Title: **5-fold cross-validation approach in evaluating the robustness of Machine learning models for prediction of oesophageal cancer**

Abstract:

Introduction:

Oesophageal cancer is a major health concern around the world, accounting for 3.1% of all cancer burdens and responsible for 5.5% of all cancer-related deaths. Due to its impact, there has been growing interest in adopting advanced methodologies. Machine learning techniques offer a promising avenue for understanding the disease better.

Methodology:

The study is based on a case-control study design, with a total of 400 case-control subjects equally sampled. The study explored different machine learning-based prediction models, and for each model, various performance metrics such as accuracy, precision, f1 score, recall, and roc-auc were evaluated. To optimize each model and determine the importance of the factors, a 5-fold cross-validation technique was employed, and the ranking of feature importance was done based on their weights in each model.

Results:

The study identified the Extra tree classifier model as the optimal approach in the prediction of oesophageal cancer, with a model accuracy of 87.50%, sensitivity of 92.5%, and specificity of 80%. When compared to the top 10 risk factors on basis of weight of feature of importance, the model showed an ROC-AUC value of 0.913, representing a substantial improvement of 10.1% over the baseline of the traditional risk prediction model (ROC-AUC 0.812; 95% CI 0.59-0.94).

The model identified Tooth loss, Smoking, Alkaline ethnic food, Smokeless tobacco, and Co-morbidity as the top 5 features of importance out of a total of 20 risk factors fitted in the model, these top five variables contributed to 34% of oesophageal cancer. If we incorporated Smoked fish, Betel nut use, Alcohol consumption, Traditional alcoholic drink, Use of fertilizer and Use of pesticide, the contribution rises to about 63%. The model algorithm predicted 36 out of 40 cases (sensitivity 0.90) and 33 out of 40 non-cases (specificity 0.825). It outperformed the traditional model by correctly predicting 3 more additional cases, resulting in the Extra Tree classifier model having 7.5% more sensitivity in the detection of oesophageal cancer cases compared to the baseline prediction model.

Conclusion:

The Extra tree classifier model exhibited higher predictability and accuracy in identifying important predictors of oesophageal cancer. The incorporation of this machine learning-based model presents exciting opportunities for policymakers to focus on specific risk factors.

Keyword: Machine learning, oesophageal cancer, etiologies, burden.