International Journal of Digital Earth

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjde20

Relationships between geo-spatial features and COVID-19 hospitalisations revealed by machine learning models and SHAP values

Lixia Chu, Jeroen Nelen, Alessandro Crivellari, Dainius Masiliūnas, Carola Hein & Christoph Lofi

To cite this article: Lixia Chu, Jeroen Nelen, Alessandro Crivellari, Dainius Masiliūnas, Carola Hein & Christoph Lofi (2024) Relationships between geo-spatial features and COVID-19 hospitalisations revealed by machine learning models and SHAP values, International Journal of Digital Earth, 17:1, 2358851, DOI: 10.1080/17538947.2024.2358851

To link to this article: https://doi.org/10.1080/17538947.2024.2358851

Published online: 27 May 2024.

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2024, VOL. 17, NO. 1, 2358851
https://doi.org/10.1080/17538947.2024.2358851

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Relationships between geo-spatial features and COVID-19 hospitalisations revealed by machine learning models and SHAP values

Lixia Chu[a,b], Jeroen Nelen[a], Alessandro Crivellari[c], Dainius Masiliūnas[d], Carola Hein[e] and Christoph Lofi[a]

[a]Computer Science, Delft University of Technology (TU Delft), Delft, the Netherlands; [b]Environmental Technology, Wageningen University & Research, Wageningen, the Netherlands; [c]Department of Geography, National Taiwan University, Taipei, Taiwan; [d]Laboratory of Geo-information Science and Remote Sensing, Wageningen University & Research, Wageningen, the Netherlands; [e]Architecture, Delft University of Technology (TU Delft), Delft, the Netherlands

ABSTRACT

Uncovering relationships between geospatial features and COVID-19 features is a comprehensive, confounding, cross-disciplinary and challenging topic, as the spread and effects of COVID-19 are related to many aspects of our lives, including socio-economic, cultural, and environmental features. Our research aims to provide an innovative data-driven method to uncover the relationships between the heterogeneous and cross-disciplinary geospatial features with COVID-19 features at the municipality scale in Germany. We exploit these relationships using supervised machine learning, explainable AI and spatial analysis in Germany from March 2020 to October 2021. First, we integrated multi-source data including social data, economic data, cultural data, air pollution data and COVID-19 features data into one spatiotemporally harmonised dataset. Second, we trained three machine learning models (a Support Vector Regressor, a Random Forest, and a Light Gradient Boosting Machine) on the integrated dataset to learn the relationships between the spatial features and the COVID-19 features. Third, we used Shapley Additive exPlanations (SHAP) to rank the relevance of each feature. After that, we illustrated the results by the visualised spatial differences within municipalities. The output delivers key information regarding the Covid hospitalisation rate with the control of NO_2 concentration and education level in Germany with transferable methods.

ARTICLE HISTORY
Received 12 February 2024
Accepted 17 May 2024

KEYWORDS
COVID-19; multi-source data; data integration; machine learning; explainable AI; SHAP value

1. Background

The outbreak of the pathogenic agent of coronavirus disease in 2019 (COVID-19) has resulted in numerous fatalities globally, prompting researchers to investigate the correlated features, as well as its effects on the environment, health, and economy (Bowe et al. 2021; Cole, Ozgen, and Strobl 2020; Dutheil, Baker, and Navel 2020; Gautam and Hens 2020; Gomathy et al. 2021; Maier et al. 2022; Muhammad, Long, and Salman 2020; Neidell 2004; Parida et al. 2021; Stojkoski et al. 2020; Tosepu et al. 2020; Urrutia-Pereira, Mello-da-Silva, and Solé 2020; Watts and Kommenda 2020;

CONTACT Lixia Chu lixia.chu@wur.nl

Xiong et al. 2020). Social scientists have conducted several studies on factors that contributed to the spread of COVID-19, death or hospitalisation rates of COVID-19, etc (Ehlert 2021; Gautam 2020; Gautam and Hens 2020; Ghahremanloo et al. 2021; Kuebart and Stabler 2020; Stojkoski et al. 2020). Some of those studies found that both weather and air pollution are correlated to COVID-19 features (Cole, Ozgen, and Strobl 2020; Godawska 2021; Li, Li, et al. 2020; Li, Xu, et al. 2020; Manne and Kantheti 2020; Naqvi et al. 2021; Ogen 2020; Pascal et al. 2013; Tosepu et al. 2020; Urrutia-Pereira, Mello-da-Silva, and Solé 2020; Vîrghileanu et al. 2020; Watts and Kommenda 2020). However, there are not enough studies that incorporate a comprehensive set of multifaceted geospatial features and automatically model the relationships throughout an entire country. The integration of high-dimensional data from diverse disciplines poses a challenge, given the varied types, temporal dynamics, and spatial scales inherent in data from each domain. To address this challenge, we collected features from socio-economic, cultural and air pollution domains and established a new harmonised dataset.

Another challenge is to model the relationships from the integrated data automatically. There have been several studies that examined the determinants of COVID by analysing a diverse set of socio-economic characteristics using Bayesian models during the first wave of the COVID-19 pandemic (Gautam and Hens 2020; Stojkoski et al. 2020; Xiong et al. 2020). There have been some studies on methods that can automatically model the relationships with limited intervention from humans (Choudary et al. 2021; Kavouras et al. 2022; Snider et al. 2021). Machine Learning and explainable AI are capable of directly training models with limited interaction from humans and exploring the hidden relationships for confounding questions. Explainable Artificial Intelligence (explainable AI, XAI) focuses on how we can interpret and explain black-box AI models, especially Deep Learning methods with complex structures and a large number of parameters (Linardatos, Papastefanopoulos, and Kotsiantis 2020; Temenos et al. 2022; Xu et al. 2019). Shapley Additive exPlanations (SHAP) by Lundberg and Lee (Lundberg and Lee 2017) is a method to explain individual predictions, which works as both a local and global explainer. Originally it was used to calculate the contributions of features in polynomial time for game theory applications (Lundberg et al. 2020). There are also some studies that performed geographic regression to analyse the spatial comparative effects of different geographic regions (Cao, Rui, and Liang 2018; Shao, Xu, and Wu 2021). There remains a scarcity of research dedicated to uncovering the global relationships among heterogeneous and cross-disciplinary features between spatiotemporal features and COVID-19 features at the municipal level for all regions. In this research, we provide a data-driven method by using three machine learning models and XAI SHAP, as well as spatiotemporal analysis, to uncover the geospatial features that are related to the COVID-19 hospitalisation rate at the municipality level.

2. Data

In this study, we aim to alleviate the bias in feature selection arising from human choice to provide an innovative data-driven method to uncover the relationships between the heterogeneous and cross-disciplinary geospatial features with COVID-19 features. To achieve this, we created a new dataset that incorporates a mix of commonly and rarely used features from existing studies. We incorporated features that exhibit correlations with COVID features, as identified from the previous studies (Cole, Ozgen, and Strobl 2020; Dutheil, Baker, and Navel 2020; Gautam and Hens 2020; Kuebart and Stabler 2020; Li, Xu et al. 2020; Linardatos, Papastefanopoulos, and Kotsiantis 2020; Maier et al. 2022; Neidell 2004; Ogen 2020; Parida et al. 2021; Shao, Xu, and Wu 2021; Stojkoski et al. 2020; Tosepu et al. 2020; Urrutia-Pereira, Mello-da-Silva, and Solé 2020; Xiong et al. 2020). We compiled and pre-processed a comprehensive array of features from various publicly available data sources in Germany (Appendix) for our harmonised dataset. These data sources encompass three categories of open-source data: statistical data including social-cultural and economic features, air pollution data and COVID-19 data.

We collected demographic data, social data, cultural data, and 50 core indicators for the Sustainable Development Goals from INKAR (Indikatoren und Karten zur Raum – und Stadtentwicklung, https://www.inkar.de/). The other social features were obtained from the Regionalatlas DE (https://regionalatlas.statistikportal.de/) and the ZENSUS Datenbank (https://ergebnisse2011.zensus2022.de/datenbank/online/). We queried and processed COVID-19 data provided by the Robert Koch Institute (https://www.corona-datenplattform.de/dataset/). For each municipality, we used the mortality rate, the infection rate, the intensive care rate, and the number of cases per 100.000 inhabitants per day as COVID-19 features.

In the previous studies which have analysed how long-term exposure to air pollution may have an impact on the death rate (Cole, Ozgen, and Strobl 2020; Ogen 2020; Strak et al. 2017; Urrutia-Pereira, Mello-da-Silva, and Solé 2020), NO_2 and respirable articulate matter turned out to be highly related to the pandemic variables (Ogen 2020). In this research, NO_2 was selected as the representative air pollutant, and we obtained NO_2 data from the offline level-3 high-resolution daily imagery of NO_2 concentrations acquired from the Sentinel-5P TROPOMI sensor. The dataset was accessed using the catalogue of Google Earth Engine (https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S5P_OFFL_L3_NO2). For the data pre-processing, we replaced the negative values of NO_2 with '0', as these values mainly occur in areas that have clean air or very low NO_2 emissions.

Ultimately, a new spatially and temporally harmonised dataset was created. The dataset encompassed the period from March 1, 2020, to October 30, 2021, spanning a total of 87 weeks. The dataset was structured with a weekly temporal resolution and a municipal-level spatial resolution.

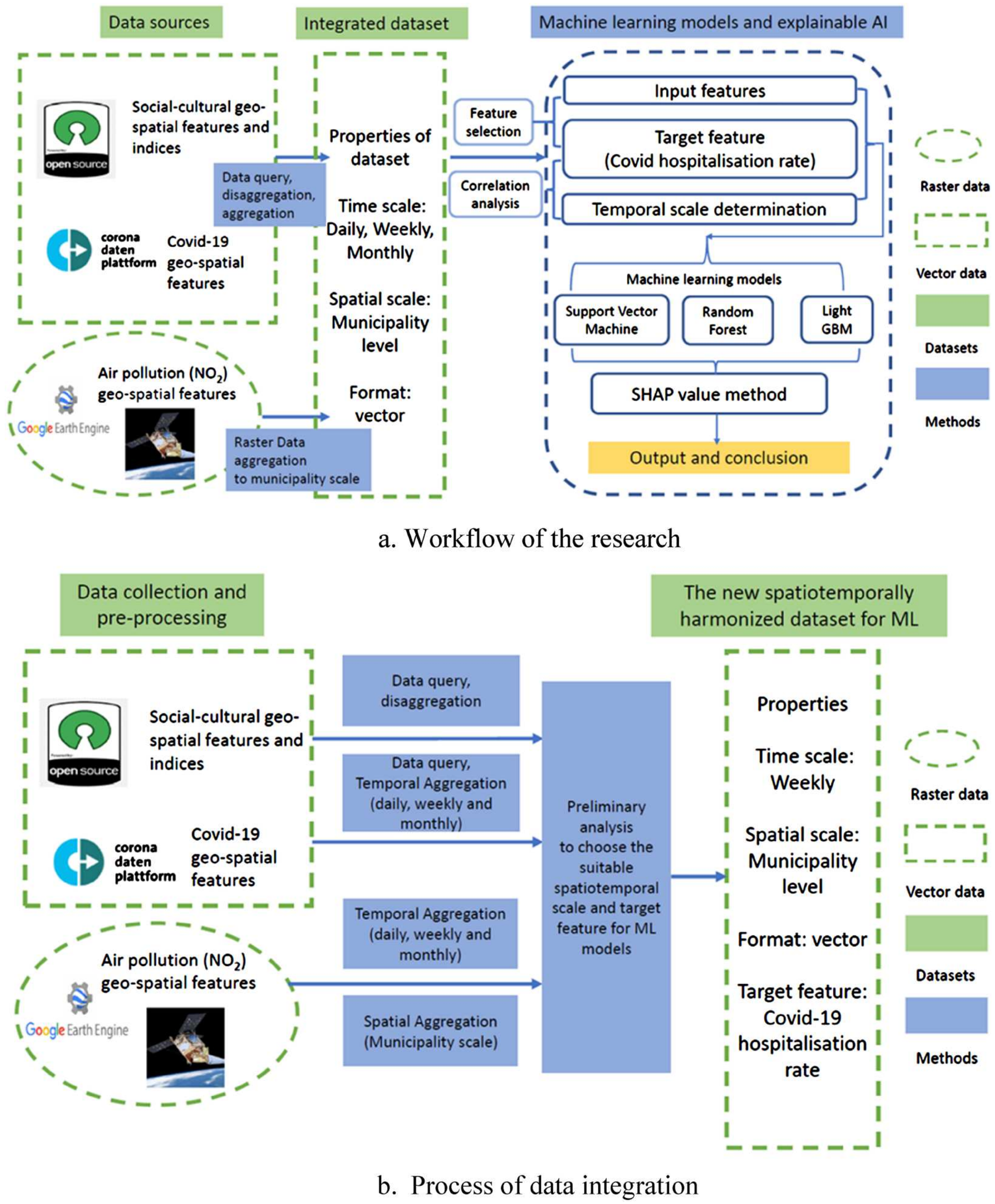
3. Methodology

Our study included four methodological steps: data integration, machine learning model training, explainable AI, and spatiotemporal visualisation. Before running machine learning models, a new spatiotemporally harmonised dataset was needed. We aggregated data into municipality scale as the municipality unit serves as the primary unit for both COVID-19 features and socioeconomic data. Conducting Pearson correlation analysis using daily, weekly, and monthly data individually, we discovered that weekly data aggregation offers the optimal balance for highlighting potential relationships, surpassing other time scales. Additionally, we observed that the hospitalisation rate exhibits a higher correlation with geospatial features compared to the other three COVID-related features, so it was selected as the target label in our research. Employing three different ML models – Support Vector Regressor (SVR), Random Forest (RF), and Light Gradient Boosting Machine (LGBM) and utilising SHAP values, we ranked the importance of correlated geospatial features with the Covid hospitalisation rate. Subsequently, we visualised the high-ranking features contributing to COVID-19 hospitalisation rates, elucidating the interpretation of correlations to the best of our knowledge. The workflow of this research can be seen in Figure 1(a) and data integration is illustrated in Figure 1(b).

3.1. Data integration

The initial step before applying the machine learning models is to integrate the heterogeneous datasets into a new spatiotemporally harmonised dataset. The new dataset operates at the municipality level and weekly scale, with the hospitalisation rate as the target feature.

Concerning the spatial scale, we aggregated data at the municipality scale to compare spatial variations among municipalities and draw overall conclusions from the time series changes observed across all municipalities. In terms of the temporal scale, the source data had been acquired at daily, weekly, monthly, and annual intervals. We plotted NO_2 concentration alongside COVID-19 features and conducted Pearson correlation analysis separately for daily, weekly, and monthly data. The comparison of correlation results across these temporal scales revealed that the weekly scale offers the most advantageous balance. It retains granularity in data while mitigating the impact of noise and addressing time lag issues arising from COVID-19 data reporting, vaccinations, or the COVID-19 incubation period. For the selection of target feature, we observed a stronger correlation


Data sources
Integrated dataset
Machine learning models and explainable AI
Social-cultural geo-spatial features and indices
open source
Data query, disaggregation, aggregation
corona daten plattform
Covid-19 geo-spatial features
Air pollution (NO_2) geo-spatial features
Google Earth Engine
Raster Data aggregation to municipality scale
Properties of dataset
Time scale: Daily, Weekly, Monthly
Spatial scale: Municipality level
Format: vector
Feature selection
Correlation analysis
Input features
Target feature (Covid hospitalisation rate)
Temporal scale determination
Machine learning models
Support Vector Machine
Random Forest
Light GBM
SHAP value method
Output and conclusion
Raster data
Vector data
Datasets
Methods
Data collection and pre-processing
The new spatiotemporally harmonized dataset for ML
Social-cultural geo-spatial features and indices
open source
corona daten plattform
Covid-19 geo-spatial features
Air pollution (NO_2) geo-spatial features
Google Earth Engine
Data query, disaggregation
Data query, Temporal Aggregation (daily, weekly and monthly)
Temporal Aggregation (daily, weekly and monthly)
Spatial Aggregation (Municipality scale)
Preliminary analysis to choose the suitable spatiotemporal scale and target feature for ML models
Properties
Time scale: Weekly
Spatial scale: Municipality level
Format: vector
Target feature: Covid-19 hospitalisation rate
Raster data
Vector data
Datasets
Methods

a. Workflow of the research

b. Process of data integration

Figure 1. (a) Workflow of the research; (b) Process of data integration.

between the hospitalisation rate and geospatial features compared to the other three COVID-related factors. We did not include data plots or correlation analyses as the primary research outputs.

Specifically, we aggregated daily COVID-19 data into weekly intervals by calculating the seven-day average. For the NO_2 data, we transformed the satellite images from a raster type into a vector type to obtain usable tabular data, which was then aggregated into weekly intervals using the average value over each week. In terms of the spatial dimension, NO_2 data was aggregated using padding interpolation based on each municipality in Germany. This method employs the nearest known value in time for the specific municipality. For the social data, we derived values for each municipality based on the national data, considering the areas and population of each municipality. We employed a MinMaxScaler for feature scaling to standardise all values to a consistent scale. Given that social data are updated annually, the values remained constant within each year.

Before employing machine learning models on the harmonised dataset, we conducted feature selection to eliminate co-dependent features and enhance model interpretability. Our process began with utilising the Pearson correlation coefficient to identify and remove highly correlated features. Additionally, we leveraged the F-Statistic to select the 10 best features, which involved calculating the cross-correlation between each feature and the target prediction feature. These correlations were transformed into an F-score. The F-score is a well-established benchmark for feature selection that highlights the important features, which is what interests us most in this experiment. However, other selection methods (Alahmadi et al. 2023; Nuutinen et al. 2017; Prinzi et al. 2023), such as the sequential forward selection, would also be a good feature selection method, which could further improve the results.

In our case, we have heuristically chosen to work with 10 features, which we've found to be sufficient for effectively training a Machine Learning algorithm while also yielding noteworthy SHAP values. The selected 10 best features were 'Labor Force at a Low Education Level', 'NO_2 concentration', 'Employment', 'Branches Agriculture', 'Low Education Level', 'Vocational Education', 'Population Aged 15–24 years', 'Primary Sector Agriculture', 'Cases', 'Branches Industry'. The definition and the calculation algorithm of these 10 features, along with all other features, are listed in Appendix.

To facilitate the training of the machine learning models, the dataset was split into training and test sets, allocating 70% for training and 30% for testing purposes. This partitioning process involved a random division of municipalities into two distinct groups. Each municipality included its complete time series data for the research period, yet the municipalities differed between the two sets. Among the 400 municipalities in total, 280 were designated for training, leaving the remaining 120 for testing.

A hyperparameter search was performed for each model's training. The most influential hyperparameters were chosen for each model. Subsequently, a grid search, which exhaustively tests all combinations within a range for each parameter, was employed to optimise each model's performance. Five-fold cross-validation was used in this hyperparameter search.

3.2. Learning the relationships using three machine learning (ML) models

In this study, we employed three different ML models: a Support Vector Regressor (SVR), a Random Forest (RF) and a Light Gradient Boosting Machine (LGBM). These ML models, trained on historical data, aim to uncover the intricate relationships among various features. Machine learning models are typically used to create predictions or to classify data into categories, but in this paper they had a different purpose. We sought to elucidate the relationships between these features to better understand their influence on the hospitalisation of COVID patients. We achieved this by utilising explaining methods for machine learning models.

All three models are supervised learning methods, involving input features and a pre-determined target label, which is the Hospitalisation Rate in our case. We provide a brief overview of the three models we used:

3.2.1. Support vector regressor

SVR is an extension of the Support Vector Machine (SVM) and employs a kernel to classify support vectors, ultimately creating a hyperplane that captures the nonlinear patterns in the data. This makes SVR a suitable choice for high-dimensional data, aligning with the characteristics of our case study. Where SVM produces a binary output, SVR excels at handling regression problems with real-valued scale estimation functions (Zhang and O'Donnell 2020).

3.2.2. Random forest regressor

RF is a regression technique that leverages multiple Decision Tree methods to predict values. During training, RF constructs many decision trees and then ensembles the results back together (Varghese et al. 2022). It also uses bagging as a bootstrap method and incorporates a feature importance learner, making it particularly useful for high-dimensional data (Rodriguez-Galiano et al. 2015).

3.2.3. Light gradient boosting machine

LightGBM is an implementation of a Gradient Boosting Decision Tree (GBDT) (Ke et al. 2017). Gradient boosting methods are constructing tree models where trees are added one at a time to the ensemble. LGBM extends this ensemble method by adding a form of feature selection and focusing on larger gradients. It is proven to be a model with high prediction accuracy, rapid computational speed and excellent capability of handling large-scale, high dimensional and imbalanced data and is therefore a valuable addition to this study (Al Daoud 2019; Alzamzami, Hoda, and Saddik 2020).

3.3. Identifying contributions and ranking relevance using SHAP values

Interpreting the output of the prediction model is a crucial aspect of this study as it allows us to uncover the relationships we seek. Explainable AI and interpretability provide insights on model enhancement, build trust among users and expand the models' applicability across various domains, including policy development.

SHAP is a unified method to explain individual predictions concerning the target feature by computing the contribution of each feature to the prediction. The method was developed by Lundberg and Lee and feature contributions are calculated using the weighted average of all possible differences (Lundberg and Lee 2017). A SHAP value is a feature importance metric that can be used with any machine learning and AI models to calculate the impact of each input feature by training the model with and without each feature. SHAP was originally used in game theory (Futagami et al. 2021) to compute the 'payout' (= prediction of target features) among 'players' (=features). In this research, our method was to predict the target feature, namely the COVID-19 hospitalisation rate, using selected socio-economic and air pollution features. Then we assessed the correlations between COVID-19 hospitalisation rate and these features by employing SHAP values to rank their importance.

4. Results

4.1. Machine learning

The R^2 and Mean Absolute Error (MAE) were used to calculate the performance of the model. As previously mentioned, the parameters were determined through cross-validation and a hyperparameter search. Table 1 presents the results with the corresponding hyperparameters employed in the analysis.

4.2. Resulting SHAP values

In Figures 2, 3 and 4, the calculated SHAP values can be seen for the three models that were used. The features are sorted on importance from top to bottom, which is the sum of absolute SHAP values (leverage). The horizontal axis of the diagram represents the SHAP value, namely, the leverage each feature value has on the target variable. The colour represents the normalised magnitude of each feature value, which becomes red (high value) as the feature increases and blue (low value) as the value of the feature decreases.

Table 1. The R^2, MAE & Hyperparameters from the three models on the test set.

Model	R^2	*MAE*	Hyperparameters

SVR	0.6606	0.0619	kernel = Radial Basis Function
			gamma = 0.1
			Regularisation parameter (C) = 1000
			epsilon = 0.1

RF	0.7621	0.0429	Number of trees = 200, Minimum samples before splitting = 2 Minimum samples at each leaf = 2 Number of features to consider at every split = 7
LightGBM	0.7642	0.0427	Boosting Type = Gradient Boosting Decision Tree Number of leaves = 20 Learning rate = 0.1 Number of trees = 100 Minimal data in leaves = 100

Compared to RF and LightGBM, the SVR model had relatively lower performance scores as shown in Table 1 and the feature importance of RF and LightGBM reveals a strong agreement with each other, as depicted in Figure 3 and Figure 4. Notably, the top four most important features identified by the two high-performance models (RF and LightGBM) remain consistent: the 'number of COVID cases', the 'NO_2 concentration', 'Low education', and 'Labor force at a low education level'.

As shown in Figures 2, 3 and 4, all three models show that the feature of COVID cases is the most important, as the more cases of COVID-19, the higher the number of patients who need care in hospitals. While this result is not surprising, it shows that an automated approach does pick up these relationships effectively.

Also, a positive relationship can be found between the hospitalisation rate and NO_2 concentration for both two high-performance models (RF and LGBM). The result shows that the weekly NO_2 concentration level is highly correlated to the rate of COVID patients who need care in hospitals. It implies that the air pollutant NO_2 concentration in the air may worsen the symptoms of COVID patients. We should note that this can hint at a correlation, not causation per se.

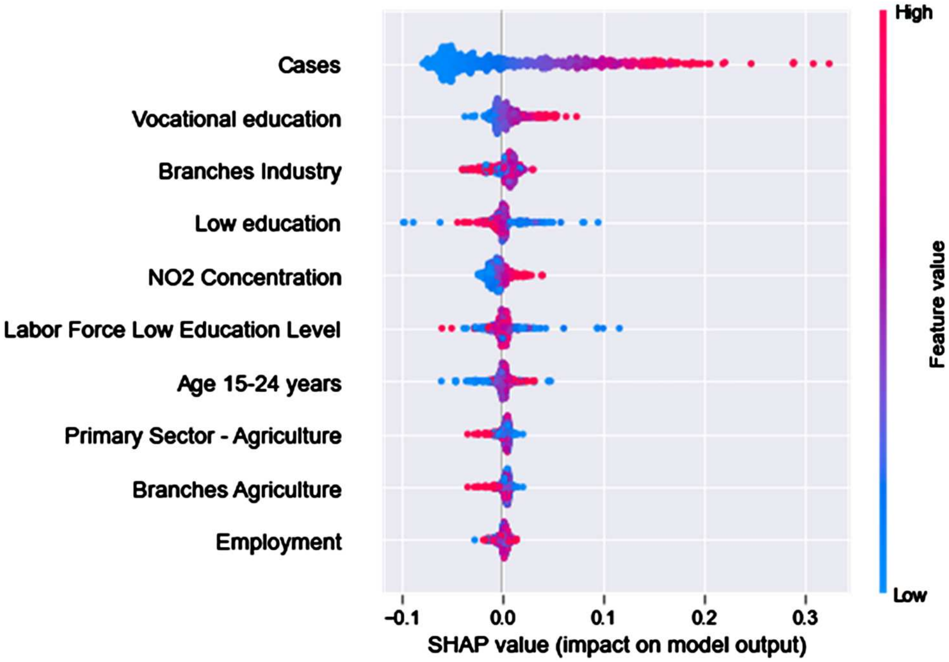


Figure 2. SHAP values SVR.

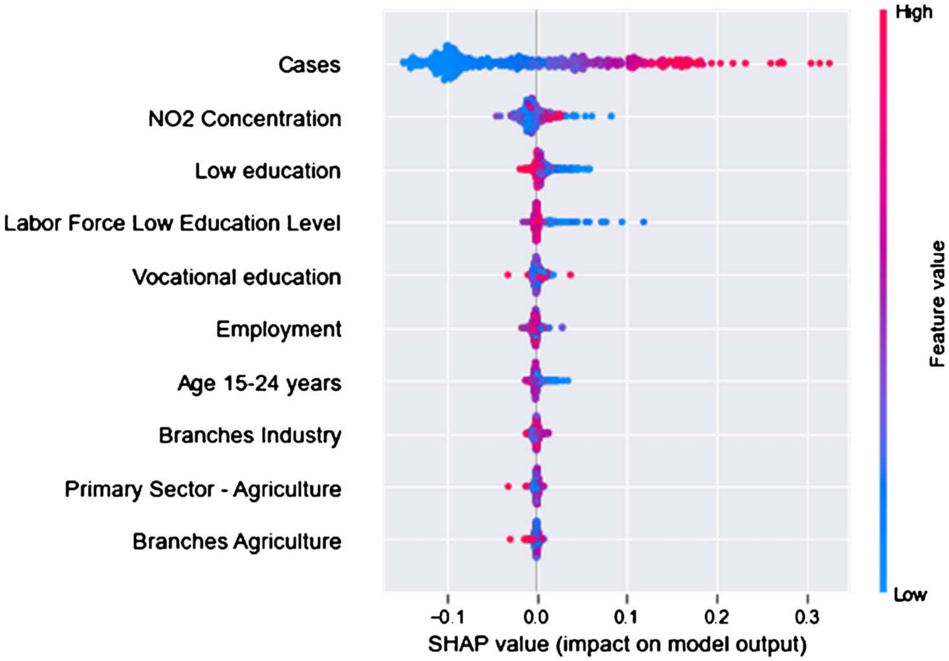


Figure 3. Shap values RF.

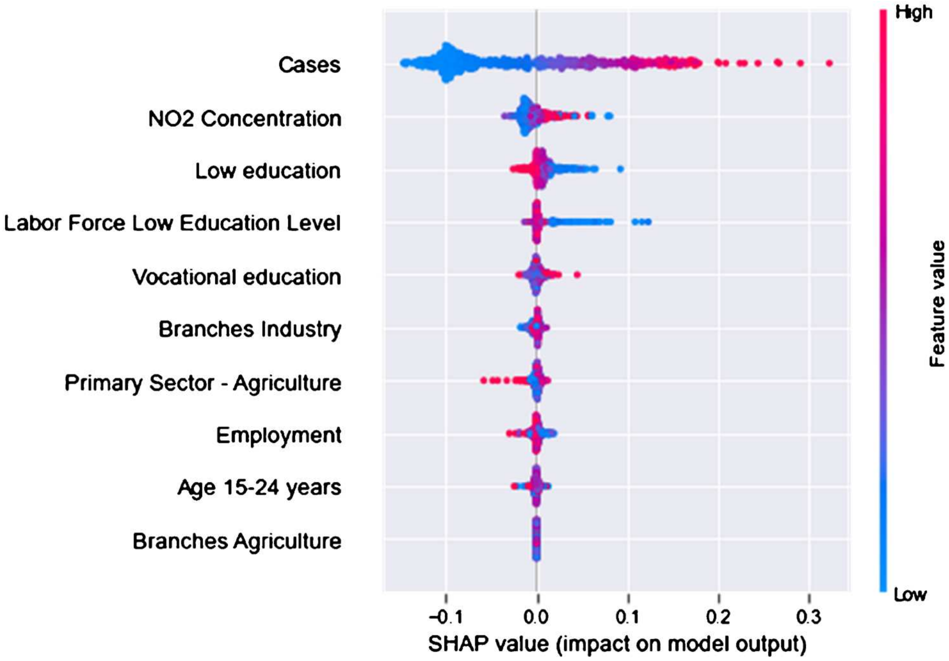


Figure 4. Shap values LightGBM.

Both ‘Labour Force at a Low Education Level’ and ‘Low Education’ demonstrate negative correlations with the COVID hospitalisation rate, given their similarity in nature. ‘Labour Force at a Low Education Level’ represents the rate of labour force without a qualification per 100 employees, while the latter feature denotes the proportion of school leavers without a certificate per 100 inhabitants. This intriguing observation warrants future investigation to elucidate the influences of education attainment on the COVID hospitalisation rate. We performed a spatiotemporal analysis to uncover additional insights into the underlying factors driving these correlations.

4.3. Spatiotemporal visualisation

Based on the result seen in Figures 2, 3 and 4 from all three models, except for the cases, among the top 6 features with high contribution to the COVID-19 hospitalisation rate are 'NO$_2$ concentration', 'low education', 'Labor force at a low education level', 'vocational education' and 'population of 15–24 years old'. As both 'low education' and 'Labor force at a low education level' are closely related, we chose to only visualise the spatial distribution of the simpler education-related feature 'low education'. Altogether, we visualised the aforementioned four high-ranking features.

To illustrate the spatiotemporal contribution to the hospitalisation rate, we analyse and map the dynamics of spatiotemporal variations of NO$_2$. It is shown (temporal changes in Figures 5 and 6) that the trend of NO$_2$ temporal density is highly linked with the trend of COVID-19 hospitalisation rate, which shows that there are high hospitalisation records among the COVID cases when the NO$_2$ concentration is higher than normal, and vice versa. The peak and valley times for both NO$_2$ and hospitalisation rate are similar with each other (as listed in Table 2). Although there is a time lag of 2 weeks, the low NO$_2$ density in the air is followed by the low hospitalisation rate and vice versa. The time lag could be influenced by the vaccination, the incubation period and other features.

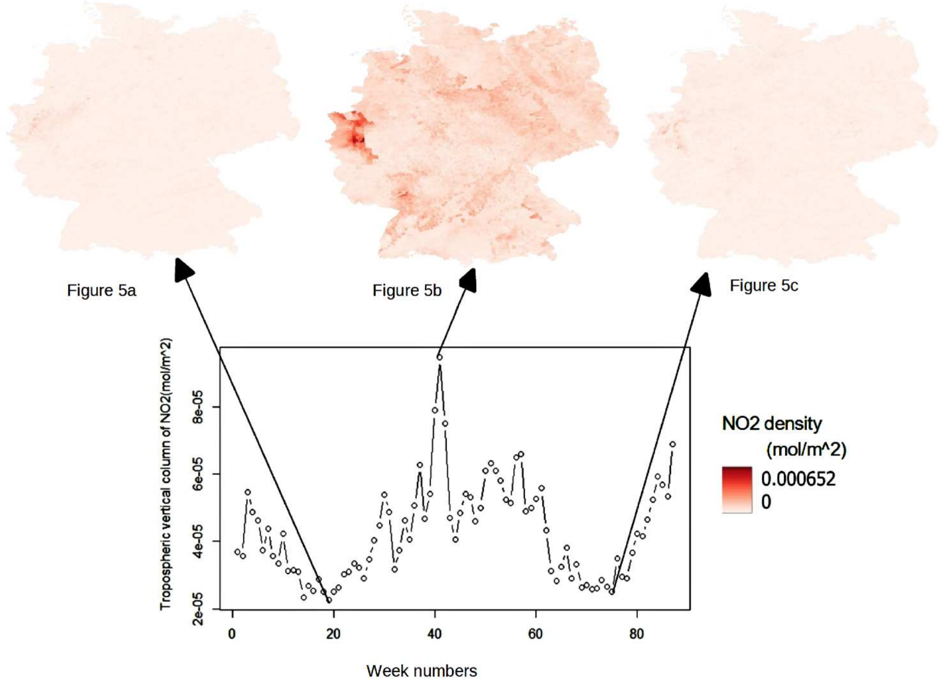


Figure 5. Spatiotemporal changes of NO$_2$ concentration in Germany from 2020.03.01 to 2021.10.31. (a) NO$_2$ concentration at Week 19 (2020.07.05–07.11); (b) NO$_2$ concentration at Week 41 (2020.12.06–12.12); (c) NO$_2$ concentration at Week 75 (2021.08.01–08.07). The subfigures at the top correspond to the NO$_2$ concentration variation in space at the time points indicated by the arrows.

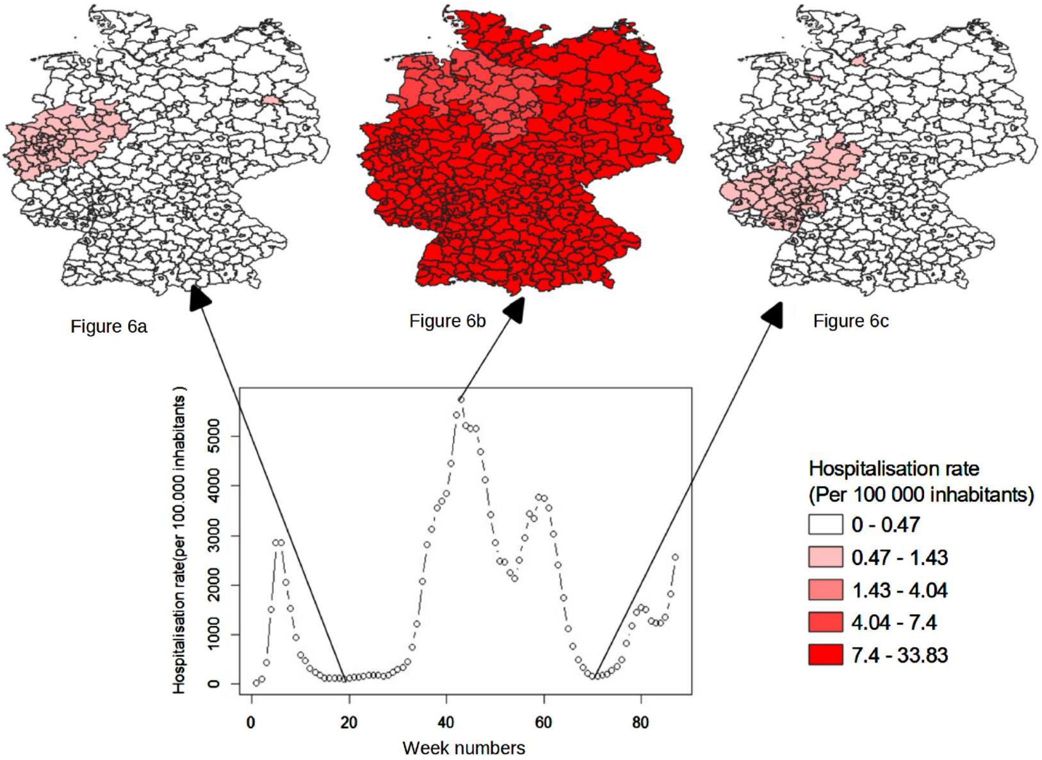


Figure 6. Temporal-spatial changes of COVID-19 hospitalisation rate in Germany from 2020.03.01 to 2021.10.31. (a) Hospitalisation rate at Week 19 (2020.07.05–07.11); (b) Hospitalisation rate at Week 43 (2020.12.20–12.26); (c) Hospitalisation rate at Week 71 (2021.07.04–07.10).

NO_2 density has a very high contribution based on the results of all three machine learning models. Combined with the result of the spatiotemporal analysis, these results imply that NO_2 may agitate the symptoms of COVID patients, which leads to more patient demand for hospitalisation care. The results in Figures 3 and 4 also support this hypothesis. More importantly, the NO_2 concentration contributes to the changes of hospitalisation rate from both temporal and spatial dimensions. The education level and the age level are linked to the changes in hospitalisation rate spatially.

To illustrate the high contribution at the spatial dimension, we map the spatial distribution of early school leavers without a certificate ('low education') at each municipality in Germany, and we also map the distribution of the population who gained vocational education ('vocational education') and the distribution of youths by municipalities ('population of 15–24 years old') in Germany (Figure 7). These features show a spatial divide between eastern and western Germany, going largely along the historical West Germany and East Germany divide (Figure 7). We could

Table 2. Comparison of peak and through time/values between NO_2 and hospitalisation rate.

Features	Unit	The peak week	The lowest week	The second-lowest week	The value of the peak week	Value of the lowest week	The value of the second-lowest week
NO_2	mol/m^2	41st	19th	75th	9.48e−05	2.26e−05	2.50e−05
Hospitalisation rate	Per 100,000 inhabitants	43rd	19th	71st	5741.46	104.61	155.75

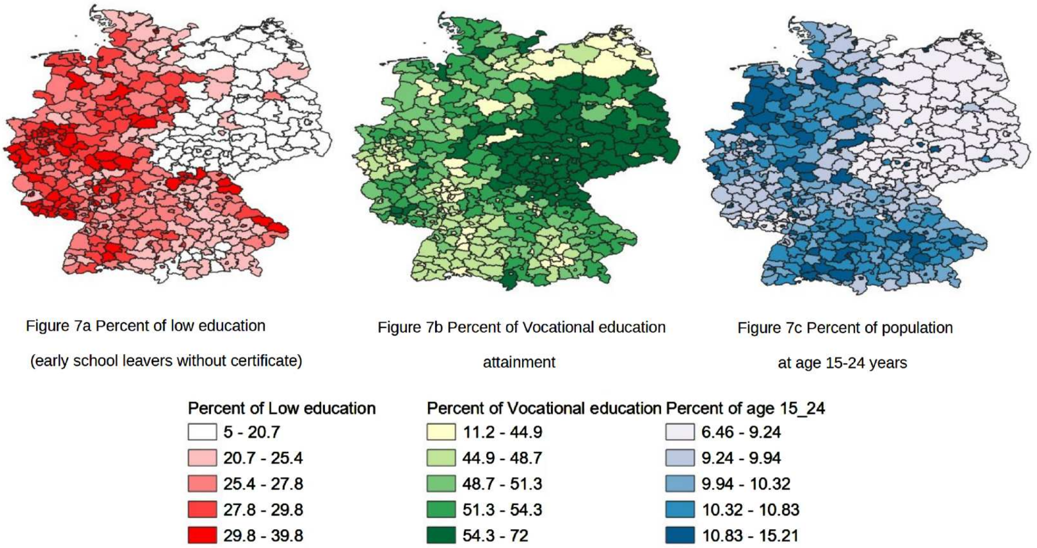


Figure 7a Percent of low education (early school leavers without certificate)

Figure 7b Percent of Vocational education attainment

Figure 7c Percent of population at age 15-24 years

Figure 7. Spatial distribution of 3 selected features in Germany. (a) The percent of low education; (b) ratio of vocational education; (c) Percent of population at aged from 15 to 24 years.

find there is a higher proportion of the low-education-level population and a higher number of youths in western Germany than in eastern Germany (Figure 7(a,c)). Conversely, the proportion of people who gained vocational education in eastern Germany is higher than that of western Germany (Figure 7(b)). Based on Figures 2, 3 and 4, we see a negative correlation between the proportion of youths as well as low education with hospitalisation rates, and a positive correlation between vocational education and hospitalisation rates. This could be explained in several ways, e.g. the larger proportion of youths in western Germany naturally being more resilient to the effects of COVID-19 and leading to fewer hospitalisations.

5. Discussion and conclusion

This research shows the potential of combining spatial data with socio-economic data to discover relationships with COVID-19 features by using ML and explainable AI models on a spatially and temporally harmonised dataset. A great effort has been put into selecting, integrating, cleaning, modelling and visualising the features but there are certainly some future improvements that can be made.

Although we have used many features, including socio-economic, demographic, and air pollution features, there are still some features that we have not included in this research. Potential additional features include social features such as a lockdown index (COVID-19 management policy), political views, compliance with rules, as well as additional environmental features, such as temperature, humidity, rainfall, water pollution, PM2.5, etc. Nevertheless, this research aims to provide an innovative and transferrable method to analyse potential features across multiple domains to discover geospatial features correlated with COVID-19, which could help other researchers to implement and further the research in this direction in other areas of the world.

Another limitation is the data availability with fine temporal and spatial dimensions. Socio-economic features only have annual data, but some features still have a high contribution, as those features contribute to the COVID hospitalisation rate by their spatial distribution. Therefore, social-economic data with finer temporal resolution could provide more evidence for the correlation analysis. In addition, other explainable AI techniques like LIME could be used to compare the current results with those of future studies.

We chose the Covid hospitalisation rate as the target feature for ML models, noting its strong correlation compared to the other three Covid-related features. This observation underscores that not all COVID-19 features exhibit equally high correlations with spatial features, likely due to time lag issues from COVID-19 reporting, vaccines, or the COVID-19 incubation period. In fact, during our data plot and correlation analysis, we detected a time lag of one to two weeks. This finding emphasises the need for additional research into the influence of time lag on correlation analysis between COVID-19 features and heterogeneous geospatial features.

To summarise, in this study, an innovative data-driven method is provided for high-dimensional and complex relationship analysis. We integrated heterogeneous and cross-disciplinary geospatial features with COVID-19 features of Germany into a new spatially and temporally harmonised dataset. The three machine learning models (RF, SVR and LGBM) perform well and the results of two high-performance models (RF and LGBM) show a high degree of agreement. In the results, both NO_2 concentration and the education level of the labour force have a higher contribution to model hospitalisation rates among COVID-19 cases in Germany.

Acknowledgements

The authors sincerely thank colleagues including Dirk Schubert, Hülya Lasch, Lukas Höller, and Joerg Pohlan, who worked in the Time Travel project. The authors thank Jie Yang for the discussion about explainable AI models. The authors thank the three anonymous reviewers for their valuable comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was financially supported by the Time Travel project of the Volkswagen Foundation in Germany under [grant number 91 913-2].

Data availability

The data that support the findings of this study are available from the corresponding author, Lixia Chu, upon reasonable request.

References

Alahmadi, A., A. Alansari, N. Alsheikh, S. Alshammasi, M. Alshamery, R. Al-abdulmohsin, L. Al Rabia, et al. 2023. "Beta Blockers may be Protective in COVID-19; Findings of a Study to Develop an Interpretable Machine Learning Model to Assess COVID-19 Disease Severity in Light of Clinical Findings, Medication History, and Patient Comorbidities." *Informatics in Medicine Unlocked* 42: 101341. https://doi.org/10.1016/j.imu.2023.101341.

Al Daoud, E. 2019. "Comparison Between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset." *International Journal of Computer and Information Engineering* 13 (1): 6–10.

Alzamzami, F., M. Hoda, and A. E. Saddik. 2020. "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation." *IEEE Access* 8: 101840–101858. https://doi.org/10.1109/ACCESS.2020.2997330

Bowe, B., Y. Xie, A. K. Gibson, M. Cai, A. van Donkelaar, R. V. Martin, R. Burnett, and Z. Al-Aly. 2021. "Ambient Fine Particulate Matter air Pollution and the Risk of Hospitalization among COVID-19 Positive Individuals: Cohort Study." *Environment International* 154: 106564. https://doi.org/10.1016/j.envint.2021.106564

Cao, Q., G. Rui, and Y. Liang. 2018. "Study on PM2.5 Pollution and the Mortality due to Lung Cancer in China Based on Geographic Weighted Regression Model." *BMC Public Health* 18 (1): 925. https://doi.org/10.1186/s12889-018-5844-4

Choudary, M. N. S., V. B. Bommineni, G. Tarun, G. P. Reddy, and G. Gopakumar. 2021. "Predicting Covid-19 Positive Cases and Analysis on the Relevance of Features Using SHAP (SHapley Additive ExPlanation)." *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*.

Cole, M. A., C. Ozgen, and E. Strobl. 2020. “Air Pollution Exposure and Covid-19 in Dutch Municipalities.” *Environmental and Resource Economics* 76 (4): 581–610. https://doi.org/10.1007/s10640-020-00491-4

Dutheil, F., J. S. Baker, and V. Navel. 2020. “COVID-19 as a Factor Influencing air Pollution?” *Environmental Pollution* 263: 114466. https://doi.org/10.1016/j.envpol.2020.114466

Ehlert, A. 2021. “The Socio-Economic Determinants of COVID-19: A Spatial Analysis of German County Level Data.” *Socio-Economic Planning Sciences* 78: 101083. https://doi.org/10.1016/j.seps.2021.101083

Futagami, K., Y. Fukazawa, N. Kapoor, and T. Kito. 2021. “Pairwise Acquisition Prediction with SHAP Value Interpretation.” *The Journal of Finance and Data Science* 7: 22–44. https://doi.org/10.1016/j.jfds.2021.02.001

Gautam, S. 2020. “COVID-19: Air Pollution Remains low as People Stay at Home.” *Air Quality, Atmosphere & Health* 13: 853–857. https://doi.org/10.1007/s11869-020-00842-6

Gautam, S., and L. Hens. 2020. “COVID-19: Impact by and on the Environment, Health and Economy.” *Environment, Development and Sustainability* 22 (6): 4953–4954. https://doi.org/10.1007/s10668-020-00818-7

Ghahremanloo, M., Y. Lops, Y. Choi, and S. Mousavinezhad. 2021. “Impact of the COVID-19 Outbreak on air Pollution Levels in East Asia.” *Science of The Total Environment* 754: 142226. https://doi.org/10.1016/j.scitotenv.2020.142226

Godawska, J. 2021. “Environmental Policy Stringency and its Impact on air Pollution in Poland.” *Ekonomia i* Środowisko(1): 52–67.

Gomathy, V., K. Janarthanan, F. Al-Turjman, R. Sitharthan, M. Rajesh, K. Vengatesan, and T. P. Reshma. 2021. “Investigating the Spread of Coronavirus Disease via Edge-AI and Air Pollution Correlation.” *ACM Transactions on Internet Technology* 21 (4): 1–10. https://doi.org/10.1145/3424222

Kavouras, I., M. Kaselimi, E. Protopapadakis, N. Bakalos, N. Doulamis, and A. Doulamis. 2022. “COVID-19 Spatio-Temporal Evolution Using Deep Learning at a European Level.” *Sensors* 22 (10): 3658. https://doi.org/10.3390/s22103658

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. 2017. “Lightgbm: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems,* 30.

Kuebart, A., and M. Stabler. 2020. “Infectious Diseases as Socio-Spatial Processes: The COVID-19 Outbreak In Germany” *Tijdschrift Voor Economische en Sociale Geografie* 111 (3): 482–496. https://doi.org/10.1111/tesg.12429

Li, L., Q. Li, L. Huang, Q. Wang, A. Zhu, J. Xu, Z. Liu, et al. 2020. “Air Quality Changes During the COVID-19 Lockdown Over the Yangtze River Delta Region: An Insight Into the Impact of Human Activity Pattern Changes on air Pollution Variation.” *Science of The Total Environment* 732: 139282. https://doi.org/10.1016/j.scitotenv.2020.139282

Li, H., X.-L. Xu, D.-W. Dai, Z.-Y. Huang, Z. Ma, and Y.-J. Guan. 2020. “Air Pollution and Temperature are Associated with Increased COVID-19 Incidence: A Time Series Study.” *International Journal of Infectious Diseases* 97: 278–282. https://doi.org/10.1016/j.ijid.2020.05.076

Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis. 2020. “Explainable ai: A Review of Machine Learning Interpretability Methods.” *Entropy* 23 (1): 18. https://doi.org/10.3390/e23010018

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. “From Local Explanations to Global Understanding with Explainable AI for Trees.” *Nature Machine Intelligence* 2 (1): 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., and S.-I. Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” *Advances in Neural Information Processing Systems,* 30.

Maier, B. F., M. Wiedermann, A. Burdinski, P. P. Klamser, M. A. Jenny, C. Betsch, and D. Brockmann. 2022. “Germany’s Fourth COVID-19 Wave was Mainly Driven by the Unvaccinated.” *Communications Medicine* 2 (1): 116. https://doi.org/10.1038/s43856-022-00176-7

Manne, R. and S. Kantheti (2020). “COVID-19 and Its Impact on air Pollution.” *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* ISSN: 2321–9653.

Muhammad, S., X. Long, and M. Salman. 2020. “COVID-19 Pandemic and Environmental Pollution: A Blessing in Disguise?” *Science of The Total Environment* 728: 138820. https://doi.org/10.1016/j.scitotenv.2020.138820

Naqvi, H. R., M. Datta, G. Mutreja, M. A. Siddiqui, D. F. Naqvi, and A. R. Naqvi. 2021. “Improved air Quality and Associated Mortalities in India Under COVID-19 Lockdown.” *Environmental Pollution* 268: 115691. https://doi.org/10.1016/j.envpol.2020.115691

Neidell, M. J. 2004. “Air Pollution, Health, and Socio-Economic Status: The Effect of Outdoor Air Quality on Childhood Asthma.” *Journal of Health Economics* 23 (6): 1209–1236. https://doi.org/10.1016/j.jhealeco.2004.05.002

Nuutinen, M., R.-L. Leskelä, E. Suojalehto, A. Tirronen, and V. Komssi. 2017. “Development and Validation of Classifiers and Variable Subsets for Predicting Nursing Home Admission.” *BMC Medical Informatics and Decision Making* 17 (1): 39. https://doi.org/10.1186/s12911-017-0442-4

Ogen, Y. 2020. “Assessing Nitrogen Dioxide (NO2) Levels as a Contributing Factor to Coronavirus (COVID-19) Fatality.” *Science of The Total Environment* 726: 138605. https://doi.org/10.1016/j.scitotenv.2020.138605

Parida, B. R., S. Bar, D. Kaskaoutis, A. C. Pandey, S. D. Polade, and S. Goswami. 2021. “Impact of COVID-19 Induced Lockdown on Land Surface Temperature, Aerosol, and Urban Heat in Europe and North America.” *Sustainable Cities and Society* 75: 103336. https://doi.org/10.1016/j.scs.2021.103336

Pascal, M., M. Corso, O. Chanel, C. Declercq, C. Badaloni, G. Cesaroni, S. Henschel, et al. 2013. "Assessing the Public Health Impacts of Urban Air Pollution in 25 European Cities: Results of the Aphekom Project." *Science of The Total Environment* 449: 390–400. https://doi.org/10.1016/j.scitotenv.2013.01.077

Prinzi, F., C. Militello, N. Scichilone, S. Gaglio, and S. Vitabile. 2023. "Explainable Machine-Learning Models for COVID-19 Prognosis Prediction Using Clinical, Laboratory and Radiomic Features." *IEEE Access* 11: 121492–121510. https://doi.org/10.1109/ACCESS.2023.3327808

Rodriguez-Galiano, V., M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas. 2015. "Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines." *Ore Geology Reviews* 71: 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Shao, Q., Y. Xu, and H. Wu. 2021. "Spatial Prediction of COVID-19 in China Based on Machine Learning Algorithms and Geographically Weighted Regression." *Computational and Mathematical Methods in Medicine* 2021: 7196492.

Snider, B., E. A. McBean, J. Yawney, S. A. Gadsden, and B. Patel. 2021. "Corrigendum: Identification of Variable Importance for Predictions of Mortality from COVID-19 Using AI Models for Ontario, Canada." *Frontiers in Public Health* 9: 759014. https://doi.org/10.3389/fpubh.2021.759014

Stojkoski, V., Z. Utkovski, P. Jolakoski, D. Tevdovski and L. Kocarev (2020). "The socio-Economic Determinants of the Coronavirus Disease (COVID-19) Pandemic." arXiv preprint arXiv:2004.07947.

Strak, M., N. Janssen, R. Beelen, O. Schmitz, I. Vaartjes, D. Karssenberg, C. van den Brink, M. L. Bots, M. Dijst, and B. Brunekreef. 2017. "Long-term Exposure to Particulate Matter, NO2 and the Oxidative Potential of Particulates and Diabetes Prevalence in a Large National Health Survey." *Environment International* 108: 228–236. https://doi.org/10.1016/j.envint.2017.08.017

Temenos, A., I. N. Tzortzis, M. Kaselimi, I. Rallis, A. Doulamis, and N. Doulamis. 2022. "Novel Insights in Spatial Epidemiology Utilizing Explainable AI (XAI) and Remote Sensing." *Remote Sensing* 14 (13): 3074. https://doi.org/10.3390/rs14133074

Tosepu, R., J. Gunawan, D. S. Effendy, L. O. A. I. Ahmad, H. Lestari, H. Bahar, and P. Asfian. 2020. "Correlation Between Weather and Covid-19 Pandemic in Jakarta, Indonesia." *Science of The Total Environment* 725: 138436. https://doi.org/10.1016/j.scitotenv.2020.138436

Urrutia-Pereira, M., C. A. Mello-da-Silva, and D. Solé. 2020. "COVID-19 and air Pollution: A Dangerous Association?" *Allergologia et Immunopathologia* 48 (5): 496–499. https://doi.org/10.1016/j.aller.2020.05.004

Varghese, R. R., D. Aiswarya, A. Roy, V. Muraly, and S. Renjith. 2022. "A Novel Approach to Predict Success of Online Games Using Random Forest Regressor for Time Series Data." *Advances in Electrical and Computer Technologies* 881: 27–40. https://doi.org/10.1007/978-981-19-1111-8_3

Vîrghileanu, M., I. Săvulescu, B.-A. Mihai, C. Nistor, and R. Dobre. 2020. "Nitrogen Dioxide (NO2) Pollution Monitoring with Sentinel-5P Satellite Imagery Over Europe During the Coronavirus Pandemic Outbreak." *Remote Sensing* 12 (21): 3575. https://doi.org/10.3390/rs12213575

Watts, J. and N. Kommenda. 2020. "Coronavirus Pandemic Leading to Huge Drop in Air Pollution." *The Guardian*, https://www.the guardian.com, Archived from the original on **4**.

Xiong, J., O. Lipsitz, F. Nasri, L. M. W. Lui, H. Gill, L. Phan, D. Chen-Li, et al. 2020. "Impact of COVID-19 Pandemic on Mental Health in the General Population: A Systematic Review." *Journal of Affective Disorders* 277: 55–64. https://doi.org/10.1016/j.jad.2020.08.001

Xu, F., H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. 2019. *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges*. Cham: Springer International Publishing.

Zhang, F., and L. J. O'Donnell. 2020. "Chapter 7 - Support Vector Regression." In *Machine Learning*, edited by A. Mechelli and S. Vieira, 123–140. Boston, MA, United States: Academic Press.

Appendix

All features used in machine learning models with feature selection are listed in the table. The 10 highest features are highlighted in bold.

MUN: Municipality level.

No.	Feature name	Definition and algorithm	Units	Data Source	Temporal resolution	Spatial resolution
Socio-demographic character						
1	Age 14 years and younger	Percentage of population 14 years and younger	%	Corona Datenplattform	Yearly	MUN
2	**Age 15–24 years**	Percentage of population at 15–24 years	%		Yearly	MUN
3	Age 25–34 years	Percentage of population at 25–34 years	%		Yearly	MUN
4	Age 35–44 years	Percentage of population at 35–44 years	%		Yearly	MUN
5	Age 45–54 years	Percentage of population at 45–54 years	%		Yearly	MUN
6	Age 55–64 years	Percentage of population at 55–64 years	%		Yearly	MUN
7	Age 65–74 years	Percentage of population at 65–74 years	%		Yearly	MUN
8	Age 75 years and older	Percentage of population at 75 years and older	%		Yearly	MUN
9	Average age of Population	Average age of population	year	INKAR	Yearly	MUN
10	Average Household size	Average household size	m^2	Regionalatlas DE	Yearly	MUN
11	Birth rate	Crude birth rate	%	INKAR	Yearly	MUN
12	Death rate	Crude death rate	%		Yearly	MUN
13	Household Density	Household Density	per km^2	Corona Datenplattform	Yearly	MUN
14	Household with Children	Household with Children	%		Yearly	MUN
15	Household without Children	Household without Children	%		Yearly	MUN
16	Inhabitants	Population of inhabitants – City size (small towns to big cities)	person	INKAR	Yearly	MUN
17	Life expectancy	Life expectancy	Year		Yearly	MUN
18	Migration background	Percentage of persons with a migratory background	%	ZENSUS Datenbank	Yearly	MUN
19	Old-age dependency ratio	Old-age dependency ratio (population at 65 years or over to population 15–64 years)	%	INKAR	Yearly	MUN
20	Population change	Crude rate of total population change	%		Yearly	MUN
21	Population density (inhabitants/km^2)	Population density Yearly	MUN			
22	Singel-person Houshold	Single-person Household	%	Corona Datenplattform	Yearly	MUN
23	Young-age dependency ratio	Young-age dependency ratio (population from 0 to 14 years to population from 15 to 64 years)	%	INKAR	Yearly	MUN
Education						
24	Distance to primary School	Distance to primary school	Metres	INKAR	Yearly	MUN
25	High education	Tertiary educational / academic education attainment per 100 inhabitants	%	ZENSUS Datenbank	Yearly	MUN

(*Continued*)

Continued.

No.	Feature name	Definition and algorithm	Units	Data Source	Temporal resolution	Spatial resolution
26	**Low education**	Early leavers from education / School leavers without certificate per 100 inhabitants	%		Yearly	MUN
27	Pupils	Total Pupils per 1,000 inhabitants	%	INKAR	Yearly	MUN
28	Secondary education	Secondary Educational Attainment per 100 inhabitants	%	ZENSUS Datenbank	Yearly	MUN
29	Students Scientific Universities	Students from Scientific Universities per 1000 inhabitants*	%	INKAR	Yearly	MUN
30	Students University of Applied Sciences	Students from University of Applied Sciences per 1000 inhabitants	%		Yearly	MUN
31	Trainees	Trainees per 1,000 inhabitants	%		Yearly	MUN
32	University students	University Students from Scientific Universities and University of Applied Science per 1,000 inhabitants	%		Yearly	MUN
33	**Vocational education**	Vocational Educational Attainment per 100 inhabitants	%	ZENSUS Datenbank	Yearly	MUN
Economics						
34	Average income	Average monthly income of households per person	Euro	INKAR	Yearly	MUN
35	**Branches Agriculture**	Branches Agriculture per 100 Establishments	%	Corona Datenplattform	Yearly	MUN
36	Branches Business and other services	Branches Business and other services per 100 Establishments	%		Yearly	MUN
37	Branches Collective Services	Branches Collective Services per 100 Establishments	%		Yearly	MUN
38	**Branches Industry**	Branches Industry per 100 Establishments	%		Yearly	MUN
39	Branches Trade	Branches Trade per 100 Establishments	%		Yearly	MUN
40	Employed labour force	Employed labour force per 100 inhabitants	%	INKAR	Yearly	MUN
41	**Employment**	(Employees / inhabitants 15 < 65 years) x 100 (%)	%		Yearly	MUN
42	Establishments	Rate of numbers of establishment per 1000 inhabitants aged 15–74	%	Corona Datenplattform	Yearly	MUN
43	Foreigners Employment	Net employment rate foreigners per 100 inhabitants	%	INKAR	Yearly	MUN
44	Labour force high education level	Labour force with an academic qualification per 100 employees	%		Yearly	MUN
45	**Labour force low education level**	Labour force without a qualification per 100 employees	%		Yearly	MUN

(*Continued*)

Continued.

No.	Feature name	Definition and algorithm	Units	Data Source	Temporal resolution	Spatial resolution
46	Labour force secondary education level	Labour force secondary education level	%		Yearly	MUN
47	Numbers of Recreation and tourism	Numbers of Recreation and tourism per 100 Establishments	%	Corona Datenplattform	Yearly	MUN
48	**Primary sector – Agriculture**	Primary sector (Agriculture) per 100 Establishments	%		Yearly	MUN
49	Secondary sector – Industry	Secondary sector (industry) per 100 Establishments	%		Yearly	MUN
50	Self Employed labour force	Self Employed labour force per 100 inhabitants	%	INKAR	Yearly	MUN
51	Tertiary sector	Tertiary sector (Trade, business, collective services and other services) per 100 establishments	%	Corona Datenplattform	Yearly	MUN
52	Unemployment	Unemployment rate	%	INKAR	Yearly	MUN
Urban-environmental factors						
53	Ratio of built-up areas	Share of built-up areas/ municipality area	%	INKAR	Yearly	MUN
54	**NO_2 concentration**	tropospheric vertical column of NO_2	mol/m^2	Dataset of Sentinel-5P sensor from Google Earth Engine	Daily	1113.2 metres
55	Ratio of Forest areas	Forest areas/municipality area		INKAR	Yearly	MUN
56	Municipality Area	Size of administrative area	Ha	Corona Datenplattform	Yearly	MUN
57	Ratio of Open public space	Open public space/size of municipality area		INKAR	Yearly	MUN
58	Ratio of Urban Green Space	Area of Green recreation Space/municipality area			Yearly	MUN
59	Ratio of water areas	Water areas/size of municipality area			Yearly	MUN
COVID-19						
60	**Cases**	COVID-19 total cases per 100.000 inhabitants (7days)		Corona Datenplattform	Weekly	MUN
61	Death rate	Death rate (Case Fatality Rate) (7days)				
62	Hospitalisation rate	The hospital admission rate of COVID-19-confirmed patients (Number of COVID-19 patients in hospital per 100.000 inhabitants) (7days)				
63	Mortality Rate	Mortality Rate (7days)				