# Improved Feature Selection and Ensemble Learning for Cervical Cancer Assessment

Noor Alani, Md Rajib M Hasan
Auckland University of Technology, New Zealand
rhasan@aut.ac.nz

## Keywords

Feature Selection, Ensemble Learning, Data Mining, Machine Learning, Models, HPV, WEKA, MATLAB

## Abstract

Choosing the right influencing feature is a challenging field in data science due to the presence and complexity of multi-dimensional data. Cervical cancer is an excellent example for such study, as well as impacting individuals and families, presents almost no-symptoms at the early stages of development of this condition. Because multi-factors may be involved, this demands a lot of research and analysis to identify causative or linked features. The researchers have applied and optimised an ensemble learning algorithm as it is the best model for multi-modal medical data when relatively high dimensionality is present. The main objective of this study was to minimize the dependency on data pre-processing techniques, whilst analysing the data (filling/ignoring missing values with the statistical method). Main factors were studied and validated using Root Mean Square Error (RSME) and Mean Absolute Error (MAE). The classification accuracy for features were obtained by 10-fold cross-validation and test (where 66% is training data and 34% test data). The data was obtained from the UCI machine learning repository. WEKA and MATLAB were used to identify features. SPSS and SAS were used for RMSE and MAE. This approach is generic, and may also be applied to any relevant dataset for other purposes, and for teaching data analytics.

This quality assured paper appeared at the 11th annual conference of Computing and Information Technology Research and Education New Zealand (CITRENZ 2020) and the 33rd Annual Conference of the National Advisory Committee on Computing Qualifications [Convened online in 2020], Oct. 7-8.