

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

¹Bikkumalla Varun, ²Maisa Sugandh, ³Malyala Swathi, ⁴Perumala Purushotham,

⁵Kumbala Pradeep Reddy, ⁶S.Dhanalakshmi

^{1,2,3,4}UG Scholar, Department of CSE (AI&ML)

⁵Associate Professor, ⁶Professor, Department of CSE (AI&ML)

CMR Institute of Technology, Hyderabad, Telangana, India-501401

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 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."

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

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.

METHOD

Work Flow:

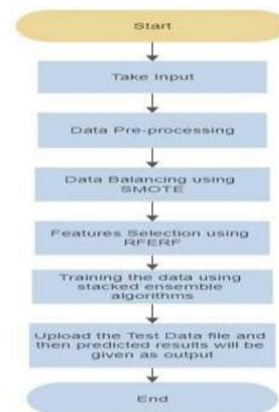


Figure 4.1 Flow chart of the proposed system

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



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.

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.00000000000002792>, PMID: 31094727.
1. S.Dhanalakshmi,,T.Prabakaran, and D.Krishna Kishore, "Content Delivery Networks - A Survey", *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)*, Volume 7, Issue 7, July 2017, and ISSN: 2277 - 128X, PP 228 – 230. (UGC JOURNALS)
 2. S.Dhanalakshmi,, D.Krishna Kishore, and T.Prabakaran, "Analysis and Performance of Divide and Conquer Methodology", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*,Volume 6, Issue 8, August 2017,PP 1295 - 1298,ISSN: 2278 – 1323 (UGC JOURNALS)
 3. S.Dhanalakshmi,, and T.Ravichandran, "Evaluating Breast Cancer using Mammogram Imaging", *Asian Journal of Information Technology*, ISSN: 1682-3915 (Print), Volume 16, Issue 2, November 2017, PP 178 – 180. (SCOPUS/Annexure II JOURNALS)
 4. S.Dhanalakshmi, D.Krishna Kishore, and T.Prabakaran, "AM - FM BASED METHOD IN RETINAL IMAGES FOR BLOOD VESSEL SEGMENTATION ", *International Journal of Electronics, Electrical and Computational Systems*, Volume 6,Issue 11,November 2017,ISSN: 2348 – 117X, PP 352 – 361. (UGC JOURNALS)
 5. S.Dhanalakshmi,, "Representation of Quad Tree Decomposition and Various Segmentation Algorithms in Information Bottleneck Method", *International Journal of Recent Trends in Engineering & Research (IJRTER)*,Volume 4,Issue2,February 2018,ISSN : 2455-1457, PP 198 -207 (UGC JOURNALS – 4.101)
 6. S.Dhanalakshmi,, K. Ramakrishna Reddy, and K. Vijaya Krupa Vatsal, "An Overview of Combinatorial Optimization Techniques", *International Journal of Scientific Research in Computer Science, Engineering and Information Technology(IJSRCSEIT)*,Volume 3,Issue1,January - March 2018,ISSN : 2456- 3307,PP 1052-1055 (UGC JOURNALS – 4.032)
 7. Sama Sindhuja Reddy, and S.Dhanalakshmi,, "CLEANING BIG DATA - AN INTERACTIVE APPROACH WITH SQOOP AND HIVE", *International Science and Technology Journal (ISTJ)*,Volume 7, Issue 5, May 2018, ISSN : 1632-2882, PP 56-60 (UGC JOURNALS – 5.3)
 2. 16. M. Satish, Prakash, S. M. Babu, P. P. Kumar, S. Devi and K. P. Reddy, "Artificial Intelligence (AI) and the Prediction of Climate Change Impacts," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg,

- Germany, 2023, pp. 660-664,
doi:
10.1109/ICCCMLA58983.2023.10346636.
3. 17. Prakash, S. M. Babu, P. P. Kumar, S. Devi, K. P. Reddy and M. Satish, "Predicting Consumer Behaviour with Artificial Intelligence," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023, pp. 638-643, doi:
10.1109/ICCCMLA58983.2023.10346963., pp. 687-692, doi:
10.1109/ICCCMLA58983.2023.10346916.
4. 18. S. M. Babu, P. P. Kumar, B. S. Devi, K. P. Reddy, M. Satish and A. Prakash, "Enhancing Efficiency and Productivity: IoT in Industrial Manufacturing," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023, pp. 693-697, doi:
10.1109/ICCCMLA58983.2023.10346807.
5. 19. Prakash, S. M. Babu, P. P. Kumar, S. Devi, K. P. Reddy and M. Satish, "Predicting Consumer Behaviour with Artificial Intelligence," *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, Hamburg, Germany, 2023, pp. 698-703, doi:
10.1109/ICCCMLA58983.2023.10346660.