Contribution ID: **5** Type: **not specified**

ON THE ADVANCED PREDICTIVE MODELLING FOR CERVICAL CANCER DIAGNOSIS USING SUPPORT VECTOR MACHINE

Tuesday, 22 July 2025 10:55 (15 minutes)

Cervical cancer is a leading cause of death among women particularly in developing countries, and often arises from abnormal cell growth in the cervix. Various risk factors contribute to its development, and statistical models can help predict its occurrence. It remains an important and public health concern, especially in regions where fundamental screening programs and diagnosis efficiency face ongoing challenges. The current diagnostic practices rely on invasive methods and subjective evaluations resulting in late-stage detection and increased mortality rates. Studies have not yet integrated Statistics methods with Support Vector Machines (SVM) and interpretable Artificial Intelligence (AI) approaches to boost reliable and understandable cervical cancer risk prediction. This study therefore aimed to fill this gap by using Principal Component Analysis (PCA) and Mutual Information (MI) for feature selection and compare the predictive performance of Logistics Regression (LR) model and SVM.

The traditional multiple linear regression is the framework for this study. Predicting qualitative responses, a process known as classification, can be a daunting challenge. To enhance model interpretability and reduce feature dimensions, feature selection was performed using the combination of MI and PCA. Two predictive models were developed using LR as the baseline classifier and SVM for nonlinear classification. The models were evaluated by using accuracy, sensitivity, specificity, Positive Predicted Value (PPV), Negative Predicted Value (NPV), Balanced Accuracy (BA), kappa statistics and area under the curve (AUC). Datasets containing diagnostic indicators of cervical cancer from UCI Repository was used. It comprised demographic information, habits, and historic medical records of 858 patients with 36 features.

Mutual Information was used to determine the most relevant predictors for the presence of cervical cancer (Biopsy). Principal Component Analysis results indicate that the first two principal components, x-axis (Dim 1: 11.55%) and y-axis (Dim 2: 8.39%), together explain 19.94% of the total variance in the dataset. The LR model shows Schiller (8.80e-08), Citology (0.035774) and Dx (0.000388) are significant in predicting the presence of cancer at 0.05 significance level. The performance metrics for LR and SVM were; accuracy (55.56%; 94.78%), sensitivity (100%; 94.86%), specificity (0%, 93.75%), PPV (55.56%; 99.51%), NPV (NaN; 57.69%), BA (50%, 94.30%), Kappa (0.00; 0.6874) and AUC (0.9221; 0.9804). These results indicate that SVM significantly outperforms LR in cervical cancer prediction, demonstrating higher accuracy, balanced ability to identify both cancerous and healthy cases, higher precision, with fewer false positives, stronger agreement, and a superior AUC.

For cervical cancer detection, Support Vector Machine is preferred to Logistic Regression as it provides a better balance between sensitivity and specificity, has a much lower false positive rate, and correctly classifies most cases. Healthcare professionals can confidently rely on model predictions due to their interpretable features, which facilitate informed medical decision-making.

Keywords: Logistic Regression, Principal Component Analysis, Balanced Accuracy, Prediction Models

Primary author: OLADOJA, Oladapo (Abiola Ajimobi Technical University, Ibadan)

Co-author: Prof. ADEPOJU, Abosede (University of Ibadan, Ibadan, Nigeria)

Presenter: OLADOJA, Oladapo (Abiola Ajimobi Technical University, Ibadan)

Session Classification: Contributed talk: Room-2 (Statistics, Biostatistics and Epidemiology)

Track Classification: Sciences: Biostatistics/Medical Statistics