Paper ID: 1570943593

Paper Title: Identifying Critical Factors in Fetal Cancer-Related Deaths Using Machine Learning Models and Principal Components Analysis

Authors: Lalita Manavibool, Phattranij Meepadung, Nanticha Supmool and Naravin Vechpanich (Satriwithaya School, Thailand); Piyaon Sapphaphab and Oraya Rinthon (Satriwithaya School, Bangkok, Thailand); Suejit Pechprasarn (Rangsit University, Thailand)

Email: 47998@satriwit.ac.th

Abstract

Fetal cancer, a rare and complex condition, exerts a significant impact on affected families and healthcare systems. Occurring in the developing fetus, this condition presents unique challenges due to its scarcity and enigmatic nature. This study uses a publicly available clinical dataset of 2126 records of features extracted from Cardiotocogram exams with 21 variables: (1) baseline value, (2) accelerations, (3) fetal movement, (4) uterine contractions, (5) light decelerations, (6) severe decelerations, (7) prolonged decelerations, (8) abnormal short-term variability, (9) mean value of short-term variability, (10) percentage of time with abnormal long-term variability, (11) mean value of long-term variability, (12) histogram width, (13) histogram min, (14) histogram max, (15) histogram number of peaks, (16) histogram number of zeroes, (17) histogram mode, (18) histogram mean, (19) histogram median, (20) histogram variance, (21) histogram tendency. Here, we use this data to ensure its unbiasedness and accuracy. We then apply principal component analysis and machine learning models to identify factors in diagnosing classes of fetal health that consist of normal, suspect, and pathological. We investigate and compare the classification accuracy of different machine learning models, including tree, linear discriminant, quadratic discriminant, logistic regression, naive Bayes, support vector machine (SVM), K-nearest neighbor (KNN), ensemble, neural network, and kernel. The best models that can achieve the highest accuracy are Boosted trees with an accuracy of 94.65%. The principal component analysis method is then performed to identify crucial components and build an accurate model with fewer parameters. The Boosted Trees has the highest cross-validation classification and requires only five predictors: fetal movement, severe decelerations, histogram number of zeroes, prolonged decelerations, and histogram number of peaks.