Synthetic Data in Evidencing Hypothesis for Grant Application to collect Large Dataset of Routine Hematology Testing: A Call for New Narrative?

Jaweria M. (1), Neeha A. (1), Areeba M. (1), Zeeshan H. (1, 2), Ikram-Uddin U. (2), Najeed K. (1)

1 NED University of Engineering and Technology, Karachi-Pakistan. 2 Liaquat University of Medical and Health Sciences, Pakistan.

Introduction: Artificially generated or Synthetic data (SD) is helping the researchers in making data more accessible for their research and development, even carrying major limitations like 'questionable replication of the content and properties of original dataset' and others.

Material and Methodology: For present study, we used synthetic data approach for hypothesis testing in routine hematology laboratory to strengthen our proposal of collecting large dataset. After, the introduction of extended analytical channels in advance CBC analyzers potential morphological parameters (cell population data) are routinely generated. Focusing more than a hundred parameters against every analysis to read any specific deviational trend (fingerprint) is a real challenge that becomes opportunity for machine learning (ML). Data were extracted, labeled, and pre processed for application of ML models through computer command language (Python).

Results: A total of 5860 cases belong 65 study groups was an original dataset. At this point, the highest accuracy for our ML model was remained as higher as just of 45%. Next, data synthesis step was performed for getting double number (11720) of total cases, and similarly ML models were called. In results, worth discussing points including a noticeable accuracy of 85.61% with 91.92% precision for random forest classifier followed by decision tree, support vector machine, logistic regression, K-nearest neighbor, stochastic decent, and Gaussian naïve bayers were noted.

Conclusion: Although for present study synthetic data bridge data access gap to evidence our hypothesis at grant application stage while the original real data remains the preferred choice as we requested in our proposal to funding agency. It is a call for new narrative to examine synthetic data's validity and its utility in research through future discussions and studies for awareness among our hematology community.