International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 6, September-October 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Prediction of Cervical Cancer using Machine Learning and Deep Learning Algorithms

Kayalvizhi. K. R, N Kanimozhi

Department of Computer Science and Engineering,
G.K.M College of Engineering and Technology, Chennai, Tamil Nadu, India

ABSTRACT

As the development of machine learning and deep learning, more and more people or organizations use multiple algorithms to analyse large collections of data to produce meaningful results that help to predict behaviour. And this kind of technology is increasingly used in medical field to predict some severe illness in their early stage, for example, cervical cancer. Cervical Cancer is one of the main reasons of deaths in countries having a low capita income. It is the second most common cancer in India in women accounting for 22.86% of all cancer cases in women. It becomes quite complicated while examining a patient on the basis of result obtained from various doctor's preferred test to determine if the patient is positive with the cancer. There were 96,922 new cases of cervical cancer diagnosed in India in 2018. Around the globe, around a quarter of million people die owing to cervical cancer. Screening and different deterministic tests confuse the available Computed Aided Diagnosis (CAD) to treat the patient correctly for the cancer.

Machine learning and Deep learning algorithms are used in this project and determine if the patient has cancer based on the analyses of the risk factors available in the dataset. Predicting the presence of cervical cancer can help the diagnosis process to start at an early stage and comparing various models will help in finding out the best prediction model for predicting the presence of cervical cancer effectively.

KEYWORDS: *Cervical Cancer, Machine learning, Deep learning, Logistic regression, SVM, Decision Tree, Random Forest, Deep Neural networks, Dataset*

***How to cite this paper**:* Kayalvizhi. K. R | N Kanimozhi "Prediction of Cervical Cancer using Machine Learning and Deep Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-6, October 2020, pp.426-430, URL: www.ijtsrd.com/papers/ijtsrd33378.pdf

IJTSRD33378

I. INTRODUCTION

Cervical cancer is one of the deadliest cancers which are threatening women's health all over the world and it is hard to observe any sign in the early stage. The uterine cervix is the lowest portion of a woman's uterus (womb), connecting the uterus with the vagina. Cervical cancer occurs when the cells of the cervix grow abnormally and invade other tissues and organs of the body. When it spreads, this cancer affects the deeper tissues of the cervix and may have spread to other parts of the body, most notably the lungs, liver, bladder, vagina, and rectum. But since cervical cancer is slow-growing, its progression through precancerous changes provides opportunities for early detection, prevention and treatment. Better means of detection have meant a decline in cervical cancer in most of the countries over the decades.

In 2015, E&Y in association with FICCI Flo has released that 19% of women has been suffering with breast cancer, 14% and 7% of women has been suffering with cervical and uterus cancers. Especially in India, highest incidences are recorded in Delhi, Kerala and Tamil Nadu due to overweight, tobacco use, low fruit and vegetable intake, alcohol use and lack of physical activity.

Early screening can reduce cancer rates by 1.5-2.5 times, increases five-year survival rate by 3 to 17 times and also decreases the treatment cost. Likewise, it extends growth rate among ladies in India to increment from 110 for each 1 Lakh populace to 190-260 for every 1 Lakh population by 2025.

Most of the Machine learning algorithms play a vital role in providing more appropriate remedy for the affected person. Among all, classification (Supervised) and regression, clustering (Unsupervised), Deep Neural Network methods are widely used in diagnosis.

The main idea behind this project is by making the utilization of machine learning and Deep Neural Network algorithms, cervical cancer can be predicted and classify in the stream of data analytics. ML is the key components in big data revolution said by McKinsey Global Institute because it can effectively find out hidden knowledge and unknown patterns in the specified datasets.

Machine learning algorithms are classified into two kinds: supervised (class label is known initially) and unsupervised (class label is not known initially). Classification (Logistic regression, Decision Tree, Naïve Bayes, Neural networks, SVM), regression comes under supervised and clustering (K-Means, K-Medoids, AGNES, DIANA) comes under unsupervised learning.

In this project, by applying supervised algorithms like Logistic Regression (LR), Decision Tree Classifier (DT), Random Forest Classifier (RF), Support Vector Machine (SVM) and Neural networks, persons with cervical cancer will be classified. It obviously reduces further proceeding tests.

II. RELATED WORK

There are many researches with the same topic for the Prediction of cervical cancer using machine learning and deep learning algorithms.

Survey 1: WEN WU "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine- Based Approaches", Department of Blood Transfusion, Jinan Military General Hospital, Jinan, China Year: 2017 [1]. An advantage of this paper is Support vector machine (SVM) approach is introduced for cervical cancer diagnosis. Two improved SVM methods, support vector machine-recursive feature elimination and support vector machine-principal component analysis (SVM-PCA), are proposed to diagnose the cancer samples.

Survey 2: Yasha Singh, Dhruv Srivastava, P.S. Chandranand & Dr. Surinder Singh "Algorithms for screening of Cervical Cancer: A chronological review" [2]. Advantages of this paper are the critical review of different research papers published that integrated AI methods in screening cervical cancer via different approaches analysed in terms of typical metrics like dataset size, drawbacks, accuracy etc.

Survey 3: Xiaoyu Deng, Yan Luo, Cong Wang "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods", School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China [3]. Advantages are some risk factors of cervical cancer are introduced and three kinds of machine learning algorithms are used to classification of the cervical cancer dataset from UCI.

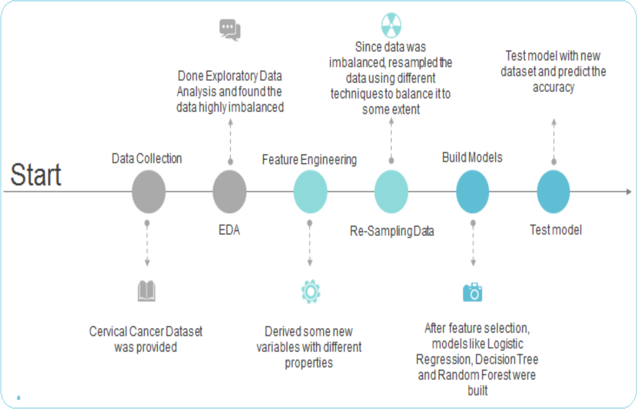
Survey 4: Prediction of Cervical Cancer using Voting and DNN Classifiers, Komala Rayavarapu, Krishna Kishore K.V, Vignan's Foundation for Science, Technology and Research [4].

Advantages are by applying supervised algorithms like Logistic Regression (LR), Decision Tree Classifier (DT), Naïve Bayes Classifier (Naïve), Random Forest Classifier (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Deep Neural Networks (DNN), persons with cervical cancer will be classified.

Drawbacks of these approaches are High Computational cost, Low prediction accuracy, combinations of machine and deep learning algorithms was not considered.

III. SYSTEM ARCHITECTURE

The process of training a Machine and Deep Learning model involves providing a Machine and Deep Learning algorithm with training data to learn from. The term Machine Learning model refers to the model artefact that is created by the training process.


Done Exploratory Data Analysis and found the data highly imbalanced
Since data was imbalanced, resampled the data using different techniques to balance it to some extent
Test model with new dataset and predict the accuracy
Start
Data Collection
Feature Engineering
Build Models
EDA
Re-Sampling Data
Test model
Cervical Cancer Dataset was provided
Derived some new variables with different properties
After feature selection, models like Logistic Regression, Decision Tree and Random Forest were built

A. Data Collection

The dataset, "Cervical Cancer Risk Factors for Biopsy" was obtained from the UCI Repository. The dataset contains habits, demographic information, and medical history of 858 patients from the hospital. There are many missing values in this dataset, due to many patients not answering questions because of privacy concerns. The dataset consists of 858 instances, with 36 attributes. The dataset consists of 36 variables and records of 858 women patients. Of the 36 variables 4 variables are the target variables.

No.	Attribute	Type
1	Age	Int
2	Number of sexual partners	Int
3	First sexual intercourse	Int
4	Number of pregnancies	Int
5	Smokes	Bool
6	Smokes(years)	Bool
7	Smokes(pack/year)	Bool
8	Hormonal Contraceptives	Bool
9	Hormonal Contraceptives (years)	Int
10	IUD	Bool
11	IUD (years)	Int
12	STDs	Bool
13	STDs(number)	Int
14	STDs: condylomatosis	Bool
15	STDs: cervical condylomatosis	Bool
16	STDs: vaginal condylomatosis	Bool
17	STDs: vulvo-perineal condylomatosis	Bool
18	STDs: syphilis	Bool
19	STDs: pelvic inflammatory	Bool
20	STDs: genital herpes	Bool
21	STDs: molluscum contagiosum	Bool
22	STDs: AIDS	Bool
23	STDs: HIV	Bool
24	STDs: Hepatitis B	Bool
25	STDs: HPV	Bool
26	STDs: Number of diagnosis	Int
27	STDs: Time since first diagnosis	Int
28	STDs: Time since last diagnosis	Int
29	Dx: Cancer	Bool
30	Dx: CIN	Bool
31	Dx: HPV	Bool
32	Dx	Bool

TARGET VARIABLES

The target variables are four tests and are characterized by '1' or '0' in our data set where '1' represents a malignant tumor and '0' indicates benign tumor.

They are detailed below:

SCHILLER

In this test, Schiller's iodine solution is applied to the cervix under direct vision. Normal cervical mucosa contains glycogen and stains brown, whereas abnormal or cancer affected areas do not take up the stain. The abnormal areas can then be biopsied and examined histologically. The composition of Schiller's iodine is the same as Lugol's iodine, but Lugol's iodine being more concentrated. When Schiller's iodine is not available, Lugol's iodine can be used as an alternative. Schiller's test is not specific for cervical cancer, as areas of inflammation and keratosis may also not take up the stain.

CYTOLOGY

The medical and scientific study of cells. Cytology refers to a branch of pathology that deals with diagnoses of diseases and conditions through the examination of tissue samples from the body.

Cytological examinations are performed on body fluids (examples are urine, blood and cerebrospinal fluid) or on material that is aspirated (drawn out via suction into a syringe) from the body. Screening is performed using cervical cytology (Pap test) or a human papillomavirus (HPV) test, or a combination of the two tests.

HINSELMANN

Colposcopic examination was almost impossible to perform because of the distance from the focus that was no more than 80 mm. Hinselmann test tried to solve this problem by pulling out the uterine cervix. The examined part is anemised by this procedure, which can prejudice the final result and a small amount of blood might also leak. Besides that, a patient can feel pain if the portio is held by thin forceps.

BIOPSY

Cervical biopsy is a procedure to remove tissue from the cervix to test for abnormal or precancerous conditions, or cervical cancer. Cervical biopsies can be done in several ways. The biopsy can remove a tissue sample for testing.

B. Exploratory Data Analysis

EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data.

Fig 1- Distribution of Target Variables

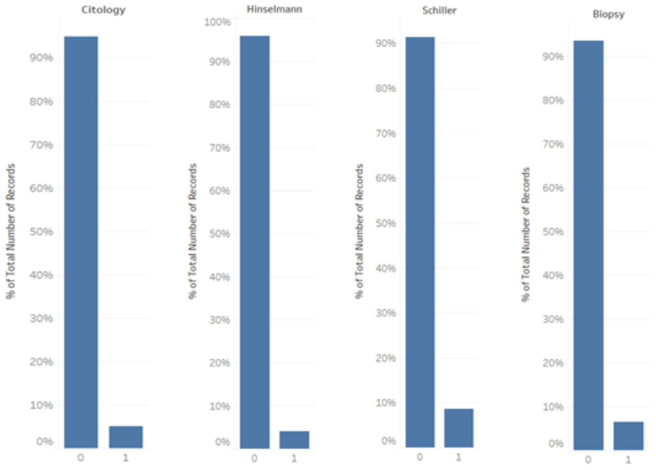
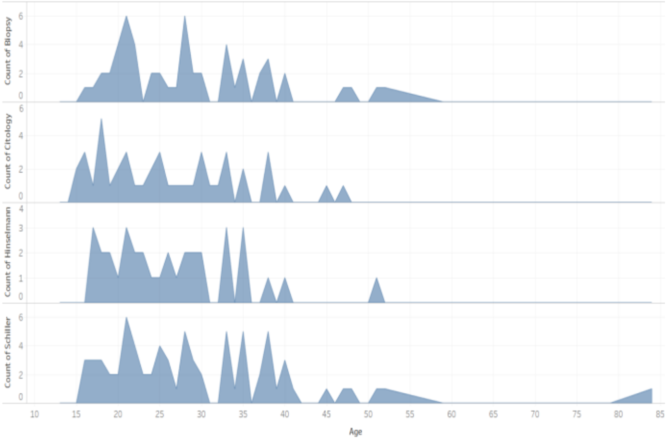


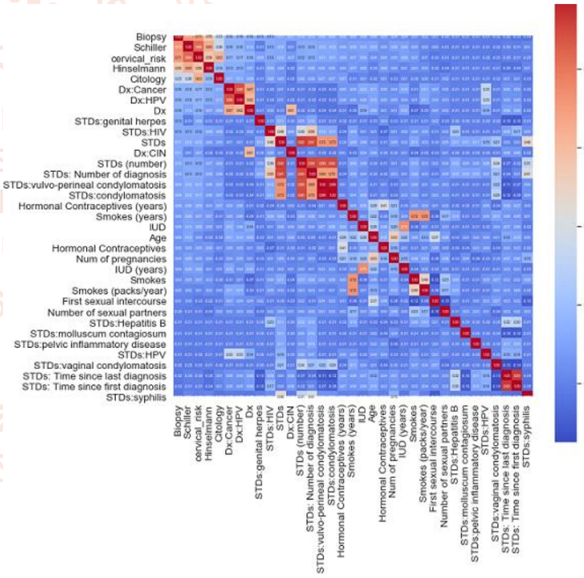
Fig 2- Count of Target Variables Vs Age



C. Feature Engineering And Data Visualization

During feature engineering process, data is classified based on the measurement levels as Numerical data and Categorical data. Target variables are classified separately. Numerical data are represented as numbers. Features which cannot be grouped are classified under numerical data. Categorical data describes categories or groups and also answers to yes or no questions.

Fig 3- Heatmap Correlation: Numerical & Categorical Data



D. Re-Sampling Data

Dealing with imbalanced datasets require strategies such as improving classification algorithms or balancing classes in the training data (data pre-processing) before providing the input data to the machine learning algorithm. The balancing classes technique is preferred as it has wider application.

The main objective of balancing classes is to either increase the frequency of the minority class or decrease the frequency of the majority class. The balancing is done in order to obtain approximately the same number of instances for both the classes.

E. Build Models

Modelling was done on the original data after default data cleaning and scaling where necessary. For all the classifier algorithms, the dataset is split 25%. 75% for training and 25% for testing. Feature selection is done and the models are built based on the corresponding features.

Logistic regression, SVM, Decision tree, Random forests and Deep neural network (DNN) were the algorithms used.

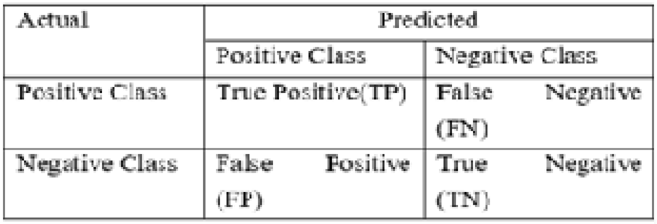
F. Model Evaluation and Testing

The conventional evaluation methods do not accurately measure model performance when faced with imbalanced datasets.

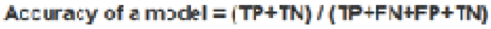
Standard classifier algorithms like Decision Tree and Logistic Regression are biased towards classes which have number of instances. They can predict only the majority class data. There is a high probability of wrong classification of the minority class as compared to the majority class.

Performance evaluation of a classification algorithm is measured by the Confusion Matrix which contains information about the actual and the predicted class.

Fig 4- Confusion Matrix



Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)



Accuracy of a model = (TP+TN) / (TP+FN+FP+TN)

However, while working in an imbalanced data accuracy is not an appropriate measure to evaluate model performance. To fully evaluate the effectiveness of our model, we must examine precision and recall as well.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

IV. RESULTS AND DISCUSSIONS

Since we are dealing with medical data, we will be focusing on **Recall** rather than Precision i.e, we will allow Type I error to creep in to our models because in our case, Type II error is far costlier.

Being diagnosed with cancer when you don't have cancer (false positive), although not desirable, is far better than being diagnosed that you don't have cancer when you do (false negative | Type II Error) – Even costing a life.

From the original dataset, we had good precision and recall but the accuracy is very less for Logistic Regression and SVM and reasonably good for Decision Tree, Random Forest and DNN. However the confusion matrix showed that the model is not able to identify the affected people at a reasonable rate except DNN.

Algorithm	Accuracy (%)	Precision 0	Precision1	Recall 0	Recall 1	F1 1	F1 2
Logistic Regression	75%	0.94	0.16	0.78	0.44	0.85	0.244
SVM	82%	0.94	0.21	0.87	0.38	0.90	0.27
Decision Tree	85%	0.93	0.21	0.91	0.25	0.92	0.23
Random Forest	91%	0.92	0.50	0.99	0.06	0.95	0.11
Deep Neural Networks (DNN)	91%	0.91	1.00	1.00	0.13	0.95	0.23

We cannot rely on these and I further go on with resampling the data using SMOTE and above is the result of resampled dataset.

Even after resampling the data, accuracy is less for Logistic Regression and SVM. Accuracy of Decision Tree has also dropped but we have good precision and recall.

We get a reasonable accuracy and recall rate for Random Forest Model with both original data and resampled data. However resampling is not applicable for DNN.

Confusion Matrix for Random Forest:

```
Accuracy: 0.912087912879121
[ [165    1]
  [15      1] ]
```

Confusion Matrix for DNN:

```
[ [192    0]
  [ 20     3] ]
```

The confusion matrix showed that both Random Forest and DNN are able to identify the affected people at a reasonable rate compared to other algorithms.

V. CONCLUSIONS

Cervical Cancer is the cancer arising from the cervix. Usually it is very difficult to identify cancer at early stages. The early stages of cancer are completely free of symptoms. It is only during the later stages of cancer that symptoms appear. Predicting the presence of cervical cancer can help the diagnosis process to start at an early stage.

We get a reasonable accuracy and recall rate for Random Forest Model with both original data and resampled data and DNN.

Hence it's been concluded that the Random Forest with SMOTE and DNN are the overall best model so far for predicting the cancer indicator if all the 4 target variables (Biopsy, Cytology, Hinselmann and Schiller) are combined together and classified as multi classifier for target variable.

VI. REFERENCES

[1] WEN WU "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine- Based Approaches", Department of Blood Transfusion, Jinan Military General Hospital, Jinan, China Year: 2017

[2] Yasha Singh, Dhruv Srivastava, P.S. Chandranand & Dr. Surinder Singh "Algorithms for screening of Cervical Cancer: A chronological review"

[3] Xiaoyu Deng, Yan Luo, Cong Wang "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods", School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[4] Upasana "Handle Imbalanced Classification Problems in machine learning", Consultant of Data & Analytics in KPMG

[5] Prediction of Cervical Cancer using Voting and DNN Classifiers Publication: Komala Rayavarapu, Krishna Kishore K.V | Vignan's Foundation for Science, Technology and Research

[6] Dhwaani Parikh, Vineet Menon "Machine Learning Applied to Cervical Cancer Data", RMIT University, 124 La Trobe St, Melbourne VIC 3000