American Journal of Mathematical and Computer Modelling
2021; 6(3): 43-49
http://www.sciencepublishinggroup.com/j/ajmcm
doi: 10.11648/j.ajmcm.20210603.11
ISSN: 2578-8272 (Print); ISSN: 2578-8280 (Online)

Predicting Important Features That Influence COVID-19 Infection Through Light Gradient Boosting Machine: Case of Toronto

Yein Choi

Envirosuite Seoul Office, Seoul, South Korea

Email address:
yeinchoi0402@gmail.com

To cite this article:
Yein Choi. Predicting Important Features That Influence COVID-19 Infection Through Light Gradient Boosting Machine: Case of Toronto. *American Journal of Mathematical and Computer Modelling*. Vol. 6, No. 3, 2021, pp. 43-49. doi: 10.11648/j.ajmcm.20210603.11

Received: June 21, 2021; **Accepted**: July 2, 2021; **Published**: July 13, 2021

Abstract: COVID-19, a disease starting from December 2019, spreads from person to person through contact, and has symptoms of cough, fever, muscle pain, etc. The diagnosis is usually done by polymerase chain reaction (PCR) test which collects samples from the nasopharyngeal area. Today, machine learning or deep learning is used to analyze data such as confirmed cases or mortality, differentiate x-ray images of COVID-19 patients and others. Not many of the researches completed before predicted important features that influence COVID-19. Therefore, we mainly address the influence of related features. Our data includes demographic, geographic, and severity information in Toronto. The experiment was developed in this order: data import, label encoding, correlation matrix, train-test split, min-max normalization, machine learning models, gridsearchcv, and feature importance. We applied a boosting algorithm and light gradient boosting machine to increase accuracy and speed, gridsearchcv, feature importance function to find the importance of the variable and best hyper parameters for models. Among two experiments, the first experiment using a feature-selected model concluded important features such as outbreak associated, FSA, and classification with 88 percent accuracy. The second experiment that did not select features but used entire features resulted in that neighborhood name, FSA, and age group as important features. The accuracy was mostly around 89 percent. The data did not include personal information but mostly geographical information, which might have influenced the result, determining geographical features as key features of infection, and the accuracy. Yet, the model for the experiment has advanced computation speed, less memory usage, and showed impressive performance.

Keywords: COVID-19, Machine Learning, Random Forest, LGBM, Artificial Intelligence

1. Introduction

1.1. Background

Covid-19, first appeared in December 2019, is the disease caused by SARS-CoV-2 (Coronavirus). Coronavirus can be spread from person to person through inhaling or having direct contact of the droplets containing the virus with the eyes, nose or mouth. Therefore, for prevention, physical distancing, wearing masks, keeping hygiene, staying at home if feeling sick is needed. The symptoms include cough, fever, chills, shortness of breath, muscle or body aches, diarrhea, nausea or vomiting, loss of taste or smell, and more. But it differs by individual. Some have severe illness while others have no symptoms at all. Nonetheless, coronavirus can cause respiratory failure, lasting lung, kidney failure, heart muscle damage, nervous system problems, or even death [1]. There are three COVID19 test available for both current and past infection. Polymerase Chain Reaction (PCR) test and antigen test are used for current infection, and antibody test is used to detect previous infection. PCR test, the most widely used test, is mostly done by Nasal pharyngeal swab since it is most sensitive and specific because nasal pharyngeal area has high concentration of viruses. It can be also done inside the nose and throat or just by collecting saliva. However, these methods are less accurate. Antigen testing is also done in a nasal swab, in the nostril, and it is most effective in the early stage of infection, when there are more viruses in the body. Antibody tests use blood samples to detect the presence of antibodies, which is produced by the immune system to

protect against the virus, after illness [2]. These days, to detect Covid or to predict numbers such as confirmed cases or death rates, machine learning or deep learning is used. For example, prediction models that uses several features such as sex, age, known contact with infected people, initial symptoms, were developed to evaluate the risk of infection [3]. Furthermore, a machine learning-based classifier was developed to differentiate the Chest-X ray images of Covid19 patients from other diseases like pneumonia, since the similarity challenges to distinguish between two [4]. In Toronto, as of June 6th, 2021, there are a total 163,063 cases, 161,272 have recovered, and 3,407 have died from covid19 [5]. Moreover, 2,022,554 people received at least one dose of vaccine, and 230,77 people have completed vaccination [6]. Some might have doubts about variants and the effectiveness of vaccines. Just like other viruses, the virus that causes COVID-19 has constantly changed and new variants occurred. However, the mutation will not make the vaccine absolutely incompetent since the vaccines evoke a broad immune respond including range of antibodies and cells [7].

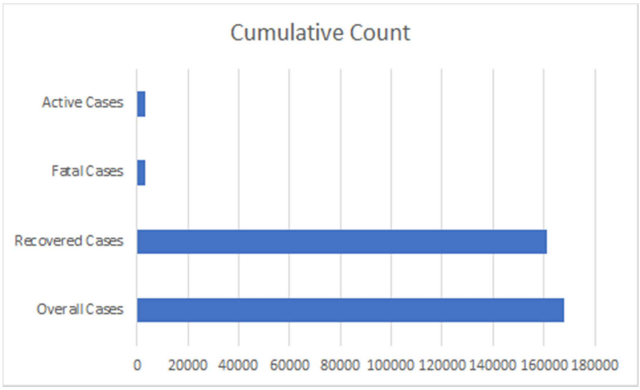

Figure 1. *Cumulative count for the COVID-19 cases.*

1.2. Objective

Prior research mainly focused on prediction of mortality, severity, diagnosis, and more. Less research was done on prediction of the important factors including environmental and personal elements. For instance, age, gender, residence, forward sortation area (FSA), and case history may be the important factors for Covid19 infection. Moreover, many other research had limitations due to insufficient datas. Considering the high level uncertainty and lack of vital data, the models developed in other studies showed low accuracy for prediction and weak generalization ability. Therefore, in this research, we primarily focused on the influences of personal and surrounding features, and the high performance of models. Below, we are going to continue on with related works, materials and methods - including description of datasets, algorithms, and models - results, discussion and conclusion.

2. Related Works

Ardabili et al. collected the data from worldometers website of five countries including Italy, Germany, Iran, USA, and China. As data were uncertain and insufficient, the standard epidemiological models could not perform well. Therefore, they proposed various machine learning models and soft computing models for COVID-19 prediction. Multi-layered perceptron (MLP) and adaptive network-based fuzzy inference system (ANFIS) yielded promising performance compared to the others [10].

Shahid et al. collected data of confirmed cases, death cases, and recovered cases of Covid19. The data got preprocessed, and were used as input data for the regression models such as autoregressive integrated moving average (ARIMA), support vector regression (SVR), gated recurrent unit (GRU), long short term memory (LSTM), and Bidirectional LSTM. Bidirectional LSTM showed the highest performance which was MAE and RMSE values of 0.007 and 0.0077, respectively [8].

Solanki and Singh collected data of India from Johns Hopkins CSSE, Worldometers website, and Kaggle. The models that have been applied to the data were: Autoregressive integrated Moving Average (ARIMA) model, the Holt-Winter model, the seasonal autoregressive integrated moving average (SARIMA) model, polynomial regression, and long short term memory (LTSM). The mean absolute percentage error value produced by the SARIMA model was 0.236, while 0.249 was produced by the Holt-Winter model. In the prediction of the number of affected cases and death, accuracy of the model estimated by the polynomial regression model is 85%. Root mean square error is calculated by the LSTM model using adaptive moment estimation optimizer. The prediction error for training is 6.45, and the calculated overall error is 5.34 [11].

Fátima Cobre et al. predicted positivity and severity of disease according to laboratory test results of patients who attended a single hospital. The four machine learning models used were Artificial neural networks (ANN), decision trees (DT), partial least squares discriminant analysis (PLS-DA), and K nearest neighbor algorithm (KNN) models. The accuracy of ANN, DT, PLS-DA, and KNN models were within 84%, and the accuracy of the classification of severe and nonsevere patients were within 86%. Hyperferritinemia, hypocalcaemia, pulmonary hypoxia, hypoxemia, metabolic and respiratory acidosis, low urinary pH, and high levels of lactate dehydrogenase were associated with the prediction [12].

Pinter et al. used data from Hungary to predict the time series of infected individuals' mortality rate. Hybrid machine learning methods such as adaptive network-based fuzzy inference system (ANFIS) and multi-layered perceptron-imperialist competitive algorithm (MLP-ICA) are used in this research. Considerable drop of the total mortality and outbreak was predicted by the model. The model accuracy is confirmed by performing validation for 9 days. Moreover, machine learning is proposed to be a capable technology to model the outbreak. The study suggests further research to enhance the quality of prediction. MLP-ICA model showed the highest performance which was rmse 8.32 while ANFIS

yielded 15.25, respectively [9].

3. Materials and Methods

3.1. Data Description

The dataset was acquired from the Kaggle, which is available at Toronto COVID-19 Cases | Kaggle [13]. It involves demographic, geographic, and severity information of both confirmed and probable cases in Toronto. The dataset consists of 14911 rows and 17 columns. "Outcome" column was set as a target column for the classification. As the "Outcome" column consists of three status which are resolved, fatal, active, multi label classification was implemented for our research. "Forward sortation area" (FSA) denotes the first three characters of postal code, "Neighborhood Name Sort" is about the divided distinct neighborhoods in Toronto and "Ever in ICU" means cases that were accepted as an intensive care unit. Columns, which are also known as features in the field of data science, were used for classifying the target.

Outbreak	Age Grou	Neighbourhood Name	FSA	Source of Infection	Classification	Client Gender	Outcome	Currently	Currently	Currently	Ever Hosp	Ever in IC	Ever Intubated
Sporadic	50-59	Malvern	M1B	Institutional	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	20-29	Malvern	M1B	Community	CONFIRMED	MALE	RESOLVED	No	No	No	Yes	No	No
Sporadic	60-69	Malvern	M1B	Travel	CONFIRMED	FEMALE	RESOLVED	No	No	No	Yes	Yes	Yes
Outbreak	50-59	Rouge	M1B	N/A - Outbreak associated	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Rouge	M1B	Close contact	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	20-29	Rouge	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	60-69	Rouge	M1B	Community	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Rouge	M1B	Close contact	PROBABLE	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	19 and yo	Malvern	M1B	Close contact	PROBABLE	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Outbreak	20-29	Malvern	M1B	N/A - Outbreak associated	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Malvern	M1B	Pending	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	30-39	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	80-89	Malvern	M1B	Community	CONFIRMED	MALE	FATAL	No	No	No	Yes	No	No
Sporadic	80-89	Malvern	M1B	Healthcare	CONFIRMED	FEMALE	RESOLVED	No	No	No	Yes	No	No
Sporadic	50-59	Malvern	M1B	Institutional	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Outbreak	30-39	Malvern	M1B	N/A - Outbreak associated	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	20-29	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Outbreak	20-29	Malvern	M1B	N/A - Outbreak associated	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	70-79	Malvern	M1B	Close contact	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	50-59	Malvern	M1B	Close contact	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	50-59	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	Yes	No	No
Sporadic	20-29	Malvern	M1B	Close contact	CONFIRMED	FEMALE	RESOLVED	No	No	No	No	No	No
Sporadic	20-29	Malvern	M1B	Unknown/Missing	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No
Sporadic	40-49	Malvern	M1B	Close contact	CONFIRMED	MALE	RESOLVED	No	No	No	No	No	No

Figure 2. *Data description of the given dataset.*

3.2. Boosting Algorithm

Boosting algorithm belongs to an ensemble algorithm, which implements more than one decision tree model on the computation Ensemble algorithm can be divided into bagging and boosting methods. While the bagging proceeds with a parallel learning and majority vote for the final decision, the boosting proceeds with a sequential learning. The boosting algorithm could be divided into two methods. In the first method, the boosting model weighs the higher gradient on important data, and Adaptive Boosting is the representative one [14]. In the second method, the boosting algorithm puts the difference between incorrect one and correct one as an input for the next classifier, which is similar to the loss function. Through this process the boosting algorithm can increase its performance and XGboost and Light Gradient Boosting ModelLGBM) belong to it [15, 16].

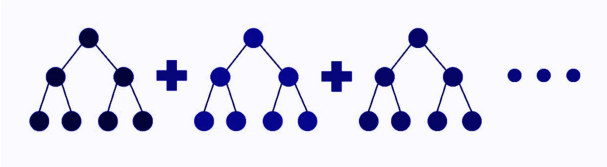

Figure 3. *Overall architecture of boosting model.*

3.3. Light Gradient Boosting Machine

As the boosting algorithm has yielded high accuracy and speed, many researchers implemented the algorithm in various research and data analysis competitions such as Kaggle. However, in the era of big data, the basic boosting machine faced the limitation on its performance compared to the deep learning algorithms. The representative drawback, which is information gain, has occurred with the rise of huge datasets. It lowered the speed of the computation and had limitations on the memory usage. Light Gradient Boosting Machine (LGBM) has occurred to solve that downside. LGBM suggested Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bounding (EFB) as solutions. For the case of normal boosting, the data with large gradient affects more on the information gain. Therefore, GOSS allows the higher accuracy of information gain with a lower dataset by excluding the data with low gradient. EFB allows LGBM to bundle mutual exclusive features in order to reduce the number of features and it allows efficient memory usage, and high performance speed. Furthermore, LGBM utilizes a Leaf-wise method while the basic boosting algorithm implements the Leve-wise method. The Leaf-wise method has an advantage on accuracy but also involves a higher probability of overfitting, which is the representative problem. To the end, LGBM supports GPU training, which could yield faster results compared to CPU training [17].

3.4. GridSearchCV

In the machine learning research, researchers implemented various methods on the same dataset to achieve the highest accuracy. For better performance, hyperparameter tuning plays a

vital role in modeling. Hyper parameter denotes the parameter whose values can be handled manually. As there is a broad range of value on each parameter, handling them with non automatic approach is inefficient. Therefore, an automatic algorithm for finding the appropriate hyperparameters is essential and RandomizedSearchCv and GridSearchCV are the representative models. Both algorithms can be implemented through the Scikit - learn's model selection package. When we put the value of predefined parameters, such as 0.1, 0.001, 0.0001 for the learning rate, the GridSearchCV calculates every predefined parameter and calculates the output through the cross validation process. The main difference between GridSearchCV and RandomizedSearchCV is that while the GridSearchCV calculates every hyper parameter, the RandomizedSearchCV calculates randomly on predefined parameters [18].

3.5. LSTM

Long short term memory (LSTM) is one of recurrent neural networks (RNN). General deep neural network (DNN) fundamentally has a one-way network. This means that the input data passes through the nodes of the neural network only once. However, RNN has a different architecture. The output from the nodes in RNN becomes an input for the same nodes. RNN models have a shortcoming of exploding and vanishing gradient problems. LSTM solves those problems by including a 'memory cell' which can conserve information for long periods of time. LSTM consists of an input gate, an output gate, and "forget" gate.

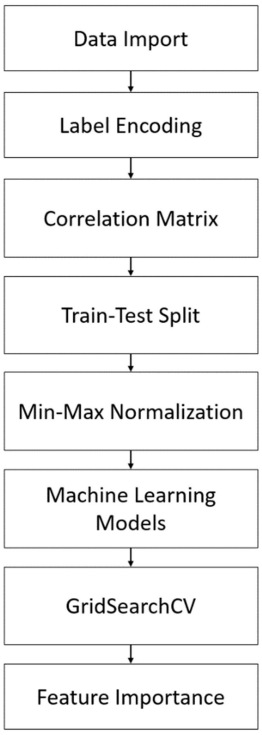

Figure 4. *Pipeline of our proposed experiment.*

First of all, the cell state receives input data and then passes it through the sigmoid layer to decide whether to update the information or forget it through (5) and (6). Secondly, the tanh layer generates a $\tilde{C}_t$, which updates the cell state (7). Thirdly, a new vector is made through (8). In this process, by multiplying f_t, the forget gates of LSTM decides whether to pass or forget the information through the previous stage. Then, it adds $i_t * \tilde{C}_t$. Lastly, output gates determine the states based on the previous cell states through (9). Through (10), the final output can be obtained through a discriminative passage of information [14].

3.6. Pipeline of Proposed Model

First, the data was imported and uploaded by the read csv function from pandas. Then, by label encoding, string data such as no or yes were converted into integers. Next, the correlation of the data was found for the feature selection. The train test split function was used for splitting the train and test size and, test size and train size were 30% and 70% per each. After MinMax normalization, which is done to adjust the range of the data, the data were put into machine learning models. Using GridSearchCV, which examines the number of cases for all combinations of hyper parameters, the model of the best fit, LGBM in this case, was searched. Lastly, applying this, we could get the important features of the infection and visualize the result.

4. Results

4.1. Confusion Matrix in Machine Learning

Table 1. *Evaluation matrix in machine learning.*

	Predicted: NO	**Predicted: YES**
Actual: NO	FN	TP
Actual: YES	TN	FP

TP, TN, FN, FP are defined as follows:

1) True Positives (TP): Data where the true label is positive and which are correctly predicted to be positive
2) False Positives (FP): Data where the true label is negative and which are correctly predicted to be positive
3) True Negatives (TN): Data where the true label is negative and which are correctly predicted to be negative
4) False Negatives (FN): Data where the true label is positive and which are correctly predicted to be negative Accuracy: Number of data correctly identified as either truly positive or truly negative out of the total number of items

4.2. Feature Selected Model

Correlation heat map showed all the correlations of the data from -1 to 1, and the features with high positive correlation with outcome were selected for the first experiment. Selected features were 'outbreak associated', 'FSA', 'classification' with correlation of 0.105228, 0.003702, and 0.029368. The data of selected features were embedded into the machine learning models, and the accuracy were

88.4 for logistic regression, 88.26 for decision tree classifier, 88.3 for random forest classifier, 88.44 for gradient boosting classifier, 88.4 for XGB classifier, and 88.51 for LGBM classifier.

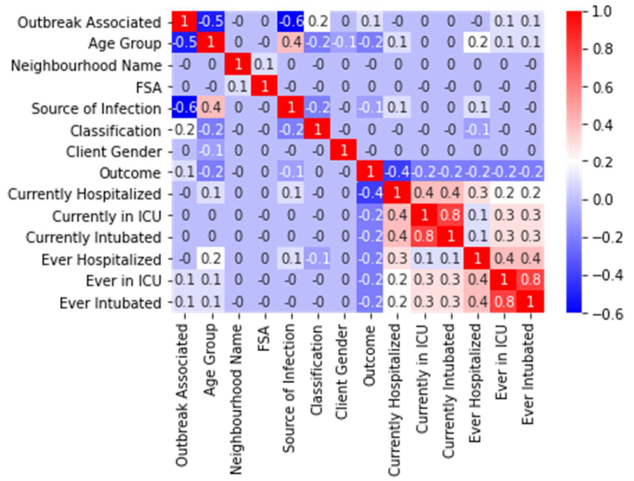

Figure 5. *Correlation results of the given dataset.*

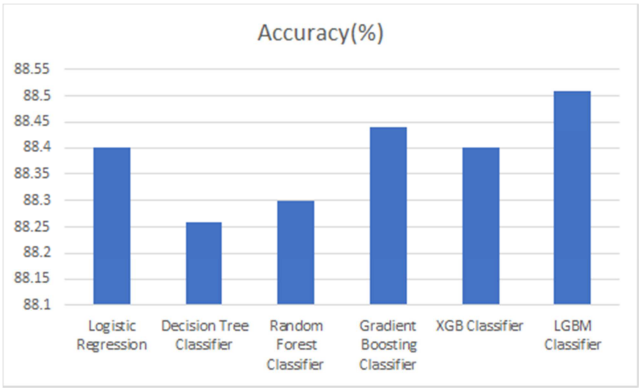

Figure 6. *Accuracy comparison among proposed models; features selected by correlation.*

4.3. Feature Unselected Model

For the second experiment which we used all of the features instead of selecting, the accuracy were 89.4, 87.49, 89.07, 89.87, 89.75, 89.91 for logistic regression, decision tree classifier, random forest classifier, gradient boosting classifier, XGB classifier, and LGBM classifier respectively. According to the models, the most important features were Neighborhood name, FSA and age group following next. From the results, we could conclude that geographical features play a large role for the COVID-19 infection. For the GridSearchCV done on the LGBM classifier, the best parameters across all searched params were 0.025 for the learning rate, 2 for depth, and 30 for iteration. The accuracy turned out to be 0.902. The best parameters for the GridSearchCV done on the random forest classifier were 10 for max dept.h, 8 for both min samples leaf and min samples split, and 100 for estimators.

Learning Rate	Depth	Iterations
0.01	2	30
0.025	4	50
0.05	6	100
0.1	8	150
	10	200
		250
		500
		750
		1000

Figure 7. *Range of hyper parameters in LGBM.*

n estimators	max depth	min samples leaf	min samples split
10	6	8	8
100	8	12	16
	10	18	20
	12		

Figure 8. *Range of hyper parameters in Random Forest.*

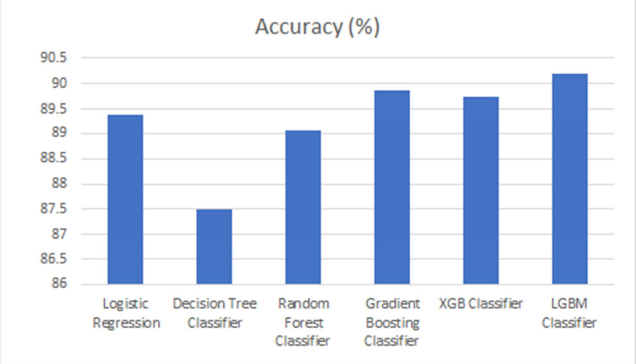

Figure 9. *Accuracy comparison among proposed models.*

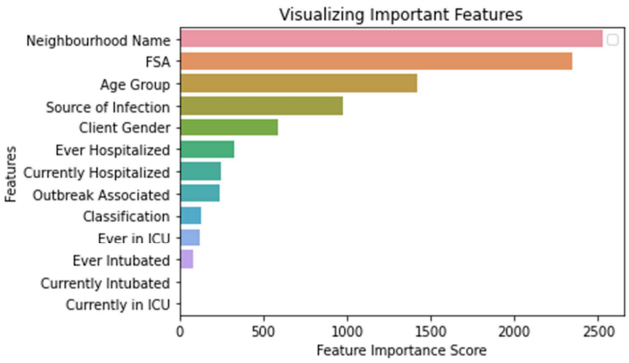

Figure 10. *Visualization of important features.*

5. Discussion

5.1. Limitation

In the research, the accuracy for machine learning models

were in a range of 87 to 90. The outcomes were all similar, making it hard to figure out the most effective model. One probability for this result is the data we used was incomplete. It might have some missing parts in the data. Moreover, the features of the data we used were a little biased on geographical information. The data did not include much information about personal characteristics. Therefore, the result mainly focused on geographical features. Furthermore, the data did not reflect increasing vaccine dissemination or vaccination rate today. Also, COVID-19 is more fatal to people who originally had underlying diseases such as cancer, chronic kidney disease, COPD, asthma, cystic fibrosis, epileptic lung disease, diabetes, immunodeficiency disease, etc [19]. However, our data did not cover personal medical histories. When selecting features, besides correlation, there are other methods like PCA, forward selection, and backward selection, but it was not used in this research.

5.2. Principal Finding

Even though we lessened the number of features based on correlation, the performance of our models were outstanding. When comparing feature selected models and the model that applied every feature of data, our models have less memory usage and high computation speed since it has a smaller number of features. In the case of deep learning, it has great performance, but the importance of the variable is unknown due to the characteristic of black box [20]. Though, we have identified the variable importance by using LGBM and GridSearchCV, and it also has high accuracy. Moreover, it does not need a high specification CPU or GPU. For medical related data, like the one in this research, it is vital to determine the features or variables that have a big influence on the outcome. In order to prevent the spread of COVID-19, ascertaining the main features that affect the propagation of the virus is a must.

6. Conclusion and Recommendation

Throughout the research, data of COVID-19 cases in Toronto, which includes demographic, geographic, severity information, was used. The experiment was done in order of importing and uploading data, encoding labels, finding correlation, doing train test split and min max normalization, putting into machine learning models, using GridSearchCV, and visualizing important features. The experiment was completed twice, one selected features based on the correlation of the data, and the other used all of the features. For the first experiment, the accuracy of the machine learning models were all around 88, and for the second experiment it was around 87 to 90. The most important features according to the model are neighborhood name, FSA, and age group. From this, we can conclude that geographical features have high influence on COVID-19 infection. Nevertheless, the accuracy of entire models were concentrated on the range of 87 to 90, and we suspect limitations in data. The data mainly focused on geographical features, and may not have included other personal features such as characteristic and medical history. However, the model itself had outstanding performance; it uses less memory and has higher computation speed.

References

[1] Sauer, L. M. (n.d.). *What Is Coronavirus?* Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus.

[2] *Which COVID test is best? Pros and cons of coronavirus detection methods: COVID: UT Southwestern Medical Center*. COVID | UT Southwestern Medical Center. (n.d.). https://utswmed.org/medblog/covid19-testing-methods/.

[3] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4* (1). https://doi.org/10.1038/s41746-020-00372-6.

[4] Zargari Khuzani, A., Heidari, M., & Shariati, S. A. (2021). COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. *Scientific Reports*, *11* (1). https://doi.org/10.1038/s41598-021-88807-2.

[5] City of Toronto. (2021, June 2). *COVID-19: Case Counts*. City of Toronto. https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-pandemic-data/covid-19-weekday-status-of-cases-data/.

[6] City of Toronto. (2021, June 2). *COVID 19: Vaccine Data*. City of Toronto. https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-pandemic-data/covid-19-vaccine-data/.

[7] World Health Organization. (n.d.). *The effects of virus variants on COVID-19 vaccines*. World Health Organization. https://www.who.int/news-room/feature-stories/detail/the-effects-of-virus-variants-on-covid-19-vaccines.

[8] [Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, *140*, 110212. https://doi.org/10.1016/j.chaos.2020.110212.

[9] Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. *Mathematics*, *8* (6), 890. https://doi.org/10.3390/math8060890.

[10] Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 Outbreak Prediction with Machine Learning. *Algorithms*, *13* (10), 249. https://doi.org/10.3390/a13100249.

[11] Solanki, A., & Singh, T. (2021). COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms. *Studies in Systems, Decision and Control*, 57–78. https://doi.org/10.1007/978-3-030-60039-6_3.

[12] Cobre, A. de, Stremel, D. P., Noleto, G. R., Fachi, M. M., Surek, M., Wiens, A., Tonin, F. S., & Pontarolo, R. (2021). Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Computers in Biology and Medicine*, *134*, 104531. https://doi.org/10.1016/j.compbiomed.2021.104531.

[13] Agrawal, D. (2020, July 17). *Toronto COVID-19 Cases*. Kaggle. https://www.kaggle.com/divyansh22/toronto-covid19-cases.

[14] AdaBoost. (2009). *Encyclopedia of Biometrics*, 9–9. https://doi.org/10.1007/978-0-387-73003-5_825.

[15] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[16] Abou Omar, K. B. (2018). XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. *Preprint Semester Project*.

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W.,... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154.

[18] Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, *14* (4), 1502.

[19] Centers for Disease Control and Prevention. (n.d.). *Certain Medical Conditions and Risk for Severe COVID-19 Illness*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

[20] London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, *49* (1), 15–21. https://doi.org/10.1002/hast.973.