Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220684

Substantive Interpretation of Machine Learning Solutions by the Example of Determining the Activity of the Tuberculosis Process in Individuals with Minimal Tuberculosis Residual Changes

Tatyana TYULKOVA [a] , Pavel CHERNAVIN [b] , Nikolai CHERNAVIN [b] , Yuri CHUGAEV [c] , Igor CHERNYAEV [d]
[a] *Ural Scientific Research Institute of Phthisiopulmonology (USRIPh) – the branch of the National Medical Research Center of Phthisiopulmonology and Infection Diseases (NMRC PhPI) of Russian Federation, Ekaterinburg, Russia*
[b] *Ural Federal University, Ekaterinburg, Russia*
[c] *Ural State Medical University, Ekaterinburg, Russia*
[d] *Minist Hlth Russia, Natl Med Res Ctr Phthisiopulmonol & Infect Dis, Ekaterinburg, Russia*

Abstract. In this article is described an application of various machine learning (ML) methods to obtain decision rules and its interpretation to a problem of recognition of activity of the tuberculosis process. The research data base included 489 patients registered in anti-tuberculosis institutions in Tyumen and Yekaterinburg. The conducted modeling by machine learning methods allowed to highlight 7 most informative features (the presence of calcifications, age, the content of leukocytes, hemoglobin, eosinophils, α2-fraction of globulins, γ-fraction of globulins) together with classification accuracy of 95% for both active and inactive patients. The research result may be interesting for medical specialists, data scientists and to all those interested in problems at the intersection of medicine and machine learning.

Keywords. tuberculosis, machine learning, committee machine

1. Introduction

In the last decade, interest to artificial intelligence (AI) in medicine has increased dramatically around the world. However, there appear more and more skeptical statements regarding both individual studies and the applicability of AI in medicine in general. In our opinion, many skeptical statements about AI were quite fair. The main reproaches from medical specialists were concerned due to the fact that in most cases the results of computations are were absolutely not interpretable. Usually, systems built on AI give the probability of the presence of a particular disease without any explanation of their conclusion. Of course, having such an answer is better than none, but questions about trust, safety, contestability and confirmation of the diagnosis by a medical

specialist immediately arise. Without a substantiated diagnosis, a self-respecting medical specialist is unlikely to make a decision to prescribe a therapy, procedure and/or surgical intervention. Moreover, without understanding of the relationship between inputed data and a diagnosis, it is quite difficult to choose a treatment procedure, so the value of a diagnosis drops sharply. The main reason for this situation is the use of uninterpreted methods. The rules intended for a decision making (decision rules - DR), built on the basis of such methods as random forest, gradient boosting and neural networks, cannot be interpreted even in any understandable form for a practical medical specialist. Consequently, medical specialists are excluded from the process of developing DR and just have to "believe or not believe" to provided DR.

In this research to show how to find a balance between accuracy and interpretability DR was built on data of patients in whom tuberculosis has gone from a latent state to an active one. The importance of such studies has been repeatedly noted by the World Health Organization and many medical specialists [1].

2. Data and methods

The research included 489 patients registered in anti-tuberculosis institutions in Tyumen and Yekaterinburg. After analysys of input data 58% of patients were excluded from the dataset, because there was no information on their blood test. Patient distribution in the final dataset is described in the Table 1 below.

Table 1. Distribution of patients with active and non-active TB status by present of calcinates

TB status	Present	Not present	Total
Active TB	17	122	139
Non-active TB	39	26	65
Total	56	148	204

The following features were taken into account: age, gender, medical history, presence of BCG vaccination, features of the early period of primary tuberculosis infection; the number of years that have passed since the change in tuberculin sensitivity to the detection of changes detected by X-ray, the presence and duration of contact with a patient with tuberculosis, the level of social adaptation of the family, indicators of the general blood test in dynamics (absolute and relative values: erythrocytes, leukocytes, platelets, hemoglobin, eosinophils, stab neutrophils, segmented neutrophils, lymphocytes, monocytes, ESR); indicators of the biochemical composition of blood in dynamics (glucose level, total protein, albumins, α1-globulin fraction, α2-globulin fraction, β-globulin fraction, globulin γ-fraction, albumin-globulin coefficient (A\G). Particular attention was paid to radiographic characteristics that were formalized and entered into the research database on a binary basis (presence/absence) For validation purpose there was taken 20% of data in random order[1]. To process data and build a model there were used a Python libraries such as MIP, SKLEARN and Keras.

[1] by sklearn.model_selection.train_test_split method

3. Results

3.1. Comparison of ML methods

Results of classification models constructed by various methods are presented in the Table 2 below.

Table 2. Comparison of different ML methods by share of identified patients with inflammation activity non-active and active, %

Method	Non-active	Active
Logistic Regression	64.9	93.3
Naive Bayes	29.8	92.2
Nearest Neighbors	61.4	97.2
Neural Network	68.4	94.4
Support Vector Machine	70.0	92.7
Random Forest	82.3	93.1

According to the Table 2 the Random Forest showed the best results, but it is quite difficult to interpret its results. Therefore, committee machine method (CM), in its geometric formulation, was used for DR construction. That approach made it possible to localize areas in the feature space that corresponded to sick and healthy patients from the training set. These areas had been unambiguously described as a system of inequalities and were easily explained to medical professionals. Geometrically, they corresponded to some convex polyhedra in the feature space.

As a result by the CM there were identified[2] 7 most informative features which are described in details below in the Table 3.

Table 3. Statistical description of the most informative features in the CM model

Feature	Min	Max	Mean	St. error, ± t*m (95%CI)	St. deviation	Moda	Median
Age	0.11	17	8.78	0.78	5.7	16	9
Content of leukocytes	3.4	16.9	7.52	0.35	2.54	5.6	7
Hemoglobin	59	154	121.9	2.05	14.91	117	120
Eosinophils	0	18	3.04	0.42	3.04	1	2
α2-fraction of globulins	4.98	15.4	9.6	0.27	1.97	10.35	9.38
γ-fraction of globulins	9.19	37.2	18.9	0.69	4.98	14.95	18.62
Presence of calcifications	Categorical feature (0/1). Distribution description is described in the Table 1						

The accuracy of this method is 95% for both active and inactive patients. A detailed description of the CM results with the DR coefficients were given in the patent [2].

3.2. Interpretation of the CM results

According to the present model of tuberculosis infection in human [3-8], the leukocytes level presented a reaction of the organism on a pathogenic agent. In the case of a nonspecific pathogen, the increase of this feature was most significant, while in a

[2] Categorical features were identified by sklearn.feature_selection method and for quantitive features were used special restriction on the CM model with separate boolean variable to identify informative features (minimization of the boolean variables sum with restriction on classification quality). The idea of such restriction was previously described in authors' earlier research [9].

case of invasion by Mycobacterium tuberculosis, it was insignificant. The relative level of eosinophils causes the activation of the Th2 pathway of the immune response. With this type of immune reaction macroorganism can not adequately respond to the appearance of the causative agent of tuberculosis, which freely spreads through the lympatic and blood vessels. On the activation of the humoral immune link, the level of the γ-fraction of globulins containing immunoglobulins in its composition is detected. In case of tuberculosis infection, the activation of this immune link does not lead to the restriction of tuberculosis infection, which also causes the wide spread of the pathogen - Mycobacteriae tuberculosis (MBT). In general, the activity of the humoral link is indirectly characterized by the level of leukocytes, eosinophils, immunoglobulins. Due to the fact that for tuberculosis is not typical the activity of the humoral link, then a change in this feature signalize the development of dysfunction in the body and the ability of the MBT out of the control of the immune system.

α2-globulin fraction brings together α2-macroglobulin and haptoglobin. The concentration of this proteins is regulated by pro-inflammatory cytokines, which are involved in the immune response to the occurrence of a pathogen. Also MBT α2-macroglobulin is involved into regulation of proteolytic effect of leukocyte collagenases. That 2 factors explain their joint application to determine a specific inflammation activity. Haptoglobin and its complexes with hemoglobin (the level of which is also taken into account to determine increased activity of the inflammatory process), play an important role in maintaining a level of iron, controlling of inflammatory processes, hydrolization of peroxides, which appear during the action of phagocytes and modulation of the activity and proliferation of leukocytes in the focus of inflammation. Haptoglobin is bacteriostatic pathogens in infections with iron-dependent bacteria, which include MBT. Long-lasting high haptoglobin values were signs of an unfavorable course of the disease. According to literature data, an increase in α2-fractions and γ-fractions of globulins was for the activity of chronic inflammation, which may to be tuberculosis.

4. Conclusions

In this research was demonstrated that machine learning methods can form DR applicable to assist in decision making on a patient state. CM allowed to move from the geometric processing of research results to the content interpretation and identified 7 features which have high influence for the model of minimal tuberculosis residual changes. Despite the fact that these features are commonly used in standard medical analysis further developing of such models together with dataset extension can allow in future to identify more informative features. In authors' opinion the obtained results as a list of coefficients to corresponding features can be independently tested and later be put into work by any medical specialist.

References

[1] Houben RM, Dodd PJ. The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. PLoS Med 2016;13:e1002152. https://doi.org/10.1371/journal.pmed.1002152)

[2] Tyulkova T.E., Vladimirsky M.A., Khabibullina N.F., Chernavin P.F., Chernavin N.P. Method for determining the activity of specific inflammation in the presence of minimal tuberculous changes in children and adolescents, patent RU 2 728 943 C1, 2020.08.03

[3] The Interactive Core Curriculum on Tuberculosis . Page last reviewed: March 1, 2021 Content source: Division of Tuberculosis Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention https://www.cdc.gov/tb/education/corecurr/

[4] Seshadri C., Sutherland J. S., Lindestam A. C. S., Burel J. G. Editorial: Exploring Immune Variability in Susceptibility to Tuberculosis Infection in Humans Frontiers in Immunology V 12, 2022 URL=https://www.frontiersin.org/article/10.3389/fimmu.2021.830920 DOI=10.3389/fimmu.2021.830920

[5] Andrews JR, Nemes E, Tameris M, Landry BS, Mahomed H, McClain JB, et al. Serial QuantiFERON Testing and Tuberculosis Disease Risk Among Young Children: An Observational Cohort Study. Lancet Respir Med (2017) 5(4):282–90. doi: 10.1016/S2213-2600(17)30060-7

[6] Behr MA, Edelstein PH, Ramakrishnan L. Is Mycobacterium Tuberculosis Infection Life Long. BMJ (2019) 367:1–7. doi: 10.1136/bmj.l5770

[7] Behr MA, Kaufmann E, Duffin J, Edelstein PH, Ramakrishnan L. Latent Tuberculosis: Two Centuries of Confusion. Am J Respir Crit Care Med (2021) 204(2):142–8. doi: 10.1164/rccm.202011-4239PP

[8] Turner CT, Gupta RK, Tsaliki E, Roe JK, Mondal P, Nyawo GR, et al. Blood Transcriptional Biomarkers for Active Pulmonary Tuberculosis in a High-Burden Setting: A Prospective, Observational, Diagnostic Accuracy Study. Lancet Respir Med (2020) 8(4):407–19. doi: 10.1016/S2213-2600(19)30469-2