Open Science Index, Biomedical and Biological Engineering Vol:17, No:05, 2023 publications.waset.org/abstracts/159836.pdf

# Automated Prediction of HIV-associated Cervical Cancer Patients Using Data Mining Techniques for Survival Analysis

**Authors :** O. J. Akinsola, Yinan Zheng, Rose Anorlu, F. T. Ogunsola, Lifang Hou, Robert Leo-Murphy

**Abstract :** Cervical Cancer (CC) is the 2nd most common cancer among women living in low and middle-income countries, with no associated symptoms during formative periods. With the advancement and innovative medical research, there are numerous preventive measures being utilized, but the incidence of cervical cancer cannot be truncated with the application of only screening tests. The mortality associated with this invasive cervical cancer can be nipped in the bud through the important role of early-stage detection. This study research selected an array of different top features selection techniques which was aimed at developing a model that could validly diagnose the risk factors of cervical cancer. A retrospective clinic-based cohort study was conducted on 178 HIV-associated cervical cancer patients in Lagos University teaching Hospital, Nigeria (U54 data repository) in April 2022. The outcome measure was the automated prediction of the HIV-associated cervical cancer cases, while the predictor variables include: demographic information, reproductive history, birth control, sexual history, cervical cancer screening history for invasive cervical cancer. The proposed technique was assessed with R and Python programming software to produce the model by utilizing the classification algorithms for the detection and diagnosis of cervical cancer disease. Four machine learning classification algorithms used are: the machine learning model was split into training and testing dataset into ratio 80:20. The numerical features were also standardized while hyperparameter tuning was carried out on the machine learning to train and test the data. Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (KNN). Some fitting features were selected for the detection and diagnosis of cervical cancer diseases from selected characteristics in the dataset using the contribution of various selection methods for the classification cervical cancer into healthy or diseased status. The mean age of patients was 49.7±12.1 years, mean age at pregnancy was 23.3±5.5 years, mean age at first sexual experience was 19.4±3.2 years, while the mean BMI was 27.1±5.6 kg/m2. A larger percentage of the patients are Married (62.9%), while most of them have at least two sexual partners (72.5%). Age of patients (OR=1.065, p<0.001**), marital status (OR=0.375, p=0.011**), number of pregnancy live-births (OR=1.317, p=0.007**), and use of birth control pills (OR=0.291, p=0.015**) were found to be significantly associated with HIV-associated cervical cancer. On top ten 10 features (variables) considered in the analysis, RF claims the overall model performance, which include: accuracy of (72.0%), the precision of (84.6%), a recall of (84.6%) and F1-score of (74.0%) while LR has: an accuracy of (74.0%), precision of (70.0%), recall of (70.0%) and F1-score of (70.0%). The RF model identified 10 features predictive of developing cervical cancer. The age of patients was considered as the most important risk factor, followed by the number of pregnancy livebirths, marital status, and use of birth control pills, The study shows that data mining techniques could be used to identify women living with HIV at high risk of developing cervical cancer in Nigeria and other sub-Saharan African countries.

**Keywords :** associated cervical cancer, data mining, random forest, logistic regression
**Conference Title :** ICCSCR 2023 : International Conference on Cancer Science and Clinical Research
**Conference Location :** London, United Kingdom
**Conference Dates :** May 15-16, 2023