

Assessing the Relative Importance of Land-Atmosphere Variables for Extreme Heat Events in Czechia

Yilmaz Volkan

Department of Geomatics Engineering, Karadeniz Technical University, Trabzon, 61080, Türkiye, volkanyilmaz.jdz@ktu.edu.tr

Abstract

The aim of this study is to investigate the key environmental drivers of extreme high-temperature events across Czechia using multi-source Earth observation data. For this purpose, extreme hot days were identified through a percentile-based approach, using daily maximum 2-m air temperature (T_{max}) from the ERA5-Land reanalysis data obtained for the period between 2005 and 2025. A day was classified as extreme if T_{max} exceeded the long-term monthly 90th percentile (p₉₀), calculated separately for each calendar month to account for seasonality. The occurrence of extreme temperature events was modelled using a Random Forest (RF) classification approach. Predictor variables included the cumulative precipitation, mean surface soil moisture, mean Normalized Difference Vegetation Index (NDVI), mean shortwave surface albedo, mean land surface temperature (LST) and mean wind speed obtained from ERA5-Land, MODIS and CHIRPS datasets using variable-specific temporal aggregation windows designed to capture short-term memory effects. RF model was trained using data from 2005-2018 and the evaluation was conducted on an independent test set covering 2019-2025. SHAP (SHapley Additive exPlanations)-based variable importance values obtained from RF model were normalized to percentage contributions in order to compare the impacts of the variables. LST was found to be the most influential predictor (23.4%), emphasizing the role of temperature dynamics in the occurrence of extreme heat. Soil moisture (18.4%) and wind speed (17.2%) followed closely. Vegetation conditions, represented by NDVI, contributed 15.6%, while precipitation accounted for 13.6%. The predictor with the least impact on extreme heat events was surface albedo (11.8%).

Keywords: Extreme heat, Land-atmosphere interactions, Machine learning, Climate change, Remote sensing

1. Introduction

Extreme high-temperature events play an important role in climate change, with profound impacts on ecosystems, agriculture, energy systems and human health. These events are even more dangerous for older people (Fastl et al., 2024). Central Europe has experienced a noticeable increase in the frequency and intensity of heat extremes in recent years, increasing concerns about regional exposure and the reasons behind these events. Several current studies reported this upward trend in extreme heat events (Lorenz et al., 2019; Weinhhammer et al., 2021; Malcheva et al., 2022; Suarez-Gutierrez et al., 2023).

Suarez-Gutierrez et al., (2023) explored how climate warming influences the likelihood of extreme heat and drought events over Europe. They reported that the heat conditions that can be expected by the end of the century could occur in the 2030s with a considerable probability. Vautard et al. (2023) focused on the reasons why extreme heat in Western Europe has intensified more rapidly than climate models have projected over the past seven decades. They showed that an increase in southerly atmospheric circulation patterns accounts for a significant portion of the observed rise in heat extremes, an effect that is largely missing from current climate models and thus introduces substantial uncertainty into future projections. Kirchengast et al., (2025) revealed five- to twenty-five-fold amplifications in the total extremity of extreme

heat in Austria and large parts of Europe, with signals clearly emerging beyond natural variability and providing robust evidence of anthropogenic climate change.

Despite the growing body of literature on heat extremes, there is still a limited understanding of how land-surface states and near-surface atmospheric conditions jointly modulate extreme high-temperature events at regional scales. In particular, disentangling the relative contributions of soil moisture, vegetation, surface radiative properties and atmospheric mixing remains challenging due to the nonlinear and interacting nature of these processes. Moreover, many existing studies rely on complex climate model simulations, which can obscure seasonal variability and limit physical interpretability. Motivated by these gaps, this study adopts a data-driven yet physically grounded framework that combines percentile-based extreme definitions with explainable machine learning (ML) techniques. By focusing on Czechia as a representative region in Central Europe, the study aims to provide transferable insights into the mechanisms governing extreme heat and to support improved climate risk assessment and adaptation planning under ongoing climate change.

2. Material and Method

2.1. Study Area

The study was conducted in Czechia, which is located in Central Europe (see Fig. 1). The country has a temperate climate with warm summers, and cold (snowy) and cloudy winters. Temperature levels are not uniform in the country and vary depending on elevation. The coldest and warmest months are generally January and July, respectively. Winter temperatures are about 20-30°C lower than those of summer. The majority of precipitation occurs in spring and summer months. Yearly precipitation varies by location. Lower-lying regions receive approximately 500-700 mm of rainfall annually, whereas precipitation in mountainous regions like Krkonoše Mountains may reach up to 1250 mm/year (URL-1, URL-2).



Fig. 1. Study area.

2.2. Data

Extreme high-temperature events and their environmental drivers over the Czechia for the period 2005-2025 were analysed using a combination of reanalysis and satellite-based datasets. Daily maximum 2-m air temperature (Tmax) was obtained from the ECMWF ERA5-Land reanalysis (ECMWF/ERA5_LAND/DAILY_AGGR). The temperature_2m_max band, provided in Kelvin, was converted to degrees Celsius and served as the primary temperature variable in the analysis.

Environmental conditions associated with extreme heat were described using multiple datasets representing land-surface, vegetation, hydrological and atmospheric processes. Surface volumetric soil water content (`volumetric_soil_water_layer_1`) and near-surface wind components were taken from the ERA5-Land reanalysis. Zonal and meridional wind components at 10 m height (`u_component_of_wind_10m` and `v_component_of_wind_10m`) were used to derive wind speed, representing atmospheric circulation and ventilation conditions near the surface.

Precipitation information was obtained from the CHIRPS daily precipitation dataset (UCSB-CHG/CHIRPS/DAILY), which combines satellite observations with in situ measurements to provide spatially consistent rainfall estimates. Vegetation conditions were characterized using the Normalized Difference Vegetation Index (NDVI) from the MODIS MOD13A2 Collection 061 product. Surface thermal properties were represented by daytime land surface temperature (LST) from the MODIS MOD11A2 product, while surface radiative characteristics were described using shortwave white-sky albedo from the MODIS MCD43A3 BRDF/Albedo product.

Topographic influences on extreme heat occurrence were assessed using elevation data from the NASADEM (NASA/NASADEM_HGT/001) dataset. Elevation was used in supplementary analyses to examine the relationship between mean annual extreme hot day frequency and terrain height across the study region.

2.3. Method

This study applies a combined percentile-based extreme detection and ML framework to investigate the drivers of extreme high-temperature events over the Czechia. Daily Tmax from the ERA5-Land reanalysis was used as the target variable. Extreme hot days were identified using a seasonally adaptive approach, whereby a day was classified as extreme if Tmax exceeded the long-term monthly 90th percentile (p_{90}), calculated separately for each calendar month over the 2005-2025 period.

Environmental predictors representing land-surface, vegetation, hydrological and atmospheric processes were derived from ERA5-Land, MODIS and CHIRPS datasets using variable-specific temporal aggregation windows. Surface soil moisture was aggregated as a 7-day mean, precipitation as a 5-day cumulative sum, NDVI and surface albedo as 16-day means, LST as an 8-day mean and 10-m wind speed as a 3-day mean derived from zonal and meridional wind components. All variables were spatially clipped to the Czechia and resampled to a common spatial resolution of 10 km. To reduce temporal autocorrelation and computational demand, the daily time series was temporally thinned by selecting one representative day every 15 days.

Balanced training and testing datasets were constructed using stratified random sampling, drawing equal numbers of samples from extreme and non-extreme classes for each selected day. Model training was performed for the 2005-2018 period, while independent evaluation used data from 2019-2025. Extreme hot day classification was carried out using a Random Forest (RF) algorithm with 300 decision trees. Model performance was assessed using overall accuracy statistic. To enhance the interpretability of the RF model, SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) was employed to quantify the contribution of individual predictors to model outputs. SHAP values were computed using the TreeExplainer algorithm, which is specifically optimized for tree-based models. Global feature importance was derived by calculating the mean absolute SHAP value of each predictor across all samples and subsequently normalizing these values to percentages so that the total contribution sums to 100%.

In addition to classification, spatial patterns of extreme hot days were analysed by mapping annual and long-term mean extreme-day frequencies, while temporal variability was examined using regionally averaged annual extreme-day counts. Supplementary analyses investigated the relationship between extreme heat frequency and elevation.

3. Results and Discussion

Fig. 2 presents the monthly mean Tmax together with the corresponding monthly p90 thresholds over the Czechia for the 2005-2025 period. Both metrics exhibit an apparent seasonal cycle, with lowest values during winter months (January-February) and highest values in summer (July-August). Mean Tmax increases from 1.6 °C in January to a peak of 23.3 °C in July, before declining toward winter. The p90 threshold follows a similar pattern, rising from 6.9 °C in January to approximately 28.4 °C in July. The consistent separation between monthly mean Tmax and p90 thresholds reflects the upper-tail behaviour of the temperature distribution and highlights the strong seasonality of extreme heat conditions. This seasonal structure underpins the use of month-specific percentile thresholds to define extreme hot days in a physically consistent manner.

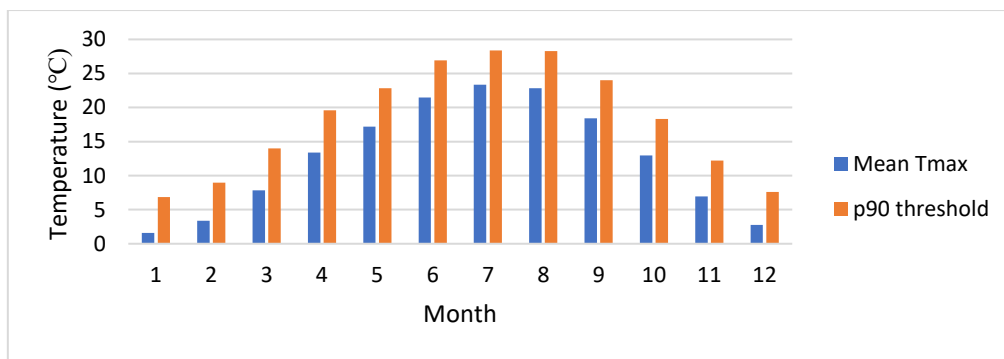


Fig. 2. Monthly mean Tmax values and p90 thresholds.

Fig. 3 shows the mean extreme days and monthly fraction of extreme hot days over Czechia. The mean annual number of extreme hot days over the Czech Republic shows clear interannual variability during 2005-2025, together with a clear increase in recent years. Early values generally remained below 40 days per year, whereas markedly higher frequencies occurred after 2014, with apparent peaks in 2015 (57.3 days) and 2018 (63.3 days). Elevated counts persist in the most recent years, including 2023 and 2024 (>54 days), indicating a sustained intensification of extreme heat conditions.

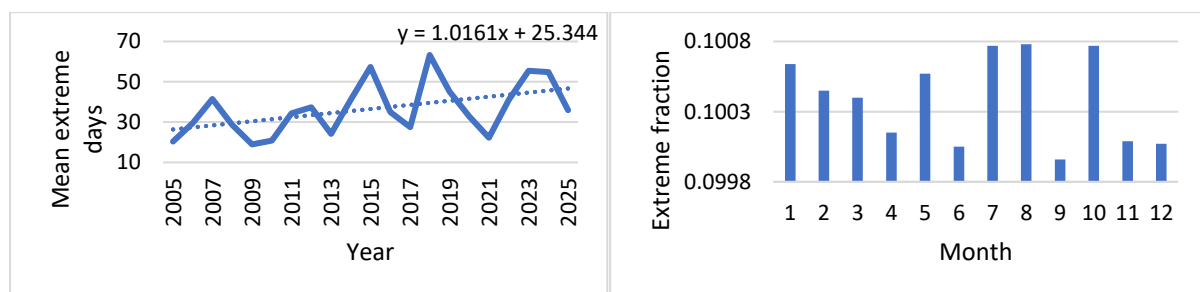


Fig. 3. Mean extreme days over Czechia (left), monthly fraction of extreme hot days (right).

Fig. 3 also shows a remarkably uniform distribution of extreme-day fractions throughout the year, with values consistently close to 0.10. Monthly fractions range from 0.09996 in

September to approximately 0.10078 in August, with slightly higher values during summer and early autumn months (July-October). This near-constant fraction confirms the effectiveness of the percentile-based definition in normalizing seasonal variability, ensuring that extreme events are identified relative to each month's climatological temperature distribution rather than absolute thresholds.

Fig. 4 presents the relationship between elevation and the mean annual number of extreme hot days and the mean SHAP-based normalized predictor importance values derived from the RF model. The scatter distribution shown in the figure indicates that mean extreme-day frequency exhibits relatively limited sensitivity to elevation over most of the altitude range. Although extreme hot days occur across a broad range of elevations, a weak negative trend is evident, suggesting a slight decrease in the mean number of extreme days with increasing altitude. This muted elevational dependence implies that, while topography modulates local thermal conditions, large-scale atmospheric forcing and land-atmosphere interactions dominate the occurrence of extreme heat events within the region. As also seen in Fig. 4, LST emerges as the most influential predictor, accounting for 23.4% of the total importance, followed by surface soil moisture (18.4%) and wind speed (17.2%). Vegetation conditions, represented by NDVI, contributed 15.6%, while precipitation accounts for 13.6%. Surface albedo exhibited the lowest relative importance (11.8%), indicating a more indirect role in extreme heat occurrence. The importance ranking highlights the dominant influence of land-atmosphere interactions and atmospheric mixing processes on extreme temperature events over the Czechia. The RF model achieved an overall classification accuracy of 0.81 in distinguishing extreme from non-extreme hot days. Precision, recall and F1-score were calculated as 0.79, 0.85 and 0.82, respectively.

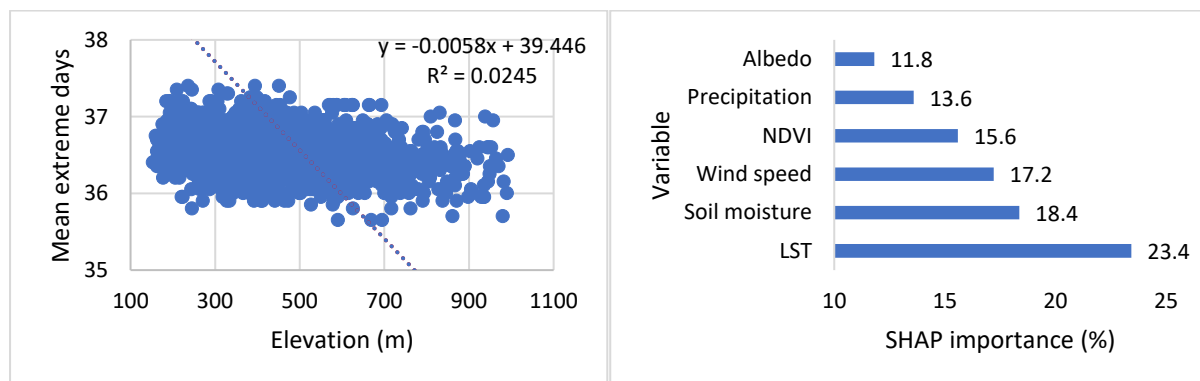


Fig. 4. Elevation vs. mean annual number of extreme hot days (left), mean SHAP-based normalized predictor importance values (right).

Fig. 5 shows scatter plots of daily Tmax against selected predictors used in the RF model, with linear trend lines providing a physical interpretation of the relationships learned by the model. The scatter plots reveal contrasting first-order relationships between Tmax and individual predictors. A positive slope is observed for LST, indicating a strong coupling between surface thermal conditions and near-surface air temperature, as expected from surface-atmosphere energy exchange processes. In contrast, the positive slopes identified for NDVI and precipitation suggest that higher vegetation activity and recent rainfall tend to co-occur with higher Tmax at the daily scale, likely reflecting their joint seasonality and the dominance of warm-season conditions rather than a direct causal warming effect. Conversely, negative slopes for soil moisture, wind speed and albedo indicate that enhanced evaporative cooling, increased turbulent mixing and higher surface reflectivity are generally associated with reduced maximum temperatures.

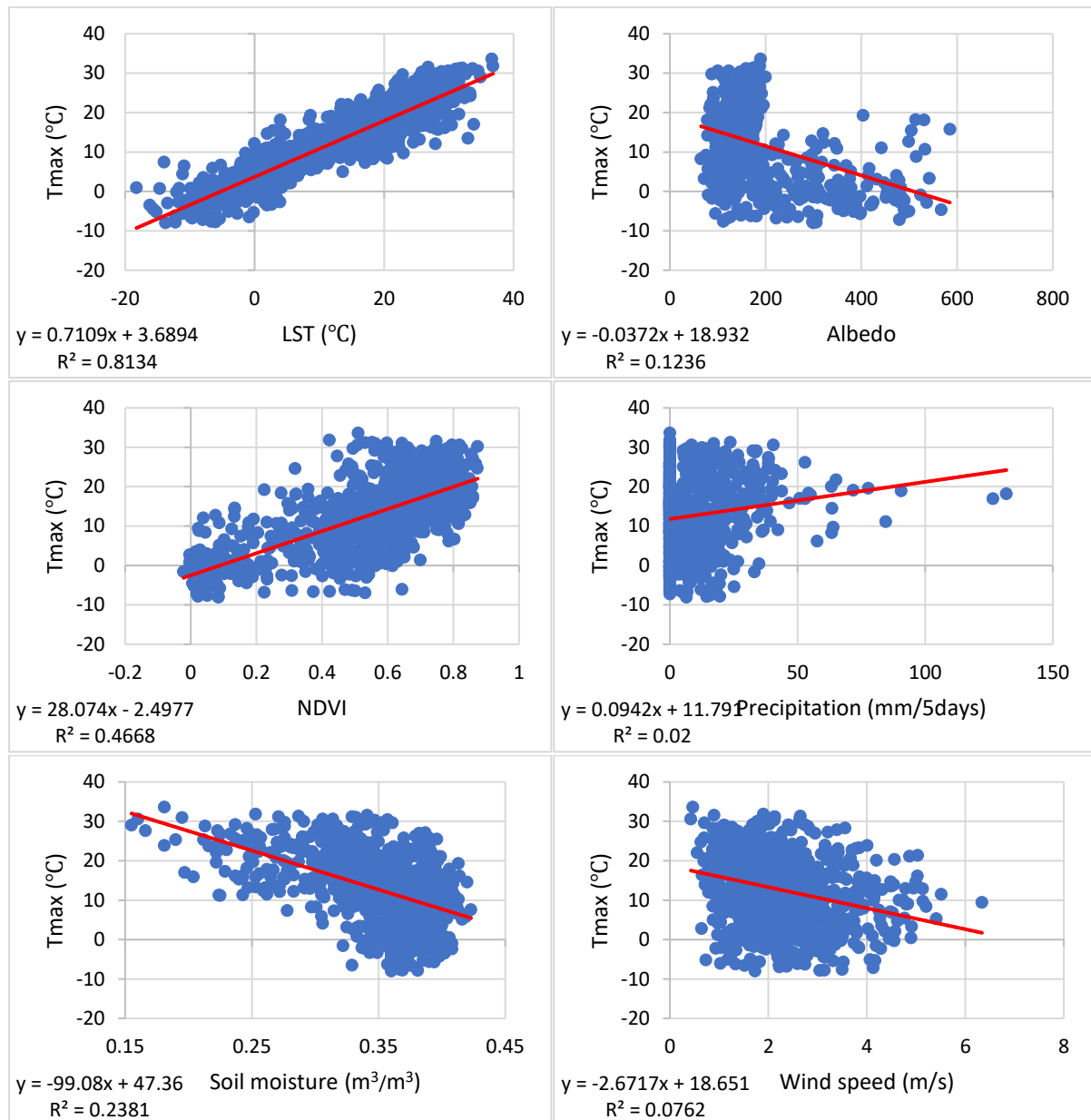


Fig. 5. Scatter plots of daily Tmax against selected predictors

It should be noted that the percentile thresholds used to define extreme hot days were calculated using the full study period. In a warming climate, temperature distributions are not strictly stationary, and percentile-based thresholds derived from long-term records may partially reflect recent warming trends. Consequently, the increasing frequency of extreme hot days observed in the later years of the study period may be influenced both by actual warming and by the shifting temperature distribution. Nevertheless, percentile-based definitions remain widely used in climate extreme studies because they provide a seasonally adaptive framework for identifying anomalies relative to the climatological temperature distribution.

It should also be noted that the reported overall accuracy reflects model performance under balanced sampling conditions. Because extreme hot days represent a relatively small fraction of all observations in reality, the reported accuracy should be interpreted as an indicator of classification skill rather than as a direct estimate of operational predictive performance.

Several predictor variables used in the analysis, including NDVI precipitation and LST, exhibit apparent seasonal cycles. Consequently, part of the relationships observed in the scatterplots may reflect shared seasonal variability rather than direct causal mechanisms. In this study, the scatterplots are intended to provide a qualitative visualization of first-order associations rather than a causal interpretation. Future research could explore anomaly-based (deseasonalized) analyses to better isolate the influence of land-atmosphere interactions independent of seasonal variability.

4. Conclusion

This study provides a comprehensive assessment of extreme high-temperature events over Czechia by integrating multi-source reanalysis and satellite observations with an explainable ML framework. Using a month-specific percentile-based definition, extreme hot days were consistently identified across seasons, enabling robust spatial and temporal analyses. The RF model demonstrated strong predictive skill (overall accuracy of 0.81) and revealed that LST, soil moisture and wind speed are the dominant drivers of extreme heat occurrence, underscoring the critical role of land-atmosphere interactions and atmospheric mixing processes. The observed increase in the frequency of extreme hot days in recent years highlights the growing intensity of heat-related climate risks in Central Europe.

References

- Fastl, C., Arnberger, A., Gallistl, V., Stein, V. K., Dorner, T. E. (2024) Heat vulnerability: health impacts of heat on older people in urban and rural areas in Europe. *Wiener Klinische Wochenschrift*, 136 (17), 507-514, doi: 10.1007/s00508-024-02419-0.
- Kirchengast, G., Haas, S., Fuchsberger, J. (2025) A new class of climate hazard metrics and demonstration based on tracking extreme heat amplification over Europe, EGU General Assembly 2025, Vienna, Austria, 27 Apr-2 May 2025, EGU25-10409, doi: 10.5194/egusphere-egu25-10409.
- Lorenz, R., Stalhandske, Z., Fischer, E. M. (2019) Detection of a climate change signal in extreme heat, heat stress, and cold in Europe from observations. *Geophysical Research Letters*, 46 (14), 8363-8374. doi: 10.1029/2019GL082062.
- Lundberg, S. M., Lee, S. I. (2017) A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Malcheva, K., Bocheva, L., Chervenkov, H. (2022) Spatio-Temporal Variation of Extreme Heat Events in Southeastern Europe. *Atmosphere*, 13 (8), 1186, doi: 10.3390/atmos13081186.
- Suarez-Gutierrez, L., Müller, W. A., Marotzke, J. (2023) Extreme heat and drought typical of an end-of-century climate could occur over Europe soon and repeatedly. *Communications Earth & Environment*, 4 (1), 415, doi: 10.1038/s43247-023-01075-y.
- URL-1, https://en.wikipedia.org/wiki/Czech_Republic, Access date: January 20th, 2026.
- URL-2, https://bluegreenatlas.com/climate/czechia_climate.html, Access date: January 20th, 2026.
- Vautard, R., et al. (2023) Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends. *Nature Communications*, 14 (1), 6803, doi: 10.1038/s41467-023-42143-3.
- Weilnhammer, V., Schmid, J., Mittermeier, I., Schreiber, F., Jiang, L., Pastuhovic, V., Herr, C., Heinze, S. (2021) Extreme weather events in europe and their health consequences-A systematic review. *International Journal of Hygiene and Environmental Health*, 233, 113688, doi: 10.1016/j.ijheh.2021.113688.