Waterborne illness cases in Turkey: an effective deep learning prediction method

Mehmet Akif Çifçi, Ozan Deniz

1*Bandirma Onyedi Eylul Univ., Dept. of Computer Eng., Balikesir, Turkey, mcifci@bandirma.edu.tr, 0000-0002-6439-8826
2Namık Kemal University, Tekirdağ Turkey, ozandeniz.1978@gmail.com

Abstract: Because of extensive growth in the population of Turkey, quality of water has been compromised and threatened by various pollutants, causing an intense increase of waterborne disease and affecting many areas of Turkey. Therefore, modeling and prediction of water quality have become very important in controlling water pollution and have become one of the hot topics for researchers. We have developed different Machine Learning models and Artificial Intelligence algorithms to predict the water quality index, water quality classification, and water quality classification forecasting. In our research, we have focused on predicting the positive case of waterborne disease and we have used Marmara Ereglisi region's data collected by us in 2018-2020 that may contain Typhoid infection and Malaria disease. To deal with the imbalance problem found in the dataset, we used an under-sampling technique. We have performed experiments on open public malaria patient records (22916) and records of typhoid (68624). We have used five Machine learning algorithms namely, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression, and K-Nearest Neighbor (KNN). The mentioned data set has 6 significant input features and the developed models were evaluated on it for the prediction of positive waterborne disease. The experimented results revealed Random Forest performed well in terms of accuracy prediction of waterborne disease 60% for the malaria dataset and 77 % for the typhoid data set as compared to alternative ML Classifier. In this research, we have also focused on the factors that are more important in the prediction of symptoms that will help in the analysis of positive cases of waterborne disease. The random Forest feature selection technique has been used and experimental results have shown that age, history, and test results play a significant role in the prediction of waterborne disease-positive cases. Finally, we have concluded that this type of promising research can contribute to health care departments for the health-relevant decisions.