**BREAST CANCER PREDICTION - A MACHINE LEARNING BASED APPROACH**

**A Project Work Report**

***Submitted in partial Fulfillment for the Award of the Degree***

***of***

**BACHELOR OF TECHNOLOGY**

**in**

**ARTIFICIAL INTELLIGANCE AND DATA SCIENCE**

## Submitted by

**ANDE GUNASHEKAR 23B91A5407**

**BETHAPUDI VIVEK 23B91A5417**

**GUBBALA KIRAN BABU 23B91A5459**

**JAMPANA PURNA SURYA CHANDRA RAJU 23B91A5470**

**Under the esteemed guidance of**

## V.SAIBABA

**Assistant Professor**

****

**DEPARTMENT OF**

**INFORMATIONTECHNOLGY**

**SAGI RAMAKRISHNAM RAJU ENGINEERING COLLEGE(AUTONOMOUS)**

(Approved by AICTE, New Delhi, Affiliated to JNTU University, Kakinada)

##### CHINNAAMIRAM::BHIMAVARAM-534202

###### May-2025

SAGI RAMAKRISHNAM RAJUENGINEERINGCOLLEGE

(AUTONOMOUS)

**(Approved by AICTE, New Delhi, Affiliated to JNTU UNIVERSITY, KAKINADA}**

**CHINNA AMIRAM, BHIMAVARAM-534204**

**DEPARTMENT OF ARTIFICIAL INTELLIGANCE AND DATA SCIENCE**

****

Certificate

**This is to certify that the Project Work report entitled "BREAST CANCER PREDICTION – A MACHINE LEARNING BASED APPROACH ”, is Bonafide work submitted by ANDE GUNASHEKAR (Regd.No:23B91A5407), BETHAPUDI VIVEK(Regd.No:23B91A5417), GUBBALA KIRAN BABU(Regd.No:23B91A5459), JAMPANA PURNA SURYA CHANDRA RAJU (Regd.No:23B91A547) in Information Technology during the Academic year 2024-2025.**

**Head of the Department**

Dr. P RAVI KIRAN VARMA

HOD-IT Dept

**GUIDE**

V. SAI BABA

Assistant Professor

**CERTIFICATION OF EXAMINATION**

This to certify that I have examined the concept and hereby accord my approval of it as a Project Work entitled “**BREAST CANCER PREDICTION – A MACHINE LEARNING BASED APPROACH**” carried out and presented in a manner required for its acceptance on partial fulfillments for the award of the degree of BACHELOR OF TECHNOLOGY in Information Technology for which it has been submitted.

This approval does not necessarily endorse or accept every statement made opinion expressed or conclusions drawn as recorded in the Project Work report it only signifies the acceptance of the report for the purpose for which submitted.

**Signature**

**Project Guide :**

**Senior Faculty :**

**External Examiner :**

**HOD:**

**DECLARATION**

This Project Work report entitled “**BREAST CANCER PREDICTION – A MACHINE LEARNING BASED APPROACH**” has been carried out by us in the partial fulfillment of the requirements for the award of the degree of B.Tech, S.R.K.R Engineering College(A). We hereby declare this project work/project report has not been submitted to any of the other university/Institute for the award of any other degree/diploma.

**Name Register No Signature**

ANDE GUNASHEKAR 23B91A5407

BETHAPUDI VIVEK 23B91A5417

GUBBALA KIRAN BABU 23B91A5459

JAMPANA PURNA SURYA CHANDRA RAJU 23B91A5470

**ACKNOWLEDGEMENT**

The victorious completion of this project would be incomplete without greeting those who made it possible and whose guidance and encouragement made efforts that taken to the success.

The thesis presented here is the work accomplished under the enviable and scholarly guidance of **V. SAI BABA**, Assistant Professor of Information Technology Department. We are grateful for the indomitable support. We express deep gratitude for her inspiring remarks and for helping with this project.

We take this opportunity to express our sincere thanks to **Dr. P RAVI KIRAN VARMA**, Head of the Department, Information Technology, for his support and encouragement.

We express our sincere thanks to **Prof. K V Murali Krishnam Raju**, PRINCIPAL, S.R.K.R Engineering College, Bhimavaram for giving us this opportunity for the successful completion of this project.

Finally, we also express our sincere thanks to other **Teaching and non-Teaching staff** for their support in accomplishing this project.

1. ANDE GUNASHEKAR 23B91A5407
2. BETHAPUDI VIVEK 23B91A5417
3. GUBBALA KIRAN BABU 23B91A5459
4. JAMPANA PURNA SURYA CHANDRA RAJU 23B91A5470

**ABSTRACT**

Breast cancer remains one of the most common and life-threatening cancers affecting women globally. With rising incidence rates, early and accurate diagnosis is critical to improving survival outcomes. Traditional diagnostic techniques, although valuable, can be limited by subjective interpretation and variability in results. As a result, machine learning has emerged as a transformative tool in medical diagnostics. Among the various algorithms explored, the **Random Forest algorithm** stands out due to its high accuracy, robustness to overfitting, and ability to handle complex datasets with multiple features. It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) of the individual trees, making it particularly effective for medical data classification tasks.

This study applies the **Random Forest classifier** to the Breast Cancer dataset to develop a reliable model for distinguishing between malignant and benign tumours. The model achieved an impressive accuracy of **94%**, outperforming several other commonly used classification algorithms. The strength of Random Forest lies in its ensemble nature, where multiple decision trees collectively contribute to improved predictive performance and reduced variance. These characteristics make it especially suitable for use in clinical settings, where diagnostic precision is vital. The results of this study underscore the potential of the Random Forest algorithm as a powerful component in computer-aided diagnosis systems, offering valuable support to healthcare professionals in making informed, timely decisions for breast cancer detection and treatment planning.

**TABLE OF CONTENTS**

**1. INTRODUCTION 1**

**2. PROBLEM DEFINITION 5**

2.1 PROBLEM STATEMENT 5

2.2 PROBLEM DEFINITION 5

**3. EXSISTING SYSTEMS 6**

**4. PROPOSED SYSTEM 7**

**5. SYSTEM REQUIREMENTS 12**

5.1. HARDWARE REQUIREMENTS 12

5.2 SOFTWARE REQUIREMENTS 12

**6. SYSTEM ARCHITECTURE 14**

**7. SYSTEM IMPLEMENTATION 15**

7.1 . DATA PREPARATION 15

7.1.1 DATA COLLECTION 15

7.1.2 PRE-PROCESSING 15

7.2 MODEL SELECTION AND TRAINING 16

7.2.1 MODEL SELECTION 16

7.2.2 TRAINING AND TESTING 16

7.3 EVALUATION METRICS 16

7.3.1 PERFORMANCE ASSESSMENT 16

7.3.2 EXPERIMENTAL RESULTS 17

7.4 . DEPLOYMENT 17

7.4.1 SYSTEM INTEGRATION 17

7.4.2 CONTINUOUS IMPROVEMENT 18

**8. CONCLUSION 19**

**APPENDIX -1**

SOURCE CODE 21

USER INTERFACE OVERVIEW 26

**LIST OF FIGURES**

**S.No Figure No Name of the Figure Page.No**

1 Figure 1 Block Diagram of proposed system 8

2 Figure 2 Pictorial representation Dataset 8

3 Figure 3 System Architecture 16

4 Figure 4 User Interface 26

5 Figure 5 Result Page 27

**LIST OF TABLES**

**S.No Table No Name of the Table Page.No**

1 Table 1 Experimental results 20

**CHAPTER 1**

**INTRODUCTION**

**Background**

Breast cancer remains one of the most prevalent and life-threatening forms of cancer

affecting women globally, with millions of new cases diagnosed each year and a high mortality rate in low-resource settings. According to the World Health Organization (WHO), early diagnosis followed by effective treatment can significantly increase survival chances, reduce healthcare costs, and improve quality of life. The disease not only burdens individual patients but also imposes substantial psychological, emotional, and financial stress on families and public health systems.

Traditional diagnostic techniques—such as physical breast exams, mammography, ultrasound imaging, and biopsies—form the cornerstone of clinical screening and diagnosis. However, these methods present critical limitations. Manual examinations can be inconsistent due to the subjective interpretation of results. Imaging techniques, though more objective, are susceptible to false positives and false negatives, potentially leading to unnecessary biopsies or missed diagnoses. Biopsies, while considered the gold standard for confirming malignancy, are invasive, time-consuming, and stressful for patients]. Furthermore, delays in diagnosis due to healthcare system inefficiencies can negatively impact treatment outcomes.

**Role of Machine Learning in Breast Cancer Diagnosis**

In recent years, the integration of artificial intelligence (AI) and machine learning (ML) into healthcare has shown transformative potential, particularly in disease diagnosis and prognosis. Machine learning algorithms, which are capable of learning patterns and making predictions based on data, can enhance diagnostic accuracy by identifying subtle correlations in medical datasets that may be imperceptible to human clinicians.

When applied to breast cancer diagnosis, ML models can analyse patient attributes (such as cell characteristics from biopsies, demographic information, or genetic data) to predict the likelihood of malignancy. These models provide decision support by minimizing human error, speeding up diagnosis, and reducing the need for invasive procedures. Importantly, ML-driven systems can be deployed in both high- and low-resource settings, making them scalable solutions for global health applications.

Moreover, explainable AI (XAI) techniques are now being incorporated to make machine learning decisions more interpretable to clinicians, thereby increasing trust and adoption in clinical workflows. By continuously learning from new patient data, ML models can also evolve over time to reflect updated clinical practices and population health trends.

**Objectives**

The primary objectives of this project are:

* To explore and apply various machine learning classification algorithms for predicting whether a breast tumour is malignant or benign.
* To design and implement an automated, data-driven system that aids healthcare professionals in making early and accurate breast cancer diagnoses.
* To evaluate and compare the performance of multiple ML models using appropriate statistical metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
* To identify the most effective model that balances predictive performance with computational efficiency and interpretability.

**Scope of the Project**

This project focuses on the design, implementation, and assessment of machine learning models using the publicly available **Wisconsin Breast Cancer Diagnostic (WBCD)** dataset, a well-known benchmark in medical machine learning research. The dataset includes key features derived from digitized images of breast mass fine needle aspirates, describing the characteristics of the cell nuclei present in the samples.

The project encompasses the following components:

* **Data Acquisition & Preprocessing**: Cleaning the dataset, handling missing values, feature selection, and normalization.
* **Model Development**: Training and tuning five distinct classification algorithms—Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Decision Tree, Random Forest, and Naive Bayes.
* **Model Evaluation**: Performing a comparative analysis based on multiple evaluation metrics and visualizing results using confusion matrices, ROC curves, and classification reports.
* **Model Deployment Considerations**: Discussing the potential for integrating the best-performing model into a clinical decision support tool or diagnostic software.

**Advantages**

1. **High Accuracy and Consistency**  
   Machine learning models like SVM can achieve very high accuracy (up to 97% in your case), reducing the chances of misdiagnosis and ensuring consistent results across patients.
2. **Early Detection**  
   ML models can identify subtle patterns in data that may not be visible to human experts, allowing for earlier detection of malignancies which leads to better treatment outcomes.
3. **Speed and Efficiency**  
   Once trained, these models can make predictions within seconds, which is especially valuable in time-sensitive medical environments.
4. **Automation of Diagnosis**  
   Reduces reliance on human interpretation, lowering the risk of diagnostic errors caused by fatigue, bias, or inexperience.
5. **Scalability**Can be deployed across hospitals and diagnostic centers with minimal cost and effort once the model is trained, especially if integrated into cloud-based healthcare systems.
6. **Support for Clinical Decision-Making**  
   Acts as a decision-support system, helping doctors to verify their diagnoses or catch conditions that might be overlooked.

**Limitations**

1. **Data Dependency**  
   Machine learning models are only as good as the data they are trained on. Poor-quality or biased datasets can lead to inaccurate predictions.
2. **Black Box Nature**  
   Some models, particularly complex ones like SVM and neural networks, can lack transparency, making it hard for clinicians to understand or trust the reasoning behind predictions.
3. **Need for Preprocessing**  
   Requires careful data cleaning, feature scaling, and handling of missing values to avoid skewed results.
4. **Generalization Issues**  
   A model trained on a specific dataset may not perform equally well on data from a different population or institution unless properly validated and retrained.
5. **Computational Cost (during training)**  
   Some algorithms can be computationally intensive during training, requiring significant processing power and memory.
6. **Regulatory and Ethical Challenges**  
   Implementing ML in clinical settings must comply with strict healthcare regulations and address concerns like patient data privacy and model accountability.

**CHAPTER 2**

**PROBLEM DEFENITION**

**2.1 Problem Statement**

Breast cancer continues to be a significant health challenge worldwide, with early diagnosis being crucial for effective treatment and improved survival rates. However, traditional diagnostic methods—such as physical examinations, mammograms, and biopsies—are often time-consuming, subjective, and may lead to misdiagnosis due to human error or limited resources. As the volume of patient data increases, the need for automated, accurate, and consistent diagnostic support systems becomes more urgent. This project seeks to address this gap by developing a machine learning-based classification system that can efficiently and accurately predict whether a breast tumour is benign or malignant based on medical features.

**2.2 Problem Definition**

The problem focuses on creating an intelligent, data-driven system that utilizes supervised machine learning algorithms to classify breast cancer tumours using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset. The objective is to implement and compare multiple models—such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Naive Bayes—to determine the most effective approach in terms of diagnostic accuracy and reliability. The system must be capable of handling real-world medical data, ensuring quick predictions, and maintaining high precision and recall to support healthcare professionals in making informed decisions. Ultimately, the goal is to provide a scalable, robust, and clinically relevant solution to improve early breast cancer detection.

**CHAPTER 3**

**EXISTING SYSTEM**

Existing methods for breast cancer diagnosis traditionally involve a combination of clinical assessments and medical imaging techniques. These methods are crucial for detecting and characterizing breast abnormalities, but they also present certain limitations. Key components of the existing system include:

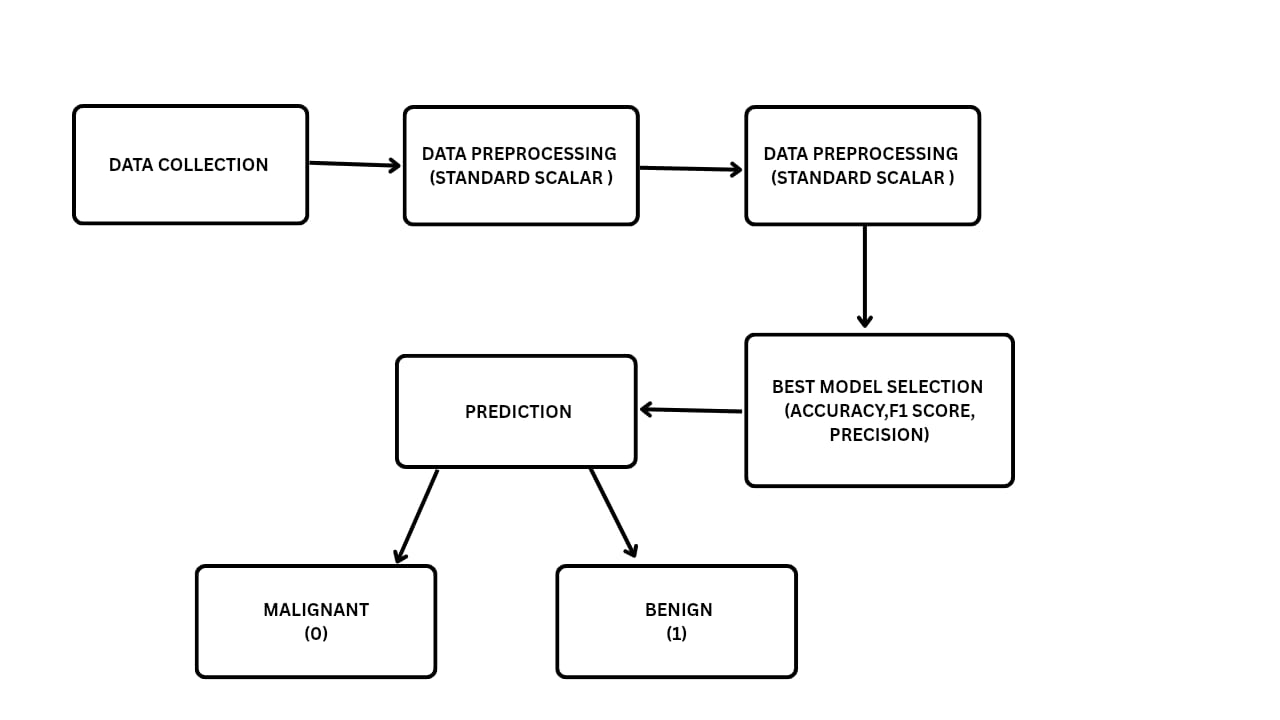
* **Clinical Breast Examination:** This involves a physical examination of the breast by a healthcare professional to check for lumps, changes in size or shape, and other abnormalities. While it's a fundamental first step, its accuracy can vary depending on the examiner's experience and the characteristics of the breast tissue.
* **Mammography:** This technique uses low-dose X-rays to create images of the breast, allowing for the detection of tumors or other anomalies. Mammography is a widely used screening tool, but its sensitivity can be reduced in women with dense breast tissue.
* **Ultrasound:** This imaging method uses sound waves to produce images of breast tissue, often used to further evaluate abnormalities found during a clinical breast exam or mammography.
* **Biopsy:** This is the process of removing a tissue sample from the breast for laboratory examination. Biopsies are typically used to confirm whether a suspicious area is cancerous and to determine the specific type of cancer. While biopsy provides definitive diagnosis, it is an invasive procedure and can involve delays in obtaining results, causing anxiety for patients.

These existing methods, while essential for breast cancer diagnosis, can be time-consuming, subjective, and may involve delays in obtaining results. There is a growing need for more efficient, accurate, and less invasive diagnostic tools to improve early detection and patient outcomes.

**CHAPTER 4**

**PROPOSED SYETEM**

The proposed methodology for this breast cancer prediction project encompasses a systematic sequence of steps designed to accurately classify tumors as either malignant or benign. This comprehensive pipeline integrates data collection, data preprocessing, machine learning model training, and rigorous evaluation techniques to ensure the development of a robust and effective prediction system.

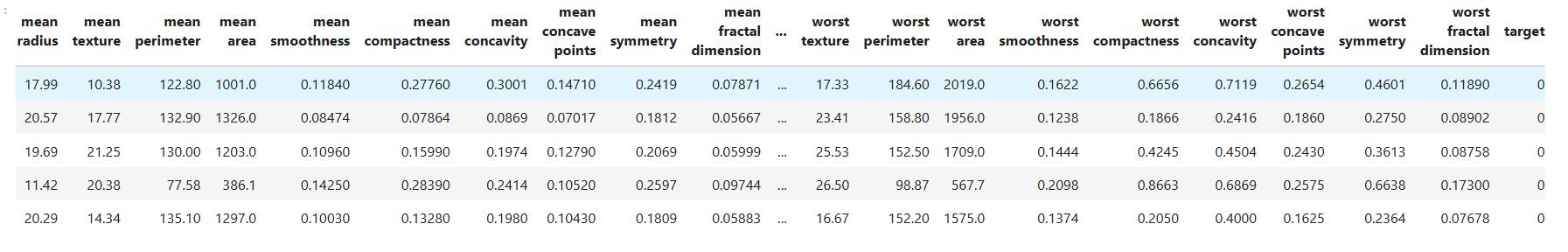


**Fig:1**

1. **Data Collection:**

The dataset used in this study is sourced from a publicly available dataset. Specifically, we have utilized the Cancer Wisconsin dataset.

* + - The dataset is the Breast Cancer Wisconsin (Diagnostic) Dataset.
    - The dataset contains features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features describe characteristics of the cell nuclei present in the images.
    - The dataset includes numerical features and a categorical label. The label indicates whether a tumor is malignant (represented as 1) or benign (represented as 0).



**Fig**: 2

1. **Data Preprocessing:**

Raw data from the Cancer Wisconsin dataset is preprocessed by:

* + **Missing Handling Values**: Checking for and addressing any missing values (NaNs) in the dataset. Removing rows or columns with excessive missing data to ensure data integrity.
  + **Scaling Features**: Standardizing the range of all features using StandardScaler. This ensures that each feature contributes equally to the model training process and prevents features with larger scales from dominating.

**3. Model Selection**

* Multiple machine learning algorithms are considered to compare their performance in breast cancer prediction:
  + **I. Support Vector Machine (SVM)**
    - Support Vector Machine is a powerful algorithm for binary classification. It finds the optimal hyperplane that best separates data points into different classes. SVMs are effective in high-dimensional spaces and can handle non-linear relationships through the use of kernel functions.
    - **Key advantages:**
      * Effective in high-dimensional spaces.
      * Versatile due to different kernel functions.
      * Relatively robust to outliers.
  + **II. K-Nearest Neighbors (KNN)**
    - K-Nearest Neighbors is a non-parametric algorithm that classifies data points based on the majority class of their k-nearest neighbors in the feature space. It is simple to implement but can be computationally intensive for large datasets.
    - **Key advantages:**
      * Simple to understand and implement.
      * No assumptions about data distribution.
      * Can be used for both classification and regression.
  + **III. Decision Tree Classifier**
    - Decision Tree Classifier is a tree-like model that partitions the data into subsets based on feature values. It is easy to interpret but can be prone to overfitting.
    - **Key advantages:**
      * Easy to interpret and visualize.
      * Can handle both categorical and numerical data.
      * Captures non-linear relationships.
  + **IV. Random Forest Classifier**
    - Random Forest Classifier is an ensemble learning method that builds multiple decision trees and combines their predictions. It is more robust than a single decision tree and reduces the risk of overfitting.
    - **Key advantages:**
      * Higher accuracy than individual decision trees.
      * Reduces overfitting.
      * Provides feature importance estimates.
  + **V. Naive Bayes**
    - Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes that features are conditionally independent given the class label. Despite its simplicity, it can be effective for text and categorical data classification.
    - **Key advantages:**
      * Simple and efficient.
      * Works well with categorical features.
      * Performs well with high-dimensional data.

**4.Training and Testing:**

The dataset is split into:

* **Training Set (80%)** – used to train the model
* **Testing Set (20%)** – used to evaluate the model

**Cross-validation** techniques are used to ensure robustness and prevent overfitting.

**5.Evaluation Metrics:**

* **Accuracy**
* Accuracy is the most straightforward metric, representing the overall proportion of correct predictions made by the model. It's calculated as:
* Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)
* While accuracy provides a general sense of model performance, it can be misleading in cases where the classes are imbalanced (i.e., significantly more benign tumors than malignant).
  + - * **Precision**
* Precision focuses on the accuracy of positive predictions. In the context of breast cancer prediction, precision measures the proportion of tumors correctly predicted as *malignant* out of all tumors that the model *classified* as malignant. It's calculated as:
* Precision = (True Positives) / (True Positives + False Positives)
* High precision is crucial in medical diagnosis to minimize the risk of falsely labeling a benign tumor as malignant, which can lead to unnecessary anxiety and invasive procedures.
  + - * **F1-Score**
* The F1-score provides a balanced measure of both precision and recall. It's the harmonic mean of precision and recall, and it's particularly useful when there is an imbalance between the number of malignant and benign tumors. The F1-score is calculated as:
* F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)
* A high F1-score indicates that the model performs well in terms of both
* minimizing false positives and minimizing false negatives.
  + - * **Confusion Matrix**
* A confusion matrix is a table that visually summarizes the performance of a classification model. It shows the number of:
* True Positives (TP): Malignant tumors correctly predicted as malignant.
* True Negatives (TN): Benign tumors correctly predicted as benign.
* False Positives (FP): Benign tumors incorrectly predicted as malignant.
* False Negatives (FN): Malignant tumors incorrectly predicted as benign.
* The confusion matrix provides a detailed breakdown of the model's errors, which can be helpful in understanding its strengths and weaknesses.

**6.Prediction:**

After training and evaluating the models, the selected best-performing model (e.g., Support Vector Machine) can be used to predict whether a breast tumor is malignant or benign for new, unseen data.

The system accepts the relevant features of a breast tumor (e.g., radius, texture, etc.) as input. The trained model then processes this input and outputs a prediction.

* If the output is "1," the model predicts that the tumor is malignant.
* If the output is "0," the model predicts that the tumor is benign.

This prediction provides valuable decision support to medical professionals in the diagnostic process.

**CHAPTER 5**

**SYSTEM REQUIREMENT**

**5.1 Hardware Requirements**

* **Processor:**
  + Minimum: Intel Core i5 or equivalent. A multi-core processor is essential for parallelizing the computations involved in training machine learning models.
  + Recommended: Intel Core i7 or better. This will significantly speed up training, especially with larger datasets.
* **RAM:**
  + Minimum: 8GB. Sufficient memory is needed to load and process the datasets and to efficiently run the machine learning algorithms.
  + Recommended: 16GB or more. This is beneficial for handling larger datasets and complex models.
* **Storage:**
  + Minimum: 100GB SSD. An SSD is highly recommended for fast read and write speeds, which are crucial for efficient data access during training.
  + Recommended: 256GB SSD or more. Additional storage allows for larger datasets, multiple models, and other project files.

**5.2 Software Requirements**

* **Operating System:**
  + Windows, macOS, or Linux. These operating systems all provide good support for Python and the necessary libraries.
* **Programming Language:**
  + Python (3.7 or later). Python is the primary language for machine learning development due to its rich ecosystem of libraries.
* **Libraries and Frameworks:**
  + scikit-learn: For machine learning algorithms, model selection, and evaluation.
  + pandas: For data manipulation and analysis.
  + NumPy: For numerical computing.
  + matplotlib/seaborn: For data visualization.
  + TensorFlow: (If you anticipate using this in the future) For deep learning model implementation.
* **Development Tools:**
  + Jupyter Notebook or JupyterLab: Interactive environments for development and experimentation.
  + Python IDE (e.g., VS Code, PyCharm): For code editing and debugging.
  + Google Colab: (Optional) A cloud-based platform for running Python code, which can be useful for resource-intensive tasks.

**CHAPTER 6**

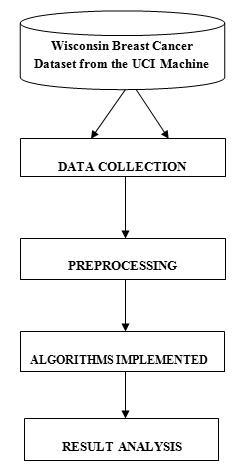
**SYSTEM ARCHITECTURE**

The architecture of the breast cancer prediction system comprises several interconnected components designed for the efficient processing and classification of tumor data. The workflow follows a structured pipeline, encompassing data input, preprocessing, feature extraction, classification, and output.

**Architecture Overview:**

* **Data Input Layer:**
  + The system receives the breast cancer dataset as input. This dataset contains features relevant to tumor characteristics (e.g., radius, texture, etc.).
* **Preprocessing & Feature Extraction Layer:**
  + The input data undergoes preprocessing, which may include handling missing values and feature scaling.
  + Feature extraction is implicitly performed during the model training process, where the machine learning models identify relevant patterns in the data.
* **Classification Engine:**
  + The pre-processed data is fed into the trained machine learning models.
  + These models (e.g., SVM, KNN, Decision Tree, Random Forest, Naive Bayes) classify the tumor as either malignant or benign.
* **Prediction Output Layer:**
  + The final classification result (malignant or benign) is provided as the system's output.
  + This output can be used to support medical decision-making.
* **Performance Analysis & Model Evaluation Layer:**
  + The models' performance is continuously evaluated using metrics such as accuracy, precision, recall, and F1-score.
  + A confusion matrix is used to visualize the classification results.
  + The best-performing model, based on the evaluation metrics, is selected for prediction.

This architecture ensures a systematic and data-driven approach to breast cancer prediction, leveraging machine learning to aid in diagnosis.



**Fig: 3**

**CHAPTER 7**

**SYSTEM IMPLEMENTATION**

The proposed breast cancer prediction system addresses the critical need for accurate and timely diagnosis, contributing to improved patient outcomes and reduced healthcare costs. By leveraging machine learning-based classification techniques, this system aims to accurately identify malignant tumors while minimizing both false positives and false negatives. The system encompasses several key phases: data collection, data preprocessing, feature consideration, model selection and training, evaluation, and prediction.

**7.1. Data Preparation**

* + 1. **Data Collection:**
* **Objective**: Gather a high-quality dataset containing features relevant to breast tumors, with labels indicating whether each tumor is malignant or benign.
* **Sources**: Publicly available datasets, such as the Breast Cancer Wisconsin (Diagnostic) Dataset.
* **Dataset Details**: The dataset consists of features describing characteristics of cell nuclei and a corresponding label (1 for malignant, 0 for benign).

**7.1.2. Preprocessing:**

Raw data from the Cancer Wisconsin dataset is pre-processed by:

* + - **Missing Handling Values**:
      * Checking for and addressing any missing values (NaNs) in the dataset. Removing rows or columns with excessive missing data to ensure data integrity.
    - **Scaling Features**:
      * Standardizing the range of all features using StandardScaler. This ensures that each feature contributes equally to the model training process and prevents features with larger scales from dominating.

**7.2. Model Selection and Training**

**7.2.1. Model Selection:**

Five classification models are evaluated for breast cancer prediction:

* **Model 1: Support Vector Machine (SVM)** - Effective for high-dimensional medical data and capturing complex relationships.
* **Model 2: K-Nearest Neighbors (KNN)** - Simple and versatile, suitable for exploring different neighborhood patterns.
* **Model 3: Decision Tree Classifier** - Provides interpretable decision rules and can handle non-linear relationships.
* **Model 4: Random Forest Classifier** - An ensemble method to improve accuracy and reduce overfitting.
* **Model 5: Naive Bayes** - Computationally efficient and performs well with categorical features.

**7.2.2. Training and Testing:**

The dataset is split into a training set (e.g., 80%) to train the models and a testing set (e.g., 20%) to evaluate their performance on unseen data. Cross-validation techniques may be employed during training to enhance robustness and prevent overfitting.

7.**3. Evaluation Metrics**

**7.3.1. Performance Assessment:**

Each model's performance is evaluated using established classification metrics relevant to medical diagnosis:

* **Accuracy:** Measures the overall correctness of the tumor classification (malignant or benign).
* **Precision, Recall, and F1-score:** Determine the reliability of the malignant tumor detection. Precision indicates how often the model is correct when it predicts "malignant." Recall indicates how well the model identifies all the actual malignant cases. The F1-score balances precision and recall.
* **Confusion Matrix:** A table visualizing the number of true positives, true negatives, false positives, and false negatives, providing a detailed view of the

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** | **Precision** | **F1 Score** | **Confusion Matrix** |
| **K-Nearest Neighbors** | 0.98 | 0.95 | 0.96 | 0.96 | [[40, 3], [3, 68]] |
| **Decision Tree** | 1.00 | 0.95 | 0.96 | 0.96 | [[40, 3], [3, 68]] |
| **Random Forest** | 1.00 | 0.96 | 0.96 | 0.97 | [[40, 3], [1, 70]] |
| **Support Vector Machine** | 0.99 | 0.97 | 0.97 | 0.98 | [[41, 2], [1, 70]] |
| **Naive Bayes** | 0.94 | 0.96 | 0.96 | 0.97 | [[40, 3], [1, 70]] |

* model's classification performance.

**7.3.2. Experimental Results:**

**Table: 1**

The Support Vector Machine (SVM) model demonstrated the highest performance, achieving an accuracy of 97% and an F1 score of 0.98. This indicates its effectiveness in accurately classifying tumors while maintaining a strong balance between precision and recall.

7.**4. Deployment**

**7.4.1. System Integration:**

* Develop a user-friendly interface for inputting patient tumor data (e.g., feature values) and obtaining breast cancer prediction results (malignant or benign).
* Integrate the prediction system into a clinical workflow to provide decision support to medical professionals. This may involve connecting to electronic health record (EHR) systems.

**7.4.2. Continuous Improvement:**

* Regularly update the prediction models with new data to maintain accuracy and adapt to evolving diagnostic knowledge.
* Incorporate feedback from medical professionals to refine the model's performance and address specific clinical needs.
* Explore advanced techniques, such as deep learning, to further enhance prediction accuracy and potentially identify new diagnostic features.

**CHAPTER 8**

**CONCLUSION**

The breast cancer prediction system presented in this project effectively utilizes machine learning techniques to classify tumors as malignant or benign with high accuracy. By employing supervised learning algorithms, the system demonstrates the potential to provide valuable decision support to medical professionals, enhancing the accuracy and efficiency of diagnosis.

Among the models evaluated, the Support Vector Machine (SVM) achieved the highest F1 score (0.98) and high accuracy, indicating its superior performance in distinguishing between malignant and benign tumors. This suggests that the SVM is a strong candidate for a reliable prediction model, demonstrating a good balance between predictive performance.

By integrating data preprocessing, feature consideration, and classification techniques, the proposed system contributes to improved breast cancer diagnosis and potentially leads to earlier and more effective treatment. While challenges such as acquiring large, diverse datasets and further refining model accuracy remain, the project demonstrates the potential of machine learning-based systems as a valuable tool in the fight against breast cancer.

In future improvements, deep learning models and techniques to incorporate medical imaging data could be explored to further enhance prediction accuracy and potentially identify new diagnostic markers. Continuous updates with new clinical data will ensure the system remains effective and adaptable to evolving medical knowledge, making it a reliable solution for supporting breast cancer diagnosis in clinical settings.

**APPENDIX – 1**

**Source code:**

**# Import Libraries**

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import SVC

from sklearn.naive\_bayes import GaussianNB

from sklearn.metrics import accuracy\_score, precision\_score, f1\_score, confusion\_matrix

import matplotlib.pyplot as plt

import seaborn as sns

import joblib

from sklearn.datasets import load\_breast\_cancer

**# Load Data**

data = load\_breast\_cancer()

df = pd.DataFrame(data.data, columns=data.feature\_names)

df['target'] = data.target

**# Data Preprocessing**

X = df.drop(columns=['target'])

y = df['target']

scaler = StandardScaler()

X = scaler.fit\_transform(X)

**# Split Data**

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

**# Define Classifiers**

classifiers = {

"K-Nearest Neighbors": KNeighborsClassifier(),

"Decision Tree": DecisionTreeClassifier(random\_state=42),

"Random Forest": RandomForestClassifier(random\_state=42),

"Support Vector Machine": SVC(random\_state=42),

"Naive Bayes": GaussianNB()

}

**# Train and Evaluate Models**

accuracy\_results\_train = {}

accuracy\_results\_test = {}

precision\_results\_test = {}

f1\_results\_test = {}

for name, model in classifiers.items():

model.fit(X\_train, y\_train)

train\_predictions = model.predict(X\_train)

test\_predictions = model.predict(X\_test)

accuracy\_results\_train[name] = accuracy\_score(y\_train, train\_predictions)

accuracy\_results\_test[name] = accuracy\_score(y\_test, test\_predictions)

precision\_results\_test[name] = precision\_score(y\_test, test\_predictions)

f1\_results\_test[name] = f1\_score(y\_test, test\_predictions)

cm = confusion\_matrix(y\_test, test\_predictions)

print(f"{name}:\n"

f" Train Accuracy: {accuracy\_results\_train[name]:.2f}\n"

f" Test Accuracy: {accuracy\_results\_test[name]:.2f}\n"

f" Precision: {precision\_results\_test[name]:.2f}\n"

f" F1 Score: {f1\_results\_test[name]:.2f}\n"

f" Confusion Matrix:\n{cm}\n")

**# Select Best Model**

best\_model\_name = max(f1\_results\_test, key=f1\_results\_test.get)

best\_model = classifiers[best\_model\_name]

best\_predictions = best\_model.predict(X\_test)

best\_cm = confusion\_matrix(y\_test, best\_predictions)

print(f"\nBest Model: {best\_model\_name}")

print(f" Test Accuracy: {accuracy\_results\_test[best\_model\_name]:.2f}")

print(f" Precision: {precision\_results\_test[best\_model\_name]:.2f}")

print(f" F1 Score: {f1\_results\_test[best\_model\_name]:.2f}")

print(f" Confusion Matrix:\n{best\_cm}")

**# Save the Model**

model\_data = {

'scaler': scaler,

'model': best\_model

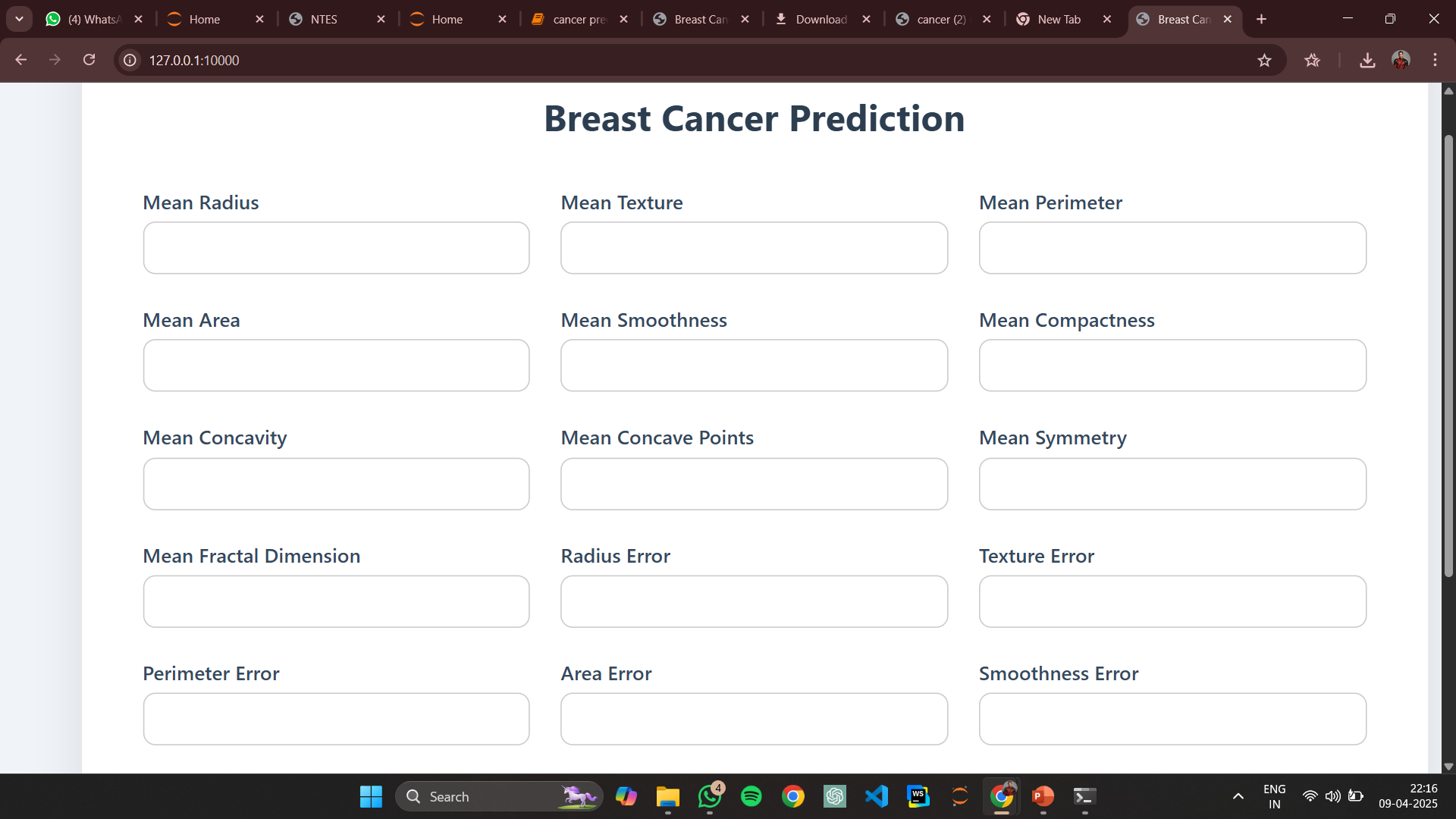
}

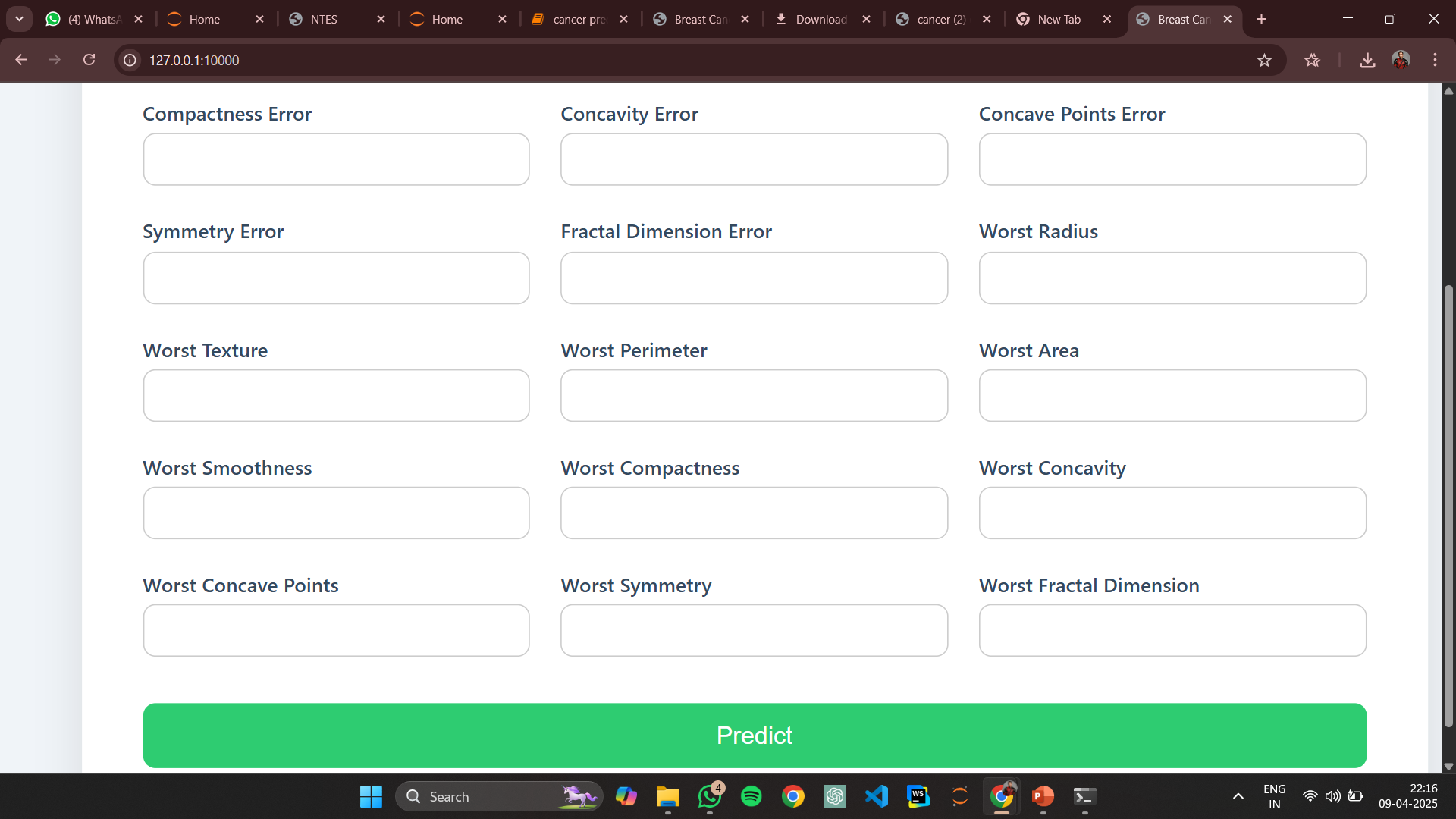
joblib.dump(model\_data, 'svm\_breast\_cancer\_model.pkl')

print("✅ SVM model and scaler saved as 'svm\_breast\_cancer\_model.pkl'")

**User Interface Overview:**

**1.UI Before Entering The values**





**Fig: 4 User interface**

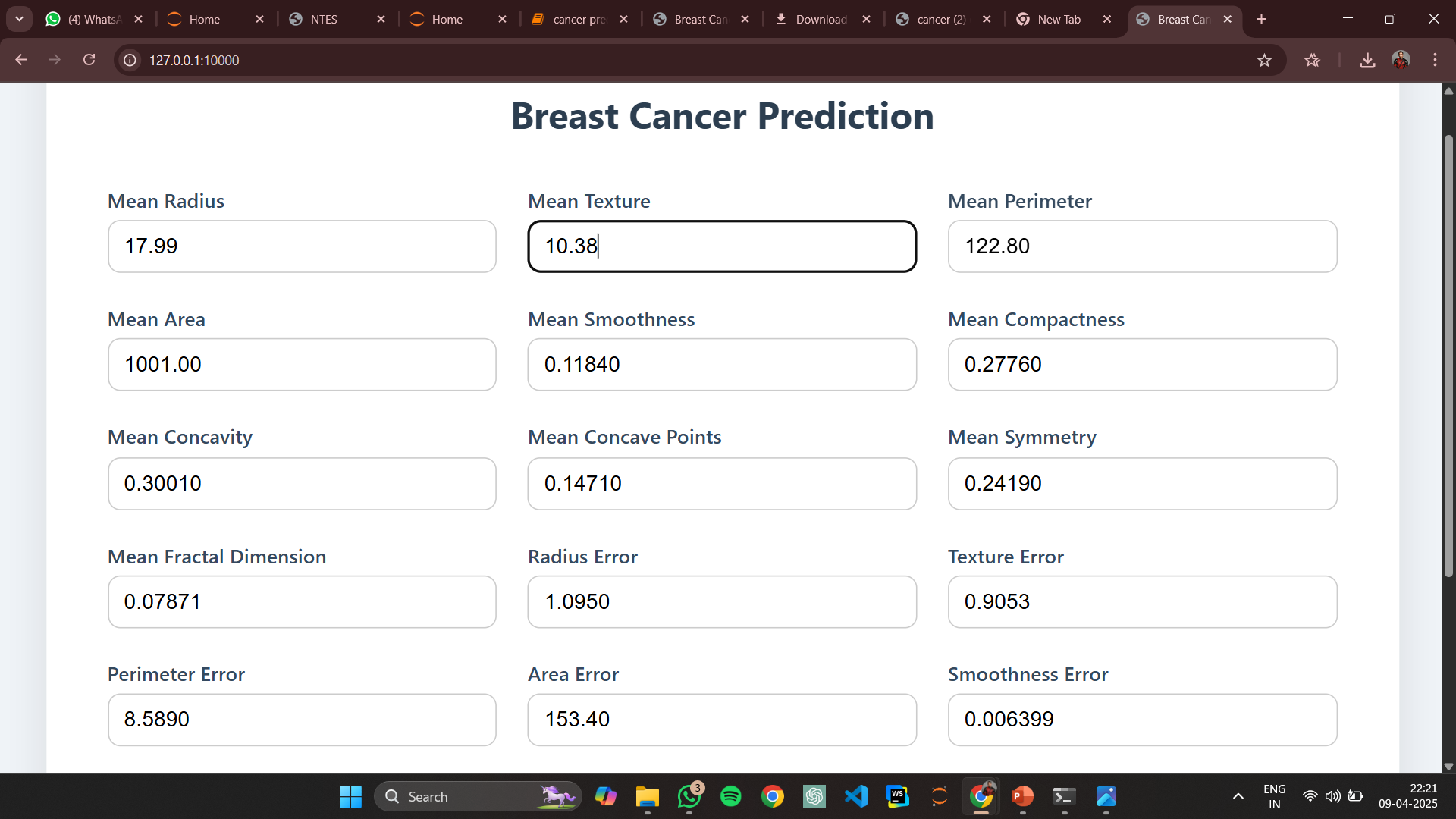
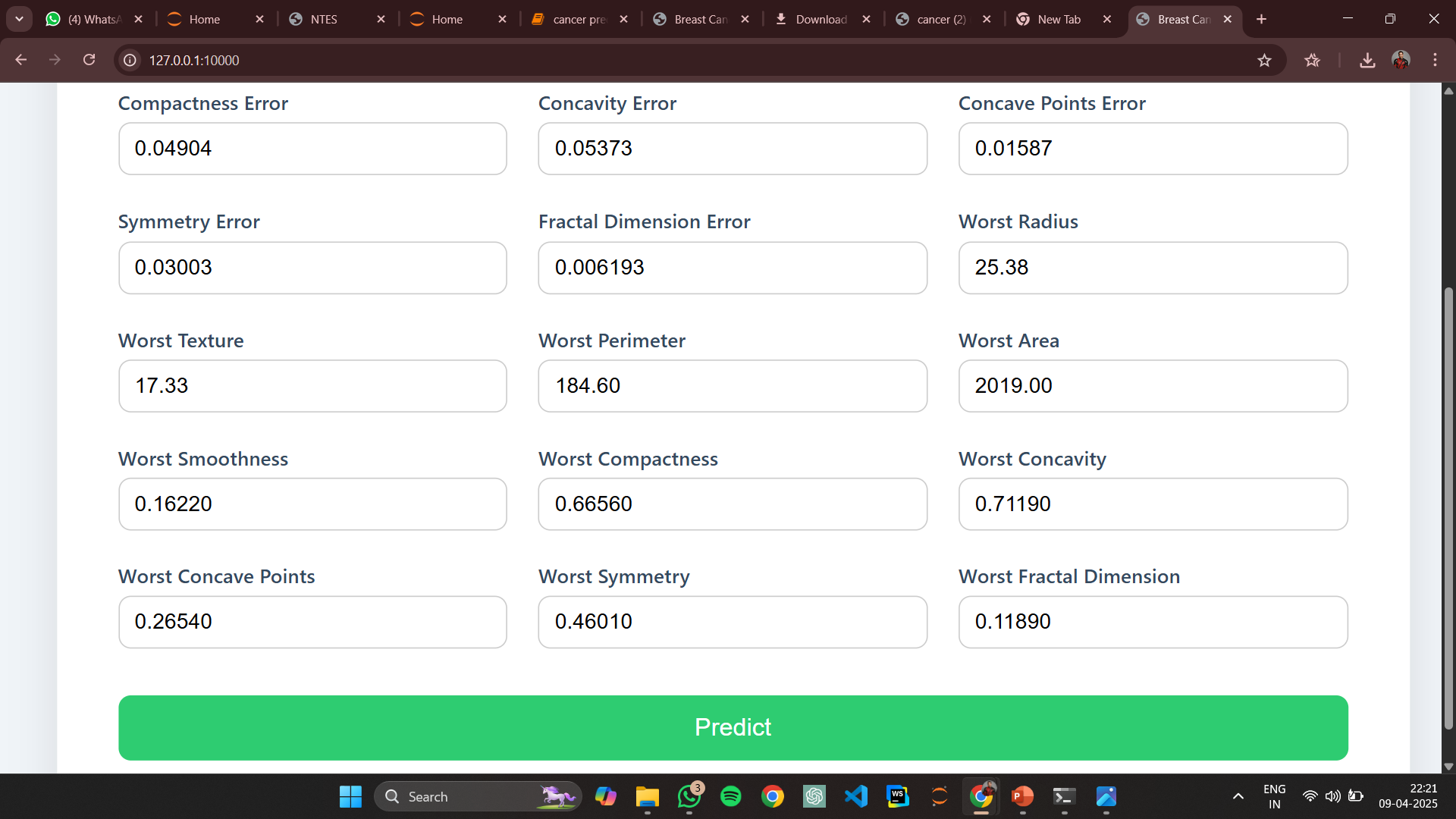
**About the UI:**

The user interface of the breast cancer prediction system is designed to be simple, intuitive, and user-friendly. It allows users—such as doctors, patients, or researchers—to easily input relevant medical data into a web form. The form typically includes fields for input features like:

* Mean Radius
* Mean Texture
* Mean Perimeter
* Mean Area
* Mean Smoothness
* (or other features from the breast cancer dataset)

Once the user fills out the form and clicks "Predict", the data is sent to the server where the machine learning model processes it and returns a result such as "Malignant" or "Benign". The result is clearly displayed on the same page or a results page.

**2.UI After Entering The values**

**The tumor is likely: Malignant**

**Fig: 5**  **Result Page**