Assiciation between biofactors, hematological factors and COVID-19 in MASHAD cohort: A machine learning technique

\*Amin Mansoori1, Nasrin Talkhi2, Habibollah Esmaily2, Somayeh Ghiasi Hafezi3, Majid Ghayour-Mobarhan1

1 International UNESCO Center for Health Related Basic Sciences and Human Nutrition, Mashhad University of Medical Sciences, Mashhad, Iran

**2** Department of Epidemiology and Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran

3 Department of Applied Mathematics, Ferdowsi University of Mashhad, Mashhad, Iran

am.ma7676@yahoo.com

***Abstract*—The SARS-CoV-2 is a novel corona virus that was unknown before the outbreak's onset and was first observed in China in late December 2019. It is now a serious global health concern. Since January 8, Iran has reported 1,431,416 total cases and 58,110 deaths. The virus has a high mortality and disability rate, particularly in some individuals, such as the elderly, those with underlying disorders, and those with immune system deficiencies. Diabetes and Covid-19 have a bidirectional connection. Diabetes is associated with a greater risk of COVID-19 infection. Individuals with diabetes are more vulnerable to infections, and diabetes has been reported as a significant risk factor for mortality in H1N1 (patients infected with Pandemic disease Influenza A 2009), SARS (Severe Acute Respiratory Syndrome) corona-virus, and MERS-CoV (Middle East Respiratory Syndrome-related corona-virus). It is also notable that more cases of early-onset diabetes and diabetic ketoacidosis have been documented in patients with SARS corona-virus. More knowledge of the specific symptoms and risk determinants of COVID-19 in different clinical settings is needed to properly treat these patients and to avoid disease complications.**

**A total of 8,791 participants who met the inclusion criteria were recruited. The models were built on the training dataset (80%, randomly) and validated using test data (20%). Decision Tree (DT) and Logistic Regression (LR) techniques were fitted into training data. Two models were considered in this study, in model I, the biochemical features and in model II, the hematological features were studied.**

**In model I, the Total Bilirubin and Magnesium variables have been identified as the most important variable and in model II, Age, smoking status, RBC, and Hemoglobin were significant by the LR model. Our DT model identified Copper, HDL, hs.CRP, CPK, age, BUN, Calcium, ALP, SBP in model I and Age, MCV, BMI, RDW, PDW, and MCH in model II as important predictors. The sensitivity, specificity and accuracy of the DT model I (model II) in the testing were 70.73% (66%), 71.27 (72.88), 70.99% (68.80%), respectively.**

Keywords—Corona virus; Biochemical factors; Hematological factors; Decision tree algorithm

##### References

1. M. M. Ghiasi and S. Zendehboudi, “Application of decision tree-based ensemble learning in the classification of breast cancer,” Computers in Biology and Medicine, vol. 128, 104089, 2021.
2. N. Al-Azzam, L. Elsalem, and F. Gombedza, “A cross-sectional study to determine factors affecting dental and medical students’ preference for virtual learning during the COVID-19 outbreak,” Heliyon, vol. 6(12), e05704, 2020.
3. M. Saberi‐Karimian, H. Safarian‐Bana, E. Mohammadzadeh, T. Kazemi, A. Mansoori, H. Ghazizadeh, S. Samadi, I. Nikbakht‐Jam, M. Nosrati, G. A Ferns, H. Esmaily, M. Aghasizadeh, and M. Ghayour‐Mobarhan, “[A pilot study of the effects of crocin on high‐density lipoprotein cholesterol uptake capacity in patients with metabolic syndrome: A randomized clinical trial](https://iubmb.onlinelibrary.wiley.com/doi/abs/10.1002/biof.1783),” BioFactors, in press.
4. M. Saberi-Karimian, Z. Khorasanchi, H. Ghazizadeh, M. Tayefi, S. Saffar, G. A Ferns, M. Ghayour-Mobarhan, “[Potential value and impact of data mining and machine learning in clinical diagnostics](https://www.tandfonline.com/doi/abs/10.1080/10408363.2020.1857681),” Critical Reviews in Clinical Laboratory Sciences, vol. 58(4), pp. 275-296, 2021.