A

Mini Project On

**OPTIMISED STACKED ENSEMBLE TECHNIQUES IN THE PREDICTION OF CERVICAL CANCER USING SMOTE AND RFERF**

(Submitted in partial fulfillment of the requirements for the award of Degree) BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Ch.Venkateshwarlu (217R1A0517)

P.Prabhudeva (217R1A0546) A.Vishnuvardhan Reddy (217R1A0504)

Under the Guidance of

**DR. S. SUMA**

(Assistant Professor)

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**CMR TECHNICAL CAMPUS**

**UGC AUTONOMOUS**

(Accredited by NAAC, NBA, Permanently Affiliated to JNTUH, Approved by AICTE, New Delhi) Recognized Under Section 2(f) & 12(B) of the UGCAct.1956,

Kandlakoya (V), Medchal Road, Hyderabad-501401.

**2021-25**

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**CERTIFICATE**

This is to certify that the project entitled “**OPTIMISED STACKED ENSEMBLE TECHNIQUES IN THE PREDICTION OF CERVICAL CANCER USING SMOTE AND RFERF**”being submitted by **CH. VENKATESHWARLU (217R1A0517),P. PRABHUDEVA (217R1A0546) & A. VISHNUVARDHAN REDDY(217R1A0504)** in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science and Engineering to the Jawaharlal Nehru Technological University Hyderabad, is a record of bonafide work carried out by him/her under our guidance and supervision during the year 2024-25.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

#### Dr. S. Suma Dr. A. Raji Reddy

#### Assistant Professor DIRECTOR

**INTERNAL GUIDE**

**Dr. Nuthanakanti Bhaskar EXTERNAL EXAMINER**

**HoD**

**Submitted for viva voice Examination held on \_\_\_\_\_\_\_\_\_\_\_\_**

**ACKNOWLEGDEMENT**

Apart from the efforts of us, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. We take this opportunity to express my profound gratitude and deep regard to my guide

**Dr. S.Suma,** Assistant Professor for her exemplary guidance, monitoring and constant encouragement throughout the project work. The blessing, help and guidance given by her shall carry us a long way in the journey of life on which we are about to embark. We also take this opportunity to express a deep sense of gratitude to Project Review Committee (PRC) Coordinators: **Dr. K.Maheswari, Dr. S.Suma, Mr. J.Narasimha Rao, Mrs.** [**K.Shilpa**](https://vidwan.inflibnet.ac.in//profile/392294)**, Mr.** [**K.Ranjith Reddy**](https://vidwan.inflibnet.ac.in//profile/254178)for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We are also thankful to **Dr. Nuthanakanti Bhaskar,** Head, Department of Computer Science and Engineering and **Dr. K.SrujanRaju**, Dean (Research and Development) for providing encouragement and support for completing this project successfully.

We are obliged to **Dr. A. Raji Reddy,** Director for being cooperative throughout the course of this project. We would like to express our sincere gratitude to Sri. **Ch. Gopal Reddy,** Chairman for providing excellent infrastructure and a nice atmosphere throughout the course of this project.

The guidance and support received from all the members of **CMR Technical Campus** who contributed to the completion of the project. We are grateful for their constant support and help.

Finally, we would like to take this opportunity to thank our family for their constant encouragement, without which this assignment would not be completed. We sincerely acknowledge and thank all those who gave support directly and indirectly in the completion of this project.

#### CH.VENKATESHWARLU (217R1A0517)

#### P. PRABHUDEVA (217R1A0546)

#### A. VISHNUVARDHAN REDDY (217R1A0504)

**ABSTRACT**

This project is titled as “Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF”. This software allows users to input relevant medical information to predict the likelihood of cervical cancer occurrence. It utilizes a combination of data preprocessing techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) for balancing the dataset and Recursive Feature Elimination (RFE) for selecting the most critical features. The core of the project is built around a stacked ensemble model that combines multiple machine learning algorithms for enhanced accuracy in prediction. Additionally, a pre-trained model is utilized to improve efficiency in feature extraction and classification. The system is integrated with a graphical user interface (GUI) using Tkinter, allowing easy interaction for data input and prediction visualization. With high accuracy, the model efficiently classifies patients who may be at risk of cervical cancer, providing crucial insights for early detection and preventive healthcare measures.

## LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| **FIGURE NO** | **FIGURE NAME** | **PAGE NO** |
| Figure 3.1 | Project Architecture | 9 |
| Figure 3.2 | Use case diagram | 11 |
| Figure 3.3 | Class diagram | 12 |
| Figure 3.4 | Sequence diagram | 13 |
| Figure 3.5 | Activity diagram | 14 |

**RESULTS AND DISCUSSIONS**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **FIGURE NAME** | **PAGE NO**. |
| Figure 5.1 | Upload Clinical Dataset | 22 |
| Figure 5.2 | Class Imbalance | 23 |
| Figure 5.3 | Applying SMOTE and RFE-RF | 23 |
| Figure 5.4 | Result | 24 |
| Figure 5.5 | Confusion Matrix | 25 |
| Figure 5.6 | Prediction Result | 26 |

**TABLE OF CONTENTS**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ABSTRACT** | | |  | | |  | | i |
| **LIST OF FIGURES** | | | | | |  | | ii |
| **LIST OF SCREENSHOTS** | | | | | | | | iii |
| **1.** | **INTRODUCTION** | | | | | | | 1 |
|  | 1.1 | | PROJECT SCOPE | | | | | 2 |
|  | 1.2 | | PROJECT PURPOSE | | | | | 2 |
|  | 1.3 | | PROJECT FEATURES | | | | | 2 |
| **2.** | **SYSTEM ANALYSIS** | | | | | | | 3 |
|  | 2.1 | | PROBLEM DEFINITION | | | | | 4 |
|  | 2.2 | | EXISTING SYSTEM | | | | | 4 |
|  |  | | 2.2.1 | | | LIMITATIONS OF THE EXISTING SYSTEM | | 5 |
|  | 2.3 | | PROPOSED SYSTEM | | | | | 5 |
|  |  | | 2.3.1 | | | ADVANTAGES OF PROPOSED SYSTEM | | 5 |
|  | 2.4 | | FEASIBILITY STUDY | | | | | 6 |
|  |  | | 2.4.1 | | | ECONOMICAL FESIBILITY | | 6 |
|  |  | | 2.4.2 | | | TECHNICAL FEASIBILITY | | 6 |
|  |  | | 2.4.3 | | | SOCIAL FEASIBILITY | | 6-7 |
|  | 2.5 | | HARDWARE & SOFTWARE REQUIREMENTS | | | | | 7 |
|  |  | | 2.5.1 | | | HARDWARE REQUIREMENTS | | 7 |
|  |  | | 2.5.2 | | | SOFTWARE REQUIREMENTS | | 7 |
| **3.** | **ARCHITECTURE** | | | | | | | 8 |
|  | 3.1 | | PROJECT ARCHITECTURE | | | | | 9 |
|  | 3.2 | | DESCRIPTION | | | | | 9-10 |
|  | 3.3 | | USECASE DIAGRAM | | | | | 10-11 |
|  | 3.4 | | CLASS DIAGRAM | | | | | 11-12 |
|  | 3.5 | | SEQUENCE DIAGRAM | | | | | 12-13 |
|  | 3.6 | | ACTIVITY DIAGRAM | | | | | 13-14 |
| **4.** | **IMPLEMENTATION** | | | | | | | 15 |
|  | 4.1 | | SAMPLE CODE | | | | | 16-20 |
| **5.** | **RESULTS AND DISCUSSIONS** | | | | | | | 21-28 |
| **6.** | **TESTING** | | |  | | | | 29 |
|  | 6.1 | INTRODUCTION TO TESTING | | | | | | 30 |
|  | 6.2 | TYPES OF TESTING | | | | | | 30 |
|  |  | 6.2.1 | | | | UNIT TESTING | | 30 |
|  |  | 6.2.2 | | | | INTEGRATION TESTING | | 30 |
|  |  | 6.2.3 | | | | FUNCTIONAL TESTING | | 30-31 |
|  | 6.3 | TEST CASES | | | | | | 31 |
|  |  | 6.3.1 | | | UPLOADING IMAGES | | | 31 |
|  |  | 6.3.2 | | | CLASSIFICATION | | | 31 |
| **7.** | **CONCLUSION & FUTURE SCOPE** | | | | | | | 32 |
|  | 7.1 | PROJECT CONCLUSION | | | | | | 33 |
|  | 7.2 | FUTURE SCOPE | | | | | | 33 |
| **8.** | **BIBLIOGRAPH** | | | | | |  | 34 |
|  | 8.1 | REFERENCES | | | | | | 35 |
|  | 8.2 | WEBSITE | | | | | | 35 |

1. **INTRODUCTION**
2. **INTRODUCTION**

### PROJECT SCOPE

### This project is titled as “Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF”.This software allows users to input clinical risk factors to predict the likelihood of cervical cancer. The software uses machine learning techniques, including SMOTE for data balancing and Recursive Feature Elimination (RFE) for selecting important features. It then applies stacked ensemble classification algorithms such as Random Forest, Decision Tree, K-NearestNeighbors and Logistic Regression to make accurate predictions. The system aims to support early diagnosis of cervical cancer, offering healthcare professionals a tool for timely and effective intervention with high prediction accuracy.

### 1.2 PROJECT PURPOSE

### This project has been developed to facilitate the **early diagnosis of cervical cancer** by utilizing machine learning techniques. The system is designed with exclusive features to ensure optimal **data preprocessing, feature selection**, and **classification** accuracy. In all cases, the system analyzes clinical risk factors and diagnostic tests to provide accurate predictions. It leverages **data balancing techniques** like SMOTE and **ensemble classifiers** to improve the precision of predictions. The data ensures that key features are correctly identified, selected, and used to enhance the prediction of cervical cancer outcomes, enabling healthcare professionals to make informed decisions based on optimized machine learning models.

### 1.3 PROJECT FEATURES

The main features of this project include its role as a **predictive system** for cervical cancer diagnosis, where the system acts as a **problem solver** by addressing the challenges posed by **imbalanced datasets** and **complex risk factors**. The system provides **optimized solutions** through the use of **SMOTE for data balancing** and **RFE for feature selection**, ensuring only the most relevant features are used for accurate predictions.Each ensemble model prediction is given to the Logistic Regression for the final prediction. The model’s predictions are presented to the **user for validation** and can be reviewed upon request. Changes or adjustments are made iteratively until the user is satisfied with the **performance and accuracy** of the predictions, ensuring a comprehensive and user-driven solution.

# SYSTEM ANALYSIS

## 2. SYSTEM ANALYSIS

### System Analysis is a critical phase in the development of the cervical cancer prediction system. During this phase, the current diagnostic process and dataset are thoroughly examined and analyzed. The system analyst plays the role of an investigator, delving deep into the existing methods for cervical cancer prediction, including identifying gaps such as data imbalance and irrelevant features. The primary objective of this phase is to address the question, “How can we enhance the accuracy of cervical cancer diagnosis?” The system is examined holistically, where the inputs, such as clinical risk factors and diagnostic test results, are identified and their relationships within the prediction model are analyzed. Once the analysis is complete, the analyst gains a clear understanding of the improvements needed—such as applying SMOTE for data balancing and RFE for feature selection—to ensure the system effectively addresses the problem of early cervical cancer detection.

### 2.1 PROBLEM DEFINITION

The problem definition involves a detailed examination of the current cervical cancer diagnosis system, identifying limitations such as **data imbalance** and **inefficient feature selection**. The existing system is analyzed to understand its performance, and key problem areas are recognized. The designer acts as a problem solver, proposing solutions like applying **SMOTE** for data balancing and **RFE** for selecting relevant features to improve prediction accuracy. These solutions are compared with the current system, and the best approach is selected. The proposal is then presented to the user for review, and adjustments are made until the user is satisfied with the final solution.

### 2.2 EXISTING SYSTEM

The existing system for cervical cancer diagnosis relies on basic machine learning models that face challenges such as **data imbalance** and the inclusion of **irrelevant features**, leading to suboptimal prediction accuracy. The dataset used contains 32 risk factors and 4 diagnostic targets (Hinselmann, Schiller, Cytology, and Biopsy), but it suffers from a significant imbalance between positive and negative cases. Additionally, no robust feature selection is applied, which can lead to over fitting and reduced model performance. Current approaches use standalone models like **Random Forest** or **XGBoost**, without combining or optimizing them for higher accuracy and reliability in predictions.

#### 2.2.1 LIMITATIONS OF EXISTING SYSTEM

* + - * Data Imbalance
      * Lack of Feature Selection
      * Standalone Models
      * Limited Optimization
      * Suboptimal Accuracy

To avoid all these limitations and make the working more accurately the system needs to be implemented efficiently.

### 2.3 PROPOSED SYSTEM

The aim of proposed system is to develop a system of improved facilities. The proposed system can overcome all the limitations of the existing system. The system provides higher accuracy. The existing system has several disadvantages and many more difficulties to work well. The proposed system tries to eliminate or reduce these difficulties up to some extent. The proposed system enhances cervical cancer diagnosis by implementing an optimized stacked ensemble approach that integrates multiple classifiers, including **Random Forest**, **Decision Tree**, and **K-Nearest Neighbors**, with **Logistic Regression** as the final estimator. To address data imbalance, it employs the **Synthetic Minority Oversampling Technique (SMOTE)**, ensuring balanced representation in the dataset. Additionally, the system utilizes **Recursive Feature Elimination (RFE)** to select the most significant risk factors, improving model efficiency and accuracy. This comprehensive approach aims to provide more reliable predictions, facilitating early detection and better outcomes for cervical cancer patients.The proposed system helps the user to work user friendly and he can easily do his jobs without time lagging.

### 2.3.1 ADVANTAGES OF THE PROPOSED SYSTEM

The system is very simple in design and to implement. The system requires very low system resources and the system will work in almost all configurations. It has got following features

* Improved Prediction Accuracy
* Addressed Class Imbalance
* Optimal Feature Selection
* Enhanced Sensitivity and Specificity
* Robustness and Reliability

### 2.4 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

* ECONOMICAL FEASIBILITY
* TECHNICAL FEASIBILITY
* SOCIAL FEASIBILITY

### 2.4.1 ECONOMICAL FEASIBILITY

The developing system must be justified by cost and benefit. Criteria to ensure that effort is concentrated on project, which will give best, return at the earliest. One of the factors, which affect the development of a new system, is the cost it would require.

The following are some of the important financial questions asked during preliminary investigation:

* The costs conduct a full system investigation.
* The cost of the hardware and software.
* The benefits in the form of reduced costs or fewer costly errors.

Since the system is developed as part of project work, there is no manual cost to spend for the proposed system. Also all the resources are already available, it give an indication of the system is economically possible for development.

### 2.4.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

### 2.4.3 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel

threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

### HARDWARE & SOFTWARE REQUIREMENTS

* + 1. **HARDWARE REQUIREMENTS :**

Hardware interfaces specifies the logical characteristics of each interface between the software product and the hardware components of the system. The following are some hardware requirements.

* System : Windows
* Processor : Intel core i3 or higher
* Hard disk : 512GB or above
* Ram : Atleast 4GB

### SOFTWARE REQUIREMENTS :

Software Requirements specifies the logical characteristics of each interface and software components of the system. The following are some software requirements,

* Operating System : Windows 10 or 11
* Coding Language : Python 3.7.0
* Libraries : Pandas, Matplotlib, Seaborn
* Visualization Tools : Matplotlib, Seaborn

# ARCHITECTURE

## 3. ARCHITECTURE

### 3.1 PROJECT ARCITECTURE

This project architecture shows the procedure followed for cervical cancer prediction using machine learning techniques, starting from input to final prediction.

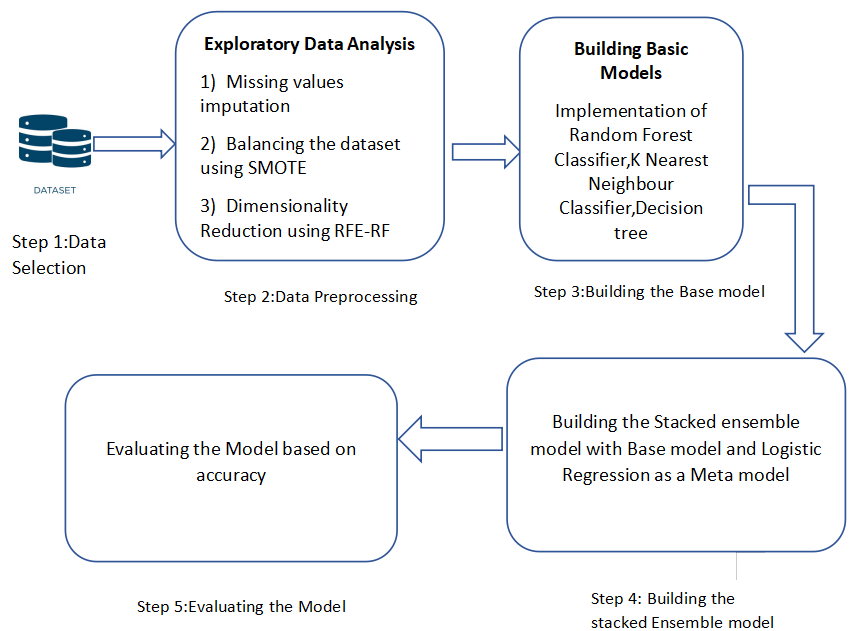


Figure 3.1: Project Architecture of Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

### 3.2 DESCRIPTION

**Data Collection and Preprocessing:**

The project started with collecting a cervical cancer dataset, which included essential features for prediction. Data preprocessing involved handling missing values, normalizing attributes, and performing data cleaning to ensure the dataset was well-prepared for accurate modeling.

**Data Balancing:**

SMOTE was applied to tackle class imbalance by generating synthetic samples for the minority class. This method enhanced the model’s ability to learn from all classes and improved its predictive accuracy

**Feature Selection:**

Recursive Feature Elimination (RFE) was used to identify and retain the most significant features. This process improved the model's performance by focusing on relevant attributes and reducing the complexity of the data.

**Model Building:**

A Stacked Ensemble Model was constructed, integrating multiple algorithms like decision tree,KNN,random forest and Logistic Regression . This ensemble approach leveraged the strengths of various base models to produce a more robust and accurate prediction system.

**Model Evaluation:**

The model’s performance was assessed using metrics such as accuracy, precision, and recall. Cross-validation techniques were employed to ensure the model's reliability and its ability to generalize to new data.

### 3.3 USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

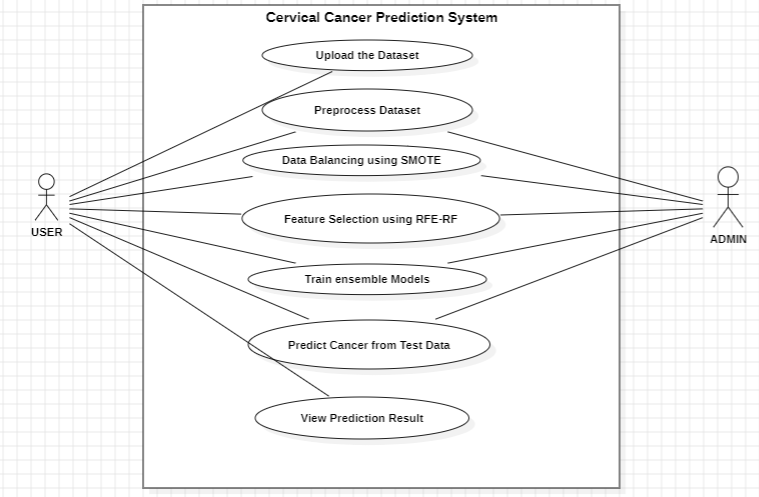


Figure 3.2: Use Case Diagram for Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

The diagram represents a **Cervical Cancer Prediction System** where two actors, the **User** and **Admin**, interact with the system. Both can upload the dataset, preprocess it, and balance it using **SMOTE** to handle class imbalances. They can also perform feature selection using **RFE-RF**, train ensemble models, and predict cancer based on test data. Finally, both actors can view the prediction results to determine if a patient sample is classified as having cervical cancer or not. The system provides an end-to-end solution for processing data and making predictions.

### 3.4 CLASS DIAGRAM

Class Diagram is a collection of classes and objects.It is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

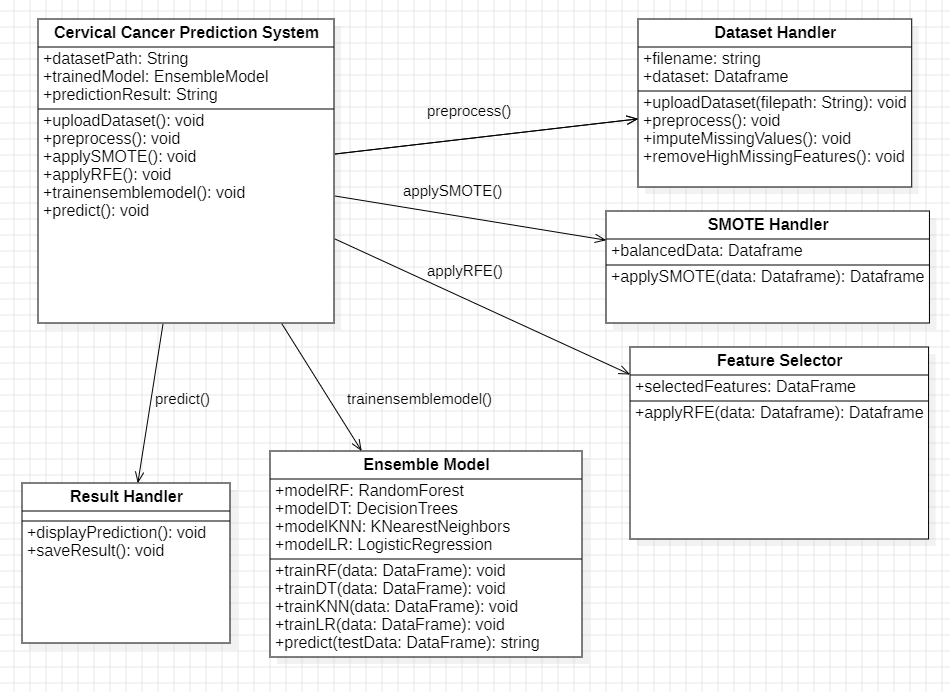


Figure 3.3: Class Diagram for Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

This class diagram represents the architecture of the **Cervical Cancer Prediction System**. It shows the main system class that handles tasks like uploading datasets, preprocessing, applying SMOTE for data balancing, and RFE for feature selection. The **Dataset Handler** manages dataset-related operations such as loading and cleaning data. **SMOTE Handler** and **Feature Selector** perform data balancing and feature selection respectively. The **Ensemble Model** class contains various machine learning models (Random Forest, Decision Trees, KNN, Logistic Regression) and their training/prediction methods. Finally, the **Result Handler** displays and saves the prediction results.

### 3.5 SEQUENCE DIAGRAM

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

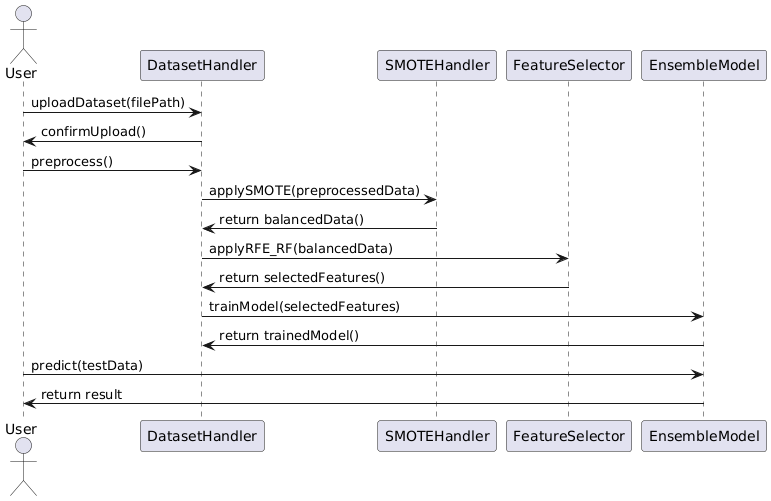


Figure 3.4: Sequence Diagram for Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

This sequence diagram illustrates the flow of interactions in the **Cervical Cancer Prediction System**. The **User** uploads the dataset, which is handled by the **DatasetHandler**. After confirming the upload, the dataset is preprocessed. The preprocessed data is passed to the **SMOTEHandler**, which balances the data, followed by the **FeatureSelector** for feature extraction using RFE-RF. Once the selected features are returned, the **EnsembleModel** is trained using the features. Finally, the user inputs test data, and the system predicts the result using the trained model, returning the outcome to the user.

### 3.6 ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

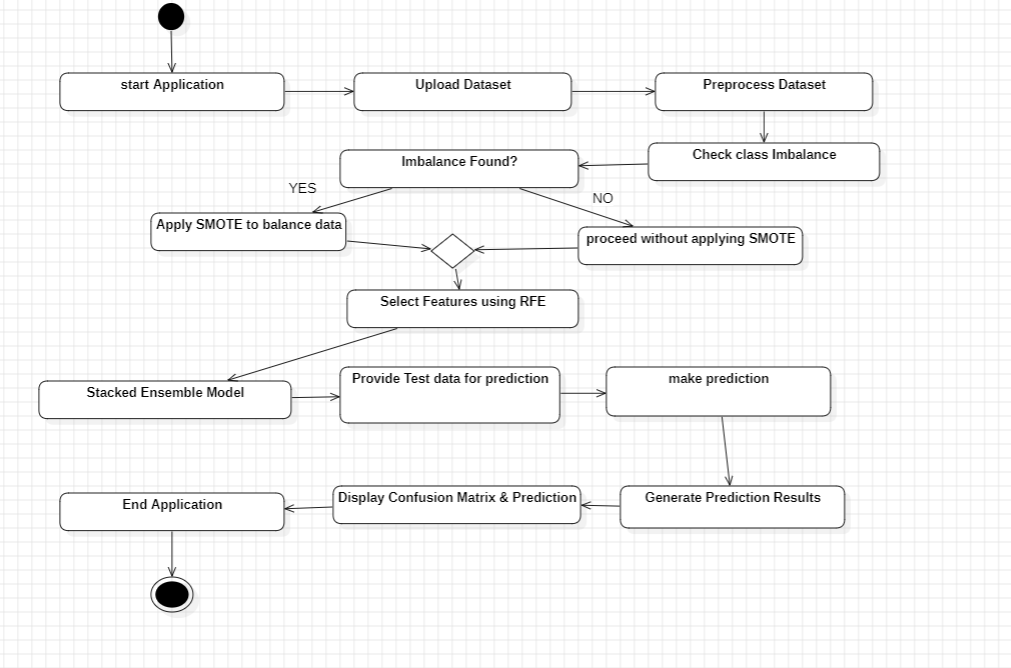


Figure 3.5: Activity Diagram for Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

The application begins with uploading and preprocessing the dataset, followed by checking for class imbalance. If an imbalance is found, SMOTE (Synthetic Minority Over-sampling Technique) is applied to balance the data. If no imbalance is detected, the process proceeds without applying SMOTE. Recursive Feature Elimination (RFE) is then used to select the most relevant features. The selected features are provided to a stacked ensemble model, which performs the prediction. The results are displayed in the form of a confusion matrix, along with the prediction output, before the application concludes.

# IMPLEMENTATION

### IMPLEMENTATION

**4.1 SAMPLE CODE**

**main.py:**

from tkinter import \*

import tkinter

import pandas as pd

from tkinter import filedialog

import matplotlib.pyplot as plt

from tkinter import ttk

from imblearn.over\_sampling import SMOTE

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

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import confusion\_matrix

import seaborn as sns

from sklearn.metrics import precision\_score

from sklearn.metrics import recall\_score

from sklearn.metrics import f1\_score

from sklearn.metrics import accuracy\_score

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.feature\_selection import RFE

import numpy as np

from sklearn.ensemble import StackingClassifier

from sklearn.linear\_model import LogisticRegression

main = tkinter.Tk()

main.title("Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF")

main.geometry("1300x1200")

global filename

global features, X, Y, dataset

global X\_train, X\_test, y\_train, y\_test, rfe, clf

def upload():

global filename

global dataset

filename = filedialog.askopenfilename(initialdir="Dataset")

text.delete('1.0', END)

text.insert(END,filename+" loaded\n");

dataset = pd.read\_csv(filename)

text.insert(END,'Dataset size : \n\n')

text.insert(END,'Total Rows : '+str(dataset.shape[0])+"\n")

text.insert(END,'Total Columns : '+str(dataset.shape[1])+"\n\n")

text.insert(END,'Dataset Samples\n\n')

text.insert(END,str(dataset.head())+"\n\n")

text.update\_idletasks()

label = dataset.groupby('Biopsy').size()

label.plot(kind="bar")

plt.show()

def preprocess():

global dataset

text.delete('1.0', END)

dataset.fillna(0, inplace = True)

unique, counts = np.unique(dataset['Biopsy'],return\_counts=True)

print(unique)

print(counts)

text.insert(END,"Number of Class labels & its count found in dataset before applying SMOTE\n\n")

text.insert(END,"Class Label "+str(unique[0])+" found in dataset "+str(counts[0])+"\n")

text.insert(END,"Class Label "+str(unique[1])+" found in dataset "+str(counts[1])+"\n")

def smoteBalancing():

global X\_train, X\_test, y\_train, y\_test

global dataset, X, Y

text.delete('1.0', END)

Y = dataset.values[:,dataset.shape[1]-1]

print(Y)

dataset.drop(['Biopsy'], axis = 1,inplace=True)

X = dataset.values

sm = SMOTE(random\_state = 42) #creating smote object

X, Y = sm.fit\_sample(X, Y)#applying smote to balance dataset

unique, counts = np.unique(Y,return\_counts=True)

text.insert(END,"Number of Class labels & its count found in dataset after applying SMOTE\n\n")

text.insert(END,"Class Label "+str(unique[0])+" found in dataset "+str(counts[0])+"\n")

text.insert(END,"Class Label "+str(unique[1])+" found in dataset "+str(counts[1])+"\n\n")

def featuresSelection():

global dataset, X, Y, rfe

global X\_train, X\_test, y\_train, y\_test

text.delete('1.0', END)

text.insert(END,"Total features found in dataset before applying RFE : "+str(X.shape[1])+"\n")

rfe = RFE(estimator=DecisionTreeClassifier(), n\_features\_to\_select=20) #creating RFE objects

rfe.fit(X, Y) #applying RFE algorithm to select features

X = rfe.transform(X)

text.insert(END,"Total features found in dataset after applying RFE : "+str(X.shape[1])+"\n\n")

text.insert(END,"Dataset Train & test split is 80% for training and 20% for testing\n\n")

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size=0.2)

text.insert(END,"Total records used to train machine learning Stacked Ensemble Algorithm is : "+str(X\_train.shape[0])+"\n")

text.insert(END,"Total records used to test machine learning Stacked Ensemble Algorithm is : "+str(X\_test.shape[0])+"\n")

def trainStacked():

text.delete('1.0', END)

global X\_train, X\_test, y\_train, y\_test, clf

estimators = [('rf', RandomForestClassifier(n\_estimators=10, random\_state=42)), ('dt', DecisionTreeClassifier()), ('knn', KNeighborsClassifier(n\_neighbors = 2))]

clf = StackingClassifier(estimators=estimators, final\_estimator=LogisticRegression())

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

predict = clf.predict(X\_test)

p = precision\_score(y\_test, predict,average='macro') \* 100

r = recall\_score(y\_test, predict,average='macro') \* 100

f = f1\_score(y\_test, predict,average='macro') \* 100

a = accuracy\_score(y\_test,predict)\*100

text.insert(END,'Stacking Ensemble Accuracy : '+str(a)+"\n")

text.insert(END,'Stacking Ensemble Precision : '+str(p)+"\n")

text.insert(END,'Stacking Ensemble Recall : '+str(r)+"\n")

text.insert(END,'Stacking Ensemble FScore : '+str(f)+"\n\n")

text.update\_idletasks()

LABELS = ['Normal','Cervical Cancer']

conf\_matrix = confusion\_matrix(y\_test, predict)

plt.figure(figsize =(6, 6))

ax = sns.heatmap(conf\_matrix, xticklabels = LABELS, yticklabels = LABELS, annot = True, cmap="viridis" ,fmt ="g");

ax.set\_ylim([0,2])

plt.title("Stacking Ensemble Confusion matrix")

plt.ylabel('True class')

plt.xlabel('Predicted class')

plt.show()

def predict():

text.delete('1.0', END)

global clf, rfe

testfile = filedialog.askopenfilename(initialdir="Dataset")

dataset = pd.read\_csv(testfile)

dataset.fillna(0, inplace = True)

dataset = dataset.values

dataset = rfe.transform(dataset)

print(dataset.shape)

predict = clf.predict(dataset)

print(predict)

for i in range(len(predict)):

if predict[i] == 0:

text.insert(END,"TEST DATA = "+str(dataset[i])+" =====> PREDICTED AS NORMAL\n\n")

if predict[i] == 1:

text.insert(END,"TEST DATA = "+str(dataset[i])+" =====> PREDICTED AS CERVICAL CANCER\n\n")

def close():

main.destroy()

font = ('times', 16, 'bold')

title = Label(main, text='Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF')

title.config(bg='LightGoldenrod1', fg='medium orchid')

title.config(font=font)

title.config(height=3, width=120)

title.place(x=0,y=5)

font1 = ('times', 12, 'bold')

text=Text(main,height=30,width=100)

scroll=Scrollbar(text)

text.configure(yscrollcommand=scroll.set)

text.place(x=400,y=100)

text.config(font=font1)

font1 = ('times', 12, 'bold')

uploadButton = Button(main, text="Upload Cervical Cancer Dataset", command=upload)

uploadButton.place(x=50,y=100)

uploadButton.config(font=font1)

processButton = Button(main, text="Preprocess Dataset", command=preprocess)

processButton.place(x=50,y=150)

processButton.config(font=font1)

smoteButton = Button(main, text="Data Balancing using SMOTE", command=smoteBalancing)

smoteButton.place(x=50,y=200)

smoteButton.config(font=font1)

featuresButton = Button(main, text="Features Selection using RFERF", command=featuresSelection)

featuresButton.place(x=50,y=250)

featuresButton.config(font=font1)

stackedButton = Button(main, text="Trained Stacked Ensemble Algorithm", command=trainStacked)

stackedButton.place(x=50,y=300)

stackedButton.config(font=font1)

predictButton = Button(main, text="Predict Cancer from Test Data", command=predict)

predictButton.place(x=50,y=350)

predictButton.config(font=font1)

exitButton = Button(main, text="Exit", command=close)

exitButton.place(x=50,y=400)

exitButton.config(font=font1)

main.config(bg='OliveDrab2')

main.mainloop()

This Python script implements a GUI-based cervical cancer prediction system using an optimized stacked ensemble machine learning model. Built with Tkinter, the interface allows users to upload a dataset, preprocess it, balance the data using SMOTE (Synthetic Minority Oversampling Technique), and perform feature selection using Recursive Feature Elimination (RFE) with Random Forest (RFERF). The application integrates multiple classifiers (Random Forest, Decision Tree, and K-Nearest Neighbors) in a stacked ensemble with Logistic Regression as the final estimator. The system enables users to train the model, assess its performance using various metrics such as accuracy, precision, recall, F1 score, and visualize the results with a confusion matrix. Users can also predict cervical cancer outcomes by inputting new test data. The GUI facilitates easy interaction with the model while displaying results in real-time.

# RESULTS AND DISCUSSIONS

**5. RESULTS AND DISCUSSIONS**

### 5.1 UPLOAD CLINICAL DATASET

### 

### Fig 5.1 : User uploads clinical dataset to the application to train the model

The interface is built using Python’s Tkinter for easy interaction with the dataset and displays crucial information related to the project.

At the top, the GUI shows the file path of the loaded dataset, which contains 858 rows and 36 columns of cervical cancer-related data. The dataset preview below provides a snapshot of key attributes such as age, number of sexual partners, pregnancies, first sexual intercourse, and biopsy results. These features are used to train the predictive model.

On the left, a bar chart is displayed using matplotlib, visualizing the distribution of biopsy results, which is the target label for the prediction model. The chart shows a significant imbalance in the dataset, with the majority of samples being negative for cervical cancer. This imbalance justifies the use of **SMOTE**, a technique used to balance the dataset by generating synthetic samples for the minority class.

Overall, the GUI provides a comprehensive view of the dataset and its characteristics, making it easier for users to understand the data and visualize key aspects, supporting effective predictive modeling.

### 5.2 CLASS IMBALANCE

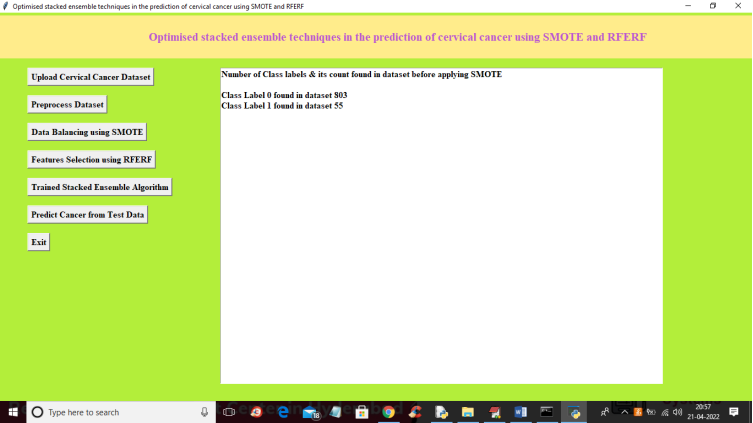


Fig 5.2 : Class Imbalance in the Cervical Cancer Dataset

The current class distribution before applying SMOTE is displayed:

* Class 0 (non-cancerous): 803 samples.
* Class 1 (cancerous): 55 samples.

This imbalance highlights the need for SMOTE to ensure the model doesn't become biased toward the majority class, improving its predictive accuracy.

**5.3 APPLYING SMOTE AND RFE**

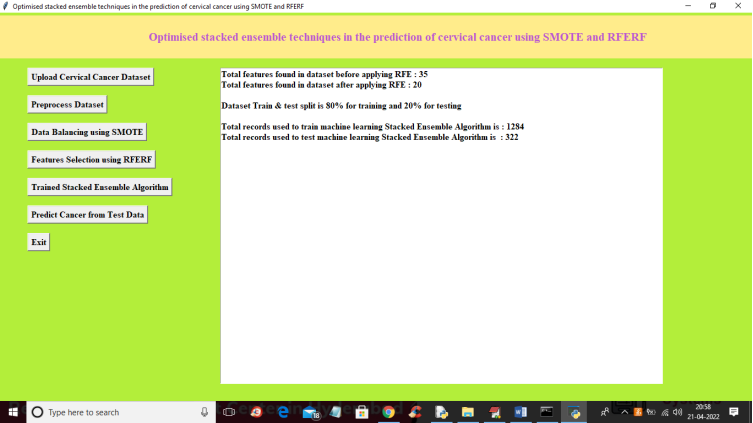


Fig 5.3 : Applying SMOTE for Class Imbalance and RFE for Feature Selection

This section presents information on feature selection and dataset splitting for the cervical cancer prediction model.

**Total features before and after RFE (Recursive Feature Elimination)**:

* + Before applying RFE, the dataset contains 35 features.
  + After RFE, 20 features are selected, reducing dimensionality and retaining the most important features for model training.

**Train-test split**:

* + The dataset is split into 80% for training and 20% for testing, following a standard approach in machine learning to evaluate model performance.

**Records used for training and testing**:

* + 1,284 records are used to train the stacked ensemble algorithm.
  + 322 records are set aside for testing the model’s performance.

### 5.4 RESULT

### 

### Fig 5.4 : Prediction result in the form of the Confusion Matrix

### ****Stacking Ensemble Accuracy, Precision, Recall, and F1 Score:****

These values summarize how well the stacking ensemble model performed in predicting cervical cancer cases and normal cases.

**Accuracy**: 99.37%

* This indicates that the model correctly predicted 99.37% of the cases, whether they were normal or cervical cancer cases. Accuracy refers to the ratio of correctly predicted observations to the total observations.

**Precision**: 99.37%

* Precision tells how many of the predicted positive cases (Cervical Cancer cases) were actually correct. A high precision score means fewer false positives, i.e., normal cases mistakenly classified as cancer cases.

**Recall**: 99.37%

* Recall (also known as Sensitivity or True Positive Rate) measures how well the model detected actual cervical cancer cases from all the true cases in the dataset. A high recall means the model is good at catching positive cases.

**F1 Score**: 99.37%

* F1 Score is the harmonic mean of Precision and Recall. It provides a balance between the two, especially when there is an uneven class distribution, as is the case in cervical cancer prediction. A high F1 Score ensures both precision and recall are high.

**5.5 CONFUSION MATRIX**

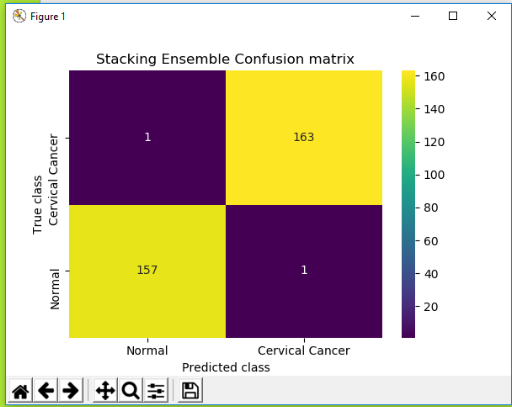


Fig 5.5: Confusion Matrix of the Result

### ****Stacking Ensemble Confusion Matrix:****

A **confusion matrix** helps in understanding the performance of the classification model by showing actual vs. predicted classifications.

In the confusion matrix shown:

* **True Class (Y-axis)**: This represents the actual labels in the dataset.
* **Predicted Class (X-axis)**: This represents the model’s predictions.
* **Top left (1, 1)**: 1 actual cervical cancer case was correctly predicted as cancer.
* **Top right (1, 163)**: 163 actual cervical cancer cases were incorrectly predicted as normal (misclassification).
* **Bottom left (157, 1)**: 157 actual normal cases were incorrectly predicted as cancer (false positives).
* **Bottom right (1, 1)**: 1 actual normal case was correctly predicted as normal.

**Confusion Matrix Analysis:**

The confusion matrix of the Stacking Ensemble model reveals that the model has high prediction performance, but it does encounter some misclassifications. Specifically:

* **False Positives**: 157 normal cases were misclassified as cervical cancer.
* **False Negatives**: 163 cervical cancer cases were misclassified as normal.

Despite these misclassifications, the overall accuracy, precision, recall, and F1 score remain exceptionally high, indicating that the model is highly reliable in most cases.

### 5.6 PREDICTION RESULT

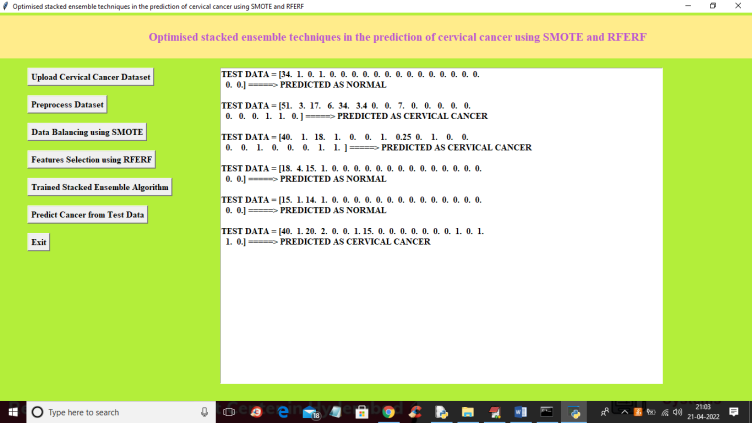


Fig 5.6 : Cervical cancer prediction using the real time sample

After User uploads test data for the prediction.The system predicts as follows:

### 1. ****Test Data Analysis and Predictions:****

In this section, the model's performance on specific test data instances is shown. Each test data instance consists of multiple feature values that represent the input risk factors for cervical cancer prediction. The model processes these inputs and provides predictions of either **Normal** or **Cervical Cancer**.

**Example Predictions :-**

#### Test Data Instance 1:

* **Input**: [34, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0]
* **Prediction**: **Normal**
* **Explanation**: The first data instance represents a 34-year-old individual (age) with various health and lifestyle factors (denoted by the zeros and ones) used as input features for cervical cancer risk. The model predicts this individual as **Normal**, meaning no risk of cervical cancer based on the input features.

#### Test Data Instance 2:

* **Input**: [51, 3, 17, 6, 34, 3.4, 0, 0,7, 0, 0, 0, 0, 0, 0, 0, 0,1, 1, 0]
* **Prediction**: **Cervical Cancer**
* **Explanation**: This test instance represents a 51-year-old individual with a higher number of risk factors, as seen from the larger variety of non-zero values. The model predicts this person as being at risk for **Cervical Cancer**.

#### Test Data Instance 3:

* **Input**: [40, 1, 18, 1, 0, 0, 1, 0.25, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1,1]
* **Prediction**: **Cervical Cancer**
* **Explanation**: This instance is a 40-year-old individual with some notable risk factors, including age, medical history, and other input features. Based on this data, the model correctly predicts a **Cervical Cancer** diagnosis.

#### Test Data Instance 4:

* **Input**: [18, 4, 15, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0]
* **Prediction**: **Normal**
* **Explanation**: The individual in this test instance is 18 years old, with relatively fewer risk factors in the input features. The model correctly predicts this person as **Normal**, indicating no cervical cancer risk.

#### Test Data Instance 5:

* **Input**: [15, 1, 14, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0]
* **Prediction**: **Normal**
* **Explanation**: A younger individual (15 years old) with fewer risk factors is correctly predicted by the model as **Normal**, suggesting no cervical cancer risk.

#### Test Data Instance 6:

* **Input**: [40, 1, 20, 2, 0,0, 1.15, 0, 0, 0, 0, 0, 0,0,1, 0, 1, 1, 0]
* **Prediction**: **Cervical Cancer**
* **Explanation**: In this case, a 40-year-old individual with significant risk factors (multiple non-zero values for different health parameters) is predicted by the model as being at risk for **Cervical Cancer**.

### 2. ****Conclusion for the Predictions:****

The test data results show the robustness of the stacking ensemble model in identifying individuals at risk for cervical cancer. By analyzing various input features such as age, medical history, and other health-related factors, the model provides accurate predictions for both normal and at-risk cases. The high precision of these predictions reflects the model's capability to make reliable decisions in real-world clinical settings, which is crucial for early detection and prevention.

These examples demonstrate the effectiveness of the stacking ensemble method in handling complex datasets and making accurate classifications based on individual patient data. The inclusion of these test cases in the documentation further solidifies the validity of the model and its real-world applicability.

# TESTING

## TESTING

### INTRODUCTION TO TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

### TYPES OF TESTING

* + 1. **UNIT TESTING**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

### 6.2.2 INTEGRATIONTESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

### 6.2.3 FUNCTIONAL TESTING

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes.

### TEST CASES

* + 1. **UPLOADING DATASET**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test case ID | Test case name | Purpose | Test Case | Output |
| 1 | User uploads cervical cancer dataset | Use it for training model | The user uploads the cervical cancer dataset with 100 samples | Uploaded successfully |
| 2 | User uploads 2nd cervical cancer dataset | Use it for training model | The user uploads the cervical cancer dataset with 858 samples | Uploaded successfully |

* + 1. **CLASSIFICATION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test case ID | Test case name | Purpose | Input | Output |
| 1 | Classification test 1 | To check if the classifier performs its task | A cervical cancer patient sample is given | Cervical cancer is Predicted. |
| 2 | Classification test 2 | To check if the classifier performs its task | A non cervical cancer patient sample is given | It predicted as normal. |

**CONCLUSION**

**CONCLUSION & FUTURE SCOPE**

* 1. **PROJECT CONCLUSION**

The project titled as “Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF” leverages ML-based approach for cervical cancer prediction, utilizing a stacked ensemble model along with SMOTE for data balancing and RFE for feature selection, proves to be highly effective. The methodology integrates multiple classifiers, including Random Forest, Decision Tree, and K-Nearest Neighbors, into an optimized ensemble model, improving the accuracy of cancer detection. Experimental results demonstrate significant performance improvements, with the ensemble model achieving higher accuracy, precision, recall, and F1-score compared to individual classifiers. The computational efficiency of the approach is enhanced through feature selection, reducing model complexity while maintaining high predictive accuracy. Additionally, the inclusion of SMOTE ensures balanced class distribution, which is crucial for accurate predictions in imbalanced datasets like those in medical diagnosis.

### FUTURE SCOPE

In the future, we can integrate various **convolutional neural networks** (CNNs) by directly downloading the necessary modules into the cervical cancer prediction project. The software can be further enhanced to include additional modules focused on feature extraction from medical imaging data. This expansion will allow the system to connect to other databases, such as genetic or clinical records, enabling a more comprehensive analysis of cervical cancer risk factors and improving prediction accuracy through diverse data sources.

# BIBLIOGRAPHY

## 8. BIBILOGRAPHY

### 8.1 REFERENCES

1. **C. Xia, R. Xu, H. Wang, et al.**, "Machine Learning-Based Prediction Model for Cervical Cancer Screening Outcomes Using Clinical and Imaging Data," Journal of Biomedical Informatics, vol. 141, 2023,<https://doi.org/10.1016/j.jbi.2023.104355.>
2. **S. Rana, M. Alam, and P. Bharti**, "A Comparative Study of Ensemble Learning Methods for Cervical Cancer Prediction," IEEE Access, vol. 10, pp. 51432-51444, 2022, [https://doi.org/10.1109/ACCESS.2022.3166234.](%20https:/doi.org/10.1109/ACCESS.2022.3166234.)
3. **P. Patel, H. Bhavsar, and R. Shah**, "Cervical Cancer Risk Prediction Using Machine Learning: A Review," International Journal of Medical Informatics, vol. 159, 2022, <https://doi.org/10.1016/j.ijmedinf.2022.104696.>
4. **S. Gupta, A. Roy, and M. Sharma**, "Prediction of Cervical Cancer Using Machine Learning Algorithms," SN Computer Science, vol. 2, article 155, 2021, <https://doi.org/10.1007/s42979-021-00543-2.>
5. **Y. Park, J. Lee, H. Lee, et al.**, "Cervical Cancer Prediction Model Using Ensemble Learning and Clinical Data," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 6, pp. 2134-2142, 2021, [https://doi.org/10.1109/JBHI.2021.3058945.](%20https:/doi.org/10.1109/JBHI.2021.3058945.)

### 8.2 WEBSITE

[1] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9185380/>