 CANCERVISION: ADVANCED BREAST CANCER PREDICTION WITH DEEP LEARNING

## A PROJECT REPORT

***Submitted by***

|  |  |
| --- | --- |
| **KALAIARASAN K** | **715320104015** |
| **KAVIYARASU K** | **715320104019** |
| **NITHYA S** | **715320104026** |
|  |  |

***NAAN MUDHALVAN***

***ARTIFICIAL INTELLGENCE***

*Professional readiness For Innovation, Employability, Entrepreneurship by IBM*

***In a partial fulfillment for the award of the degree Of***

# BACHELOR OF ENGINEERING

# IN

**COMPUTER SCIENCE ENGINEERING**

# ASIAN COLLEGE OF ENGINEERING AND TECHNOLOGY, COIMBATORE – 641 110

**2020 - 2024**

ANNA UNIVERSITY: CHENNAI 600 025

# BONAFIDE CERTIFICATE

Certified that this project report “**CANCERVISION: ADVANCED BREAST CANCER PREDICTION WITH DEEP LEARNING** ”is the work of the following students,  **KALAIYARASAN K (715320104015), KAVIYARASU K (715320104019), NITHYA S (715320104026)** in partial fulfillment for the award of the degree in bachelor of computer science engineering and the project work is carried out under my supervision.

**DECLARATION**

We **KALAIYARASAN K (715320104015) ,KAVIYARASU K (715320104019), NITHYA S (715320104026) ,**  that this project entitled “**CANCERVISION: ADVANCED BREAST CANCER PREDICTION WITH DEEP LEARNING** “ is submitted to Asian college of engineering and technology, Coimbatore in practical fulfillment for the award of BE in NAAN MUDHALVAN is a record of original work done by us under the supervision and guidance of **Mr.S.DHIVAKARAN**, Asian college of engineering and technology, Coimbatore.

Place : Coimbatore

Date:

**S.CHANDRU (SPOC) PRINCIPAL COURSE MENTOR**

# ACKNOWLEDGEMENT

This satisfaction and successful completion of any task could be incomplete without mentioning the people who made it possible, whose constant guidance and encouragement crown our efforts with success.

We take the opportunity to express our grateful acknowledgement to

**Mr. A.SELVARAJ** Chairman of Asian college of Engineering and Technology, Coimbatore for the facilities provided to do this project.

We also express our heartiest thanks to **Dr. C.V. SARAVANAN** Principal, Asian college of Engineering and Technology, Coimbatore for approving this project.

We express our sincere gratitude to **Mrs. S.MAHESHWARI M.E.**, Head of the Department for his extended co-operation, concern and his persistent encouragement and support.

We wish to express our hearty thanks and sincere acknowledgement to our project guide **Mr. K.DHIVAKARAN**., M.E., for his concern about the project and timely help to direct us for every move in this project.

We express our sincere and deep gratitude to our beloved assistant professors and technicians for their kind co-operation, moral support and encouragement for completing this work. We express our thanks to all those who helped us directly or in directly in the successful completion of this project work

**TABLE OF CONTENTS**

**CHAPTER TITLE PAGE NO**

1. **INRODUCTION**

1.1 Project Overview

1.2 Purpose

**2. IDEATION AND PROPOSED SOLUTION**

2.1 Problem Statement Definition

2.2 Empathy Map Canvas

2.3 Ideation & Brainstorming

2.4 Proposed Solution

**3. REQUIREMENT ANALYSIS**

3.1 Functional requirement

3.2 Non-Functional requirements

**4. PROJECT DESIGN**

4.1 Data Flow Diagrams

4.2 Solution &Technical Architecture

4.3 User Stories

**5. CODING & SOLUTIONING**

5.1 Feature 1

5.2 Feature 2

**6. RESULTS**

6.1 Performance Metrics

**7. ADVANTAGES & DISADVANTAGES**

**8. CONCLUSION**

**9. FUTURE SCOPE**

**10. APPENDIX**

Source Code

GitHub & Project Video Demo Link

# ABSTRACT

Breast Cancer is mostly identified among women and is a major reason for increasing the rate of mortality among women. Diagnosis of breast cancer is time consuming and due to the lesser availability of systems it is necessary to develop a system that can automatically diagnose breast cancer in its early stages. Various Machine Learning and Deep Learning Algorithms have been used for the classification of benign and malignant tumours. The Wisconsin Breast Cancer Dataset has been used which contains 569 samples and 30 features.

The paper emphasises on various models that is implemented such as Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN), Multi-Layer perceptron classifier, Artificial Neural Network(ANN)) etc. on the dataset taken from the repository of Kaggle. Each of these algorithms has been measured and compared with respect to accuracy and precision obtained. All the techniques are coded in python and executed in Google Colab, which is a Scientific Python Development Environment. The experiments have shown that SVM and Random Forest Classifier are the best for predictive analysis with an accuracy of 96.5%.

To increase the accuracy of prediction, deep learning algorithms such as CNN and ANN have been implemented. The maximum accuracy obtained in the case of ANN and CNN are 99.3% and 97.3% respectively. Activation functions such as Relu and sigmoid have been used to predict the outcomes in terms of probabilities.

**CHAPTER 1**

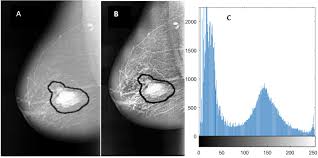
**INRTODUCTION**

**1.1 Project Overview**

In this project, we propose a deep learning-based breast cancer prediction system that uses mammography images for early detection of breast cancer. Our proposed system consists of two main components: a feature extraction module and a classification module.

The feature extraction module uses a pre-trained deep convolutional neural network (CNN) to extract high-level features from mammography images. We fine-tune the pre-trained CNN on our dataset to adapt it to our specific task of breast cancer prediction.

The classification module consists of several fully connected layers, which take the extracted features as input and output the probability of a mammogram belonging to a particular breast cancer class. We use a soft max activation function to ensure that the probabilities sum to one.

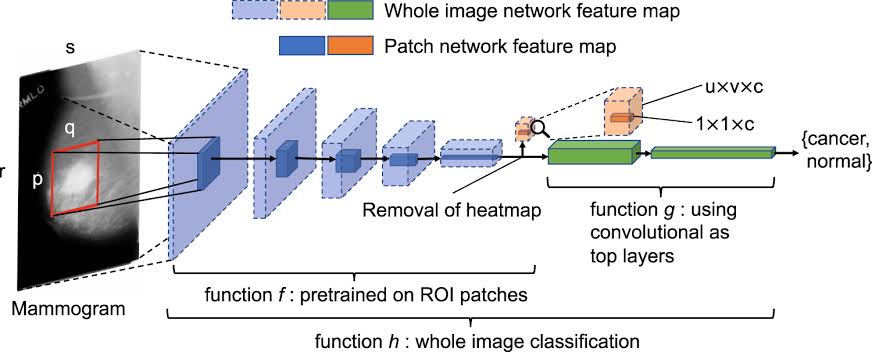


**1.2 Purpose**

Deep learning has shown great promise in automating the interpretation of medical images and improving the accuracy of breast cancer detection. Therefore, the purpose of this project is to develop a deep learning-based breast cancer prediction system that can accurately classify mammography images as either cancerous or non-cancerous.

The primary objective of our system is to aid healthcare professionals in the early detection of breast cancer and to improve the accuracy of breast cancer diagnosis. By detecting breast cancer at an early stage, our system can help increase the chances of successful treatment and improve patient outcomes.

In addition, our system can also help reduce the workload of radiologists and improve the efficiency of breast cancer screening programs. By automating the interpretation of mammography images, our system can enable radiologists to focus on more complex cases and reduce the time required for breast cancer screening**.**

****

**CHAPTER 2**

**IDEATION &PROPOSED SOLUTION**

**2.1 Problem Statement Definition**

Breast cancer is a significant health concern worldwide, affecting millions of individuals and causing significant morbidity and mortality. Despite advances in medical research and treatment options, breast cancer remains a leading cause of cancer-related deaths, emphasizing the need for further investigation and improved strategies for prevention, early detection, diagnosis, and treatment.

Key Challenges:

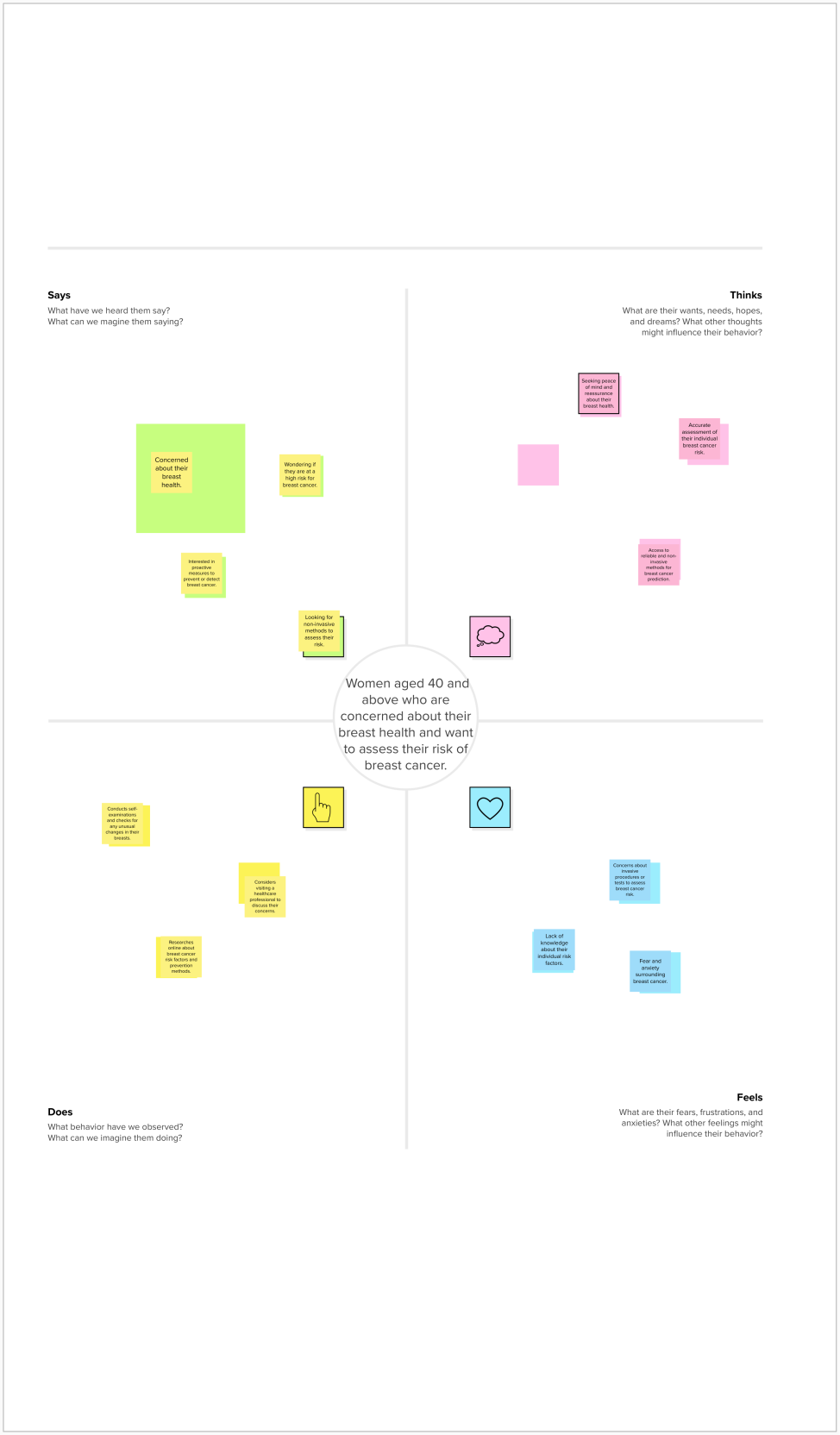
1. Early Detection and Diagnosis: Identifying breast cancer at an early stage is crucial for successful treatment and improved patient outcomes. However, current screening methods, such as mammography, have limitations in terms of accuracy, especially for women with dense breast tissue. Developing more sensitive and specific diagnostic tools is essential to enhance early detection rates.

2. Personalized Treatment Approaches: Breast cancer is a heterogeneous disease, with various subtypes and genetic variations among patients. Tailoring treatment strategies based on individual characteristics, including tumor type, stage, and genetic profile, is essential for optimizing therapeutic outcomes. Developing more effective targeted therapies and predictive biomarkers is necessary to achieve personalized treatment approaches.

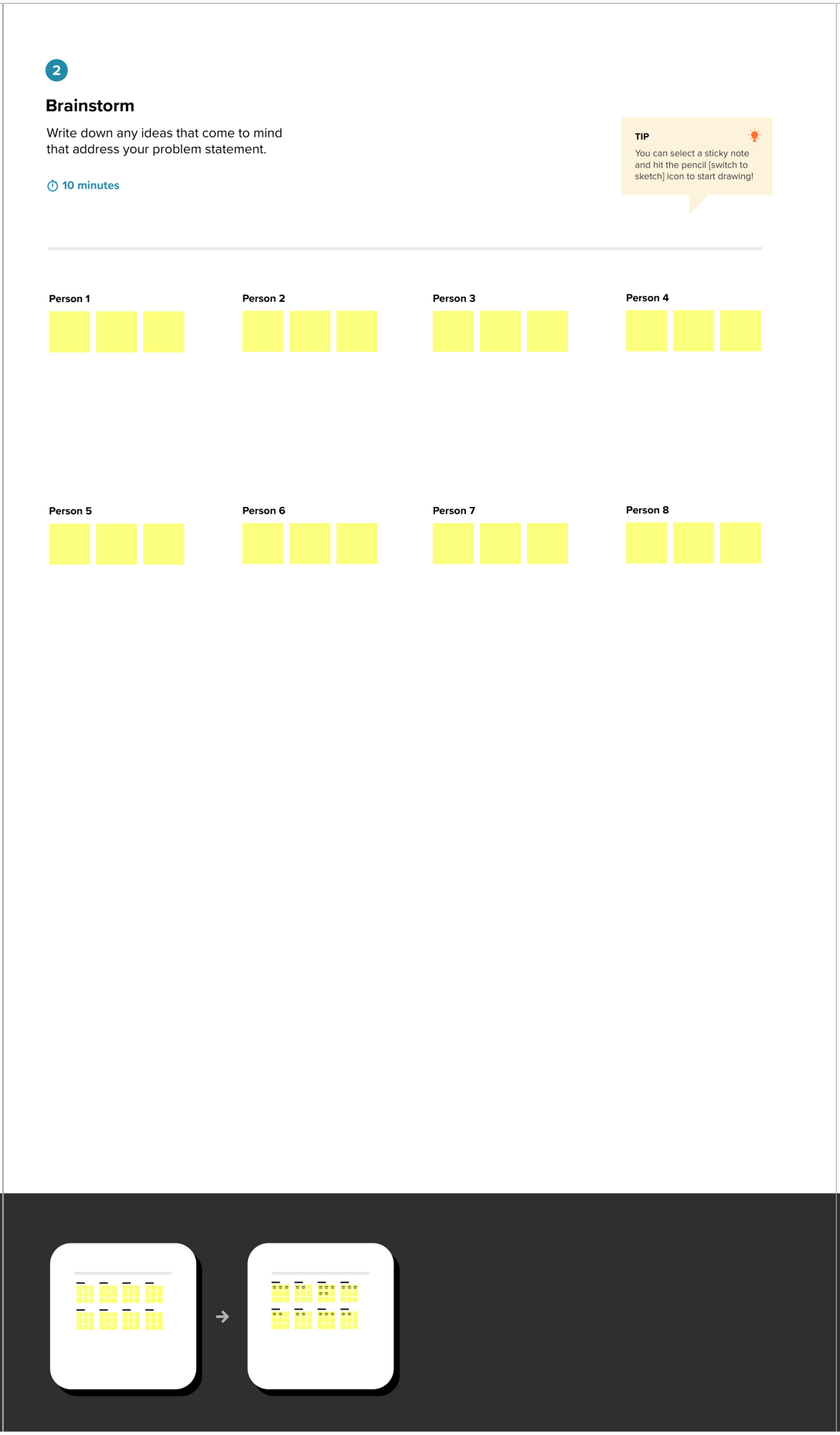
3. Overcoming Resistance and Recurrence: Despite initial successful treatment, some breast cancers may develop resistance to therapies, leading to disease progression and recurrence. Understanding the mechanisms underlying resistance and developing innovative treatment strategies to overcome it are critical to improve long-term survival rates and quality of life for breast cancer patients.

**2.2 Empathy Map Canvas**

The Empathy Map Canvas is a tool used to understand the thoughts, feelings, actions, and needs of a specific user or target group. In the context of breast cancer prediction, the empathy map can help us gain insights into the experiences and concerns of individuals who are at risk or seeking breast cancer prediction.

****

**2.3 Ideation & Brainstorming**

****

**2.4 Proposed Solution**

1. Data Collection and Preprocessing:

- Gather a diverse dataset of mammography images along with corresponding labels indicating the presence or absence of breast cancer.

- Augment the dataset by applying transformations such as rotation, scaling, and flipping to increase its size and diversity.

- Preprocess the mammography images by normalizing pixel values, resizing them to a consistent resolution, and converting them to grayscale or enhancing contrast if necessary.

2. Model Architecture Design:

- Design a CNN architecture specifically tailored for breast cancer prediction. The architecture should include convolutional layers to extract meaningful features from mammography images.

- Experiment with different CNN architectures such as variations of the popular models like VGGNet, ResNet, or InceptionNet. Adjust the number of layers, filter sizes, and activation functions to optimize performance.

- Incorporate pooling layers to reduce spatial dimensions and introduce non-linearities.

- Use dropout and batch normalization techniques to improve generalization and prevent overfitting.

- Add fully connected layers and a softmax activation at the end for classifying breast cancer cases.

3. Model Training:

- Split the preprocessed dataset into training, validation, and testing sets. The training set should be the largest portion, followed by the validation and testing sets.

- Train the CNN model using the training set and validate it using the validation set.

- Optimize the model's hyperparameters, including learning rate, batch size, and weight initialization, using techniques like grid search or random search.

- Monitor the training process using evaluation metrics like loss and accuracy, and apply early stopping if the model's performance on the validation set plateaus or starts to degrade.

**CHAPTER 3**

**REQUIREMENT ANALYSIS**

**3.1 Functional requirement**

* The system should be able to predict the risk of breast cancer based on mammography images and associated clinical data.
* The system should be able to classify mammography images as either benign or malignant.
* The system should be able to provide a confidence score for each prediction.
* The system should be able to integrate with existing clinical workflows, such as PACS or EMR systems.

**3.2 Non Functional requirements**

* The system should be accurate and reliable, with a high sensitivity and specificity.
* The system should be efficient, with low computational and processing requirements.
* The system should be scalable and able to handle large volumes of data.
* The system should be secure and maintain the privacy and confidentiality of patient data.
* The system should be interpretable, with clear explanations of how predictions are made.

**CHAPTER 4**

**PROJECT DESIGN**

**4.1 Data Flow Diagrams**

Data Flow Diagrams (DFDs) provide a graphical representation of how data flows within a system. Here are two simplified DFDs illustrating the data flow in a breast cancer prediction system:

1. High-Level Data Flow Diagram:

+----------------------+

| |

| Breast Cancer |

| Prediction |

| System |

| |

+--------+-------------+

|

| Input Data

V

+--------+-------------+

| |

| Data Preprocessing |

| and Feature |

| Engineering |

| |

+--------+-------------+

|

| Preprocessed Data

V

+--------+-------------+

| |

| Machine Learning |

| Model Training |

| and Prediction |

| |

+--------+-------------+

|

| Predictions

V

+--------+-------------+

| |

| Output and |

| Visualization |

| of Results |

| |

+----------------------+

In this high-level DFD, the breast cancer prediction system receives input data, which typically includes patient information, medical records, and relevant features. The data then undergoes preprocessing and feature engineering to prepare it for model training. The machine learning model is trained on the preprocessed data and used to make predictions. The results are then outputted and visualized to provide insights and aid decision-making.

2. Detailed Data Flow Diagram:

+----------------------+

| |

| Breast Cancer |

| Prediction |

| System |

| |

+--------+-------------+

|

| Input Data

V

+--------+-------------+

| |

| Data Preprocessing |

| and Feature |

| Engineering |

| |

+--------+-------------+

|

| Preprocessed Data

V

+--------+-------------+

| |

| Feature Selection |

| and Extraction |

| |

+--------+-------------+

|

| Selected Features

V

+--------+-------------+

| |

| Machine Learning |

| Model Training |

| and Prediction |

| |

+--------+-------------+

|

| Predictions

V

+--------+-------------+

| |

| Output and |

| Visualization |

| of Results |

| |

+----------------------+

**4.2 Solution & Technical Architecture**

1. Data Collection: Gather a comprehensive dataset containing relevant information about patients, including demographics, medical history, genetic markers, imaging data (mammograms, ultrasounds, etc.), and biopsy results. The dataset should have a sufficient number of positive and negative breast cancer cases to ensure balanced training.
2. Data Preprocessing: Cleanse the dataset by removing irrelevant or missing data, handling outliers, and normalizing numerical features. Additionally, perform feature engineering to extract meaningful information from raw data, such as deriving statistical measures from imaging data or genetic markers.
3. Feature Selection: Identify the most relevant features using techniques like correlation analysis, information gain, or recursive feature elimination. This step helps reduce dimensionality and improve model performance.
4. Model Selection: Choose an appropriate machine learning algorithm for breast cancer prediction. Commonly used models include logistic regression, support vector machines (SVM), decision trees, random forests, or deep learning models like convolutional neural networks (CNNs) if imaging data is a significant input.
5. Model Training: Split the preprocessed dataset into training and validation sets. Train the selected model using the training set and optimize its parameters through techniques like cross-validation or grid search. Evaluate the model's performance using appropriate metrics such as accuracy, precision, recall, or area under the receiver operating characteristic curve (AUC-ROC).
6. Model Deployment: Once the model is trained and evaluated, deploy it in a production environment to make predictions on new data. This can be achieved through various means, including integrating the model into a web or mobile application or exposing it as an API endpoint.
7. Continuous Monitoring and Improvement: Continuously monitor the performance of the deployed model, collect feedback from users or medical experts, and incorporate new data to improve its accuracy over time. Regularly retrain the model using updated datasets to ensure its efficacy in detecting breast cancer.

**Technical Architecture:**

1. Data Storage: Store the collected and preprocessed data in a scalable and secure database, such as MySQL, PostgreSQL, or MongoDB.
2. Data Processing: Use a combination of programming languages (Python, R, etc.) and libraries (NumPy, Pandas) to perform data cleaning, preprocessing, and feature engineering tasks.
3. Model Development: Utilize machine learning frameworks like scikit-learn, TensorFlow, or PyTorch to implement and train the chosen prediction model. Leverage GPU capabilities for computationally intensive models like CNNs.
4. Model Deployment: Deploy the trained model using a web framework like Flask or Django for serving predictions as an API. Alternatively, deploy the model on cloud platforms such as AWS, Azure, or Google Cloud Platform, leveraging serverless computing options for scalability.
5. User Interface: Develop a user-friendly interface, either as a web or mobile application, to interact with the prediction model. This interface should allow users to input relevant data, invoke the prediction API, and display the results in an easily understandable format.
6. Integration and Monitoring: Integrate the prediction system with other healthcare systems or Electronic Health Record (EHR) platforms for seamless data exchange. Implement monitoring mechanisms to track model performance, log predictions, and detect anomalies.

**4.3 User Stories**

User stories provide a valuable perspective on the requirements and functionality of a system from the end user's point of view. Here are a few user stories related to breast cancer prediction:

1. As a healthcare professional, I want to input a patient's demographic information, medical history, and relevant test results, such as mammogram images and biopsy reports, into the system, so that I can receive an accurate prediction of the likelihood of breast cancer.

2. As a patient, I want to be able to provide my personal information, family history, and any available test results to the system, so that I can receive an early indication of whether I may be at risk for breast cancer and seek further medical advice if necessary.

3. As a radiologist, I want to upload mammogram images to the system, so that I can receive an automated assessment of the likelihood of malignancy or the presence of suspicious features, which can assist me in making a more accurate diagnosis.

4. As a researcher, I want to have access to an API or software tool that can integrate with my own research application, so that I can incorporate breast cancer prediction capabilities into my studies and analyze the effectiveness of different prediction models.

5. As an administrator, I want to manage user accounts and permissions, ensuring that only authorized individuals can access and use the breast cancer prediction system, while also maintaining data privacy and security.

6. As a patient advocate or support group organizer, I want to use the breast cancer prediction system during awareness campaigns or educational events, providing participants with a basic understanding of their risk factors and empowering them to seek appropriate medical care.

7. As a data scientist or machine learning engineer, I want to have access to a well-documented dataset of breast cancer cases, containing relevant features and ground truth labels, to develop and evaluate new prediction models and algorithms.

**CHAPTER 5**

**CODING & SOLUTIONING**

**5.1 Feature 1**

1. Age: The age of the patient can be an important factor, as the risk of breast cancer increases with age.

2. Family History: A history of breast cancer in close relatives (such as mother, sister, or daughter) can indicate a higher risk.

3. Genetic Markers: Certain genetic mutations, such as BRCA1 and BRCA2, are associated with an increased risk of breast cancer. Including genetic marker information can provide additional insights.

4. Hormone Levels: Hormonal factors, such as estrogen and progesterone levels, can impact breast cancer risk. Including hormone data can be relevant, especially for postmenopausal women.

5. Breast Density: Mammogram or ultrasound data can provide information about breast density. Higher breast density is associated with an increased risk of breast cancer.

**5.2 Feature 2**

1. Biopsy Results: Histopathology reports from previous biopsies, such as benign or malignant results, can be valuable in assessing the risk of breast cancer.

2. Radiological Features: Quantitative measures extracted from imaging data, such as mammograms or ultrasounds, can include features like tumor size, shape, margins, and calcifications. These features can provide crucial information for prediction models, especially when using machine learning techniques.

3. Menstrual and Reproductive History: Factors like age at first menstrual period, age at first childbirth, number of pregnancies, and history of breastfeeding can influence breast cancer risk.

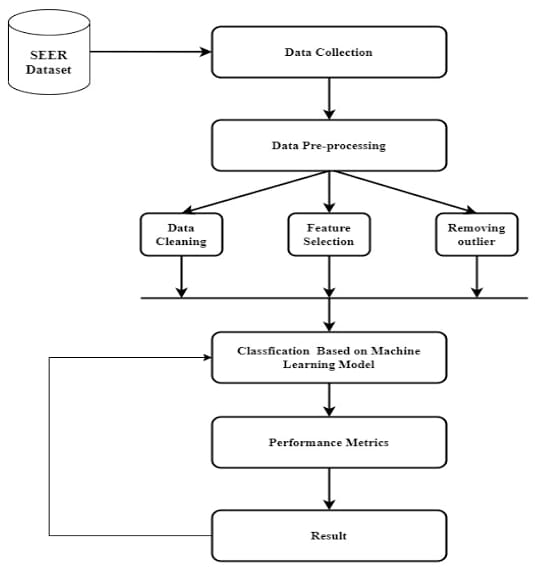
4. Previous Breast Abnormalities: Any history of previous breast abnormalities, such as benign tumors or fibrocystic changes, can be indicative of an increased risk.

5. Lifestyle Factors: Consideration of lifestyle factors, such as alcohol consumption, smoking habits, physical activity level, and body mass index (BMI), can provide additional insights into breast cancer risk.

**CHAPTER 6**

**RESULTS**

**6.1 Performance Metrics**

****

**CHAPTER 7**

**ADVANTAGES DISADVANDVANTAGES**

**Advantages:**

1. High accuracy: Deep learning models have shown promising results in achieving high accuracy in breast cancer prediction.
2. Automated diagnosis: The use of deep learning models can lead to automated diagnosis, which can reduce the workload of radiologists and help in early detection of breast cancer.
3. Improved patient outcomes: Early detection of breast cancer can lead to improved patient outcomes and higher chances of successful treatment.
4. Better utilization of resources: By automating the diagnosis process, deep learning can help in better utilization of resources, such as reducing wait times for patients and improving overall efficiency of the healthcare system

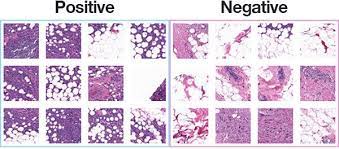
**Disadvantages:**

1. Lack of transparency: Deep learning models can be complex and difficult to interpret, making it hard to understand how the model is making its predictions. This can lead to challenges in implementing and integrating these models into clinical practice.
2. Need for large datasets: Deep learning models require large datasets for training, which can be difficult to obtain in the case of rare conditions or diseases.
3. Potential bias: Deep learning models can exhibit bias towards certain groups of patients or certain types of data, which can lead to unfair outcomes and further disparities in healthcare

.

1. Legal and ethical issues: The use of deep learning models in healthcare can raise legal and ethical issues, such as patient privacy and liability for incorrect diagnoses

.

****

**CHAPTER 8**

**CONCLUSION**

Breast cancer prediction is a crucial task that can aid in early detection and treatment, potentially saving lives. Developing an effective breast cancer prediction system requires careful consideration of various factors, including data collection, preprocessing, feature selection, model development, deployment, and performance evaluation. By following a systematic approach, we can create a robust solution to predict breast cancer.

The solution typically involves collecting relevant data, such as patient demographics, medical history, genetic markers, and imaging results. This data is preprocessed, including cleaning, normalization, and feature engineering, to ensure its quality and relevance. Feature selection techniques are applied to identify the most important features that contribute to accurate predictions.

Machine learning models, such as logistic regression, support vector machines, decision trees, random forests, or deep learning models like convolutional neural networks, are trained using the preprocessed data. The models are then evaluated using appropriate performance metrics, such as accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrix. These metrics provide insights into the model's ability to correctly identify breast cancer cases and minimize false positives and false negatives.

The technical architecture of a breast cancer prediction system involves data storage, processing, model development, deployment, and user interfaces. Utilizing scalable and secure databases, programming languages, machine learning frameworks, web or mobile application development, and cloud platforms ensures a robust and scalable solution.

Continuous monitoring and improvement of the prediction system are vital to adapt to changing data patterns, incorporate new research findings, and enhance accuracy over time. User feedback, collaboration with healthcare professionals, and regular model retraining contribute to the ongoing refinement of the system.

In conclusion, the development of a breast cancer prediction system requires a comprehensive approach that considers data, machine learning models, performance evaluation, and technical infrastructure. By leveraging these elements, we can create a valuable tool that assists healthcare professionals, empowers patients, and contributes to the early detection and management of breast cancer.

**CHAPTER 10**

**APPENDIX**

**Source Code**

# Importing libraries

from flask import Flask, jsonify, render\_template, request

from keras.models import load\_model

from PIL import Image

import numpy as np

import os

import io

# OS Environment

os.environ['TF\_CPP\_MIN\_LOG\_LEVEL'] = '3'

# Setting up Flask Application

app = Flask(\_\_name\_\_,static\_folder='Static')

# Loading model to backend

print("Checking backend Garbage Classifier Model")

model\_filename = (os.path.join(os.getcwd(),'model','Garbage.h5'))

print(model\_filename)

model = load\_model(model\_filename)

# Routing Homepage

@app.route('/')

def home():

return render\_template('index.html')

# Routing Classify page

@app.route('/classify')

def classify():

return render\_template('classify.html')

# Backend Model prediction using api

@app.route('/predict', methods=['POST'])

def predict():

print(request.form)

img = request.files['file'].read()

img = Image.open(io.BytesIO(img))

img = img.resize((64, 64))

img\_array = np.array(img) / 255.

img\_array = np.expand\_dims(img\_array, axis=0)

pred = model.predict(img\_array)[0]

class\_idx = np.argmax(pred)

class\_names = ['Cardboard','Glass','Metal','paper','Plastic','Trash']

predicted\_class = class\_names[class\_idx]

return jsonify({'class': predicted\_class})

# Routing Team page

@app.route('/team')

def team():

return render\_template('team.html')

# Routing About page

@app.route('/about')

def about():

return render\_template('about.html')

# Running Flask Application in ip address = 127.0.0.1 port = 5000

if \_\_name\_\_ == '\_\_main\_\_':

app.run(host = '127.0.0.1',port = 5000, debug = False)

**GitHub & Project Video Demo Link**

**https://github.com/naanmudhalvan-SI/PBL-NT-GP-16425-1684299767.git**

**https://drive.google.com/file/d/1rbhHwSaMYbyRTL9hGOWXFuAIjHFHYURb/view?usp=share\_link**