International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 08 | Aug 2021 www.irjet.net p-ISSN: 2395-0072

Early Prediction of Cervical Cancer Using Machine Learning Algorithms

Naveen N Mugad[1], K R Sumana[2]

[1]PG Student, Department of MCA, National Institute of Engineering, Mysuru, Karnataka, India
[2]Assistant Professor, National Institute of Engineering, Mysuru, Karnataka, India

---

Abstract - In almost all countries, precautionary measures are less expensive than medical treatment. The early prediction of any disease gives a patient better chances of successful treatment than disease discovery at an advanced stage of its development. If we do not know how to treat patients, any treatment we can provide would be useful and would provide a 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, K-N Neighbor and Random Forest.

KeyWords: Machine Learning(ML), Decision Tree(DT), Random Forest(RF), K-Nearest Neighbor(KNN).

1. INTRODUCTION

Cervical cancer is a type of cancer that begins in the uterus. It forms as a result of abnormal cell proliferation and spreads throughout the body. It is, in the vast majority of cases, fatal. HPV is responsible for the majority of instances (90 percent). Outliers have been eliminated from the data, which has been sanitised. Cervical cancer can also be caused by several pregnancies. Cancer is notoriously difficult to detect in its beginning stages. In the beginning stages of cancer, there are no symptoms. Symptoms do not occur until late in the course of cancer. Machine learning algorithms can be used to predict whether or not a person has cancer. Cervical cancer kills over a quarter of a million individuals every year around the world. The available Computed Aided Diagnosis (CAD) to appropriately treat the cancer patient is muddled by screening and numerous deterministic testing. like Age, no. of sexual partners, age from which first sexual intercourse has been occurred, no. of pregnancies, smoking habits, hormones, STDs discovered in the patient, and any cancer or disease diagnosis conducted, such as HPV (Human papillomavirus) and CIN, are all risk factors for cancer (Cervical intraepithelial neoplasia). Cervical cancer screening procedures include cytology, Schiller, Hinselmann, and the standard biopsy test. Each test is performed to see if cervical cancer is present. Many models are tested on the dataset. The parameters have been fine-tuned. Different models' performance is compared. Finally, the most effective models for cancer prediction are suggested. Implementing the aforementioned model achieves the goal of building a system that is more accurate in calculating the percentage of patients who will have a cervical attack while also resolving the shortcomings of the current approach.

2. LITERATURE REVIEW

The study of literature is a methodology to identify existing system problems by investigating and proposing the development of a system to deal with current political system problems .This present section in a project report is a literature survey or the literature review comprising multiple analyses and studies conducted in the sphere of the interest and the ear1ier conducted resu1ts taking into account the project management parameters and the design of the research.

The neural networks to predict [1] adverse events in cervical cancer patients. MLP is a type of neural network where the input signal is fed forward through a number of layers. MLP contains input layer, hidden layer and output layer. The GEP classifier delivered efficient results in the prediction of the adverse events in cervical cancer as compare to other methods.

The predicted cervical cancer [2] 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.

An automated method [3] for predicting the effect of the patient biopsy for the diagnosis of cervical cancer by using medical history of patients. Their technique allows a joint and fully supervised optimization method for high dimensional reduction and classification. They discovered certain medical results from the embedding spaces and confirmed through the medical literature.

Data mining [4] is helpful where large collections of healthcare data are available. Several data mining techniques like support vector machine (SVM), kernel learning methods as well as clustering techniques were used in healthcare. With the rise of computing methods for disease prediction, WHO and other international organizations are working together for effective screening method to detect the cervical cancer.

3. PROPOSED SYSTEM

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.

4. IMPLEMENTATION

4.1 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 36 columns, four of which are the target columns for the four tests used to establish whether the patient has cancer. The remaining 32 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 colums 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.

Features selected	
Age	STDs (number)
Number of sexual partners	STDs:condylomatosis
First sexual intercourse	STDs:cervical condylomatosis
Num of pregnancies	STDs:vaginal condylomatosis
Smokes	STDs:vulvo-perineal condylomatosis
Smokes (years)	STDs:syphilis
Smokes (packs/year)	STDs:pelvic inflammatory disease
Hormonal Contraceptives	STDs:genital herpes
Hormonal Contraceptives (years)	STDs:molluscum contagiosum
IUD	STDs:AIDS
IUD (years)	STDs:HIV
STDs	STDs:Hepatitis B
	STDs:HPV

Table-1: Paramters Selected

- Age: It defines the Age of the person
- Number of Sexual Partners: It represents the number of Sexual Partners of the person
- Number of Pregnancies: It represents the number of Pregnancies of the person
- Smokes: It represents the number of Smokes of the person
- Smoke (In Years): It represents the number of Smokes of a person per year
- Hormonal Contraceptives: These are the pills which are used to prevent pregnancy by blocking the release of eggs.
- Hormonal Contraceptives (In Years): It is used to represent the number of pills consumed by the person per year.
- IUD (Intrauterine Device) : This the device used for birth control.
- STD (Sexually Transmitted Disease): This is the disease which is transmitted sexually.
- STD (Condylomatosis): It is an infection due to virus, the papillomavirus.
- Dx: Cancer: The Oncotype Dx is test that may predict how likely it is that Breast Cancer will return.
- Dx: HPV (Human Papilloma Virus): A Vinegar solution applied to HPV infected genital areas turns them white.
- Dx: CIN: Abnormal cells are found on the surface of the cervix.
- STDs (AIDS): This is the disease transmitted sexually by Human Immuno Deficiency Virus.

4.2 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.

4.3 Prediction

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

4.4 Comparison and Visualization

The data features obtained from test is compared. Machine Learning algorithms can only be fairly compared if they are assessed on the same data. When testing the algorithms, we may force them to be assessed on a uniform test harness. The graphical representation of data will be done, which gives a user-friendly way to explore and comprehend data trends, outliers, and patterns in the data.

5. RESULT ANALYSIS

The three distinct classifier techniques: Decision tree, K-nearest Neighbour, and Random Forest are employed to build the model. The dataset was found to be biased during the initial analysis, so k-fold cross-validation was performed. For the KNN, dataset split-up used is 50-50. The feature selection process has been completed, and the features that have been chosen are used for prediction. The data is split 20-80 for Decision Tree and Random Forest. Parameter tuning is carried out get the good predictions with the highest evaluation points. K-fold cross validation is used to find the nearest neighbour. The appropriate k value is chosen using the formula sqrt(N)/2, which yields 10.3. The sample size for the training dataset is set at 429. Following the execution of several models, the value of K is determined to be 5. The Euclidean distance method is used to calculate the distance between two values. The lower the value of k, the more skewed the result. Higher values of k are less biased, but they can exhibit variance. Because k=5, which is neither less nor greater.

Feature selection is performed to assist us in selecting the features that will improve prediction. The random forest algorithm generates a decision tree with multiple levels based on the training data set. To make the tree less complex, the depth is limited to ten. After several runs of the algorithm, the maximum sample split is determined to be 75. Class weight is once again used as 'balanced' for automatic adjustment based on the inverse of input data frequency. The dataset is divided by 25% for the decision tree classifier. The pre sort function is used for quick algorithm implication. To keep the tree as small as possible, the minimum split for sample is set to one hundred ten .Class weight is taken as balanced, which means it adjusts the weight in inverse proportion to the class frequencies in the input data.

Accuracy is one of the measurements used to evaluate classification models. Informally, accuracy refers to our model's percentage of correct predictions as shown in Table-2 and Fig-3 shows the graph of the same and accuracy is calculated as

Accuracy = (TP + TN)/Total

Sl. No.	ML Algorithms	Accuracy
1	Decision Tree Classifier	85.11%
2	Random Forest Algorithm	87.90%
3	K-Nearest Neighbor Algorithm	95.30%

Table-2 Comparison Table of All Algorithms

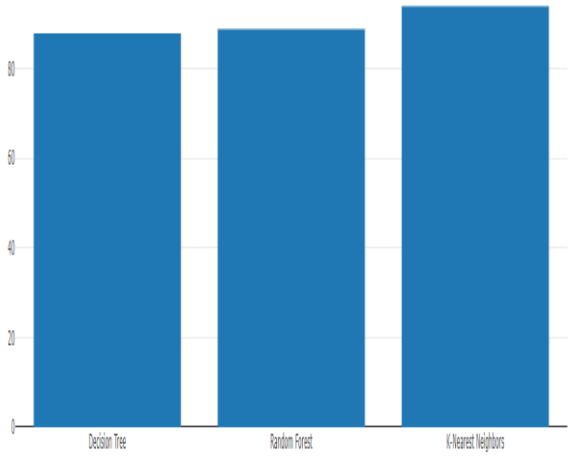

Fig-3: Graphical Visualisation of Algorithms

7. CONCLUSIONS

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. The results for algorithms based on accuracy are like K-Nearest Neighbor 95.3%, Decision Tree Classifier 85.11% and Random Forest 87.90%.

REFERENCES

[1] Python For Everybody: Exploring Data in Python 3 by Charles Severance.
[2] Introduction to Machine Learning with Python: A Guide for Data Scientists Book by Andreas C.Miller.
[3] M. Hejmadi, Introduction to cancer biology: Bookboon, 2009.
[4] N. Kamil and S. Kamil, "Global cancer incidences, causes and future predictions for subcontinent region," Systematic Reviews in Pharmacy, vol. 6, p. 13, 2015.
[5] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," CA: a cancer journal for clinicians, vol. 67, pp. 7-30, 2017.
[6] S. Subramanian, R. Sankaranarayanan, P. O. Esmy, J. V. Thulaseedharan, R. Swaminathan, and S. Thomas, "Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low-and middle-income countries," Journal of Cancer Policy, vol. 7, pp. 4-11, 2016.
[7] K. U. Petry, "HPV and cervical cancer," Scandinavian Journal of Clinical and Laboratory Investigation, vol. 74, pp. 59-62, 2014.
[8] https://www.kaggle.com/mlg-ulb/cervicalcancer
[9] https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0
[10] https://docs.python.org/3/tutorial/
[11] https://jupyter.org/