers published in the last ten years is reviewed and analyzed based on the methodologies used.

Breast cancer detection using Convolutional Neural Networks (CNNs) has garnered significant attention due to its potential to improve accuracy and efficiency in diagnosing this prevalent form of cancer. Here's a review of the existing systems:

Deep Learning Architectures: Various CNN architectures have been explored for breast cancer detection, including AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet. These architectures are often adapted or fine-tuned for specific datasets and tasks related to breast cancer detection.

Datasets: Several publicly available datasets, such as the Digital Database for Screening Mammography (DDSM), the Digital Database for Breast Cancer Screening (DDMamm), and the Breast Cancer Histopathological Database (BreakHis), have been widely used for training and testing CNN models. These datasets contain annotated mammography images or histopathological images that facilitate model training.

Preprocessing Techniques: Preprocessing steps like normalization, resizing, and augmentation are often applied to input images to enhance the performance of CNN models. Techniques such as data augmentation help in enlarging the dataset and improving model generalization.

Transfer Learning: Transfer learning, where pre-trained CNN models (e.g., ImageNet pretrained models) are fine-tuned on breast cancer datasets, has shown promising results. This approach allows leveraging features learned from large-scale datasets and adapting them to the task of breast cancer detection, even with limited labeled data.

Architectural Modifications: Researchers have proposed modifications to standard CNN architectures to better suit the task of breast cancer detection. For instance, attention mechanisms, skip connections, and feature fusion layers have been incorporated to enhance the discriminative power of CNNs for identifying cancerous regions in images.

Ensemble Methods: Ensemble methods, such as model averaging and stacking, have been employed to combine predictions from multiple CNN models, leading to improved overall performance and robustness.

Interpretability: Efforts have been made to increase the interpretability of CNN-based breast cancer detection systems. Techniques such as class activation mapping (CAM) and gradientweighted class activation mapping (Grad-CAM) help visualize which regions of the input image are most influential in the model's decision-making process, aiding clinicians in understanding the model's predictions.

Clinical Integration and Validation: Many studies focus on the integration of CNN-based systems into clinical workflows and the validation of their performance in real-world settings. Collaboration with medical experts ensures that the developed models are clinically relevant and trustworthy.

Challenges: Despite the advancements, challenges such as class imbalance, data heterogeneity, and generalization to diverse populations remain significant hurdles in developing robust and reliable CNN-based breast cancer detection systems.

Overall, CNN-based approaches for breast cancer detection hold great promise in improving diagnostic accuracy, efficiency, and accessibility, but continued research and validation efforts are necessary to address existing challenges and ensure the reliability of these systems in clinical practice.

b) Limitations of Existing Systems

While CNN-based systems for breast cancer detection have shown promise, they also have several limitations:

Data Availability and Quality: Access to large, high-quality datasets with diverse representations of breast cancer cases is crucial for training effective CNN models. However, many existing datasets suffer from issues like class imbalance, inconsistent labeling, and variability in image quality, which can hinder model performance and generalization.

Interpretability: CNNs are often regarded as black-box models, meaning that it can be challenging to understand how they arrive at their predictions. Lack of interpretability can be a significant limitation in clinical settings where explanations for model decisions are essential for trust and acceptance by medical professionals.

Generalization to Different Populations: CNN models trained on data from one population may not generalize well to other populations with different demographics, genetics, and healthcare practices. Ensuring the robustness and generalizability of CNN-based systems across diverse populations remains a challenge.

False Positives and False Negatives: Like any diagnostic tool, CNN-based systems are prone to false positives (incorrectly identifying non-cancerous regions as cancerous) and false negatives (missing cancerous regions). Mitigating these errors is critical for ensuring the clinical utility and reliability of breast cancer detection systems.

Limited Spatial Resolution: The spatial resolution of mammography images may not always be sufficient for detecting subtle or early-stage abnormalities, leading to challenges in accurate diagnosis. Improving the spatial resolution of imaging modalities or developing complementary techniques to enhance image quality could address this limitation.

Computational Resources: Training and deploying CNN models for breast cancer detection can be computationally intensive, requiring substantial resources in terms of hardware, memory, and processing power. Deploying these models in resource-constrained environments, such as clinics with limited computational infrastructure, may pose challenges.

Ethical and Privacy Concerns: CNN-based systems raise ethical considerations regarding patient privacy, data security, and potential biases embedded in the data used for training. Ensuring responsible data handling practices and addressing biases in training data are essential for ethical deployment of breast cancer detection systems.

Clinical Validation and Adoption: Despite promising results in research settings, the clinical validation and adoption of CNN-based breast cancer detection systems remain ongoing challenges. Establishing the clinical efficacy, safety, and cost-effectiveness of these systems through rigorous validation studies is essential for their widespread adoption in clinical practice.

Addressing these limitations requires interdisciplinary collaboration between computer scientists, clinicians, medical researchers, and policymakers to develop robust, interpretable, and ethically sound CNN-based systems for breast cancer detection.

III. Proposed Systems

Proposing a system for breast cancer detection using CNN involves several critical steps. Firstly, collecting a diverse dataset of mammography or histopathological images is essential, followed by preprocessing techniques like resizing, normalization, and augmentation to enhance dataset quality and variability.

Selecting an appropriate CNN architecture, such as ResNet or DenseNet, based on task complexity and computational resources, is vital. Leveraging transfer learning with pretrained weights (e.g., from ImageNet) and fine-tuning on the breast cancer dataset can improve detection accuracy.

The dataset should be split into training, validation, and testing sets, using techniques like cross-validation or stratified sampling to ensure representative splits. Defining an appropriate loss function (e.g., binary cross-entropy) and selecting optimizers (e.g., Adam, RMSprop) with tuned hyperparameters is crucial for effective training. Regularization techniques such as L2 regularization and dropout should be applied to prevent overfitting and improve generalization.

Evaluating model performance with metrics like accuracy, precision, recall, F1-score, and AUC is necessary, considering the clinical implications of false positives and negatives. Techniques like class activation mapping (CAM) or Grad-CAM can be used for interpretability, providing clinicians with insights into the model's decisions.

Collaborating with medical experts for clinical validation and conducting prospective studies to assess the system's accuracy, reliability, and impact on decision-making is essential. Ethical considerations related to patient privacy, data security, and dataset biases must be addressed, ensuring compliance with standards like GDPR and HIPAA.

Following these steps, a CNN-based system for breast cancer detection can be developed, validated, and integrated into clinical practice, improving early detection and treatment outcomes.

a) Working of the System

A working system for breast cancer detection using CNN involves several key steps. First, gather a dataset of mammography or histopathological images with corresponding labels indicating the presence or absence of breast cancer. Preprocess the images by resizing them to a uniform size, normalizing pixel values, and applying data augmentation techniques to enhance dataset diversity.

Select an appropriate CNN architecture for breast cancer detection, such as VGG, ResNet, or a custom-designed architecture tailored to breast cancer imaging data. Split the dataset into training, validation, and testing sets. The training set is used to train the CNN model, the validation set tunes hyperparameters and monitors performance during training, and the testing set evaluates the final model performance.

Train the selected CNN model using the training dataset. During training, the model learns to extract relevant features from the input images and predict the presence of breast cancer, with parameters optimized using techniques such as stochastic gradient descent (SGD) or Adam optimization.

Validate the trained model using the validation dataset, monitoring metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess performance and identify potential overfitting. Fine-tune hyperparameters like learning rate, batch size, and regularization strength based on validation set performance to optimize the model's generalization ability.

Evaluate the final trained model on the testing dataset to obtain unbiased performance estimates. Calculate metrics such as accuracy, sensitivity, specificity, and area under the ROC curve to assess the model's breast cancer detection capability.

Deploy the trained and evaluated model into a production environment for breast cancer detection tasks. Integrate the model into existing clinical workflows, radiology systems, or diagnostic tools to assist healthcare professionals in making informed decisions.

Continuously monitor the performance of the deployed model in real-world settings and perform regular maintenance to ensure its effectiveness and reliability. Periodically update the model as new data becomes available or as improvements are made to the underlying CNN architecture.

By following these steps, a working system for breast cancer detection using CNN can be developed and deployed to assist healthcare professionals in accurately and efficiently diagnosing breast cancer.

b) Algorithm

The algorithm for breast cancer detection using CNN involves several steps. Below is a highlevel outline of the algorithm:

i. Data Collection and Preprocessing: Collect a dataset of mammography images or histopathological images along with corresponding labels indicating the presence or absence of breast cancer. Preprocess the images by resizing them to a uniform size, normalizing pixel values, and applying data augmentation techniques to increase the diversity of the dataset.

ii. Data Splitting: Split the dataset into training, validation, and testing sets. Typically, the data is divided into approximately 70-80 percent for training, 10-15 percent for validation, and 10-15 percent for testing.

iii. Model Architecture Selection: Choose an appropriate CNN architecture for the task of breast cancer detection. This could be a pre-existing architecture such as VGG, ResNet, DenseNet, or a custom-designed architecture tailored to the specific characteristics of breast cancer imaging data.

iv. Model Training: Initialize the chosen CNN model with random weights or pretrained weights (e.g., weights pretrained on ImageNet). Train the model using the training dataset. During training, the model learns to extract relevant features from the input images and make predictions about the presence of breast cancer. Use techniques such as stochastic gradient descent (SGD), Adam optimization, or other optimization algorithms to update the model parameters iteratively.

v. Validation and Hyperparameter Tuning: Validate the trained model using the validation dataset. Monitor metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess the model's performance and identify any potential issues such as overfitting. Fine-tune hyperparameters such as learning rate, batch size, and regularization strength based on the performance of the model on the validation set.

vi. Model Evaluation: Evaluate the final trained model on the testing dataset to obtain unbiased estimates of its performance. Calculate evaluation metrics such as accuracy, sensitivity, specificity, and area under the ROC curve to assess the model's ability to detect breast cancer.

vii. Deployment and Integration: Deploy the trained model into a production environment where it can be used for breast cancer detection tasks. Integrate the model into existing clinical workflows, radiology systems, or diagnostic tools to assist healthcare professionals in making informed decisions.

viii. Monitoring and Maintenance: Continuously monitor the performance of the deployed model in real-world settings and perform regular maintenance to ensure its effectiveness and reliability. Update the model periodically as new data becomes available or as improvements are made to the underlying CNN architecture.

This algorithm outlines the main steps involved in developing and deploying a breast cancer detection system using CNN. However, the specifics of implementation may vary depending on factors such as the dataset, computational resources, and clinical requirements.

1. Software Requirements:

Sr. No.	Name of Resource	Specifications
1	Operating System	Windows 10/Windows 11
2	Python	version 3.0 or more
3	Flask	Python(Framework)
4	HTML/CSS	HTML5/CSS3
5	IDE	Google Colab / VS Code / Jupyter Notebook

Table 3.1 Software Requirements

2. Hardware Requirements:

Sr. No.	Name of Resource	Specifications
1	Processor	Intel core i5/AMD Ryzen 5
2	RAM	8GB or more
3	Storage	512GB SSD
4	CPU	3.1 GHz or faster

Table 3.2 Hardware Requirements

IV. Methodology

a) Dataset

We found the breast cancer detection data sets on kaggle containing mammogram or histology images. CNN are then trained on these data sets to automatically learn patterns that differentiate benign and malignant tissues this can potentially lead to faster and more accurate breast cancer screening.

CNNs are image analysis experts that can learn from mammogram data to identify breast cancer. public data sets with labeled images train the CNN to recognize patterns between healthy and cancerous tissue. This helps improve breast cancer screening by providing a fast and potentially more accurate analysis.

We have code to import the dataset from Kaggle, which is stored in Google Drive. The dataset contains a total of 277,524 images, divided into two classes, 0 and 1. The code provided by Kaggle facilitates the extraction of data from the dataset, allowing it to be used for image processing by the model.

b) Prediction and Classification

Many techniques have been proposed for Breast cancer detection. Mainly CNN is used These techniques are outlined as:

Convolutional Neural Network(CNN):

A convolutional neural network (CNN) is a type of artificial neural network used primarily for image recognition and processing, due to its ability to recognize patterns in images. A CNN is a powerful tool but requires millions of labelled data points for training. CNNs must be trained with high-power processors, such as a GPU or an NPU, if they are to produce results quickly enough to be useful.

CNN is a powerful algorithm for image processing. These algorithms are currently the best algorithms we have for the automated processing of images. Many companies use these algorithms to do things like identifying the objects in an image.

Images contain data of RGB combination. Matplotlib can be used to import an image into memory from a file. The computer doesn't see an image, all it sees is an array of numbers. Color images are stored in 3-dimensional arrays. The first two dimensions correspond to the height and width of the image (the number of pixels). The last dimension corresponds to the red, green, and blue colors present in each pixel.

Three Layers of CNN:

Convolutional Neural Networks specialized for applications in image and video recognition. CNN is mainly used in image analysis tasks like Image recognition, Object detection and Segmentation.

There are three types of layers in Convolutional Neural Networks:

Convolutional Layer:

justify In a typical neural network each input neuron is connected to the next hidden layer. In CNN, only a small region of the input layer neurons connect to the neuron hidden layer.

Pooling Layer:

The pooling layer is used to reduce the dimensionality of the feature map. There will be multiple activation and pooling layers inside the hidden layer of the CNN.

Fully-Connected layer:

Fully Connected Layers form the last few layers in the network. The input to the fully connected layer is the output from the final Pooling or Convolutional Layer, which is flattened and then fed into the fully connected layer.

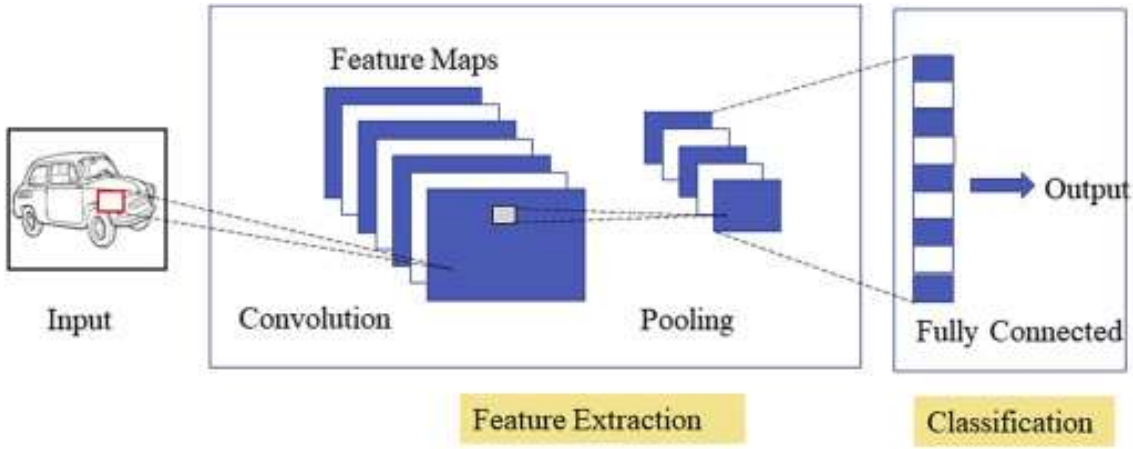

Figure 4.1 CNN

V. Implementation

At first, we will import basic libraries and some required Machine Learning libraries which are given below.

Figure 5.1 gives all the libraries used in the project

```
import os
import numpy as np
import shutil
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import itertools


from skimage.filters import gaussian
from skimage.util import random_noise
import matplotlib.image as mpimg

from sklearn.model_selection import train_test_split



import tensorflow
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from tensorflow.keras.models import Sequential
from tensorflow.keras.models import model_from_json
from tensorflow.keras.layers import Dense
from tensorflow.keras.layers import Conv2D
from tensorflow.keras.layers import MaxPool2D,Dropout,MaxPooling2D
from tensorflow.keras.layers import Flatten
from tensorflow.keras.callbacks import EarlyStopping,ModelCheckpoint
```

Figure 5.1 Libraries

The `shutil` library in Python provides utility functions for data manipulation, which is useful during data preparation for breast cancer detection using CNNs. Although not directly involved in building the CNN model, it assists in organizing and preprocessing data. For instance, `shutil.move()` can relocate disorganized images into structured folders for training, validation, and testing, while `shutil.copy()` can create copies of training images for further processing and augmentation. Additionally, `shutil.move()` can standardize filenames, ensuring consistent data processing.

The `itertools` library offers functions to create iterators, beneficial for data manipulation and custom generator creation during breast cancer detection with CNNs. While not core to the CNN model, `itertools` can enhance data handling and training efficiency. For example, `itertools.islice` can create mini-batches from a larger dataset iterator for efficient training, and it can be combined with other libraries to create custom data generators performing on-the-fly transformations during training.

The `scikit-image` (skimage) library is a powerful tool for image processing, crucial for preparing data for CNN-based breast cancer detection. It provides functionalities for image loading and saving, manipulation (resizing, cropping, rotating, flipping, and color space conversions), filtering (noise reduction, edge detection, sharpening, and smoothing), feature extraction, segmentation, color processing, and mathematical morphology. Skimage is user-friendly, versatile, integrates well with other scientific Python libraries, and is open-source and freely available. In breast cancer detection, it can be used for noise reduction, normalization, segmentation of regions of interest in mammograms, and creating variations of training data for augmentation.

The Gaussian function (normal distribution) is fundamental in statistics and machine learning, influencing breast cancer detection using CNNs in various ways. In data preprocessing, Gaussian distribution-based methods (e.g., `StandardScaler` from Scikit-learn) can standardize pixel values. In feature engineering, Gaussian filters from libraries like Scikit-image or OpenCV can smooth images. In model architecture, Gaussian distribution is used for weight initialization in CNNs to ensure centered weights and avoid biases. Theoretically, Gaussian distribution could model pixel intensity distribution within tissue regions in mammograms.

Overall, `shutil`, `itertools`, `skimage`, and Gaussian functions play crucial roles in data preparation and preprocessing, enhancing the performance and efficiency of CNN-based breast cancer detection systems.

Data Acquisition and Preprocessing:

Data Path: The dataset is organized within a directory named `cancer-rays-dir`.

Patient Subdirectories: Within `cancer-rays-dir`, each patient has a subdirectory named after their patient ID. Each patient subdirectory contains two subdirectories, "0" and "1," representing the absence and presence of Invasive Ductal Carcinoma (IDC), respectively.

Image Files: The subdirectories "0" and "1" contain image files of patches extracted from breast mammograms. The script iterates through each patient directory, copying these image files to a central directory named `all-rays-dir`.

Dataframe Creation: A Pandas dataframe, `data`, is created with a column named `image-id` populated with filenames from `all-rays-dir`.

Target Extraction: The `extract-target` function parses filenames to extract a label indicating the presence (1) or absence (0) of IDC, which is added as a new column, `target`, to the dataframe.

Patient ID Extraction: The `extract-patient-id` function extracts the patient ID from filenames and adds it as a new column, `patient-id`, to the dataframe.

Data Split: The dataframe `data` is explored to visualize the data distribution using Seaborn library functions, showing the number of patches per patient and the percentage of patches with IDC for each patient.

Coordinate Extraction: The `extract-coords` function extracts the x and y coordinates of patch locations within the original mammogram image from filenames. These coordinates are added as columns `x` and `y` to the dataframe.

Feature Engineering:

Function Definitions: Two functions, `get-cancer-dataframe` and `get-patient-dataframe`, are defined to efficiently retrieve data for a specific patient or cancer type (0 or 1) based on the patient ID. These functions create a dataframe containing image metadata (filename, path, x and y coordinates) and the target label.

Patient Visualization: The script iterates through a subset of patient IDs, utilizing the `get-patient-dataframe` function to retrieve corresponding data. It then visualizes the distribution of patches on a scatter plot, with colors representing the presence or absence of IDC.

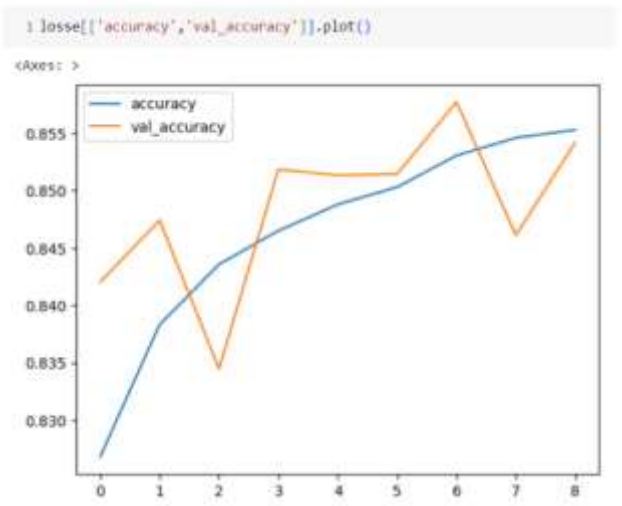

Figure 5.2 accuracy vs val-accuracy

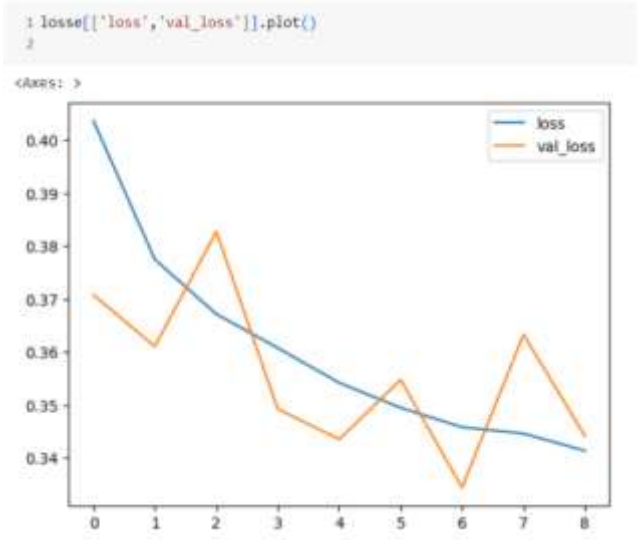

Figure 5.2 loss vs val-loss

Acknowledgments

We extend our sincere gratitude to Prof. S. M. Pardeshi Sir., our guide, for his valuable suggestions and constant encouragement during this project. Special thanks to Dr. Priti Sanjekar and Dr. R. B. Wagh for their support. Thanks to classmates for discussions, and heartfelt appreciation to our family members for their unwavering moral support.

References

[1] G. Muhammad, M. S. Hossain, and N. Kumar, “EEG-based pathology detection for home health monitoring,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 2, pp. 603–610, 2021.

[2] M. Chen, J. Yang, L. Hu, M. S. Hossain, and G. Muhammad, "Urban healthcare big data system based on crowdsourced and cloud-based air quality indicators," IEEE Communications Magazine, vol. 56, no. 11, pp. 14–20, 2018.

[3] M. S. Hossain, "Cloud-supported cyber-physical localization framework for patients monitoring," IEEE Systems Journal, vol. 11, no. 1, pp. 118–127, 2017.

[4] S. A. Alanazi, M. M. Kamruzzaman, M. Alruwaili, N. Alshammari, S. A. Alqahtani, and A. Karime, "Measuring and preventing COVID-19 using the SIR model and machine learning in smart health care," Journal of Healthcare Engineering, vol. 2020, Article ID 8857346, 12 pages, 2020.

[5] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive internet of things: improving sensitivity and interactivity," IEEE Network, vol. 33, no. 3, pp. 58–64, 2019.

[6] M. M. Kamruzzaman, "Architecture of smart health care system using artificial intelligence," in Proceedings of the 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 1–6, London, UK, July 2020.

[7] W. Min, B.-K. Bao, C. Xu, and M. S. Hossain, "Cross-platform multi-modal topic modeling for personalized inter-platform recommendation," IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1787–1801, 2015.

[8] M. M. Kamruzzaman, "Arabic sign language recognition and generating Arabic speech using convolutional neural network," Wireless Communications and Mobile Computing, vol. 2020, Article ID 3685614, 9 pages, 2020.