Validated ML analytical approaches will be applied in identifying early detection of dysplasia and malignant transformation associated with oral cavity and oro-pharyngeal cancers.

Potential informatics approaches to conduct analysis includes key supervised ML approaches including Naïve Bayes, Logistic Regression, Random Forest, Multi-layer perceptron-Artificial Neural Network and others to see which algorithms perform optimally or categorization applying support vector machines. ML algorithm will be developed and executed by using machine learning libraries (for example WEKA, Matlab or R).

ML models will be trained, tested and evaluated using 10 fold cross validation. Predicting will be treated as a classification problem, where the datasets will be sorted into two categories based on patient diagnosis of oral cavity cancer or oropharyngeal cancer. Performance of various ML algorithms will be evaluated and compared using the following metrics: total accuracy, recall, sensitivity and specificity. To select the best performing model Receiver Operating Characteristics (ROC) curve, area under ROC (AUC) and Recall-Precision curves for each ML algorithm will be plotted.