A machine learning approach to diagnose blood diseases

Alexander Luchinin, MD, PhD[1], Oleg Kolupaev, PhD[2]

[1] The Federal State-Financed Scientific Institution Kirov Research Institute of Hematology and Blood Transfusion under the Federal Medical Biological Agency

[2] UNC Chapel Hill, Lineberger Comprehensive Cancer Center

Background. Machine learning (ML) is a powerful tool for conducting complex scientific and clinical tasks, for instance, diagnostics and prognosis of human diseases. The benefit of machine learning applications is based on fast calculations and big data manipulations. However, in the field of hematology successful applications of ML tools have been mostly limited to blood cell imaging tasks or less often, laboratory results interpretation. With this in mind, we set out to develop a more effective and efficient tool that can assist in clinical decisions. Here we present the results of our work in this area, which is aimed at creating a machine learning model for predicting several types of blood disorders based on common blood counts.

Methods. The study collected anonymized common blood count (CBC) data on 8241 cases of adults (18+ years of age) who had or did not have a blood disorder. There were 14 CBC parameters, as well as age and gender, in each case. These parameters are hemoglobin level, red blood cell count, white blood cell count (WBC), platelet count, neutrophils, lymphocytes, monocytes, eosinophils, and basophils. Percentages and absolute counts were provided for all types of WBC. Six distinct categories of patients were identified: normal blood count, abnormal blood count (non-clinical significance), abnormal blood count (clinical significance), chronic lymphocytic leukemia (CLL), chronic myeloproliferative disease (CMD, and iron deficiency anemia (IDA). Expert hematologists validated all blood tests (supervised ML method). We have used random forest approach, a special type of ensemble method designed for use with structured data. Due to their lack of significance, we excluded sex and a relative number of eosinophils and basophils from our final model. All statistical analyses were undertaken using R version 3.4.2 and the "randomForest", "caret", "pROC", "ROSE" packages.

Results. During the preparation of the dataset, we encountered problems due to imbalanced classes, so synthetic balanced samples were generated. In total, 160628 variables were processed for both the training and test datasets (80/20 split). We generated a predictive model with 13 parameters by using a random forest algorithm. Based on the test dataset, our model accurately classified patients with blood disorders with an AUC of 0.982 (95%CI 0.978-0.985). The prediction accuracy of each class of cases varied: normal blood count - AUC 0.993, sensitivity 0.989, specificity 0.996; non-clinical significance abnormal blood test - AUC 0.999, sensitivity 1, specificity 0.999; clinical significance abnormal blood test - AUC 0.986, sensitivity 0.977, specificity 0.995; CLL - AUC 1, sensitivity 1, specificity 1; CMD - AUC 0.971, sensitivity 0.948, specificity 0.995; IDA - AUC 0.984, sensitivity 0.978, specificity 0.99.

Conclusion. CBC is the most common laboratory tests, and it can be challenging to interpret. Application of ML approaches, including ones using a random forest algorithm to this type of data can be an effective tool increasing speed and quality of CBC interpretation. The use of ML models in medical laboratory diagnosis has also shown promise in automated blood disorders screening systems. When implemented in large reference laboratories, ML-based algorithms can be used to assist general practitioners in detecting blood problems early, providing a great clinical value to both physicians and patients.