Abstract

Machine learning and predictive analysis have become an integral part of society. It has sped up the transformation of various fields, especially Medical Science. The main objective of this report is to identify the relapse of antineutrophil cytoplasmic antibody-associated vasculitis (AAV) based on different biomarkers. We are using machine learning algorithms to determine the relapse early to create a personalized treatment plan for the patient. Every case is unique, leading to a non-standard treatment process resulting in some missing biomarkers. These missing biomarkers cause the machine learning models to often fails due to it being a statistical formula. One of the most common methods of resolving the missing values problem is the Multivariate Imputation by Chained Equations, more commonly referred to as the MICE package in R programming language. Another hindrance to using Machine Learning is that Rare diseases such as ANCA-Associated Vasculitis have a meagre data count. Many restrictions are due to federal use rules like HIPAA and GDPR, adding multiple more restrictions to the available data and making the process of the analysis complex. To circumvent the issue of insufficient data, we are using statistical methods to synthesize new data points with similar functionality as the original Data. A constant evaluation using the pairwise correlation comparison and log cluster ensures the integrity of the new synthetic data. Sallow machine learning algorithms like Decision Tree and Random Forest are trained on the newly synthesized data and then tested on the original data along with the added imputed data points. Cross Validation and Random Search help identify the parameters for creating an unbiased model with the best performance.