CERVICAL CANCER DETECTION USING VARIOUS MACHINE LEARNING MODELS

Seema Jagannath Aher

ABSTRACT

Cervical cancer remains a significant cause of mortality in low-income countries. As in many other diseases, the existence of several screening/diagnosis methods and subjective physician preferences creates a complex ecosystem for automated methods.

Cervical Cancer is cancer arising from the cervix. It arises due to the abnormal growth of cells and spreads to other parts of the body. It is fatal most of the time. HPV causes most of the cases (90 %). In Phase I, the data is cleaned, and visualisations of the data are shown. Smoking is also considered as one of the main causes for cervical cancer. Long-term use of Oral contraceptive pills can also cause cancer. Also having multiple pregnancies can cause cervical cancer. 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. We can use machine learning techniques to predict if a person as cancer or not. Different factors such as smoking, pregnancies, habits etc can be used to predict cancer.

INTRODUCTION

About 11,000 new cases of invasive cervical cancer are diagnosed each year in the U.S. However, the number of new cervical cancer cases has been declining steadily over the past decades. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide. In the United States, cervical cancer mortality rates plunged by 74% from 1955 - 1992 thanks to increased screening and early detection with the Pap test. AGE

Fifty percent of cervical cancer diagnoses occur in women ages 35 - 54, and about 20% occur in women over 65 years of age. The median age of diagnosis is 48 years. About 15% of women develop cervical cancer between the ages of 20 - 30. Cervical cancer is extremely rare in women younger than age 20. However, many young women become infected with multiple types of human papilloma virus, which then can increase their risk of getting cervical cancer in the future. Young women with early abnormal changes who do not have regular examinations are at high risk for localized cancer by the time they are age 40, and for invasive cancer by age 50.

Socioeconomic and ethnic factors

Although the rate of cervical cancer has declined among both Caucasian and African-American women over the past decades, it remains much more prevalent in African-Americans -- whose death rates are twice as high as Caucasian women. Hispanic American women have more than twice the risk of invasive cervical cancer as Caucasian women, also due to a lower rate of screening. These differences, however, are almost certainly due to social and economic differences. Numerous studies report that high poverty levels are linked with low screening rates. In addition, lack of health insurance, limited transportation, and language difficulties hinder a poor woman's access to screening services.

HIGH SEXUAL ACTIVITY

Human papilloma virus (HPV) is the main risk factor for cervical cancer. In adults, the most important risk factor for HPV is sexual activity with an infected person. Women most at risk for cervical cancer are those with a history of multiple sexual partners, sexual intercourse at age 17 years or younger, or both. A woman who has never been sexually active has a very low risk for developing cervical cancer. Sexual activity with multiple partners increases the likelihood of many other sexually transmitted infections (chlamydia, gonorrhea, syphilis).Studies have found an association between chlamydia and cervical cancer risk, including the possibility that chlamydia may prolong HPV infection.

FAMILY HISTORY Women have a higher risk of cervical cancer if they have a first-degree relative (mother, sister) who has had cervical cancer.

USE OF ORAL CONTRACEPTIVES Studies have reported a strong association between cervical cancer and long-term use of oral contraception (OC). Women who take birth control pills for more than 5 - 10 years appear to have a much higher risk HPV infection (up to four times higher) than those who do not use OCs. (Women taking OCs for fewer than 5 years do not have a significantly higher risk.) The reasons for this risk from OC use are not entirely clear. Women who use OCs may be less likely to use a diaphragm, condoms, or other methods that offer some protection against sexual transmitted diseases, including HPV. Some research also suggests that the hormones in OCs might help the virus enter the genetic material of cervical cells.

HAVING MANY CHILDREN

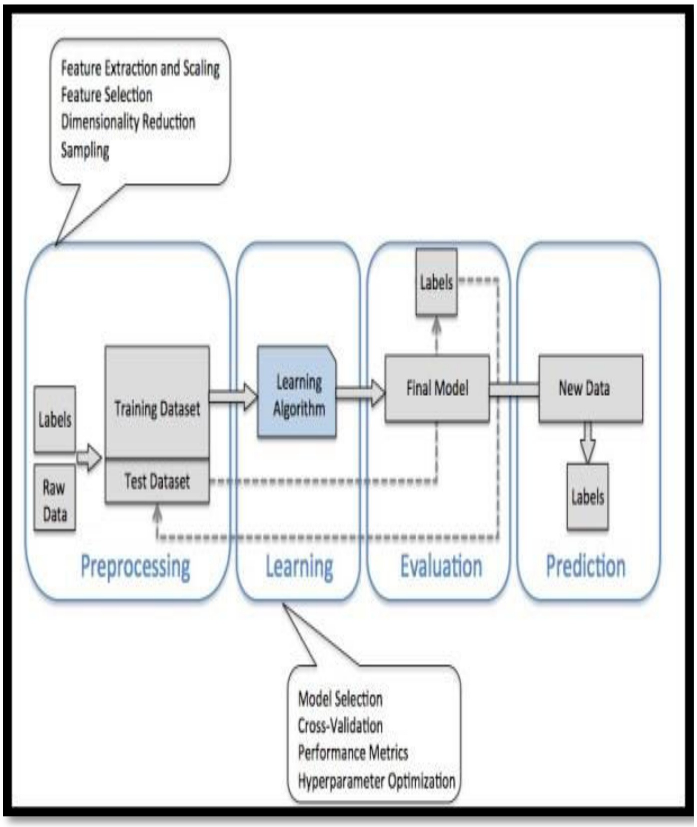
Studies indicate that having many children increases the risk for developing cervical cancer, particularly in women infected with HPV. SMOKING Smoking is associated with a higher risk for precancerous changes (dysplasia) in the cervix and for progression to invasive cervical cancer, especially for women infected with HPV. IMMUNOSUPPRESSION

Women with weak immune systems, (such as those with HIV / AIDS), are more susceptible to acquiring HPV. Immunocompromised patients are also at higher risk for having cervical precancer develop rapidly into invasive cancer.

DIETHYLSTILBESTROL (DES) From 1938 - 1971, diethylstilbestrol (DES), an estrogen-related drug, was widely prescribed to pregnant women to help prevent miscarriages. The daughters of these women face a higher risk for cervical cancer. DES is no longer prescribed.

What is Machine Learning:

Machine Learning is a branch of **artificial intelligence** that gives systems the ability to learn automatically and improve themselves from the experience without being explicitly programmed or without the intervention of human. Its main aim is to make computers learn automatically from the experience.

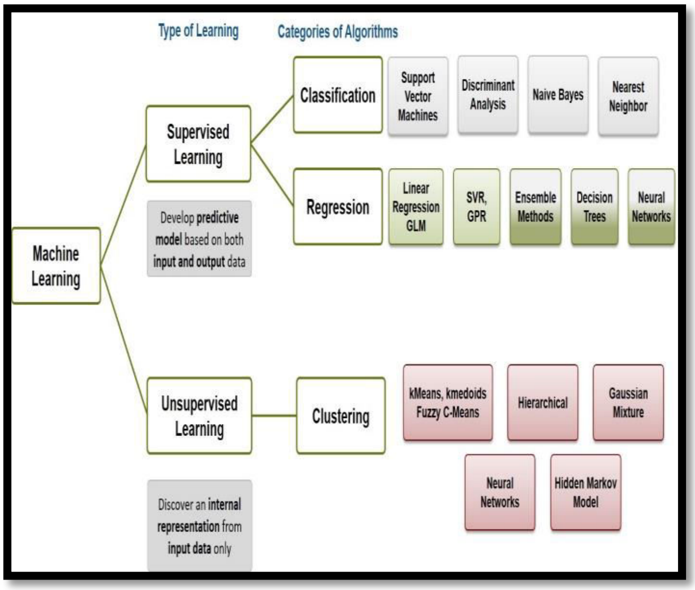


Requirements of creating good machine learning systems

So what is required for creating such machine learning systems? Following are the things required in creating such machine learning systems:

- **Data** – Input data is required for predicting the output.
- **Algorithms** – Machine Learning is dependent on certain statistical algorithms to determine data patterns.
- **Automation** – It is the ability to make systems operate automatically.
- **Iteration** – The complete process is iterative i.e. repetition of process.
- **Scalability** – The capacity of the machine can be increased or decreased in size and scale.
- **Modeling** – The models are created according to the demand by the process of modeling.

Methods of Machine Learning

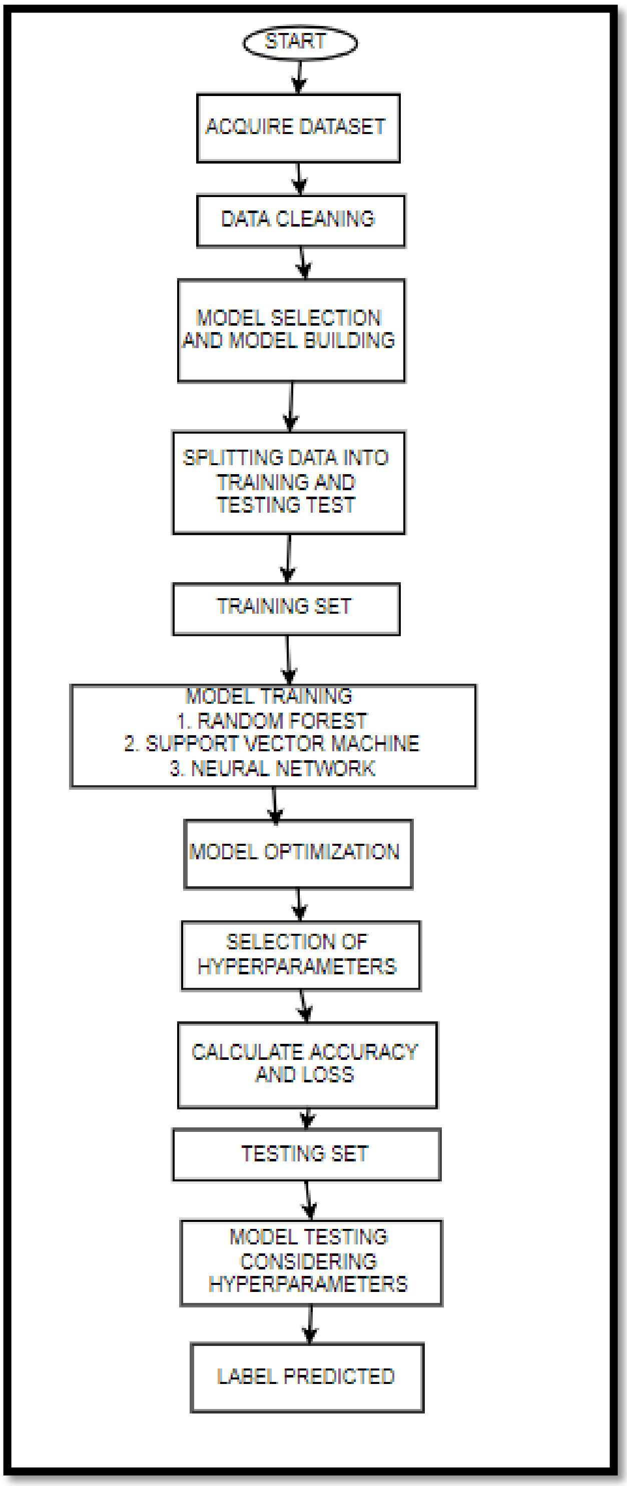


Machine Learning methods are classified into certain categories. These are:

- **Supervised Learning** – In this method, input and output is provided to the computer along with feedback during the training. The accuracy of predictions by the computer during training is also analyzed. The main goal of this training is to make computers learn how to map input to the output.
- **Unsupervised Learning** – In this case, no such training is provided leaving computers to find the output on its own.
- **Reinforcement Learning** – This type of learning uses three components namely – agent, environment, action. An agent is the one that perceives its surroundings, an environment is the one with which an agent interacts and acts in that environment. The main goal in reinforcement learning is to find the best possible policy.

DATASET: AI applications require maximum volumes of data, which must be well structured i.e. collate, stored securely,normalized, annotated, analysed, and accessible to endusers in a meaningful, intuitive fashion. Cancer detectionusing intelligent systems involves organised data set,labelled properly to assist accurate training of models.The dataset has been obtained from the dataset archive belongs to the University of California, Irvine. The dataset has been collected at *Hospital Universitario de Caracas* in Caracas, Venezuela.

PROPOSED METHODOLOGY


START
ACQUIRE DATASET
DATA CLEANING
MODEL SELECTION AND MODEL BUILDING
SPLITTING DATA INTO TRAINING AND TESTING TEST
TRAINING SET
MODEL TRAINING
1. RANDOM FOREST
2. SUPPORT VECTOR MACHINE
3. NEURAL NETWORK
MODEL OPTIMIZATION
SELECTION OF HYPERPARAMETERS
CALCULATE ACCURACY AND LOSS
TESTING SET
MODEL TESTING CONSIDERING HYPERPARAMETERS
LABEL PREDICTED

Random Forest

Random forest is a concept falling under the general technique of random decision. This algorithm operates by creating a group of decision trees at training time and outputting the class that represents the mode of classes or the mean prediction of the individual trees.

Individual decision trees are generated using a random selection of attributes at each node to determine split. During classification, each tree casts a vote and the most popular class is returned. Using the Random forests, the variance can be reduced by averaging the deep decision trees trained with different parts of the training set. To form Random forests, tree predictors should be integrated in a way that each tree be dependent on the values of a random vector sampled independently and uniformly from all trees in the forest. We use this approach to predict flight delays in our database.

Support Vector Machines (SVMs)

SVMs are supervised machine learning algorithmsmajorly used in classification problems. For the case ofpap-smear cell classification, this algorithm has provento yield considerably high accuracy [1]. During trainingprocess, it constructs a decision boundary orhyper-planes based on classified input set and divides thepoints into different classes. In case of high dimensional

data, multiple decision boundaries are constructed whichseparate the data. In real world application where theproblem is not linearly separable then such nonlinearproblem is mapped to a linearly separable problem. SVM decision boundary separates the data-points byminimizing the margin value, i.e. the distance betweenthe data-points and the decision boundary, which isbased on parameters like kernel, gamma, C, degree etc.

Neural Network

Neural Network is built by stacking together multiple neurons in layers to produce a final output. First layer is the input layer and the last is the output layer. All the layers in between is called hidden layers. Each neuron has an activation function. Some of the popular Activation functions are Sigmoid, ReLU, tanh etc. The parameters of the network are the weights and biases of each layer. The goal of the neural network is to learn the network parameters such that the predicted outcome is the same as the ground truth. Back-propagation along loss-function is used to learn the network parameters.

Conclusion

In this paper, we proposed a new method to identify cervical cancer using Machine Learning.

Three different models were used. Models used are Random Forest, Support Vector Machine and Neural Network Each yields good results about95% accuracy. We evaluated the performance of each model according to dataset.

REFERENCES

[1] *Cancer Report*. February 2017 [cited WHO World Health Organization 25.09.2017]; Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.

[2] Fernandes, K., J.S. Cardoso, and J. Fernandes, *Transfer Learning with Partial Observability Applied to Cervical Cancer Screening*, in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings*, L.A. Alexandre, J. Salvador Sánchez, and J.M.F. Rodrigues, Editors. 2017, Springer International Publishing: Cham. p. 243-250.

[3] Sen, T. and S. Das, *An Approach to Pancreatic Cancer Detection using Artificial Neural Network*, in *Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering (CEEE2013)*. 2013. p. 56-60.

[4] Fernandes, K., 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: Faro, Portugal. p. 61-69.

[5] Kalyankar, M.A. and N.R. Chopde, *Cancer Detection: Survey.* International Journal of Advanced Research in Computer Science and Software Engineering, 2013. **3**(11): p. 4.