TABLE OF CONTENTS

ABSTRACT

CANCER – a 6 lettered word, which is common to the world nowadays, is a major cause of death among the population. Cancer is a disease that occurs when cells in the body grow and divide uncontrollably, leading to the formation of abnormal masses of tissue called tumors. These tumors can be either benign (non-cancerous) or malignant (cancerous) and can grow and spread to other parts of the body. Cancer has identified a diverse condition of several various subtypes. The timely screening and course of treatment of a cancer form is now a research requirement, because it supports the medical treatment of patients.

One such forms of cancer is the Breast Cancer, which is a common mostly among women. Breast Cancer symptoms include changed genes, excruciating pain, size and shape, variations in the color (redness) of the breasts, and changes in the texture of the skin. For the prediction, machine learning methods are employed. Consequently, high accuracy in prediction is an important standard for patients' survivability. Machine learning techniques can largely contribute to the process of prediction and early diagnosis of breast cancer. It became a research hotspot and has been proved as a strong technique.

Aim of the Project

The project aims to develop a prediction model for the early diagnosis of breast cancer using Deep Learning and Stacked Long Short Term Memory (LSTM). And, to find out the accuracy with respect to confusion matrix, accuracy and precision.

Objective of the Project

- Loading Dataset
- Pre-processing the dataset
- Designing the predictive LSTM model
- Visualization
- Testing and Visualizing accuracy
- Deployment on streamlit

1. INTRODUCTION

1.1 PURPOSE

The purpose of the project is to devise an effective way to create a prediction model in order to determine, according to the dataset, which values are malignant and which are benign.

This would in turn, pave the way for early diagnosis of the tumor, there by reducing the chances of death rate and the costs of treatment.

1.2 PRODUCT SCOPE

The project works by loading the dataset into the model, extracts the features, applies the Artificial Neural Network and Transfer Learning – stacked LSTM (Long Short Term Memory). Model deployed on streamlit, accesses the API and returns the result to the user.

1.3 PRODUCT FEATURES

1.3.1 Functional Requirements

- The system shall be able to pre-process and clean the dataset to ensure accuracy and consistency; and, be able to perform transfer learning with LSTM to develop a model.

- The system shall be able to accept new patient data as input and use the model to predict the likelihood of the patient developing breast cancer.

- The system shall be able to display the prediction results in a clear and understandable format, with visualizations.

1.3.2 Non-Functional Requirements

- **Performance Requirements**

  - The system should have a response time of less than 5 seconds when making predictions.
  - The system should have a prediction accuracy of at least 80%.

- **Software Quality Attributes**
    - Maintainability - The web app should be easy to modify and maintain over time.
    - Reliability - The system should be able to handle new feature vectors and errors with minimum disruption to operations.
    - Usability - The web application should be easy to use and has a friendly UI.
    - Scalability - The system should be able to handle various types of data without compromising performance.
    - Testability - The system should have a testing framework to validate the accuracy and performance of the model.
- **Interface Requirements**
    - **User Interfaces:**

      The website interface gives 30 input boxes displaying the possible option for input feature vectors. Along with these, it also contains two buttons predicting the data. After clicking on one of the button, a text is displayed whether the tumour is cancerous or non-cancerous.
    - **Software Interfaces :**
        - The system should be developed using Python programming language.
        - The system should be able to integrate with ML/DL libraries such as Keras, TensorFlow and streamlit for web app.
        - The system should be compatible with the operating systems such as Windows, MacOS, or Linux.

2. METHODOLOGY

2.1 Deep Learning

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. It eliminates some of data pre-processing that is typically involved with machine learning. These algorithms can ingest and process unstructured data, like text and images, and it automates feature extraction, removing some of the dependencies on human experts. Then, through the processes of gradient descent and backpropagation, the deep learning algorithm adjusts and fits itself for accuracy, allowing it to make predictions about a new feature vector with an increased precision and accuracy value.

2.2 Artificial Neural Network (ANN)

ANNs or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

ANNs are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.

2.3 Transfer Learning

Transfer learning is a machine learning technique where a pre-trained model developed for one task is reused as a starting point for a different but related task. Instead of training a new model from scratch, the pre-trained model can be used as a starting point to extract useful features and patterns that are relevant to the new task. This can help to improve the performance and reduce the training time of the new model, especially when the amount of data available for the new task is limited. Transfer learning is widely used in image and speech recognition, natural language processing etc.

2.4 Long Short Term Memory (LSTM)

It is a type of Recurrent Neural Network (RNN) that is designed to address the limitations of traditional RNNs, which have difficulty retaining long-term dependencies in sequential data. LSTM achieves this by using a memory cell that can store information for an extended period, along with a set of gates that control the flow of information into and out of the cell. These gates allow LSTM to selectively forget or remember information from the past, and update the memory cell with new information as needed. It is generally used with textual data for NLP.

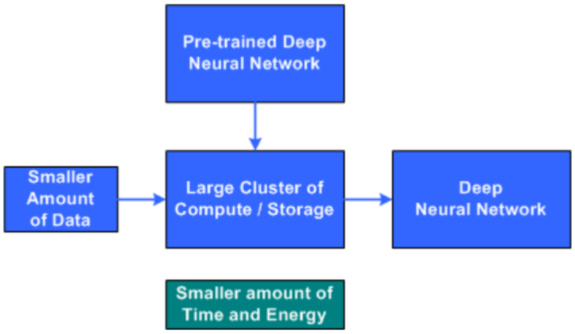

Pre-trained Deep Neural Network
Smaller Amount of Data
Large Cluster of Compute / Storage
Deep Neural Network
Smaller amount of Time and Energy

Figure1: Transfer Learning

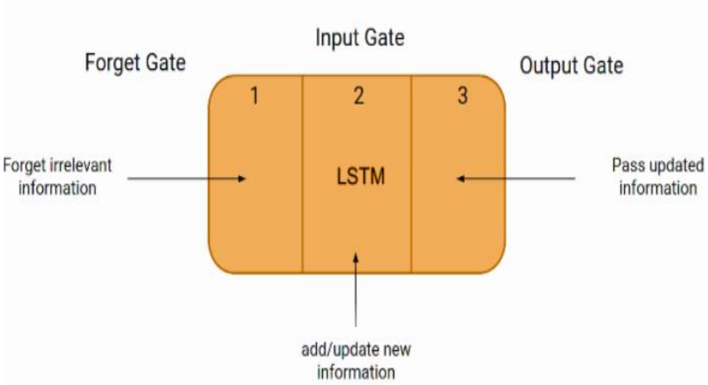


Figure 2: LSTM

2.5 DataSet

The Wisconsin Breast Cancer Dataset (WBCD) is taken from the UCI Machine Learning Repository. This dataset utilized in this project was created by Dr. William H. Wolberg of the University of Wisconsin Hospital, Madison, Wisconsin, United States. The dataset contains 357 benign and 212 malignant breast cancer data respectively. It comprises of 32 columns, with the ID number being the first column and the diagnosis outcome (0-benign and 1-malignant) being the second column. These features represent the shape and size of the target cancer cell nucleus. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure in biopsy test. For each cell nucleus, these features are determined by analyzing under a microscope in a pathology laboratory.

The 10 real-valued features are described in the Table 1:

FEATURE NAME	FEATURE DESCRIPTION
Radius	Average distance from centre of the cancer nucleus to circumference points.
Texture	Standard Deviation of Gray Scaled Value.
Perimeter	Gross distance between Snake Points.
Area	Total pixels on inside of snake along with one half of pixels in the circumference.
Smoothness	Local variance in radii length, quantified by calculating difference length.
Compactness	Perimeter^2 / Area.
Concavity	Intensity of Contour Concave Points.
Concave Points	Number of contour concavities.
Symmetry	Difference in length between lines perpendicular to major axis in both directions of cell boundary.
Fractal Dimensions	Coastline Estimation. Higher value means Higher risk of being cancerous

Table 1: Feature Description

3. SYSTEM ANALYSIS

3.1 HARDWARE REQUIREMENTS:

1. Processor: Intel core i5 or higher

2. RAM: 8 GB or more

3. ROM: 1 TB or more

4. CPU Frequency: 2.0 GHz or more

3.2 SOFTWARE REQUIREMENTS:

1. Operating System: Windows 10 or higher

2. Integrated Development Environment (IDE):

   For Model Building: Jupyter Notebook / Google Colab

   For Model Deployment: Visual Studio Code

3. Libraries:

   For Model Building: Tensorflow, Keras, Seaborn, Matplotlib

   For Model Deployment: Streamlit

SYSTEM DESIGN & SPECIFICATIONS

ACTIVTY DIAGRAM

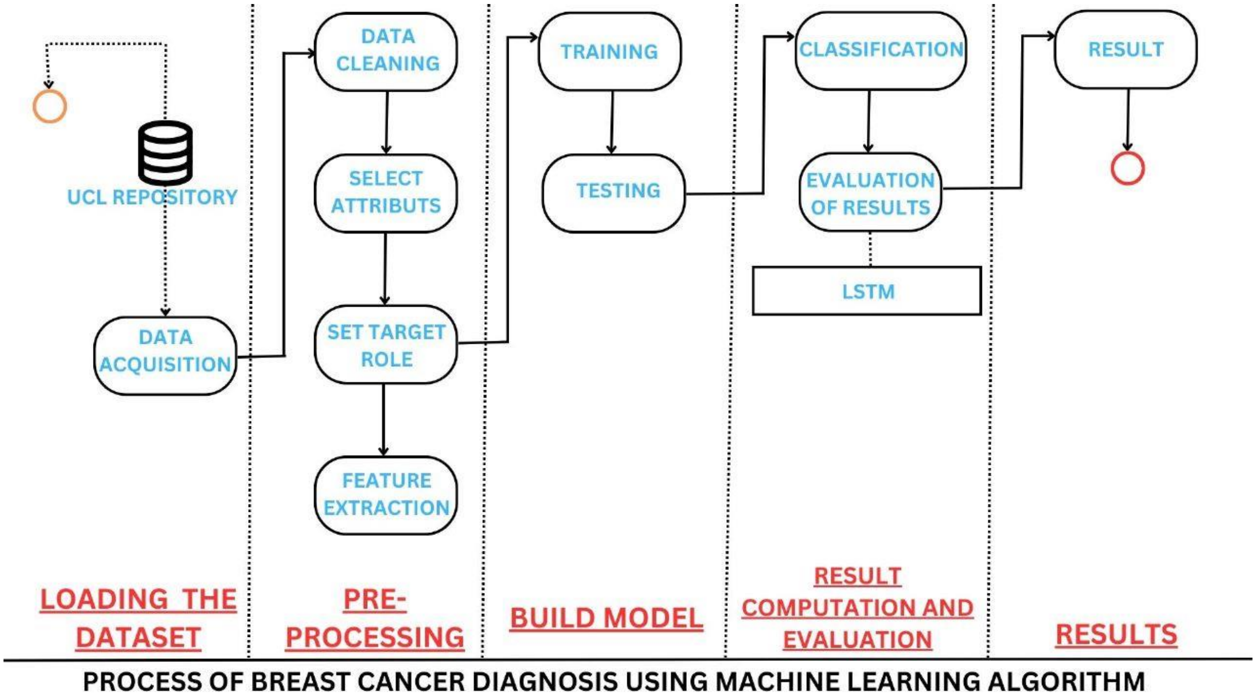
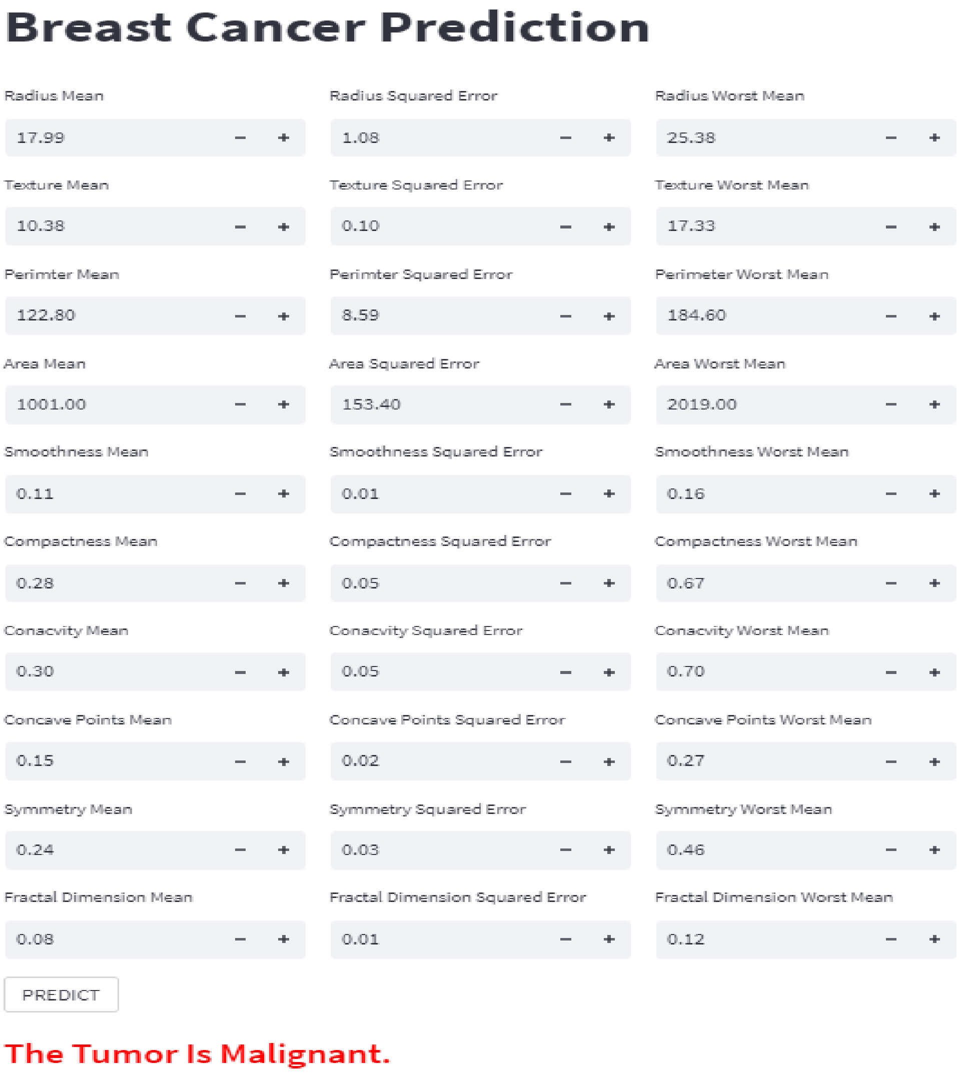


Figure 3: Activity Diagram

TESTING

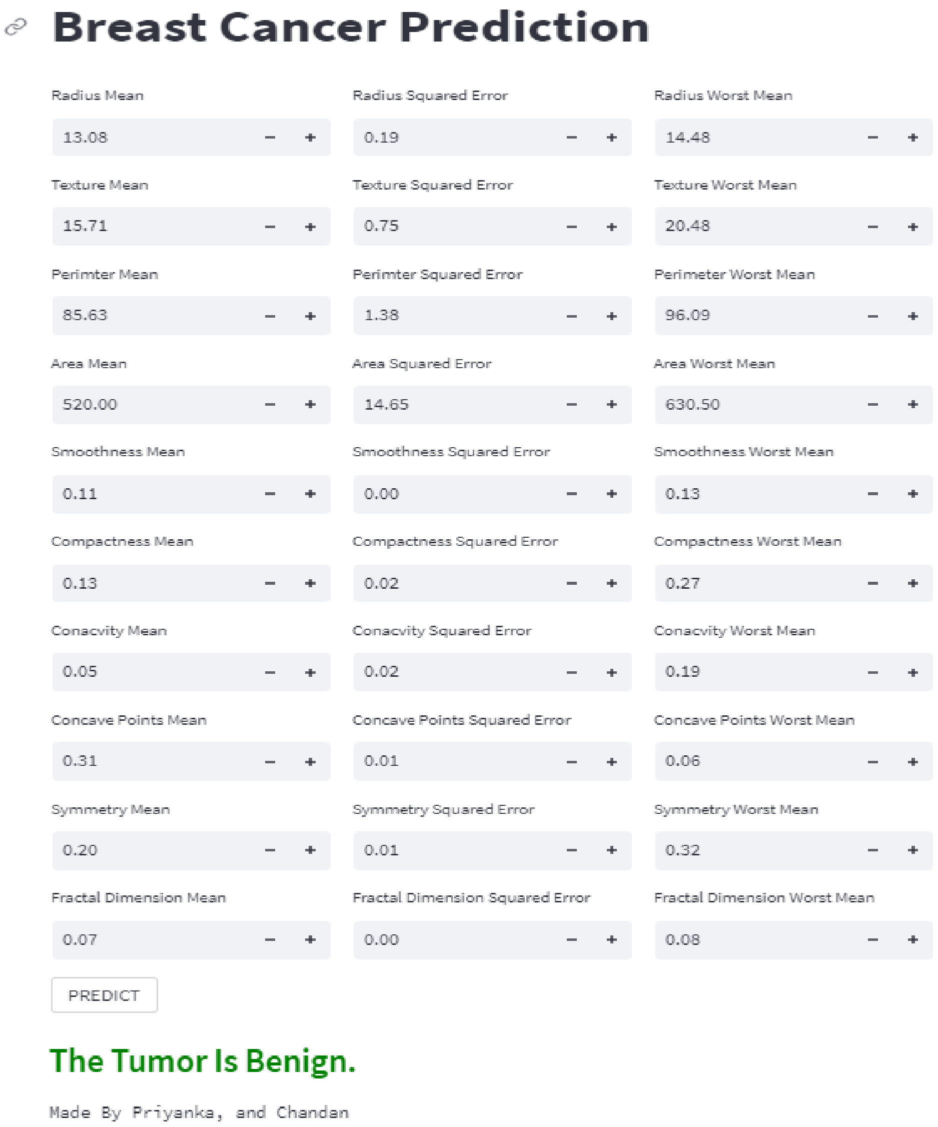


Breast Cancer Prediction

Radius Mean 17.99	Radius Squared Error 1.08	Radius Worst Mean 25.38
Texture Mean 10.38	Texture Squared Error 0.10	Texture Worst Mean 17.33
Perimter Mean 122.80	Perimter Squared Error 8.59	Perimeter Worst Mean 184.60
Area Mean 1001.00	Area Squared Error 153.40	Area Worst Mean 2019.00
Smoothness Mean 0.11	Smoothness Squared Error 0.01	Smoothness Worst Mean 0.16
Compactness Mean 0.28	Compactness Squared Error 0.05	Compactness Worst Mean 0.67
Conacvity Mean 0.30	Conacvity Squared Error 0.05	Conacvity Worst Mean 0.70
Concave Points Mean 0.15	Concave Points Squared Error 0.02	Concave Points Worst Mean 0.27
Symmetry Mean 0.24	Symmetry Squared Error 0.03	Symmetry Worst Mean 0.46
Fractal Dimension Mean 0.08	Fractal Dimension Squared Error 0.01	Fractal Dimension Worst Mean 0.12

The Tumor Is Malignant.

Figure 4: Malignant Detection Output



Breast Cancer Prediction

Radius Mean	13.08	Radius Squared Error	0.19	Radius Worst Mean	14.48
Texture Mean	15.71	Texture Squared Error	0.75	Texture Worst Mean	20.48
Perimter Mean	85.63	Perimter Squared Error	1.38	Perimeter Worst Mean	96.09
Area Mean	520.00	Area Squared Error	14.65	Area Worst Mean	630.50
Smoothness Mean	0.11	Smoothness Squared Error	0.00	Smoothness Worst Mean	0.13
Compactness Mean	0.13	Compactness Squared Error	0.02	Compactness Worst Mean	0.27
Conacvity Mean	0.05	Conacvity Squared Error	0.02	Conacvity Worst Mean	0.19
Concave Points Mean	0.31	Concave Points Squared Error	0.01	Concave Points Worst Mean	0.06
Symmetry Mean	0.20	Symmetry Squared Error	0.01	Symmetry Worst Mean	0.32
Fractal Dimension Mean	0.07	Fractal Dimension Squared Error	0.00	Fractal Dimension Worst Mean	0.08

The Tumor Is Benign.

Made By Priyanka, and Chandan

Figure 5: Benign Detection Output

CONCLUSIONS AND LIMITATIONS

In conclusion, the breast cancer prediction model using LSTM and transfer learning is a promising approach for accurate prediction of breast cancer. The model was trained on the Wisconsin Breast Cancer dataset and was able to achieve high accuracy and sensitivity in detecting breast cancer. The use of LSTM in the transfer learning approach helped in capturing temporal dependencies in the dataset and improving the model's performance.

The successful implementation of this model has significant implications for early detection of breast cancer and can aid in better patient outcomes. Future work can focus on expanding the dataset to include more diverse patient populations and incorporating additional features to improve the model's performance. Overall, this project highlights the potential of machine learning and deep learning approaches in the field of medical diagnosis and prediction.

BIBILIOGRAPHY

UCI ML Repository: https://archive.ics.uci.edu/ml/datasets.php

Streamlit Documentation: https://docs.streamlit.io/

ChatGPT: https://chat.openai.com/

Keras Documentation: https://keras.io/api/layers/recurrent_layers/lstm/

Reference Video: https://youtu.be/WGNI-k20GNo