ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 12 (2022) No. 5, pp. 1-20
DOI:10.33168/JSMS.2022.0501

Geospatial Features Influencing the Formation of COVID-19 Clusters

Radiah Haque[1], Choo-Yee Ting[1+], Yeo-Keat Ee[1], Keng-Hoong Ng[1], Mohamed Najib Shaaban[1], Chih-Yang Pee[1], Lai-Kuan Wong[1], Dhesi Baha Raja[2]

[1] Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
[2] AINQA Health Sdn Bhd, 59000 Kuala Lumpur, Malaysia

cyting@mmu.edu.my

Abstract. Machine Learning methods have been used to combat COVID-19 since the pandemic has started in year 2020. In this regard, most studies have focused on detecting and identifying the characteristics of SARS-CoV-2, especially via image processing. Some studies have applied machine learning for contact tracing to minimise the transmission of COVID-19 cases. Limited work has, however, reported on how geospatial features have an influence on the transmission of COVID-19 and formation of clusters at local scale. Therefore, this paper has aimed to study the importance of geospatial features that had resorted to COVID-19 cluster formation in Kuala Lumpur, Malaysia in year 2021. Several datasets were used in this work, which have included the address details of confirmed positive COVID-19 cases and the details of nearby residential areas and Points of Interest (POI) located within the federal territory of Kuala Lumpur. The datasets were pre-processed and transformed into an analytical dataset for conducting empirical investigations. Various feature selection methods were applied, including the Boruta Algorithm, Chi-square (Chi2) Test, Extra Trees Classifier (ETC), Recursive Feature Elimination (RFE) method, and Deep Learning Autoencoder (DLA). Detailed investigations on the top-n features were performed to elicit a set of optimal features. Subsequently, several machine learning models were trained using the optimal features, including Logistic Regression (LR), Random Forest Classifier (RFC), Naïve Bayes Classifier (NBC), and Extreme Gradient Boosting (XGBoost). It was revealed that Boruta produced the optimal number of features with $n = 96$, whereas RFC achieved the best prediction results compared to other classifiers, with around 95% accuracy. Consequently, the findings in this paper help to recognize the geospatial features that have impacts on the formation of COVID-19 and other infectious disease clusters at local scale.

Keywords: machine learning, geospatial analytics, feature importance, COVID-19 clusters

1. Introduction

The novel SARS-CoV-2 virus (COVID-19) first broke out in December 2019 in Wuhan, China, and gradually turned into a pandemic (Zhu et al., 2020). It has affected many countries around the world with almost 600 million confirmed cases and more than 6.4 million deaths. Countries like the United States has recorded more than 93 million confirmed cases, while India recorded more than 44 million, and France has more than 34 million cases, as of August 2022 (Worldometer, 2022). Malaysia is a country located in Southeast Asia that has one of the highest numbers of COVID-19 positive cases. In fact, COVID-19 is considered the greatest infectious disease outbreak that hit Malaysia since the 1918 influenza epidemic (Hashim et al., 2021). The growth rate of positive cases in Malaysia became significant in March 2020, when a group of delegates attended a religious event named Tabligh (Ting et al., 2021). This event had led to the detection of a large number of COVID-19 positive cases, particularly in Kuala Lumpur, capital of Malaysia.

In order to mitigate the transmission of COVID-19 cases and minimise the spread of the virus in densely populated areas, the federal government of Malaysia had enforced a series of national quarantines and sanitary cordon measures as a form of Movement Control Order (MCO), Conditional MCO (CMCO), and Enhanced MCO (EMCO) that lasted for around 2 years (Tang, 2022). Moreover, an outbreak management framework was implemented with an emphasis on Active Case Detection (ACD) and contact tracing, followed by cluster identification to link positive cases of COVID-19 (Hashim et al., 2021). As a result, numerous clusters have formed, especially in densely populated areas and localities (Danial et al., 2020). As of 6 August 2022, 7,047 COVID-19 clusters were detected in Malaysia, with 24 active clusters (COVIDNOW, 2022). Typically, a cluster emerges when a significant number of disease cases occur in a specific geographic location around the same period of time (Hassan et al., 2021). As such, geospatial variables, such as location data and population data, can be used to predict the likelihood of cluster formation at local scale. This, in turn, can help to mitigate COVID-19 transmission in densely populated areas.

During the initial MCO, the state government had employed geospatial analytics to detect and monitor the dynamics of COVID-19 clusters. There were, however, three challenges faced when deploying geospatial analytics: (i) how to identify the geospatial features that had resorted to COVID-19 transmission and cluster formation at local scale? (ii) how to rank the importance of those identified features? and (iii) what would be the optimal feature set for accurately predicting the likelihood of cluster formation?

Therefore, the aim of the work in this paper is to study the importance of geospatial features that had resorted to COVID-19 transmission and formation of clusters in Kuala Lumpur, Malaysia, in 2021. The main objectives of this work are:

(i) to apply feature selection methods for identifying the geospatial features that contribute to COVID-19 cluster formation at local scale, and (ii) to train machine learning models using the optimal feature set and predicting if a COVID-19 case will form a cluster (or be added to an existing cluster) or will become a sporadic (or unlinked) case. Consequently, several experiments were conducted based on geospatial analytics and feature importance. The findings in this paper help to recognise the geospatial features that have impact on the formation of COVID-19 and other infectious disease clusters at local scale.

2. Literature Review

2.1. Machine learning

Machine learning methods have been widely utilised in healthcare (Haque et al., 2021a and Haque et al., 2021b), especially to fight against infectious diseases, such as COVID-19 (Alimadadi et al., 2020, Kwekha-Rashid et al., 2021, Mohan et al., 2022, Malik et al., 2022, Adnan et al., 2022 and Alyasseri et al., 2022). It has been observed that machine learning plays a crucial role in COVID-19 investigations, discriminations, and accurate predictions (Assaf et al., 2020 and Schaar et al., 2021). One of the main contributions of machine learning and deep learning is for medical image processing to automatically detect COVID-19 cases based on the images from X-rays (Nasiri & Hasani, 2022), chest CT scans (Ardakani et al., 2020), and screening coronavirus pneumonia (Wu et al., 2020). Some studies have applied machine learning to control the spread of the virus by predicting COVID-19 positive cases (Arora et al., 2020) and estimating the number of upcoming cases (Rustam et al., 2020). Furthermore, machine learning algorithms have been employed for COVID-19 growth estimation (Tuli et al., 2020), transmission dynamic forecasting (Ravinder et al., 2020) and outbreak prediction (Bala, 2021).

Overall, machine learning algorithms demonstrate high efficacy in solving COVID-19 prediction problems (Assaf et al., 2020). In effect, accurate forecast analysis assists healthcare systems and policymakers in managing COVID-19 effectively (Schaar et al., 2021). Moreover, accurate predictive models can help identify the specific geographical locations and residential areas where the chances of cluster formation are high. As a result, targeted interventions can be applied and CMCOs and EMCOs can be enforced on those particular areas to mitigate COVID-19 transmission. However, limited work in the literature has reported on how machine learning methods can be applied for identifying the geospatial features that have an influence on the formation of COVID-19 clusters at local scale. Therefore, in this work, machine learning methods were applied for geospatial analytics and predictive modelling.

2.2. Geospatial analytics

Geospatial analytics are based on gathering and manipulating geospatial features, which often combine location information (e.g., coordinates) and location attribute information (e.g., population profile) (Ting et al., 2021). Geospatial analytics have been employed by researchers in various domains, such as retail business (Ting et al., 2018), real-estate (Muggenhuber, 2019), and disaster monitoring and prevention (Atif et al., 2020). One of the earliest uses of geospatial analytics for COVID-19 is the dashboard developed by the World Health Organisation (WHO, 2021) and Johns Hopkins University's centre for Systems Science and Engineering (Freitag et al., 2020). These websites allow users to follow up-to-date information of countries impacted with COVID-19 outbreak to support surveillance, preparedness, and response (Boulos & Geraghty, 2020). A study by (Mollalo et al., 2020) highlighted the use of geospatial features in modelling the incident and spread of COVID-19 in the United States. The utilisation of geospatial features has also been reported in Iran where epidemiological maps of cases were developed to monitor the incident locations and rates (Jesri et al., 2021). In Italy, geospatial analytics have been employed to identify the spread of COVID-19 based on data collected from social media (e.g., Facebook) (Fernandez et al., 2021). Meanwhile, geospatial analytics have been conducted for COVID-19 ACD in Selangor state, Malaysia (Ting et al., 2021). Geospatial features, such as population density information, have been considered to predict the next most probable outbreak location.

Apart from the location and population data, geospatial analytics can be used to extract Points of Interest (POI) and nearby residential information (Capanema et al., 2021). These features are capable of determining the central points and residential types with large gatherings that can cause rapid transmission of COVID-19 infections. Overall, geospatial analytics help to recognise the spatial features that have impact on the transmission of COVID-19 cases to the surrounding areas, leading to cluster formation. Nevertheless, there is a lack of empirical studies in the literature based on geospatial analytics to mitigate COVID-19 cluster formation at local scale. This is largely because there are two main challenges when deploying geospatial analytics for COVID-19 cluster formation detection and monitoring: (i) what geospatial features are required to construct the analytical dataset? and (ii) what is the importance score of each identified feature in the analytical dataset? To tackle these challenges, in this work, feature selection methods were applied to address feature importance.

2.3. Feature importance

Feature importance refers to the assigned score (or rank) to the independent variables in the experimental dataset based on their fitness at predicting the dependent variable (Razmjoo et al., 2019). Generally, predictive modelling deployed for disease related problems is considered critical that requires proper selection of the relevant features

for justified performance (König et al., 2020). Nonetheless, manually selecting the features and determining the optimal set for machine learning and deep learning applications can be challenging and time consuming, especially for large datasets that include geospatial features. Studies show that the work on feature importance have provided positive contributions towards various domains, such as Internet of Things (IoT) (Shafiq et al., 2020), biometrics (Mendes et al., 2020), medicine (Shah et al., 2020), healthcare (Figueroa et al., 2021), and security (Khammassi & Krichen, 2020). The study of feature importance plays a crucial role in the process of knowledge discovery by primarily removing noisy, redundant, and irrelevant features (Razmjoo et al., 2019). Having important features in the dataset would allow a better understanding of the influence of independent variables towards the dependent variable.

In this regard, interpretable machine learning methods, or feature selection methods, can be applied to obtain insights into the relevance of input features in the dataset (König et al., 2020). More specifically, applying a feature selection method on the experimental dataset before training the predictive model helps to recognise the features that have strong correlation with the target output. This is achieved by assigning a feature importance score for each input variable. Variables with low scores are then discarded from the final dataset. As such, feature selection is regarded as the process of reducing the number of input features when developing a predictive model (Liu et al., 2021). Feature selection is often considered crucial to improve the prediction accuracy of a classification model (Mendes Junior et al., 2020). Feature selection methods can be divided into three types: wrapper methods (Khammassi & Krichen, 2020), filter methods (Thaseen et al., 2019), and embedded methods (Shah et al., 2020). Typically, a feature selection method needs to be carefully selected based on the experimental dataset. Thus, in this study, various feature selection methods were applied, and feature importance scores were obtained to identify the geospatial features that have contributed to the formation of clusters based on COVID-19 cases in Kuala Lumpur.

3. Research Methods

3.1. Raw datasets

In this work, four datasets were used to construct the analytical dataset with relevant geospatial features, which was later employed for conducting feature importance and classification experiments. Table 1 shows the different datasets used in this paper. The first dataset, Dcase, consists of 17,842 COVID-19 case addresses in Kuala Lumpur. These cases were accumulated from 7 July 2021 to 20 July 2021 and were obtained from the Ministry of Health (MOH) Malaysia. The second dataset, Dpop, consists of population density for children under five, elderly over 60, men, woman, women of reproductive age, and youth at the level of individual latitude and longitude (DFG, 2022).

To obtain the population density for a particular area, pre-processing of the raw dataset was required. The population density was calculated at a 2 km radius from a case address. The third dataset, D_{res}, consists of the residential areas (i.e., localities) located within the federal territory of Kuala Lumpur. The dataset was obtained from the Valuation and Property Services Department Malaysia and can be found at Brickz.my (Brickz, 2022). The purpose was to extract the name, type, and price of nearby residential areas where a case was detected. The fourth dataset, D_{poi}, comprises of POI categories and specific POIs located within the federal territory of Kuala Lumpur. The dataset was obtained from Telekom Malaysia (TM, 2022). Examples of POIs are KFC, McDonald's, and KK Super Mart. Each POI is tagged to only one particular category. Examples of the POI categories are Hospital, Bank, School, Convenient Store, Construction Company, and Hypermarket. The raw datasets do not, however, allow feature selection algorithms to be applied directly. Therefore, pre-processing and transformation of the raw datasets were required for constructing the analytical dataset.

Table 1: Raw datasets.

Dataset	Description
Case Address Dataset D_{case}	A list of 17,842 COVID-19 case location address detected in Kuala Lumpur.
Population Density Dataset D_{pop}	Details about population density in case location, based on age, nationality, and ethnicity.
Residential Area Dataset D_{res}	Details about residential areas located within 2 km radius from case location.
Points of Interest Dataset D_{poi}	A list of categories for POIs and specific POI names located within 2 km radius from case location.

3.2. Analytical dataset

A well-designed data structure for the analytical dataset can help in addressing the research challenges. As such, the analytical dataset D_{alx} was formed by aggregating the transformed datasets D_{case}, D_{pop}, D_{res}, and D_{poi}. The process flow of constructing D_{alx} is demonstrated in Fig. 1, and Table 2 summarises the variables in the dataset. There are 145 features in D_{alx}, which include case address details, population density information, nearby residential area information, and POIs. Geocoding was applied to extract the population and property information for the 3 major nearby residential areas located within 2 km radius from the coordinates of case location. Similarly, the number of 50 POI categories and 50 specific POIs were calculated.

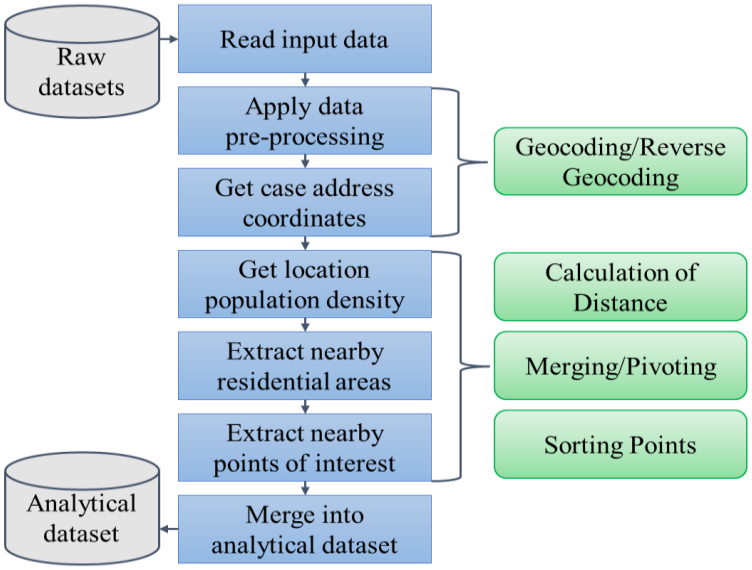

Fig. 1: Process flow of analytical dataset construction.

The target output can have one of the two classes: True, which denotes the cases that formed a cluster (or were added into a cluster), and False, which denotes the cases that became a sporadic case. Consequently, from D_{alx}, feature importance scores were extracted using feature selection methods with respect to (i) cases that belong to clusters and (ii) sporadic cases.

Table 2: Analytical dataset variables.

Variable No.	Variable Type	Description
1 - 2	Location variables (e.g., District name)	Input features with address info obtained from D_{case} dataset.
3 - 15	Population density variables (e.g., elderly_population)	Input features with population info extracted from D_{pop} dataset.
16 - 45	Nearby residential area variables (e.g., property_type)	Input features with nearby residential info extracted from D_{res} dataset.
46 - 95	POI category variables (e.g., convenience_store)	Input features with POI categories extracted from D_{poi} dataset.
96 - 145	Specific POI variables (e.g., 7_eleven)	Input features with specific POIs extracted from D_{poi} dataset.
146	Form_cluster variable	Output class: “True” for cases in clusters and “False” for sporadic cases.

3.3. Feature selection methods

In this work, four feature selection methods were applied, and feature importance scores were obtained from each method. The goal was to identify the features that have contributed to the formation of clusters based on the confirmed COVID-19 cases detected in different residential areas within the federal territory of Kuala Lumpur. The first method is the Boruta algorithm, which was designed as a wrapper around the random forest algorithm. Boruta aims to find all relevant variables and recursively removes features proved to be less relevant by a statistician rather than random probes (Anand et al., 2021). Another wrapper method was applied for feature selection is Recursive Feature Elimination (RFE), which was designed to select the optimal features by recursively applying smaller sets of features in each iteration than the previous ones (Lian et al., 2020). In addition, the Chi-square (Chi^2) test was used as the filter method for feature selection, which was designed to measure the independence of each input feature in the dataset (Thaseen et al., 2019). For the final feature selection method, Extra Tree Classifier (ETC) was applied, which is a model-based embedded approach. It was designed to select the optimal features using multiple tree-based classification models (Sharaff & Gupta, 2019). In this case, the methods were applied on D_{alx} to sort out the features that have been considered important during the training and testing process based on feature importance scores.

In addition to the feature selection methods, the Deep Learning Autoencoder (DLA) was applied, which is a type of unsupervised algorithm based on neural networks. In this case, DLA was not used to extract feature importance, rather it was trained as part of the machine learning classifiers for automatic feature extraction. The autoencoder is based on the estimation of how much each feature in the dataset contribute to the target output prediction (Xu et al., 2019). As such, DLA was used to transform the original input features in D_{alx} to its output (as encoded features). Consequently, six datasets were constructed for predictive modelling.

- D_{alx}^{Tot} includes all features from the original dataset.
- D_{alx}^{Bor} includes the optimal set of features from Boruta.
- D_{alx}^{Rfe} includes the optimal set of features from RFE.
- $D_{alx}^{Chi^2}$ includes the optimal set of features from Chi^2.
- D_{alx}^{Etc} includes the optimal set of features from ETC.
- D_{alx}^{Dla} includes the encoded features from DLA.

Subsequently, four machine learning classification algorithms were applied on these datasets, including Logistic Regression (LR), Random Forest Classifier (RFC), Naïve Bayes Classifier (NBC), and Extreme Gradient Boosting (XGBoost). The goal was to train different models with the optimal set of features in order to predict the output class with high accuracy. For validation, confusion matrix was generated to

summarise the prediction performance made by each classifier. In this case, the number of correct and incorrect output class predictions was calculated for each model based on True Positive (TP), which denotes positive values correctly predicted as actual positive, False Positive (FP), which denotes negative values incorrectly predicted as positive, False Negative (FN), which denotes positive values incorrectly predicted as negative, and True Negative (TN), which denotes negative values correctly predicted as an actual negative. Evaluation of the classification models was based on four metrics: accuracy score, precision score, recall score, and F1 score. The following equations are used to calculate the values of each evaluation metric.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4. Results and Discussion

In this work, empirical investigations were done in two stages. In the first stage, two experiments ($\mathcal{E}_{Fi}^{1}$ and $\mathcal{E}_{Fi}^{2}$) were conducted for feature importance. In the second stage, two more experiments ($\mathcal{E}_{Cl}^{1}$ and $\mathcal{E}_{Cl}^{2}$) were conducted for classification. The key findings from these experiments are described in the following sections.

4.1. Feature Importance Results

In this section, the results from $\mathcal{E}_{Fi}^{1}$ and $\mathcal{E}_{Fi}^{2}$ are analysed and discussed. In $\mathcal{E}_{Fi}^{1}$, the four feature selection methods, Boruta, RFE, Chi^2, and ETC, were applied on D_{alx} independently. For each method, feature importance scores were extracted with respect to (i) cases that form clusters (or added into the existing clusters) and (ii) sporadic cases, i.e., cases that do not form clusters (or added into the existing clusters). The feature importance scores range from 0.001 to 1.000. As such, all features in D_{alx} were sorted according to the feature importance scores generated by each feature selection method.

Table 3 shows the top-10 features obtained from the feature selection methods. It can be seen that multiple geospatial features have appeared in the top ten list of more than one method, such as total number of *Convenience Store*, *Construction Company*, and *Corporate Office*. Interestingly, all these features are POI categories. In fact, almost all features in the top-10 list of each feature selection method are POIs. This is probably due to the fact that places with high human traffic and frequently visited areas are often identified as sources of infectious disease transmission. Logically, for

highly contiguous COVID-19, these places are vulnerable, and it is very likely that a person carrying the virus will cause its spread to other nearby people, leading to the formation of clusters. Therefore, these geospatial features are essential for determining whether a particular case will form a cluster or become a sporadic case. The results further demonstrate the importance of social distancing (Qian & Jiang, 2022), which is considered the most effective measure to mitigate the transmission of COVID-19 cases and formation of clusters.

Table 3: Top ten important geospatial features.

Boruta Features	RFE Features	Chi^2 Features	ETC Features
Hair Salon	Engineer	Estee Lauder	Mukim
Car Dealer	Manufacturer	JT Express	Restaurant
Architect	Convenience Store	Nando's	Construction Company
Industrial Equipment Supplier	Company	Revenue Valley	Engineer
School	No. of Elderly Population	Secure Parking Corporation	Convenience Store
Café	Construction Company	Ayam Penyet AP	Hospital
Engineer	Financial Institution	Nelsons	Manufacturer
Accountant	Business Management Consultant	Domino's Pizza	Hardware Store
Auto Repair Shop	Corporate Office	Fos Apparel Group	Corporate Office
Convenience Store	Advertising Agency	Berjaya Roasters	Residential Property Price

In $\mathcal{E}_{Fi}^2$, further investigations were conducted on the ranked features of each method by iteratively training machine learning models with top-n features, where $n \in \{5 \ldots 100\}$. In this case, four predictive models were developed using LR, RFC, NBC, and XGBoost. In the first iteration, top-5 features were selected from each method to train the predictive models. In the subsequent iterations, features were added with step size = 1. The average accuracy generated in each iteration was recorded. The goal was to identify the optimal *n* value for each method that gives the highest average accuracy rate for all models, as seen in Fig. 2. It was observed that the four feature selection methods exhibited similar pattern with the accuracy rate fluctuating at around 94%. However, the optimal *n* value differs for each method. In case of Boruta, the optimal number of features that provides the average highest accuracy for all predictive models is $n = 96$. Meanwhile, the predictive models have demonstrated a good prediction performance with $n = 88$, 85, and 89 for RFE, Chi^2, and ETC, respectively.

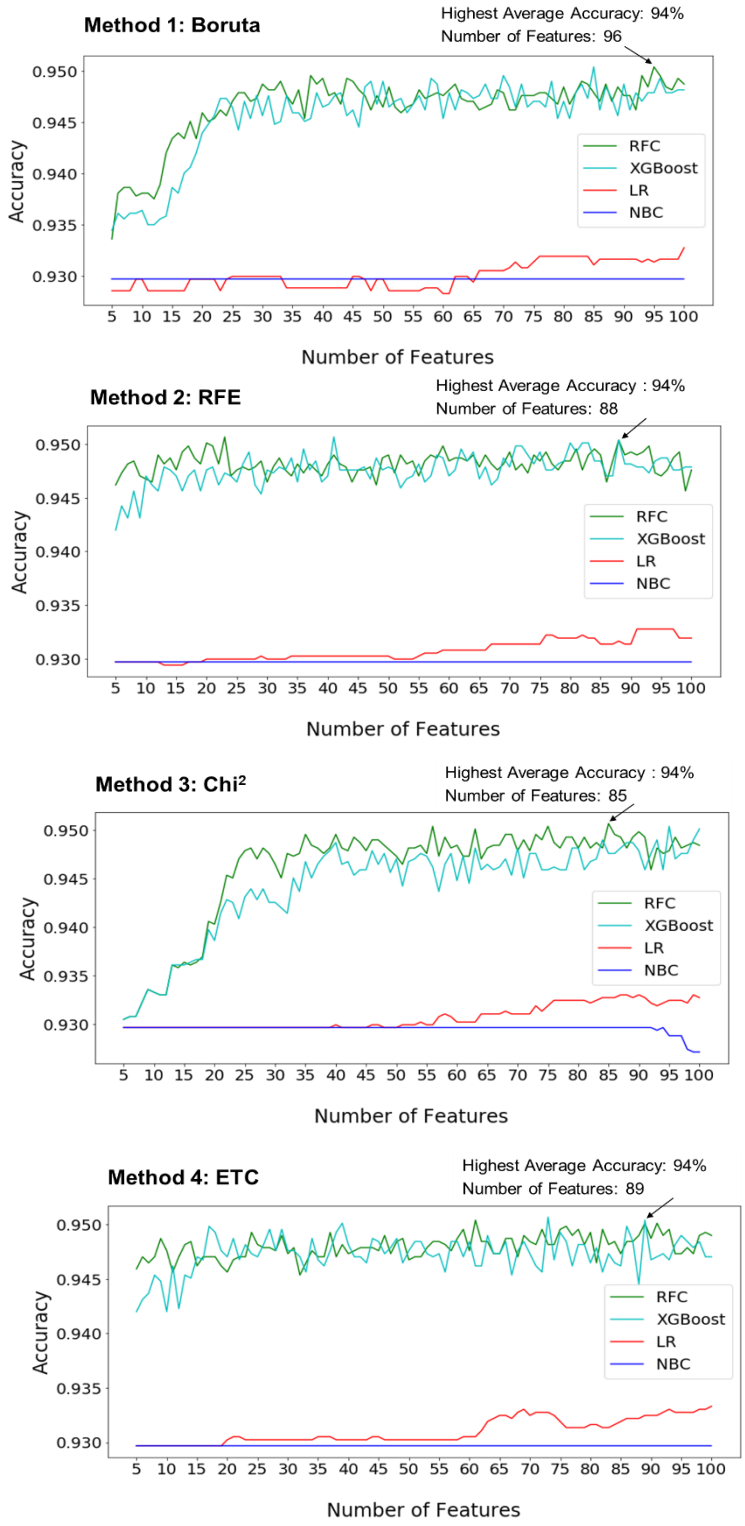

Fig. 2: Optimal number of features for all methods.

Overall, the optimum number of features for all feature selection methods is within the range of 85 and 96. The results also revealed that the prediction accuracy of the models increased gradually as the n value increased for each method. This was clearly observed with RFC and XGBoost while training. However, NBC did not provide significant improvements when the n value increased. Subsequently, the optimal n values were utilised to construct the datasets [D_{alx}^{Bor}, D_{alx}^{Rfe}, $D_{alx}^{Chi^2}$, D_{alx}^{Etc}] for conducting further classification modelling and analysis.

4.2. Classification results

In this section, the results from $\mathcal{E}_{Cl}^{1}$ and $\mathcal{E}_{Cl}^{2}$ are analysed and discussed. In $\mathcal{E}_{Cl}^{1}$, the four classification models were trained using six different datasets. It should be noted that each of these datasets consist of different numbers of features. For instance, D_{alx}^{Tot} has all the original features from D_{alx}. Meanwhile, D_{alx}^{Bor} contains top 96 features sorted by Boruta. Similarly, D_{alx}^{Rfe} has 88 features, $D_{alx}^{Chi^2}$ has 85 features, and D_{alx}^{Etc} has 89 features. Another dataset D_{alx}^{Dla} was constructed in this experiment with the encoded features generated by DLA. The average accuracy values achieved by the classification models for each feature selection method were computed, as seen in Table 4. The goal of this experiment was to identify which feature selection (or feature extraction) method generates the highest average accuracy for the predictive models based on the optimal feature list.

Table 4: Average model accuracy for analytical datasets.

Model	D_{alx}^{Tot}	D_{alx}^{Bor}	D_{alx}^{Rfe}	$D_{alx}^{Chi^2}$	D_{alx}^{Etc}	D_{alx}^{Dla}
LR	0.9330	0.9493	0.9504	0.9476	0.9468	0.9344
RFC	0.9504	0.9316	0.9316	0.9328	0.9325	0.9465
NBC	0.8902	0.9297	0.9297	0.9297	0.9297	0.9297
XGBoost	0.9498	0.9496	0.9479	0.9493	0.9484	0.9456
Average	0.9309	0.9401	0.9399	0.9399	0.9394	0.9391

Comparing the performance of the classification models, it was recognised that, on average, the models were able to provide predictions with the highest accuracy when they were trained using D_{alx}^{Bor}. This indicates that the Boruta algorithm was able to produce the most optimal feature list compared to the other feature selection methods. The results also show that the lowest average accuracy was generated when the predictive models were trained using the original features in the analytical dataset. Hence, feature selection has a positive impact on the prediction accuracy of the classifier. Meanwhile, it is worth noting that training the models with the optimal feature lists obtained from the feature selection methods with Boruta, RFE, Chi^2, and ETC has provided better classification results than the feature extraction method with DLA. This indicates that, for classification models, feature selection is effective, as it was previously suggested in the literature (Mendes et al., 2020).

During the final experiment $\mathcal{E}_{Cl}^{2}$ in this work, classification analysis was performed using LR, RFC, NBC, and XGBoost to predict the class of the target outputs. In this case, the optimal feature list produced by the Boruta algorithm was used to train the four machine learning models. The prediction performance of the models was evaluated using the confusion matrix and evaluation metrics. Fig. 3 illustrates the percentage of "True" and "False" predictions for each model, whereas Table 5 compares the evaluation metric values achieved by each model, trained with the optimal feature set. It was revealed that RFC outperformed all other models with an accuracy rate of around 95%. Moreover, precision, recall, and F1 scores generated by RFC are 0.9435, 0.9496, and 0.9443, respectively. Generally, it can be said that RFC algorithm is capable of preventing overfitting of the model by applying bagging-type ensemble of multiple trees. This is due to the fact that the final decision of RFC is based on the majority decision and features built on different decision trees (Chaabane et al., 2020), thus providing a high accuracy rate.

XGBoost also showed promising outputs with a similar accuracy rate. Meanwhile, both LR and NBC provided predictions with lower accuracy rates than RFC and XGBoost. This is probably due to the existence of linear decision surface in LR that can only model linear relationships between the dependent variable and independent variables. On the other hand, NBC produced the worst prediction accuracy because of the existence of imbalance distributions of the output classes in the dataset. For instance, the number of cases that formed clusters is significantly higher than the sporadic cases. Consequently, the NBC model was trained with a higher probability that a case will form a cluster (or added to an existing cluster), thus affecting its performance. In summary, the empirical investigations revealed that applying Boruta for feature importance, along with RFC for classification, is beneficial for conducting geospatial analytics and identifying the significant spatial features that resorted to COVID-19 transmission and formation of clusters at local scale.

Fig. 3: Confusion matrix for all classifiers.

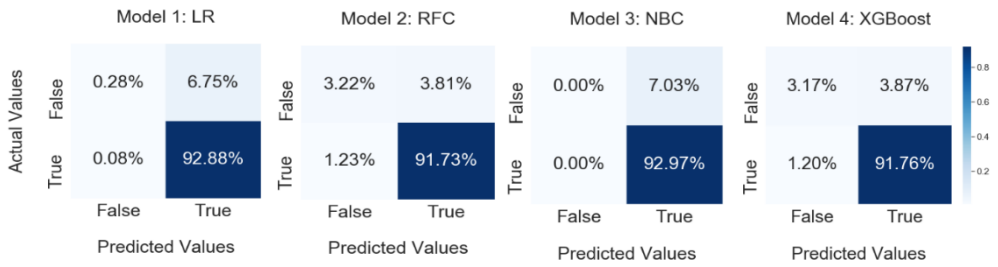

Table 5: Classification results.

Model	Accuracy	Precision	Recall	F1 Score
LR	0.9316	0.9208	0.9316	0.9020
RFC	0.9496	0.9435	0.9496	0.9443
NBC	0.9297	0.8643	0.9297	0.8958
XGBoost	0.9493	0.9430	0.9493	0.9437

5. Conclusion

One of the challenges when deploying analytics solution for COVID-19 in Malaysia especially at the initial stage was to provide a sound method for cluster detection at local scale. Kuala Lumpur, as the capital city of Malaysia with a large number of populations, could have been the source of major COVID-19 outbreak if the transmission was not monitored and put under control. One of the main challenges was to investigate and determine the geospatial features that lead to COVID-19 cluster formation at local scale. In this study, geospatial datasets from various sources (i.e., geographical, population, daily confirmed cases) were obtained and transformed into an analytical dataset. The analytical dataset was then used in the study of feature importance to investigate the relations between the geospatial features and output classes, which represent (i) cases that form clusters (or added into the existing clusters) and (ii) sporadic cases, i.e., cases that do not form clusters (or added into the existing clusters).

In the empirical study, several feature selection methods were applied, and the lists of important features were obtained. It was recognised that the Boruta algorithm generates the optimal set of features, where the top-10 geospatial features include Hai Salone, Car Dealer, Architect, Industrial Equipment Supplier, School, Café, Engineer, Accountant, Auto Repair Shop, and Convenience Store. The findings suggest that POIs have primarily contributed to COVID-19 transmission at local scale, which led to the formation of numerous clusters in Kuala Lumpur. There was, however, no feature related to population within the top-10 feature list. This is largely because there is no strong correlation between population size and number of cases, suggesting that transmission of disease is caused by human traffic rather than population density. The only limitation of the current dataset is that it does not consist of detail population information such as job type, education level, and salary range. Therefore, no investigation could be performed on finding the correlation between those features and COVID-19 cluster formation. This paper ends with a use case on how machine learning methods based on geospatial analytics can be used for accurately predicting the formation of COVID-19 clusters.

Acknowledgments

This project is funded by the Ministry of Science, Technology & Innovation Malaysia under the MOSTI Combating COVID-19 Grant (CV1220M1022).

References

Adnan, M., Altalhi, M., Alarood, A. A. & Uddin, M. I. (2022). Modeling the spread of COVID-19 by leveraging machine and deep learning models. *Intelligent Automation and Soft Computing*, *31*(3). DOI:10.32604/IASC.2022.020606.

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B. & Cheng, X. (2020). Artificial intelligence and machine learning to fight covid-19. In *Physiological Genomics*, 52(4). DOI:10.1152/physiolgenomics.00029.2020.

Alyasseri, Z. A. A., Al-Betar, M. A., Doush, I. A., Awadallah, M. A., Abasi, A. K., Makhadmeh, S. N., Alomari, O. A., Abdulkareem, K. H., Adam, A., Damasevicius, R., Mohammed, M. A. & Zitar, R. A. (2022). Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. *Expert Systems*, *39*(3). DOI:10.1111/exsy.12759.

Anand, N., Sehgal, R., Anand, S., & Kaushik, A. (2021). Feature selection on educational data using Boruta algorithm. *International Journal of Computational Intelligence Studies*, *10*(1). DOI:10.1504/ijcistudies.2021.113826.

Ardakani, A. A., Kanafi, A. R., Acharya, U. R., Khadem, N., & Mohammadi, A. (2020). Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*, *121*. DOI:10.1016/j.compbiomed.2020.103795.

Arora, P., Kumar, H. & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons and Fractals*, *139*. DOI:10.1016/j.chaos.2020.110017.1.

Assaf, D., Gutman, Y., Neuman, Y., Segal, G., Amit, S., Gefen-Halevi, S., Shilo, N., Epstein, A., Mor-Cohen, R., Biber, A., Rahav, G., Levy, I., & Tirosh, A. (2020). Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and Emergency Medicine*, *15*(8). DOI:10.1007/s11739-020-02475-0.

Atif, I., Cawood, F. T., & Mahboob, M. A. (2020). Modelling and analysis of the Brumadinho tailings disaster using advanced geospatial analytics. *Journal of the Southern African Institute of Mining and Metallurgy*, *120*(7). https://doi.org/10.17159/2411-9717/1196/2020

Bala, S. (2021). COVID-19 outbreak prediction analysis using machine learning. *International Journal for Research in Applied Science and Engineering Technology*, *9*(1). DOI:10.22214/ijraset.2021.32690.

Brickz. (2022). True property prices, brickz.my. https://www.brickz.my/.

Chaabane, I., Guermazi, R. & Hammami, M. (2020). Enhancing techniques for learning decision trees from imbalanced data. *Advances in Data Analysis and*

Classification, *14*(3). DOI:10.1007/s11634-019-00354-x.

COVIDNOW. (2022). *COVIDNOW in Malaysia*. Ministry of Health Malaysia.

Danial, M., Arulappen, A. L., Hock Ch'ng, A. S. & Looi, I. (2020). Mitigation of COVID-19 clusters in Malaysia. *Journal of Global Health*, *10*(2). DOI:10.7189/jogh.10.0203105.

DFG. (2022). High Resolution Population Density Maps. Facebook. https://dataforgood.facebook.com/dfg/tools/high-resolution-population-density-maps#accessdata.

Fernandez, G., Maione, C., Zaballa, K., Bonnici, N., Spitzberg, B. H., Carter, J., Yang, H., McKew, J., Bonora, F., Ghodke, S. S., Jin, C., De Ocampo, R., Kepner, W. & Tsou, M. H. (2021). The geography of Covid-19 Spread in Italy using social media and geospatial data analytics. *International Journal of Intelligence, Security, and Public Affairs*, *23*(3). DOI:10.1080/23800992.2021.1994813.

Figueroa B. J., López Droguett, E., & Martins, M. R. (2021). Towards interpretable deep learning: A feature selection framework for prognostics and health management using deep neural networks. *Sensors (Basel, Switzerland)*, *21*(17). DOI:10.3390/s21175888.

Freitag, M. O., Schmude, J., Siebenschuh, C., Stolovitzky, G., Hamann, H. F. & Lu, S. (2020). Critical mobility, a practical criterion and early indicator for regional COVID-19 resurgence. *MedRxiv*.

Haque, R., Ho, S.-B., Chai, I., & Abdullah, A. (2021). Optimised deep neural network model to predict asthma exacerbation based on personalised weather triggers. *F1000Research*, *10*. DOI:10.12688/f1000research.73026.1.

Haque, R., Ho, S.-B., Chai, I., Teoh, C.-W., Abdullah, A., Tan, C.-H., & Dollmat, K. S. (2021). Intelligent health informatics with personalisation in weather-based healthcare using machine learning. In *Lecture Notes on Data Engineering and Communications Technologies*, 72. DOI:10.1007/978-3-030-70713-2_4.

Hashim, J. H., Adman, M. A., Hashim, Z., Mohd Radi, M. F. & Kwan, S. C. (2021). COVID-19 epidemic in Malaysia: Epidemic Progression, challenges, and response. In *Frontiers in Public Health*, 9. DOI:10.3389/fpubh.2021.560592.

Hassan, B. A., Rashid, T. A. & Hamarashid, H. K. (2021). A novel cluster detection of COVID-19 patients and medical disease conditions using improved evolutionary clustering algorithm star. *Computers in Biology and Medicine*, *138*. DOI:10.1016/j.compbiomed.2021.104866.

Jesri, N., Saghafipour, A., Koohpaei, A., Farzinnia, B., Jooshin, M. K., Abolkheirian, S. & Sarvi, M. (2021). Mapping and spatial pattern analysis of COVID-19 in Central Iran using the local indicators of spatial association (LISA). *BMC Public Health*,

21(1). DOI:10.1186/s12889-021-12267-6.

Kamel Boulos, M. N. & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics. *International Journal of Health Geographics*, *19*(1). DOI:10.1186/s12942-020-00202-8.

Khammassi, C. & Krichen, S. (2020). A NSGA2-LR wrapper approach for feature selection in network intrusion detection. *Computer Networks*, *172*. DOI:10.1016/j.comnet.2020.107183.

König, G., Molnar, C., Bischl, B., & Grosse-Wentrup, M. (2020). Relative feature importance. *Proceedings - International Conference on Pattern Recognition*. DOI:10.1109/ICPR48806.2021.9413090.

Kwekha-Rashid, A. S., Abduljabbar, H. N. & Alhayani, B. (2021). Coronavirus disease (COVID-19) cases analysis using machine-learning applications. *Applied Nanoscience (Switzerland)*. DOI:10.1007/s13204-021-01868-7.

Lian, W., Nie, G., Jia, B., Shi, D., Fan, Q., & Liang, Y. (2020). An intrusion detection method based on decision tree-recursive feature elimination in ensemble learning. *Mathematical Problems in Engineering*, *2020*. DOI:10.1155/2020/2835023.

Liu, X., Dai, J., Chen, J. & Zhang, C. (2021). A fuzzy α-similarity relation-based attribute reduction approach in incomplete interval-valued information systems. *Applied Soft Computing*, *109*. DOI:10.1016/j.asoc.2021.107593.

Malik, M., Iqbal, M. W., Shahzad, S. K., Mushtaq, M. T., Naqvi, M. R., Kamran, M., Khan, B. A. & Tahir, M. U. (2022). Determination of COVID-19 patients using machine learning algorithms. *Intelligent Automation and Soft Computing*, *31*(1). DOI:10.32604/IASC.2022.018753.

Mendes, J. J. A., Freitas, M. L. B., Siqueira, H. V., Lazzaretti, A. E., Pichorim, S. F. & Stevan, S. L. (2020). Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels armband approach. *Biomedical Signal Processing and Control*, *59*. DOI:10.1016/j.bspc.2020.101920.

Mohan, S., John, A., Abugabah, A., Adimoolam, M., Kumar Singh, S., kashif Bashir, A. & Sanzogni, L. (2022). An approach to forecast impact of Covid-19 using supervised machine learning model. *Software - Practice and Experience*, *52*(4). DOI:10.1002/spe.2969.

Mollalo, A., Vahedi, B. & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Science of the Total Environment*, *728*. DOI:10.1016/j.scitotenv.2020.138884.

Muggenhuber, G. (2019). Data mining and analytics for real-estate applications. In *Lecture Notes in Geoinformation and Cartography*. Springer International Publishing. DOI:10.1007/978-3-319-72434-8_11.

Nasiri, H. & Hasani, S. (2022). Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. *Radiography*. DOI:10.1016/j.radi.2022.03.011.

Qian, M. & Jiang, J. (2022). COVID-19 and social distancing. In *Journal of Public Health (Germany)*, 30(1), DOI:10.1007/s10389-020-01321-z.

Ravinder, R., Singh, S., Bishnoi, S., Jan, A., Sharma, A., Kodamana, H. & Krishnan, N. M. A. (2020). An adaptive, interacting, cluster-based model for predicting the transmission dynamics of COVID-19. *Heliyon*, *6*(12). DOI:10.1016/j.heliyon.2020.e05722.

Razmjoo, A., Xanthopoulos, P. & Zheng, Q. P. (2019). Feature importance ranking for classification in mixed online environments. *Annals of Operations Research*, *276*(1–2). DOI:/10.1007/s10479-018-2972-2.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W. & Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access*, *8*. DOI:10.1109/ACCESS.2020.2997311.

Capanema, C. G., A. Silva, F., M. Braga Silva, T. R. & F. Loureiro, A. A. (2021). DCluster: Geospatial Analytics with PoI Identification. *Journal of Information and Data Management*, *12*(2). DOI:10.5753/jidm.2021.1952.

Shafiq, M., Tian, Z., Bashir, A. K., Du, X. & Guizani, M. (2020). IoT malicious traffic identification using wrapper-based feature selection mechanisms. *Computers and Security*, *94*. DOI:10.1016/j.cose.2020.101863.

Shah, S. M. S., Shah, F. A., Hussain, S. A. & Batool, S. (2020). Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods. *Computers and Electrical Engineering*, *84*. DOI:10.1016/j.compeleceng.2020.106628.

Sharaff, A. & Gupta, H. (2019). Extra-tree classifier with metaheuristics approach for email classification. *Advances in Intelligent Systems and Computing*, *924*. DOI:10.1007/978-981-13-6861-5_17.

Tang, K. H. D. (2022). Movement control as an effective measure against Covid-19 spread in Malaysia: An overview. In *Journal of Public Health (Germany)*, 30(3). DOI:10.1007/s10389-020-01316-w.

Thaseen, I. S., Kumar, C. A. & Ahmad, A. (2019). Integrated INTRUSION DETECTION MODEL USING CHI-SQUARE FEATURE SELECTION AND ENSEMBLE OF CLASSIFIERs. *Arabian Journal for Science and Engineering*,

44(4). DOI:10.1007/s13369-018-3507-5.

Ting, C. Y., Ho, C. C., Yee, H. J. & Matsah, W. R. (2018). Geospatial analytics in retail site selection and sales prediction. *Big Data*, *6*(1). DOI:10.1089/big.2017.0085.

Ting, C. Y., Zakariah, H., Kamaludin, F., Cheng, D. L. W., Tan, N. Y. Z., & Yee, H. J. (2021). Geospatial analytics for COVID-19 active case detection. *Computers, Materials and Continua*, *67*(1), 835–848. DOI:10.32604/cmc.2021.013327.

TM. (2022). *Telekom Malaysia*, *Tm.com.my*. Telekom Malaysia. https://www.tm.com.my/Pages/Home.html.

Zhu, H., Wei, L. & Niu, P. (2020). The novel coronavirus outbreak in Wuhan, China. *Global Health Research and Policy*, *5*(1). DOI:10.1186/s41256-020-00135-6.