Cervical Cancer Prediction

Introduction

The use of machine learning in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine (Weston and Hood 2004). This movement towards predictive medicine is important not only for patients but also for physicians in making informed treatment decisions. As a result of its capability machine learning has been used primarily as an aid to cancer diagnosis and detection (McCarthy et al. 2004). Embarking on cervical cancer prediction using machine learning would greatly impact the health sector and aid in alleviating the burden caused by cervical cancer in Southern Africa and Zimbabwe. This birthed the researcher's proposal of cervical cancer prediction using machine learning to be developed and deployed in all health centers in Zimbabwe.

Cervical cancer prediction using machine learning encompasses on sieving through the provided data that contains key risk factors that aid in the determination of cervical cancer development. Some of the key risk factors incorporate Age, Oral contraceptives, Smoking, the STIs one succumbed to and whether they once contracted HPV virus which is the main contributor to cervical cancer development. The system will analyse the captured medical data by employing the trained algorithm which will assign a risk score percentage. This percentage is used to predict the likelihood of cervical cancer development. This venture will result in close monitoring of the high-risk clients by the clinicians and further investigations being undertaken to probe and diagnose the client using conventional methods.

Background

Cervical Cancer is the second most common cancer that occurs in women in all age groups worldwide. In Zimbabwe, it is the most common cancer in women aged 15 to 44 years. According to the Zimbabwe Human Papillomavirus and Related Disease, Summary Report 2019 about 3186 new cervical cancer cases are diagnosed annually in Zimbabwe. The Annual number of cervical cancer deaths being at 2151. It's a deadly disease that is hard to detect in its earliest stages as it presents no symptoms. When compared to industrialized countries, developing world countries carry a disparate portion of disease burden; eighty-six percent of all cervical cancer cases and 88% of all cervical cancer deaths worldwide (Jemal et al. 2011; Ferlay et al. 2008).

The prevalence rate in Zimbabwe is very high as there are very few centers that carry out screening. Three tests can be done in the prediction and dictation of cervical cancer. The first being Pap smear which is a manual screening method where cervical cells from the cervix region of the uterus are extracted with a brush or spatula. There are approximately five pathologists in the country. This method suffers from a high rate of false positives due to human error. The most used method in Zimbabwe is Visual inspection with Acetic acid and Cervicography (VIAC). VIAC is less precise than Pap testing. It is an inexpensive screening method which reveals the presence of abnormal cells that might not be precancerous. The third less common method in Zimbabwe is liquid-based cytology (LBC) which immerses the collected cell samples in the liquid of 5% acetic acid. It is quite expensive and only one laboratory in

Bulawayo is known to carry out this method. These three conventional methods are currently being used worldwide.

The embarked system analyses the patterns in the dataset provided for training and testing it until it reaches an optimal accuracy rate of prediction. The algorithm will encompass on supervised learning where labeled data is fed into the algorithm to train and test it. Once a high accuracy rate is reached the system will be deployed. This system's sole focus will be on predicting the likely hood of cervical cancer development in the captured data. It will help in identifying high-risk females. The clinician will counsel and availing the limited conventional methods to further probe the high-risk profiles until a diagnosis is reached. The low-risk females will be sensitized on lifestyle changes necessary to avoid cervical cancer by the clinician.

Statement of the Problem

The current manual methods are prone to human error and limitation of resources makes the most adopted method VIAC only available to a few provincial hospitals and clinics hence denying the rural folks access to the services. There are very few experienced pathologists to perform these tasks and this method suffers from a high rate of false positives due to human error. Another setback is the fact that pathologists can classify only up to 5 slides per day. Power outages are also a major cause of concern as all the three screening methods are dependent on power. Due to outrageous power outages, the methods come to a halt as most facilities do not have solar panels to power the whole facility. Low socio-economic factors result in scanty resources being deployed nationwide most rural folks have no access to these services as they are in less than five provincial hospitals and clinics nationwide.

The proposed system seeks to address all these by complementing the conventional methods being used. The proposed system will predict the clients' risk of developing cancer by using a machine learning algorithm which is trained in classifying the risk level of the client, that is low, medium and high. The data fed into the system will contain crucial risk factors that contribute to the development of cervical cancer. The algorithm will look for patterns in the presented data and come up with a risk score percentage which is then used to classify the risk level. Any female client can get to know their risk status by simply providing the necessary data required by the system.

Aim

The aim of this study is to design and develop a Cervical Cancer prediction system using Machine learning approach. The system will analyse the medical data presented and assign a risk score percentage that will be used to classify the client's risk of developing cancer.

Objectives

The objectives of the system are:

- To develop an algorithm will analyse and predict a percentage of developing cervical cancer. using presented data.
- To capture patient's demographic data and medical data .
- To classify the patients' risk level based on the computed percentage.

- To export an excel document for statistical purposes for the Cancer Registry Association.

Justification

The burden of cervical cancer is still very high in Zimbabwe mainly as a result of the late presentation of the disease, poor screening, and diagnosis and treatment facilities. The current coverage of the current screening programs it is unlikely that the screening will be adequate to make an impact on cervical cancer mortality figures. Poor access to health facilities such as rural areas and lack of general education about the disease is not aiding in the fight against the disease. With a very high unemployment rate, most women fail to access quality health care.

To produce accurate results in the prediction of cervical cancer the machine learning approach will be used to predict the risk level of the client developing cervical cancer. It will classify the non-risk from the high-risk patients. The system will also aid in the detection of cervical cancer at its earliest stages as most high-risk clients will be referred for further investigations using the conventional methods. This will result in decreased mortality rate as most cases will be detected. The machine learning approach will focus on creating a model that will be able to predict easily and faster if a client has a higher risk or lower risk of developing cervical cancer just by analyzing the data entered. Machine learning methods, relative to simple statistical methods could substantially (15-25%) improve the accuracy of cancer susceptibility and cancer outcome prediction(Joseph A. Cruz and David S. Wishart, 2006).

Deployment of the cervical cancer prediction system at local health care centers will greatly impact service delivery and aid in the reduction of the epidemic as the general population will easily access the service. It is a cost-effective solution as most high-risk clients will be referred for further investigation using conventional methods. It will also increase sensitivity amongst the communities thus solving the information gap about the epidemic. Zimbabwe is susceptible to the most outrageous power shortages, this results in all the screening methods taking a huge halt as they are dependent on electricity to function. The cancer prediction system can be operated from a mobile device which enables high availability of the system even in outrageous power shortages as power banks and solar chargers can be employed to power up the system. This system will also aid in bridging the skilled labour gap as many clinicians will be able to classify the clients' risk level in their respective communities.

Scope

The system will be centered on predicting the risk level of developing cervical cancer. It is not a diagnosis system. The captured data will be fed into the system and the machine learning algorithm will analyse the data and compute a risk score percentage that will be used by the system to rank the risk level of the client. The ranks provided are Low, Medium and High-risk levels. The capturing module involves capturing of patient's demographic data and the patient's relevant medical information that is necessary for the prediction module. The prediction module will engage the machine learning algorithm which analyses the captured data, and notify the clinician the risk score percentage and classify the clients' rank.

Expected Results

The final system will be deployed on an Android platform. It will capture the patient's information which the system will then analyze to predict if the patients' risks level of developing cervical cancer and alert the clinician of the result. It will also aid in the compilation of statistical data for the National Cancer Research facility as this will aid in the statistical gathering process.