

Classification of Cervical Cancer Detection Using Machine Learning Algorithms

Manohar Zirwal¹ Dhannjay Yeola²

^{1,2}Department of Computer Engineering

^{1,2}Loknete Gopinathji Munde Institute of Engineering Education & Research, Nashik, Maharashtra, India

Abstract— Cervical cancer remains an important reason of deaths worldwide because effective access to cervical screening methods is a big challenge. Cervical cancer is one of the most deadly diseases in the world among women. It is caused by long term infection in skin cells and mucous membrane cells of the genital area. The worrisome aspect of this cancer lies in the fact that it does not show any symptoms when it starts occurring. Women in the world are suffering from many diseases among those diseases Cervical Cancer is also mentioned globally. Every year many cancer more comfortable life. Cervical cancer is one such disease, considered to be fourth among the most common types of cancer in women around the world. There are many factors that increase the risk of cervical cancer, such as age and use of hormonal contraceptives. Early detection of cervical cancer helps to raise recovery rates and reduce death rates. In our proposed work, we used Machine Learning algorithms to find a model capable of diagnosing cervical cancer with high accuracy and sensitivity. The cervical cancer risk factor dataset was used to construct the classification model through a voting method that combines three classifiers: Decision Tree, ANN, K-N Neighbor and Random Forest.

Keywords: Machine Learning, Early Detection, Machine Learning, Cervical Cancer, ANN, Naïve Bayes

I. INTRODUCTION

Many centuries have been passing for the discovery of cancer disease, yet it's been a challenge to the Medical Sciences to abolish cancer cases throughout the world. But they succeeded in the prediction and cure of the disease in its early stages up to some extent. Cancer is the collection of diseases. Human body consists of trillions of cells. Normally, cells divide and form new cells when the cells become older they die and those dead cells are replaced by the new cells. This cells replacement is done even when the cells are damaged. The uncontrollable count of new cells generation leads to cancer in human body, production of unnecessary new cells lead to tumors. There are many types of cancers present in this world. Some cancers are malignant they spread to other tissues in the body. Thus, it's better to predict the cancerous cells at the early stages. There are many medical methods to discover the cancer in its early stages which helps the patients to avoid the worst case. When compared to the olden days the Medical Sciences are much more advanced. As there is a saying, "Prevention is better than cure" it's better to prevent human from cancer disease by vaccinating early. Cervical cancer has been a major cause for death worldwide from the last few decades. It is a third main type of cancer after the lungs and breast cancer among women. It is only cured when detected and treated at its early stage. The stage is a description of cancer's growth and spread in the body of a woman. Physicians find out the stage of cancer after

diagnostic test. By knowing the correct stage of cervical cancer physician can follow a right kind of treatment. There are four stages of a cervical cancer developed/Suggested by IFOG (International Federation of Obstetrics and Gynecology). Doctors evaluate the tumor and its spread to lymph nodes or to other parts of the body and then assign the stage. Cervical cancer is, "the fourth most common cancer in women according to the survey by World Health Organization". This is dangerous cancer when compared to other cancers. Cervical cancer belongs to the carcinoma family. The occurrence of this cancer is by getting infected with HPV virus. HPV virus is found by group of scientists it is mostly transmitted through sexual contact. There are many types of HPVs present, cancer is caused by type16 and type18. These types are said to high-risk HPVs as these cause cancerous cells in the body whereas type6 and type11 are low-risk HPVs they cause genital warts on the skin. Normally, in more number of cases HPV infection disappears in two years without causing any abnormalities but in some cases, they are responsible for causing precancerous cells in the human body. Cervical cancer starts from the cervix this cancer doesn't have any symptoms till its early stages. This cancer differs there are two types it is observed when seen under a microscope. The first type is Squamous cell carcinoma starts in the cells that line the bottom of cervix. Another one is Adenocarcinoma it develops in the gland related cells that are present at the upper part of the cervix. As this cancer doesn't have any symptoms regular screening tests are said to be done to stay free from the panic of having the cancerous cells. This screening includes Pap-smear test which results the presence or absence of precancerous cells. When the results are obtained if the result shows us the presence of abnormal cells they must be further treated or may require some additional tests to be done. If there are no such abnormal cells it doesn't require any further examination. A Cervical cancer risk factor is any means that changes the possibility of contacting cervical cancer. Cervical cancer is a type of cancer that develops from the cervix. Other types of cancer includes breast cancer, lungs cancer, sarcoma cancer, leukaemia cancer, liver cancer (Idikio, 2011). It is due to the abnormal growth of cells that have the ability to spread to other parts of the cervix. Early on, typically no symptoms are seen. Some of the symptoms that precede later in the growth of abnormal cells include vaginal discharge of blood, pelvic pain. In as much as bleeding after sex is not really considered a symptom of cancer because it can be caused by several diseases, it can also indicate the presence of cervical cancer. Many medical methods have been invented in the prediction and cure of cancer. As the technology has wide spread its wings in every area many novel techniques are being implemented for faster results. Same way Medical Sciences are combined with Artificial Intelligence for better improvement in Medicine. Introduction of modern

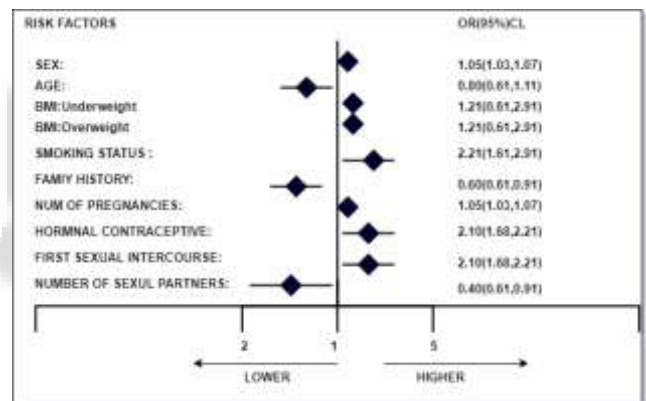
computing technologies can reduce the human errors and also help in faster results. Machine Learning has originated from Artificial Intelligence which mostly deal with Oncology department in Medicine and Deep learning is subpart of Machine Learning, fusion of Artificial Intelligence with human is said to lead a fruitful result in the ongoing centuries. Machine Learning techniques including decision tree algorithms are used in biomedical research for predictive analysis. In addition, the ability of machine learning tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Artificial Neural Networks (ANN), Naïve Bayes, Support Vector Machine (SVM), Random Forest Classifier and Decision Tree have been widely applied in cancer research for the development of models, resulting in effective and accurate decision making. In this study, we used an ensemble learning technique in the prediction of cervical cancer using risk factors. This technique was selected because it combines several machine learning techniques into one model to decrease variance, bias, and improvement in performance. Classification methods were selected because the interest of this study was to solve a classification problem. Therefore these algorithms could work well within our problem domain. The prediction model was trained and validated, and our experimental results revealed that our model.

II. LITERATURE REVIEW:

This section discusses in brief about some of the work that has already been done on creating ML models using various algorithms. Many algorithms are being used and also many more tools and techniques are used and being in use for the faster and less false – positive rate in the prediction of cancerous cells. Many researchers have predicted the cervical cancer by using different algorithms. [1] A paper reviewed the classification of cervical cancer using artificial neural networks. It uses artificial neural network to detect cervical cancers by classifying the cells either as normal or abnormal cells in the cervix area. This classification produces an accurate result compared to Pap smear test which is a manual screening method. Analysis was made using all the network architecture which includes single neural network, deep and shallow neural network. [2] A research work reviewed the manual method of screening cervical cancer using Pap smear test, the need for carrying out feature selection and the classification of the features. The classification are done using K-NN and Artificial neural network. The analysis shows that the classification accuracy for K-NN was 88.04% and 54% for artificial neural network. [3] The study by used data mining algorithms in the prediction of cervical cancer. During this study, Boosted decision tree, decision forest, and decision jungle algorithms are used by the researchers. Experimental results revealed that the Boosted decision tree algorithm performed better with 0.978 on Area under Receiver operating characteristic. However, the study did not disclose the performance of decision forest and decision jungle algorithms. Moreover, there is a need for further research to determine whether the Boosted decision tree algorithm can perform better on the dataset with more variables since the dataset used during the study had only five variables. [4] To determine the maximum risk factors

associated with cervical cancer, the study by explored the use of machine learning classification techniques on cancer datasets. The study used C5.0, random forest, rpart, KNN, and SVM classifiers. The performance accuracy of the algorithms was 97%, 96%, 96.9%, 93% and 89% for C5.0, rpart, random forest, KNN, and SVM. This performance revealed that C5.0 and random forest classifiers performed better than the rest of the classifiers in the prediction of women with higher chances of cervical cancer. [5] R. Vidya and G. M. Nasira predicted cervical cancer using random forest with K-means learning and implemented the techniques in MATLAB tool. These experiments were performed with the help of NCBI dataset to construct decision tree using classification methods. [6] Yulia et al. predicted cervical cancer using Pap smear test results. The Pap smear test results were divided into two categories: cancerous and non-cancerous patients. Three classification methods Naïve Bayes, support vector machine and random forest were used to compute the results in which random forest tree was given better results. [7] Kelwin Fernandes et al. used transfer learning technique for cervical cancer screening. Their study consists on linear predictive models. Positive results were obtained in most experiments as compared to other methods.

III. METHODOLOGY



Data collection refers to the process of gathering data on a particular area of interest that are measurable and can be statically fashioned. The main goal of data collection is to capture data that is of high quality and its analysis would lead to the formation of answers that are convincing and credible. Figure 1 explains the methodology required for the predicting cervical cancer which includes data collection, preprocessing, classification and analysis. The attributes of the cervical cancer data set can be viewed in the table. Attributes of the Cervical Cancer data set:

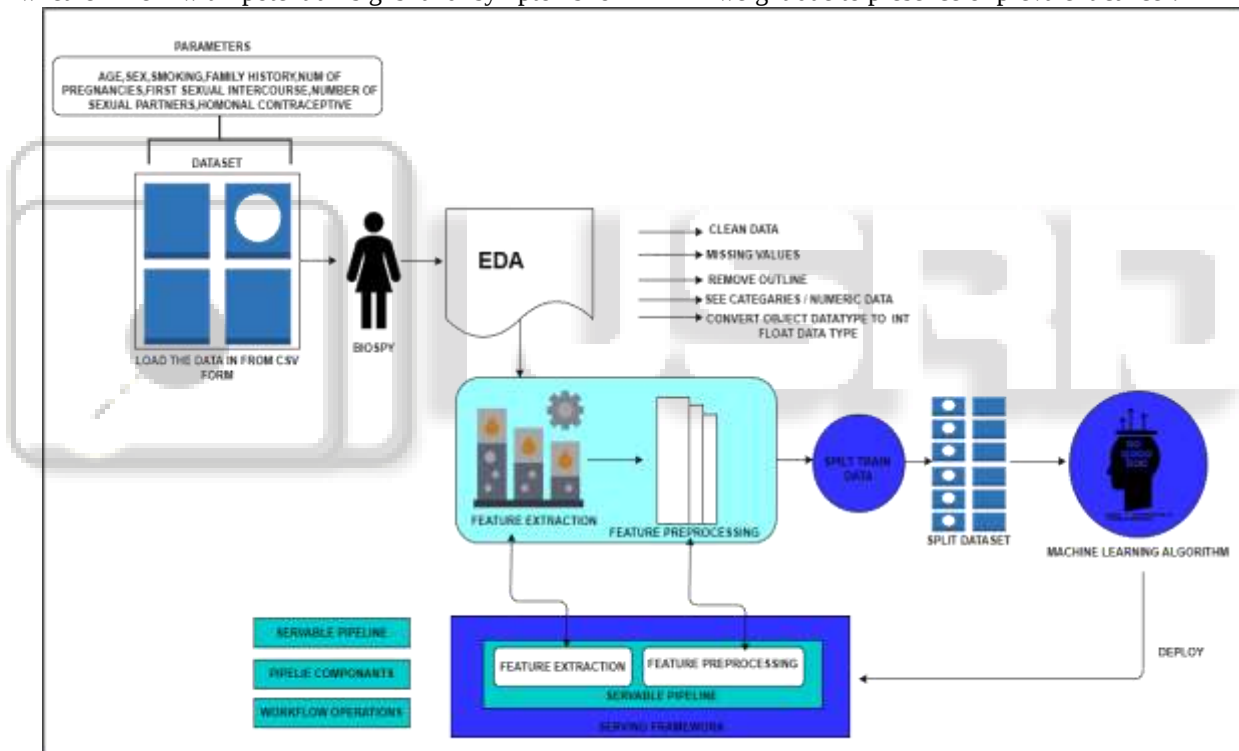
In the World, more than half (55%) of all CRCs are attributable to lifestyle factors, including a sex, age, smoking, BMI & Family History, Num of Pregnancies: First Sexual Intercourse: Number of Sexual Partners, Hormonal Contraceptive and smoking.

- 1) Age: It defines the Age of the person
- 2) Number of Sexual Partners: It represents the number of Sexual Partners of the person
- 3) Number of Pregnancies: It represents the number of Pregnancies of the person
- 4) Smokes: It represents the number of Smokes of the person

- 5) Smoke (In Years): It represents the number of Smokes of a person per year
- 6) Hormonal Contraceptives: These are the pills which are used to prevent pregnancy by blocking the release of eggs.
- 7) Hormonal Contraceptives (In Years): It is used to represent the number of pills consumed by the person per year.
- 8) Family History: Family history of pancreatic cancer has been associated with an increased risk of pancreatic cancer, but little is known regarding the role of inherited/environmental factors or that of FH of other comorbidities in pancreatic cancer risk. We aimed to address these issues using multiple methodological approaches.
- 9) Sex: Consistent with prior work, we found men had a 2.2-fold increase in EOCRC odds compared with women. To date, guidelines have been consistent in suggesting the same age of initiation for screening for men and women. Future research should determine whether men with potential signs and symptoms of

EOCRC such as iron deficiency anemia and hematochezia have substantially higher risk for CRC diagnosis than women.

- 10) BMI: In our analysis we found that increased BMI and body weight had a protective effect on the odds of EOCRC. This relationship was unexpected given that CRC has been identified as an obesity-related cancer, and given that several biologic correlates of being overweight or obese, particularly insulin resistance and increased circulating estrogens, are believed to increase cancer risk.⁴¹ Specific to EOCRC, the Nurses' Health Study II found a significant association between weight gain since early childhood among women and EOCRC risk.¹³ Women with a BMI of 30 or more had the highest risk of EOCRC compared with women with normal BMI, with an EOCRC relative risk of CRC of 1.93 (95% CI 1.15–3.25).¹³ One potential explanation for our contrasting findings is that we used data near subjects' baseline colonoscopy, a time in which cases were being diagnosed with EOCRC and were potentially losing weight due to presence of prevalent cancer.



1) Dataset and Preprocessing

The data is collected and put into our database. Gathering and analyzing information from many different sources is known as data collection. This means that the data we collect must be acquired and kept in a way that makes sense for the business challenge at hand. The information is organised into 10 columns, four of which are the target columns for the four tests used to establish whether the patient has cancer. The remaining 6 columns show the various cervical cancer risk factors. These factors are crucial in determining whether or not a patient has cancer. We created several data visualisations to examine the occurrences of these factors. The data obtained will be cleansed with data null or not applicability and the unwanted columns from the dataset will be discarded. As part of the data mining process, data

preparation is a crucial stage. Projects involving data mining and machine learning are particularly susceptible to the adage "garbage in, garbage out." Sometimes, data-collection techniques are weakly regulated, resulting in unreliable results such as out-of-range or non-existent numbers.

2) Training and testing

With the use of data mapping, the collected dataset is separated into two parts: 80 percent training data, and 20 percent testing data. In order to allocate data points to the former and the latter in the modelling dataset, the data has been separated into training and testing sets. A model is therefore trained using a training set, then applied to a test set. Our application may be evaluated in this manner.

3) Prediction

We eventually our model is ready to predict the early detection of cervical Cancer Using Machine Learning algorithms based on the given dataset.

IV. CLASSIFICATION

The proposed work sets itself apart by harnessing the powers of Machine Learning and Data Mining. In the project work, a system, with a strong prediction algorithm, which implements powerful classification steps with a comprehensive report generation module. The project aims to implement a self-learning protocol such that the past inputs of the disease outcomes determine the future possibilities of the cervical disease to a particular user. The proposed model makes use of strong preprocessing tools so that the classification and prediction do not show any errors relating to the dataset. A huge number of training sets will be used to make the prediction more and more accurate. Not only does the datasets but also the attributes to be used are selected taking into consideration the various important parameters and attributes.

V. CONCLUSION:

Cervical cancer is a widely spread condition nowadays, and screening often entails lengthy clinical examinations. In this regard, ML can provide good approaches for accelerating the diagnosing process. Furthermore, data mining technologies, particularly tree-based algorithms, are used in this study to provide accurate predictions for cervical cancer patients. The SMOTE approach was used to handle an imbalanced data set problem in which cancerous patients were too small in comparison to non-cancerous patients. The decision tree outperforms decision forest and decision jungle in terms of prediction ability as assessed by the AUROC curve value. The decision forest and decision jungle approaches were rejected as best due to their low AUROC curve values. With the expanding amount of cervical cancer patient data and the quickly advancing tools for evaluating this data, we hope we will be able to discover the optimum screening approach for cervical cancer patients that will be useful for patient treatment. This study could be utilised as a model for developing a healthcare system for cervical cancer patients in the future.

REFERENCES:

- [1] Devi, M. A., Ravi, S., Vaishnavi, J., & S, P. (2016). Classification of Cervical Cancer using Artificial Neural Networks. 465-472. doi:<https://doi.org/10.1016/j.procs.2016.06.105>
- [2] Malli, P. K., & Nandyal, S. (2017). Machine learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network. 6(4).
- [3] T. M. Alam, M. Milhan, A. Khan, M. A. Iqbal, A. Wahab, and M. Mushtaq, "Cervical Cancer Prediction through Different Screening Methods using Data Mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 388–396, 2019.
- [4] B. Ilango and V. Nithya, "Evaluation of machine learning-based optimized feature selection approaches

- and classification methods for cervical cancer prediction," *SN Appl. Sci.*, vol. 1, no. 6, pp. 1–16, 2019.
- [5] R. Vidya and G. Nasira, "Prediction of cervical cancer using hybrid induction technique: A solution for human hereditary disease patterns," *Indian Journal of Science and Technology*, vol. 9, 2016.
- [6] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results," in *Biomedical Engineering (IBIOMED), International Conference on*, 2016, pp. 1-5.
- [7] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*, 2017, pp. 243-250.
- [8] B. Obrzut, M. Kusy, A. Semczuk, M. Obrzut, and J. Kluska, "Prediction of 5-year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods," *BMC cancer*, vol. 17, p. 840, 2017.
- [9] N, T. . B. D. . J. (2017). Classification of Cervical Cancer Using Pap-Smear Images: A Convolutional Neural Network Approach. Department of Electrical and Computer Engineering, 1(d), 698–706. <https://doi.org/10.1007/978-3-319-60964-5-23>
- [10] S., A., & M.V., S. (2016). Classification of Cervical Cancer Cells in Pap Smear Screening Test. *ICTACT Journal on Image and Video Processing*, 06(04), 1234–1238. <https://doi.org/10.21917/ijivp.2016.0179>
- [11] *Python For Everybody: Exploring Data in Python 3* by Charles Severance.
- [12] *Introduction to Machine Learning with Python: A Guide for Data Scientists* Book by Andreas C.Miller.