Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 4, April : 2023

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

[1]CHUNDRU YASWANTH SAI KIRAN, [2]JAJIMOGGALA HARITHA VIJAYA BAI, [3]BALIREDDY APPALA DHANUSH, 4th year B. Tech Students, Dept. of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam

MR P NARASIMHA RAJU Assistant Professor, Dept. of Computer Science and Engineering, Raghu Engineering College, Visakhapatnam

19981a0535@raghuenggcollege.in , 19981a0560@raghuenggcollege.in ,

19981a0519@raghuenggcollege.in , narasimharaju.p@raghuenggcollege.in

ABSTRACT

Cervical cancer, which effects many women, is frequently fatal. The mortality rate and other related consequences can be decreased with the help of early identification of this disease. Related factors which leads to cervical cancer can help in early identification of this disease. Using this factors feature set and 2 ensemble-based classification algorithms, such as the Recursive Feature Elimination Technique along with Random Forest (RFE-RF) and Synthetic Minority Oversampling Techniques (SMOTE). The Cervical Cancer required Factors data consists of 32 possible causes and 4 targets, is utilized by the study (Hinselmann, Schiller, Cytology, and Biopsy). The 4 targets are utilized in the widely used cervical cancer detection test. Moreover, the Firefly features selection technique was applied to improve results while using fewer characteristics. When compared to the other three diagnostic tests, experimental results for the Biopsy test showed the proposed model to be significant and to have the best success rate.

Keywords: Synthetic Minority Oversampling Technique(SMOTE), Recursive Feature Elimination Technique(RFE-RF), 4 target tests.

1. INTRODUCTION

Millions of new cells are often produced by the human body to replace old and dead ones. Tumors will form if this cell production rises in an unpredictable way. Not every tumour is cancerous and does not leave the body. Some tumours that are left in the body develop into malignant ones [1]. A type of cancer called cervical cancer develops at the cervix of pregnant women. If detected at an early stage, this kind of cancer is treatable. These cells gradually spread through the tissues, affect the neighbouring organs, and progress to an advanced stage of cancer. This form of cancer is primarily caused by the HPV [2].

Cervical cancer is the second most frequent disease in women in India, and it comes fourth in the league among all female cancers worldwide, according to the WHO [3][4]. One-third of all cervical cancer deaths occur in India, where the cumulative risk is 1.6% and the cumulative fatality rate is 1% [4].The

DOI: 10.36893.IEJ.2023.V52I04.196-201

main causes of this type of cancer were believed to be a lack of awareness, a lack of skilled medical professionals and equipment, and a lack of early identification, especially in middle- and low income countries [5].

Although machine learning techniques are already widely used and essential to the diagnosis and prognosis of all diseases, manual disease diagnosis will continue to be very difficult [6]. By educating computers from past or prior examples, machine learning allows us to use complicated data to predict the future [7]. Machine learning makes use of a range of probability, statistical, and optimizing methodologies to classify as different events like one event with cancer and other with no cancer [8]. In the previous work, they used the stacked ensemble technique with diverse base classifiers, comprising SVM and RF [9].

The dataset was collected from the "Hospital Universitario de Caracas."

2. LITERATURE REVIEW

Ensemble methods like SVM and Random Forest algorithms were used in the literature review [10] that gives different machine learning techniques for identifying this cervical cancer disease.

The SVM algorithm is used by the current system. Support Vector Machine is a classification method that is used for classification, regression, or novelty detection in many different applications. After training in terms of classification, SVM is capable of categorizing the entering examples into separate categories. Minimizing the penalty factor and increasing the marginal width are the primary objectives of the SVM's hyperplane construction.

The necessary dataset was obtained from the "Hospital Universitario de Caracas." This one has 858 records, some of which contain missing values since some patients opt not to respond to some questions out of respect for their privacy. The data collection includes 32 required factors and 4 targets test values, or the cervical disease diagnosis tests.

3. PROPOSED APPROACH

To overcome the imbalanced data problem and efficiency problem we came with two different algorithms those are SMOTE and RFE-RF.

SMOTE: It stands for **Synthetic Minority Oversampling Technique:** It fixes issues brought on by utilizing an unbalanced data collection. It is a technique that adds new artificial data to the original data to overcome the problem of under-sampling. This Technique can be seen as a form of balancing the imbalanced data. The benefit of this technique is that it doesn't generate duplicate data; instead, it generate artificial data that are marginally different from the original data [11].

RFE-RF: It stands for RECURSIVE FEATURE ELIMINATION USING RANDOM FOREST. First to know about RFE, RF should be explained that is Random Forest.

Random Forest: Both classification and regression applications use this machine learning ensemble technique, often known as the bagging technique. It is a technique that takes different decision trees upon different sub categories of the given data and sum them to boost the expected accuracy of the data. The random forest takes results of identification or predictions from each tree and gives the result based on the more number of projections[12], as opposed to relying just on one decision tree.

Recursive Feature Elimination: As the name suggests, is a method for repeatedly eliminating the least important feature up until the desired number of features [13].

The number of features needed for the prediction will be taken into account after employing random forest, and any additional features will be removed using recursive feature elimination.

4. METHOD

Work Flow :

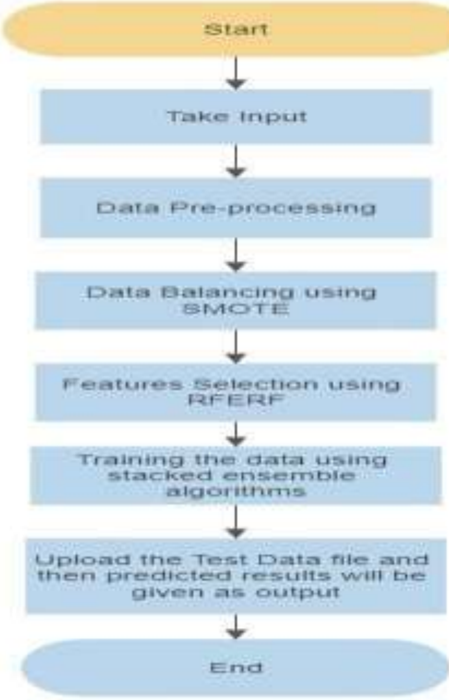
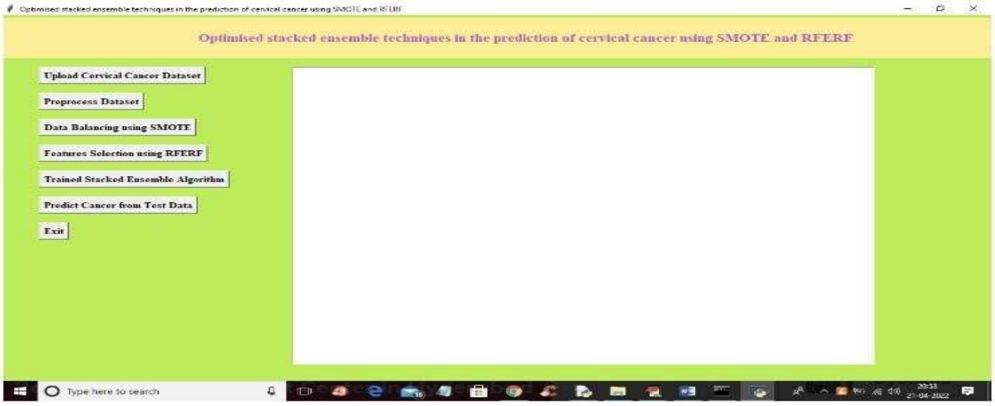

Start
Take Input
Data Pre-processing
Data Balancing using SMOTE
Features Selection using RFERF
Training the data using stacked ensemble algorithms
Upload the Test Data file and then predicted results will be given as output
End

Figure 4.1 Flow chart of the proposed system

System implementation:

GUI that is developed for the Prediction of Cervical Cancer, which consists of several functionalities is in below Figure 4.2


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF

Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search

Figure 4.2 GUI of Prediction of Cervical Cancer

Step-1: Uploading Cervical Cancer Dataset:

The necessary dataset was obtained from the "Hospital Universitario de Caracas" . This one has 858 records, some of which contain missing values since some patients opt not to respond to some questions out of respect for their privacy. The data collection includes 32 required factors and 4 targets test values.

Step-2: Pre-processing the dataset:

Since the dataset has some missing values the dataset has to be pre-processed to remove noise and dimensionality. This can be done by clicking the “pre-process dataset” button.

Step-3: Data Balancing using SMOTE:

To balance the imbalanced dataset SMOTE is used. This can be done by clicking the “Data Balancing using SMOTE” button.

Step-4: Feature Selection using RFERF:

Only the required Features for the prediction will be taken into consideration others will be eliminated by using RFERF algorithm. This can be done by clicking “Feature Selection using RFERF” button.

Step-5: Trained Stacked Ensemble Algorithm:

This is used to train the algorithm and to give the accuracy of the algorithm along with the confusion matrix. This can be done by clicking “Trained Stacked Ensemble Algorithm” button.

Step-6: Predict Cancer from Test Data:

Now a test dataset has to uploaded here by clicking the “Predict Cancer from Test Data” button. After uploading the data, the prediction will be made for the members present in the test dataset.

RESULTS

Output:

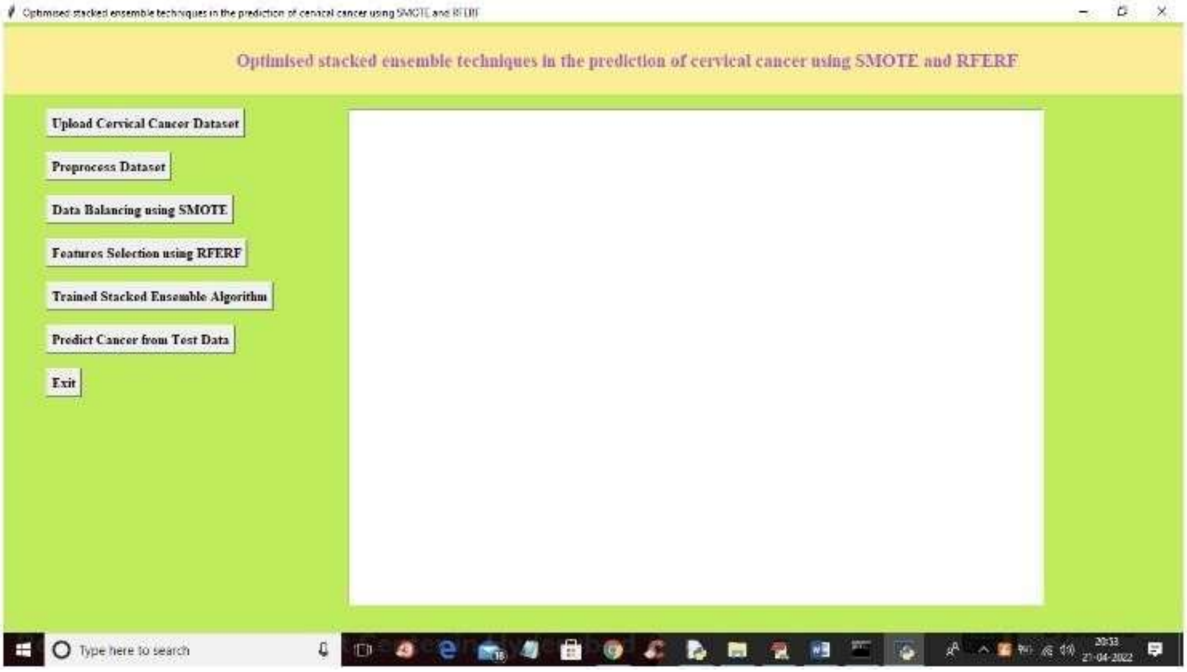

Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search
20:33
21-04-2022

Figure 5.1 of Outputs

In above image select the 'Upload Cervical Cancer Dataset' button to upload the dataset and to get below image.

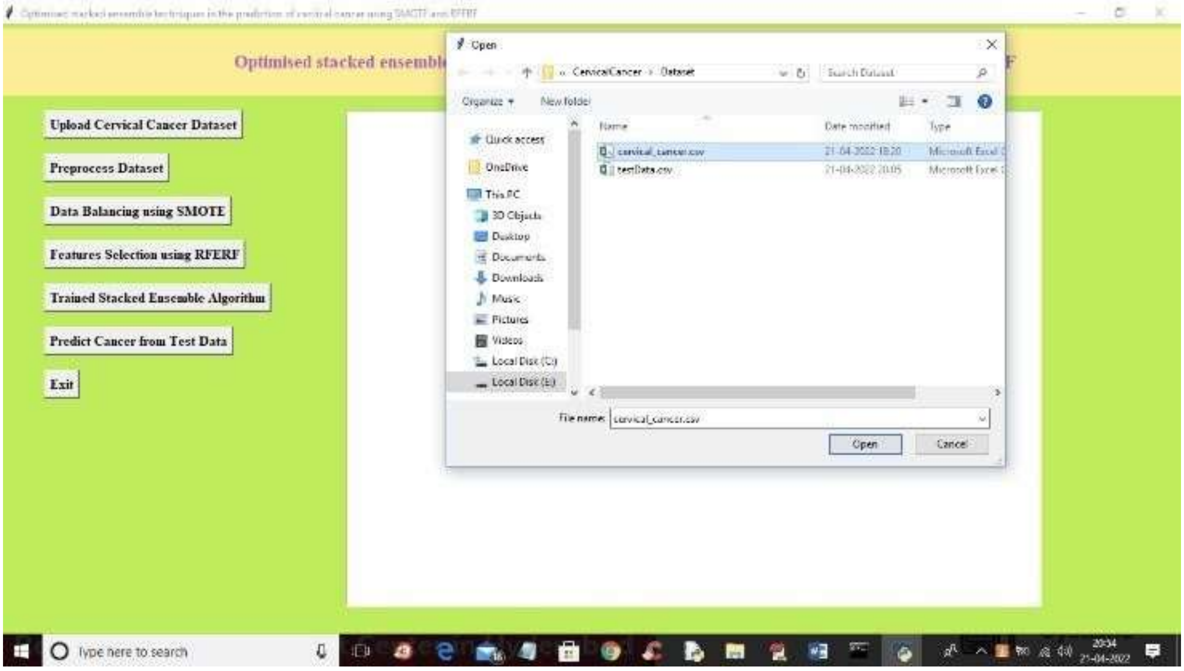
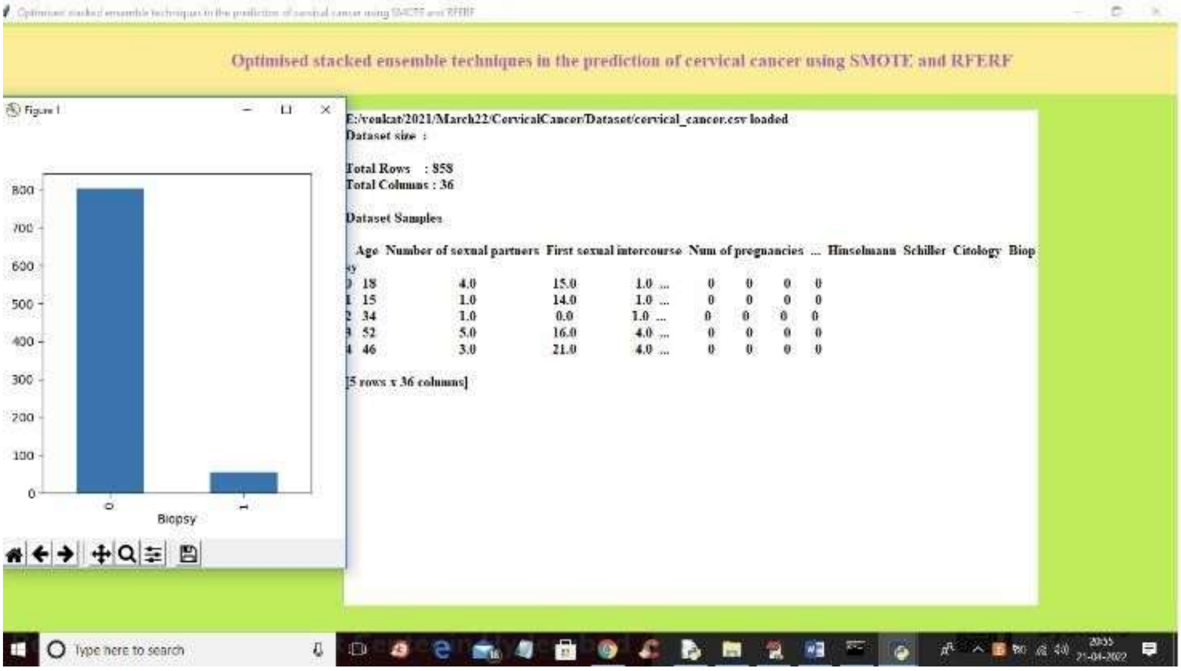


Figure 5.2 of Outputs

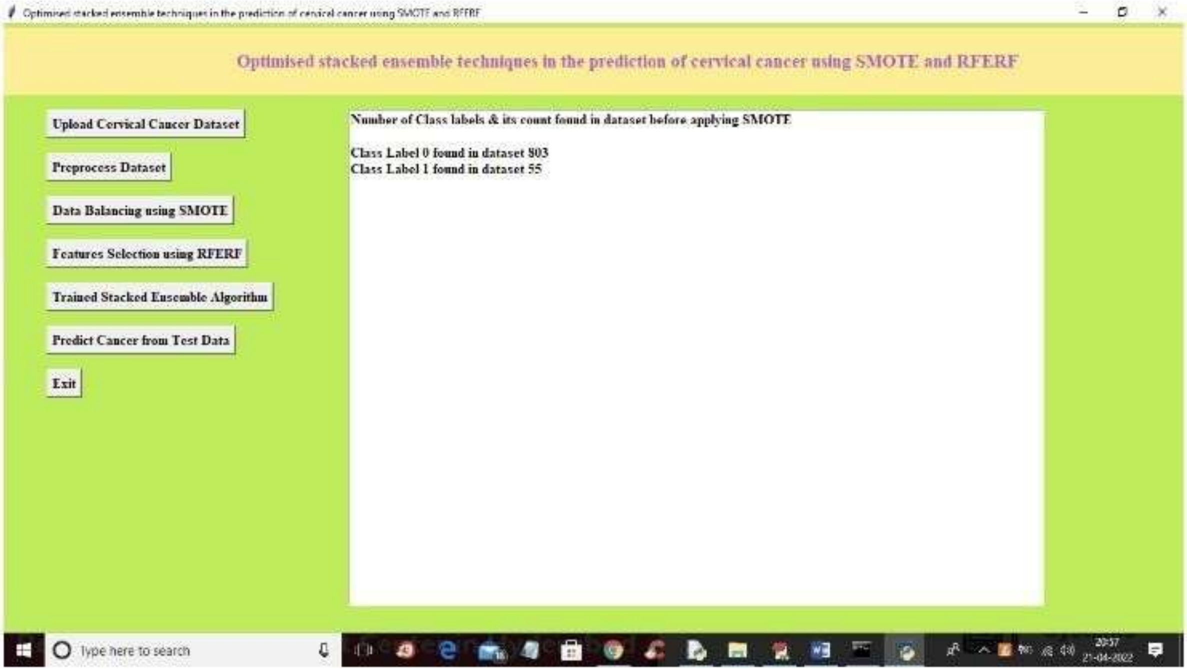
In above image you can see the uploading dataset file and then select the 'Open' button to upload data and to get below image.


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Figure 1
800
700
600
500
400
300
200
100
0
0
1
Biopsy
Type here to search
20:35
21-04-2022

```
E:/venkat/2021/March22/CervicalCancer/Dataset/cervical_cancer.csv loaded
Dataset size  :

Total Rows    : 858
Total Columns : 36

Dataset Samples

  Age  Number of sexual partners  First sexual intercourse  Num of pregnancies  ...  Hinselmann  Schiller  Citology  Biop
sy
0  18              4.0                   15.0                    1.0  ...         0         0         0     0
1  15              1.0                   14.0                    1.0  ...         0         0         0     0
2  34              1.0                    0.0                    1.0  ...         0         0         0     0
3  52              5.0                   16.0                    4.0  ...         0         0         0     0
4  46              3.0                   21.0                    4.0  ...         0         0         0     0

[5 rows x 36 columns]
```

Figure 5.3 of Outputs

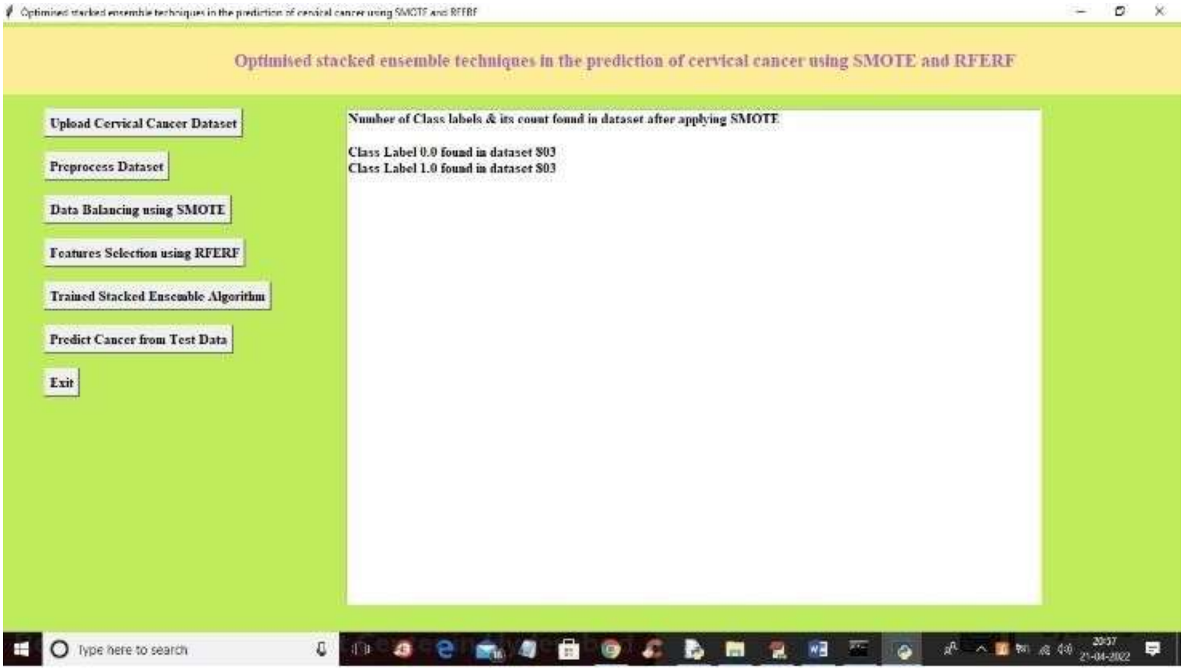
The dataset is loaded in the image above. The graph's x-axis shows the class labels 0 (normal) and 1 (cervical cancer), and the y-axis shows the number of records. This shows that the dataset is highly unbalanced because class label '0' includes more than 800 records.**and the "1" class label only has 53 entries. Close the aforementioned graph now, and then select "Pre-process Dataset" to fill in any missing numbers and normalise the dataset**.


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search
20:37
21-04-2022

```
Number of Class labels & its count found in dataset before applying SMOTE

Class Label 0 found in dataset 803
Class Label 1 found in dataset 55
```

Figure 5.4 of Outputs

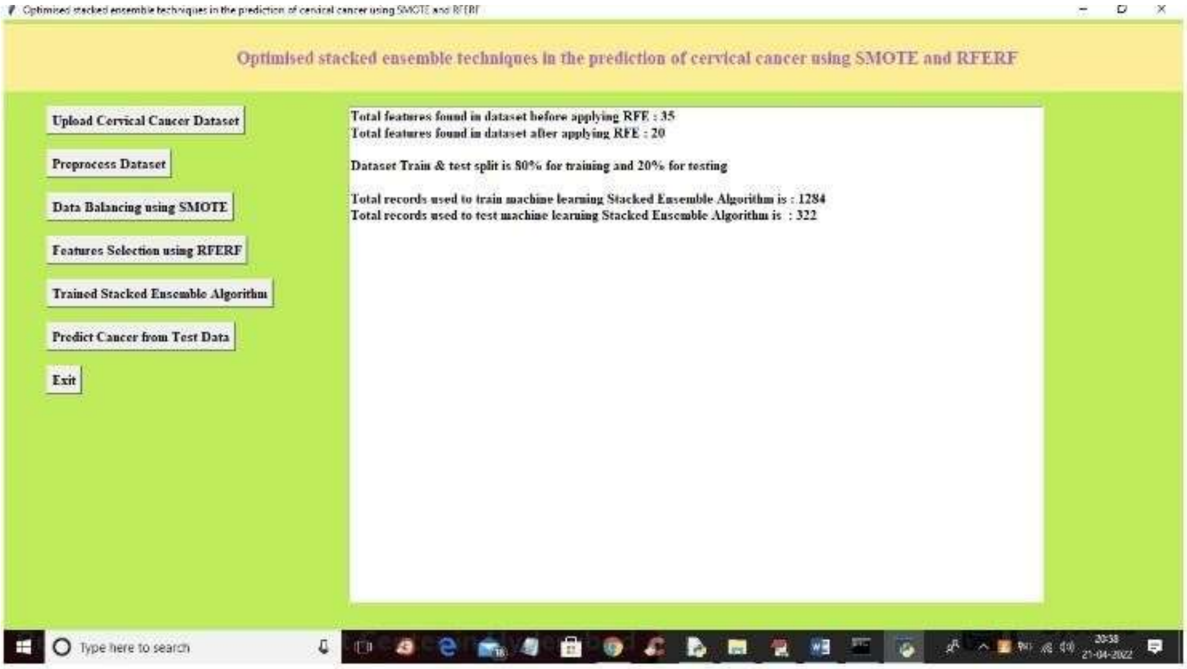
In above screen dataset is processed and class 0 contains 805 records and 1 contains 55 records so click on 'Data Balancing using SMOTE' button for balancing the data.


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search
20:37
21-04-2022

```
Number of Class labels & its count found in dataset after applying SMOTE

Class Label 0.0 found in dataset 803
Class Label 1.0 found in dataset 803
```

Figure 5.5 of Outputs

In above screen after applying smote both classes contains 803 records so dataset is balanced and now click on ‘Features Selection using RFERF’ button to get below output


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search
20:38
21-04-2022

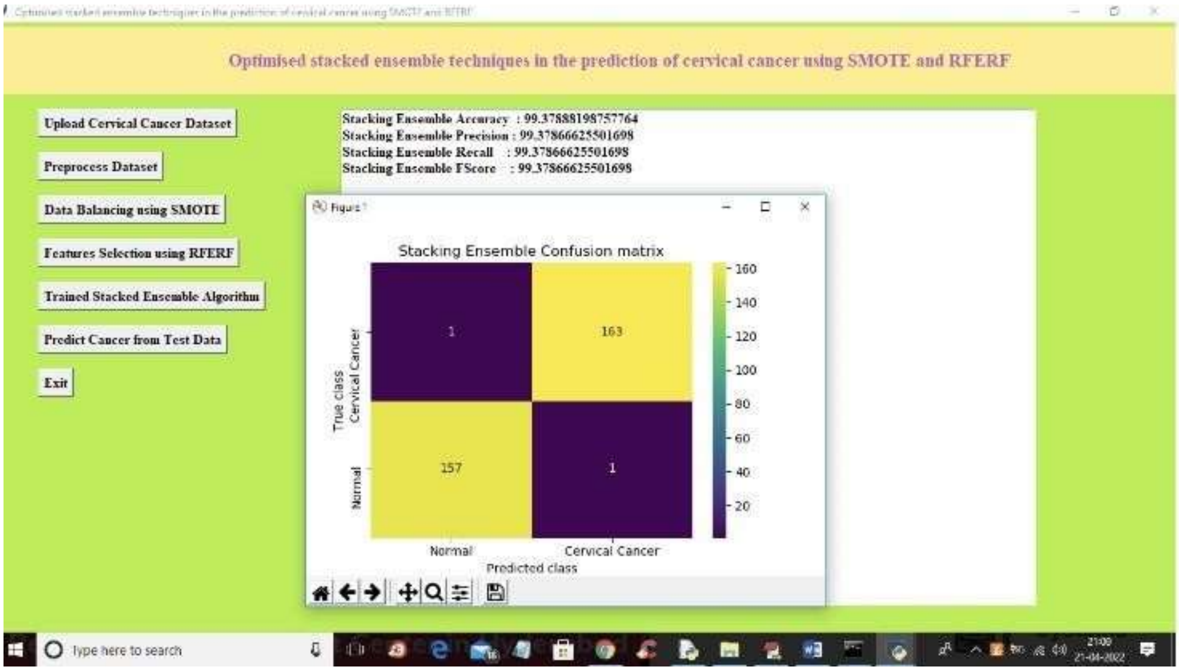
```
Total features found in dataset before applying RFE : 35
Total features found in dataset after applying RFE : 20

Dataset Train & test split is 80% for training and 20% for testing

Total records used to train machine learning Stacked Ensemble Algorithm is : 1284
Total records used to test machine learning Stacked Ensemble Algorithm is  : 322
```

Figure 5.6 of Outputs

In above screen in first 2 lines we can see dataset contain 35 attributes and after applying RFE, attributes size reduced to 20 and then we can see dataset train and test split details. Now dataset is ready to go and next select the ‘Trained Stacked Ensemble Algorithm’ button

DOI: 10.36893.IEJ.2023.V52I04.196-201

to train the data.

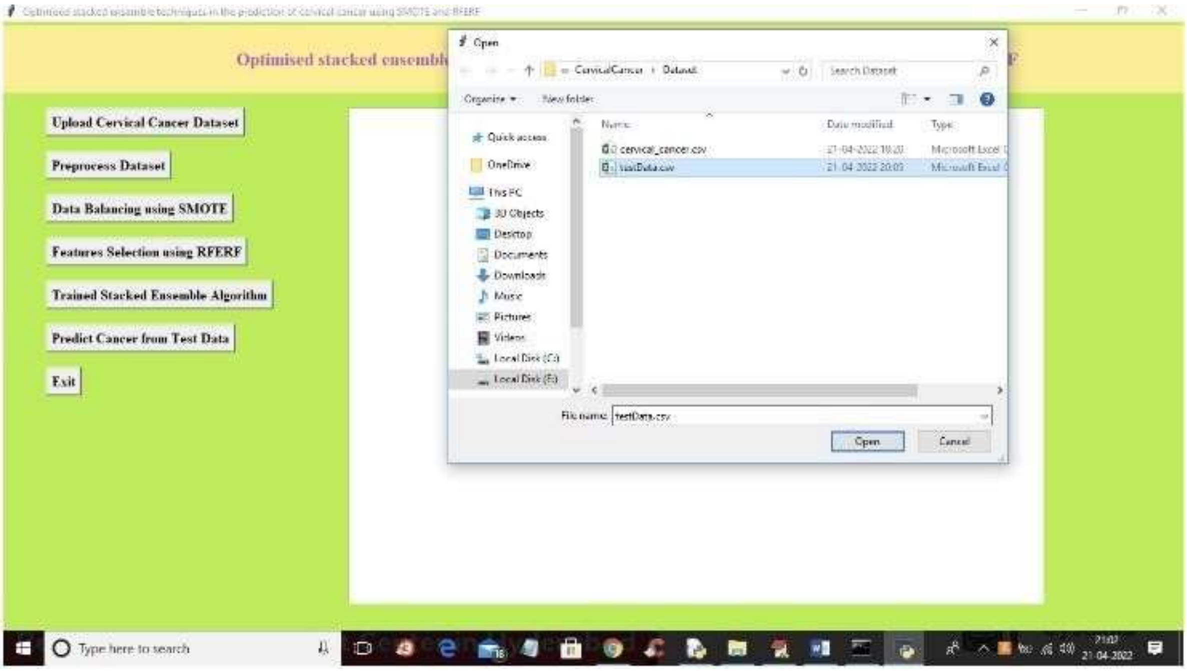

Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Figure 1
Stacking Ensemble Confusion matrix
True class
Cervical Cancer
Normal
Predicted class
Normal
Cervical Cancer
1
163
157
1
160
140
120
100
80
60
40
20
Type here to search

```
Stacking Ensemble Accuracy  : 99.37888198757764
Stacking Ensemble Precision : 99.37866625501698
Stacking Ensemble Recall    : 99.37866625501698
Stacking Ensemble FScore    : 99.37866625501698
```

Figure 5.7 of Outputs

With the stacked ensemble algorithm, we achieved 99.37% accuracy in the image above. In the above graph from the image, we can see that the x-axis gives the predicted classes information, the y-axis gives the true classes, and that only 1 class was incorrectly predicted.

The prediction was accurate. Close the previous graph, then select "Predict Cancer from Test Data" to submit a test file and obtain the output shown below.



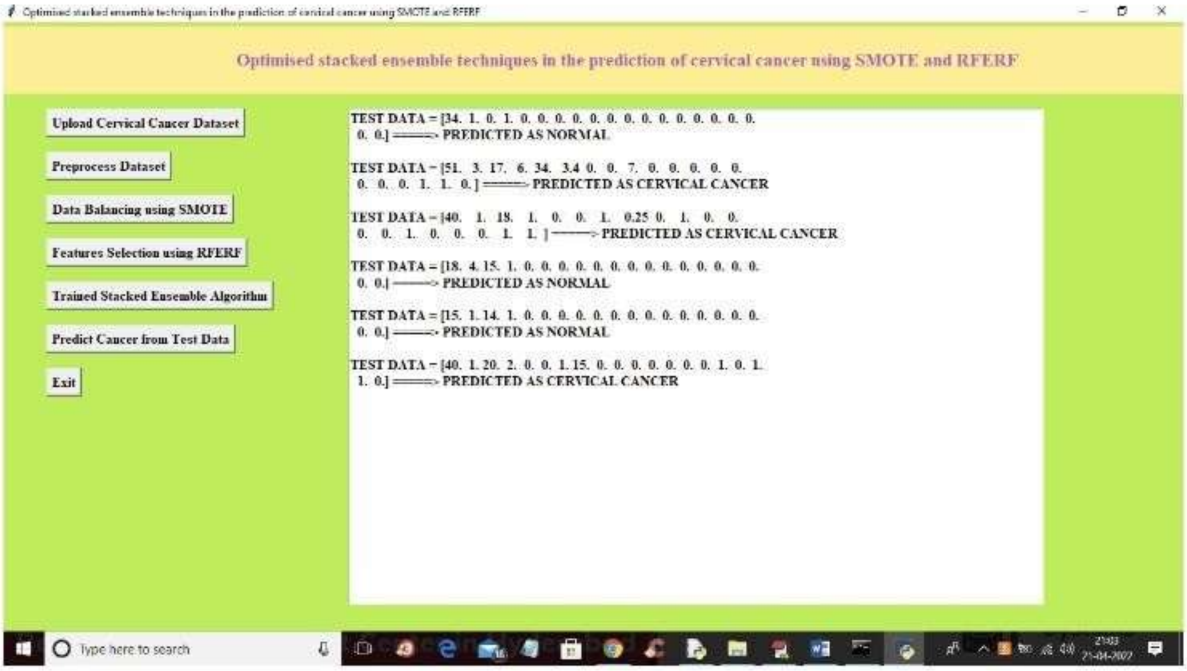
Name	Date modified	Type
cervical_cancer.csv	21-04-2022 19:20	Microsoft Excel C
testData.csv	21-04-2022 20:09	Microsoft Excel C

File name: testData.csv

Figure 5.8 of Outputs

DOI: 10.36893.IEJ.2023.V52I04.196-201

In above image select and upload test Data file.


Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Optimised stacked ensemble techniques in the prediction of cervical cancer using SMOTE and RFERF
Upload Cervical Cancer Dataset
Preprocess Dataset
Data Balancing using SMOTE
Features Selection using RFERF
Trained Stacked Ensemble Algorithm
Predict Cancer from Test Data
Exit
Type here to search

```
TEST DATA = [34. 1. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0.] ====> PREDICTED AS NORMAL

TEST DATA = [51.  3. 17.  6. 34.  3.4 0.  0.  7.  0.  0.  0.  0.  0.
 0.  0.  0.  1.  1.  0.] ====> PREDICTED AS CERVICAL CANCER

TEST DATA = [40.   1.  18.   1.   0.   0.   1.   0.25 0.   1.   0.   0.
 0.   0.   1.   0.   0.   0.   1.   1. ] ----> PREDICTED AS CERVICAL CANCER

TEST DATA = [18. 4. 15. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0.] ----> PREDICTED AS NORMAL

TEST DATA = [15. 1. 14. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.
 0. 0.] ====> PREDICTED AS NORMAL

TEST DATA = [40. 1. 20. 2. 0. 0. 1. 15. 0. 0. 0. 0. 0. 0. 0. 1. 0. 1.
 1. 0.] ====> PREDICTED AS CERVICAL CANCER
```

Figure 5.9 of Outputs

Conclusion

This project provides an application with a user-friendly graphical user interface (GUI). Where prediction can be made very easily with high accuracy and with high efficiency unlike other prediction systems as there is no need to take every feature into consideration for the prediction every time, only the required features will be considered and remaining will be eliminated using RFERF algorithm which increase efficiency. And imbalanced data will be balanced by using SMOTE so that accuracy will be more.

REFERENCES

[1] American Cancer Society. Cancer Facts & Figures, 2018.

[2] M. Schiffman, P.E. Castle, J. Jeronimo, A.C. Rodriguez, S. Wacholder, Human papillomavirus and cervical cancer, Lancet 370 (9590) (2007) 890–907.

[3] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer 136 (5) (2015) E359–E386.

[4] Monica, Mishra, R. An epidemiological study of cervical and breast screening in India: district- level analysis. BMC Women's Health 20, 225 (2020). Doi: 10.1186/s12905-020-01083-6.

[5] G.A. Mishra, S.A. Pimple, S.S. Shastri, An overview of prevention and early detection of cervical cancers, Indian J. Med. Paediat. Oncol.: Off. J. Indian Soc. Med. Paediatric Oncol. 32 (3) (2011) 125–132, https://doi.org/10.4103/0971- 5851.92808.

[6] M. Rowe, An Introduction to Machine Learning for Clinicians, Acad Med. 94 (10) (2019 Oct) 1433–1436, https://doi.org/10.1097/ACM.0000000000002792, PMID: 31094727.

[7] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Science & Business Media, New York, 2009.

[8] Y.S. Abu-Mostafa, M. Magdon-Ismail, H.-T. Lin, AMLbook.com, Learning from Data, 2012.

[9] J. De Fauw et al., Clinically Applicable Deep Learning For Diagnosis and Referral in Retinal Disease, Nat Med. 24 (9) (Sep 2018) 1342–1350.

[10] Ch. Bhavani, Dr. A. Govardhan. Supervised Algorithms of Machine Learning in the Prediction of Cervical Cancer: A Comparative Analysis. Annals of the Romanian Society for Cell Biology, 1380–1393. 2021, Retrieved from https:// www.annalsofrscb.ro/index.php/journal/article/view/4580.

[11] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[12] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[13] https://machinelearningmastery.com/rfe-feature-selection-in-python/