A Machine Learning Approach for Modeling the Spatial-temporal Propagation Pattern of COVID-19

Mohsen Foruzandeh [1], Najmeh Neysani Samany [1], Bigard Khodakaramian [1]

[1] Faculty of Geography, University of Tehran, Iran - (forouzandeh.m, nneysani, bighard.karami)@ut.ac.ir

Keywords: Health-GIS, Prediction, Random Forest, Decision tree, Simple Bayes, K-nearest neighbor

Abstract

Predicting infections resulting from epidemiological contagions involves a complex therapeutic endeavor that necessitates the consideration of various factors within the available data. COVID-19, a globally recognized highly dangerous disease, presents an opportunity to effectively understand its patterns and spatial determinants to combat similar ailments. This study seeks to employ a deep neural network and several machine learning techniques to identify the crucial factors influencing COVID-19 infections and to forecast the spatial-temporal spread of the disease. Previous research has focused on predicting outcomes using a limited set of 8 patient variables. To contribute to this field, the current investigation analyzes a dataset comprising 47,029 records of infected individuals who sought medical attention, with 18,433 of them testing positive for the coronavirus and requiring hospitalization. The study explores the prediction of illness and hospitalization needs based on geographical locations by utilizing machine learning algorithms such as Random Forest, Naive Bayes, Decision Tree, and K-Nearest Neighbor. The machine learning process commences with input data (comprising the 47,029 records), enabling the machine to identify patterns within the dataset and subsequently make informed decisions based on these patterns and insights gained. The primary objective is to allow the machine to autonomously learn and refine its predictions without human intervention. Applying these algorithms to COVID-19 data for spatial-temporal prediction revealed that the Random Forest algorithm achieved the highest accuracy of 0.84, while the Simple Bayes algorithm exhibited the lowest accuracy at 0.50. Additionally, the study compared random forest, decision tree, simple bayes, and K-nearest neighbor algorithms to predict the severity of road accidents.

1. Introduction

The global impact of the COVID-19 pandemic, stemming from the novel COVID-19 SARS-CoV-2, has presented unprecedented challenges to healthcare systems, economies, and societies worldwide. With over 600 million confirmed cases and millions of deaths reported by August 2023, the highly contagious nature of COVID-19 has resulted in significant morbidity and mortality rates. The rapid transmission of the virus and the varying severity of symptoms experienced by patients have strained medical resources, underscoring the necessity for advanced predictive models to identify individuals at high risk and effectively manage healthcare interventions. (Moulaei et al., 2022)

Conventional epidemiological models, such as the susceptible-infected-recovered (SIR) models, have been commonly utilized to estimate virus spread and forecast outcomes. However, these models are subject to multiple assumptions and uncertainties, particularly long-term projections, which can compromise their precision and dependability. (Pinter et al., 2020)

In response to these challenges, researchers have explored the potential of machine learning (ML) techniques to enhance the prediction of COVID-19 outcomes. ML algorithms can analyze extensive patient data to pinpoint crucial disease severity and mortality indicators, facilitating informed decision-making in clinical settings. (Moulaei et al., 2022; Pinter et al., 2020)

The application of ML models in forecasting COVID-19 mortality and other outcomes has yielded promising results. Studies have shown that algorithms like random forest (RF) can accurately identify patients at risk of death based on clinical and demographic data collected upon hospital admission (Moulaei et al., 2022). These predictive models play a critical role in optimizing the allocation of limited medical resources, prioritizing high-risk patients for intensive care, and devising targeted public health strategies to mitigate the pandemic's impact. (Pinter et al., 2020)

In conclusion, the fusion of ML methodologies with traditional epidemiological strategies presents a robust framework for enhancing the prediction and management of COVID-19. This integrated approach offers more precise and comprehensive insights into the disease dynamics, ultimately bolstering the global response to the pandemic. (Buvana & Muthumayil, 2021)

2. Research method

The methodology employed in forecasting COVID-19 test outcomes through machine learning encompasses various essential stages crucial for ensuring the model's precision and dependability (Figure 1). This approach can be delineated into distinct phases, including data collection, data preprocessing, model training, model evaluation, and model interpretation.

The initial phase of this research methodology involves the comprehensive compilation of patient data. This dataset should encompass demographic particulars (e.g., age, gender, ethnicity), medical background (inclusive of pre-existing ailments and prior hospital visits), laboratory findings (such as blood tests and imaging results), and Covid-19 test outcomes (positive or negative results). This data is fundamental as it is the foundation for the machine learning model to be trained to make prognostications.

Preprocessing is necessary before putting data into a machine-learning model. This encompasses data cleansing to rectify or eliminate any inaccuracies or disparities, managing missing values through imputation or elimination of incomplete data, and standardizing and normalizing the data to ensure the equitable contribution of all features to the model. Furthermore, feature engineering may be conducted to generate novel features or modify existing ones to enhance model efficacy.

After preprocessing the data, the subsequent step involves training the machine learning model. This process entails selecting an appropriate model from a range of options, such as random forests, decision trees, K-nearest Neighbour, and Naive Bayes, contingent on the specific attributes of the data and the issue at hand. The model is then trained to utilize the training dataset to grasp the correlations between the input features (e.g., demographics, medical history) and the target variable (Covid-19 test outcome).

To ascertain the model's efficacy, evaluating it on a distinct test dataset not utilized during training is imperative. This encompasses gauging performance using accuracy, precision, recall, and F1-score metrics to evaluate the model's ability to predict COVID-19 test outcomes. Validation methodologies, such as cross-validation, ensure the model generalizes effectively to novel, unseen data.

Deciphering the model's outcomes is pivotal for comprehension and trust. This involves scrutinizing feature significance to ascertain which features (risk factors) wield the most influence in predicting the COVID-19 test outcome and ensuring that the model's decisions can be comprehended and expounded upon, particularly when employing intricate models that function as black boxes.

Various machine learning techniques can be leveraged to predict COVID-19 test outcomes. Random Forest is an ensemble technique amalgamating multiple decision trees to enhance predictive performance and mitigate overfitting. A Decision Tree is a tree-like model that segments data based on feature values to make predictions. K-Nearest Neighbour (KNN) is a straightforward algorithm that categorizes data points based on the most prevalent class among their nearest Neighbour. Naive Bayes is a probabilistic model that employs Bayes' theorem to predict the likelihood of diverse outcomes.

Machine learning presents numerous advantages for disease analysis, including identifying intricate relationships among various risk factors and outcomes, uncovering previously unidentified risk factors, and generating personalized predictions based on individual patient data. Nevertheless, there are drawbacks to consider. Some machine learning models, particularly deep learning models, may pose challenges regarding interpretability. Moreover, models could exhibit overfitting issues, performing well on training data but poorly on new data. Furthermore, training effective models often necessitate substantial amounts of high-quality data.

In summary, machine learning provides a robust method for predicting COVID-19 test results by leveraging intricate patterns in patient data to yield precise predictions. However, it is crucial to carefully assess factors such as model interpretability, susceptibility to overfitting, and data requirements to ensure dependable and actionable results. By adhering to the prescribed guidelines, researchers can develop machine learning models that enhance disease analysis and facilitate informed healthcare decision-making.

3.1. Research background

The primary objective of this investigation is to evaluate various machine learning (ML) algorithms for predicting COVID-19 mortality rates by analyzing a comprehensive dataset containing clinical characteristics of hospitalized patients upon admission. By identifying the most effective algorithm and significant predictors, this study aims to create a dependable predictive tool to aid clinical decision-making and enhance patient outcomes in the ongoing battle against COVID-19.(Moulaei et al., 2022)

This research emphasizes the crucial role of machine learning and data analytics in comprehending and responding to public health crises such as the COVID-19 pandemic. Researchers can acquire insights into disease dynamics, recognize high-risk populations, and construct predictive models to support decision-making processes and resource distribution efforts by utilizing existing data and sophisticated algorithms. (Prakash, 2020)

The investigation delved into the effectiveness of machine learning methodologies in predicting COVID-19 outcomes, showcasing the superior performance of the hybrid MLP-ICA model compared to the ANFIS model in projecting case numbers and mortality rates in Hungary. Machine learning models, which can identify intricate data patterns with fewer assumptions than mechanical models, exhibited promise. Nevertheless, challenges like limited training data and the potential for overfitting were acknowledged. It is imperative to incorporate contextual information regarding public health interventions, population mobility, and other dynamic factors during an evolving outbreak to formulate robust and widely applicable pandemic prediction models. (Pinter et al., 2020)

Researchers devised a machine learning model utilizing a distinct dataset to forecast COVID-19 diagnoses based on eight binary features, including gender, age above 60, known exposure to an infected individual, and five primary symptoms. The model accurately predicted COVID-19 positivity, aiding in prioritizing testing and optimizing resource allocation. This study highlights the potential of utilizing straightforward, readily available clinical indicators to facilitate pandemic screening and triage, particularly in restricted RT-PCR testing capacity. The authors emphasize the significance of continuous robust data sharing to enhance predictive tools as our comprehension of COVID-19 symptoms advances. (Zoabi et al., 2021)

Nonetheless, the study relied on data until early April 2020, when the pandemic developed in numerous regions worldwide. Given the rapid evolution of the COVID-19 situation, with variations in case trajectories across different areas, ongoing refinement and assessment of forecasting models on larger, updated datasets that reflect the evolving landscape of the pandemic over time are necessary.(Rustam et al., 2020)

Many current LSTM models used for COVID-19 prediction primarily utilize historical case data. Still, there is potential for

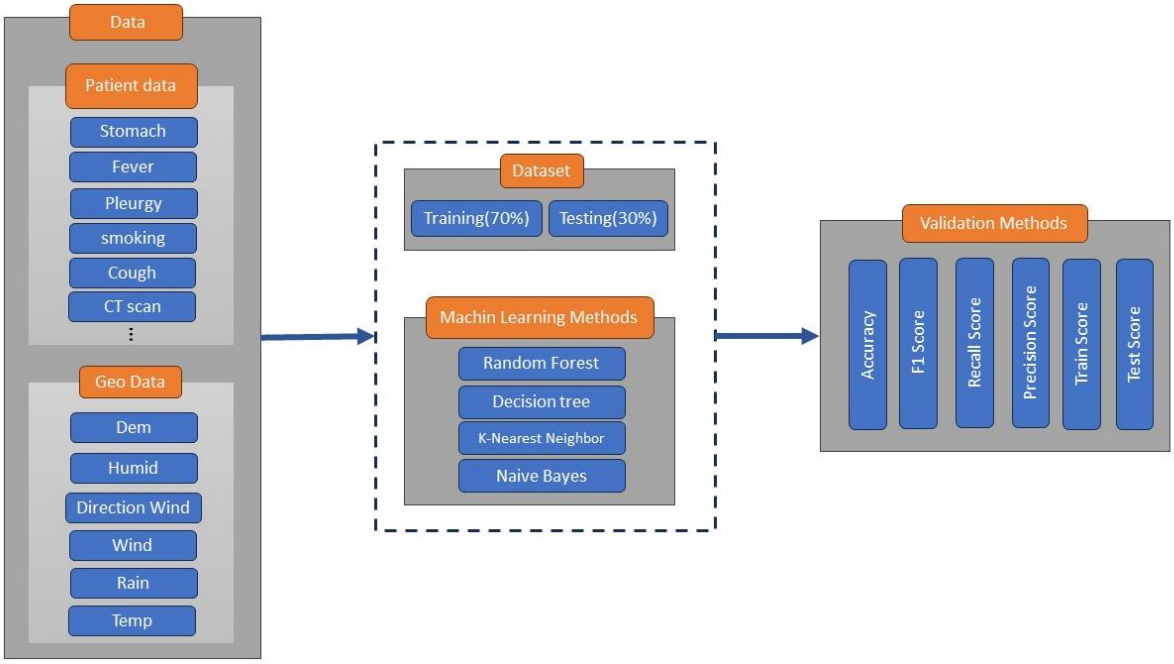

Figure 1. Research method

improved performance by integrating supplementary factors such as demographics, weather conditions, and population density. This research introduces a new K-means-LSTM framework that merges K-means clustering with LSTM neural networks. Noteworthy advancements include the utilization of XGBoost to ascertain the significance of features for K-means clustering, training distinct LSTM models for clusters of similar days or locations, and integrating demographic, meteorological, and location-specific risk elements alongside historical case figures. This methodology aims to more effectively capture spatial and temporal trends in the transmission of COVID-19. The K-means-LSTM model exhibited superior performance to the conventional SEIR model in predicting case numbers in Louisiana, showcasing the potential of data-driven techniques for real-time monitoring of outbreaks and allocation of resources during public health emergencies.(Vadyala et al., 2021)

The research emphasizes the significance of clinical characteristics such as contact with COVID-19 patients, age, respiratory symptoms, and fever in forecasting the likelihood of infection. Applying these predictive models can assist healthcare providers in prioritizing testing and resource distribution, thereby aiding in the management of pandemics. Nevertheless, the study is constrained by its reliance on a specific dataset and a limited set of features. Future investigations should encompass additional variables like comorbidities, demographics, and laboratory findings to enhance predictive precision. Integrating these models into clinical decision support systems can facilitate prompt monitoring and early identification of COVID-19 cases, enabling timely interventions and improving patient outcomes.(Buvana & Muthumayil, 2021)

The research aimed to establish robust models for predicting symptoms and mortality solely based on age, gender, and medical history. If validated, these models could serve as potent decision-support tools during the ongoing pandemic and future disease outbreaks. (Jamshidi et al., 2021)

3.2. Study area

This study leverages extensive COVID-19 infection data from Tehran, Iran, from the onset of the pandemic in 2020 through 2021 to develop a spatial-temporal prediction model for the virus's spread. By incorporating various demographic and environmental factors, the research aims to identify the primary determinants of COVID-19 infections and forecast their transmission patterns. Tehran, the capital and most populous city of Iran, serves as a critical focus due to its dense population and significant political, economic, and cultural influence. The findings from this research could be instrumental in shaping public health strategies and interventions not only in Tehran but also in other urban centers facing similar challenges.

Iran, an independent Western Asian country with a land area of 1.65 million square kilometers and a population exceeding 85 million as of 2023, operates under a unitary Islamic republic governance system. The country shares borders with Armenia, Azerbaijan, Turkmenistan, Afghanistan, Pakistan, Turkey, Iraq, the Persian Gulf, and the Gulf of Oman. The outcomes of this study, depicted in Figure 2, offer a predictive framework for understanding the spatial-temporal spread and severity of COVID-19. Such a framework, reliant on machine-learning methodologies, can be adapted to other research settings and global contexts where relevant data are available, potentially benefiting cities worldwide in their public health responses.

3.3. Dataset

The data collected for patients in Tehran from 21 November 2020 to 16 April 2021 was analyzed in this study, encompassing a total of 47,000 individuals over five months. Of these patients, 18,433 individuals tested positive for COVID-19 and were subsequently hospitalized. The research identified the key characteristics

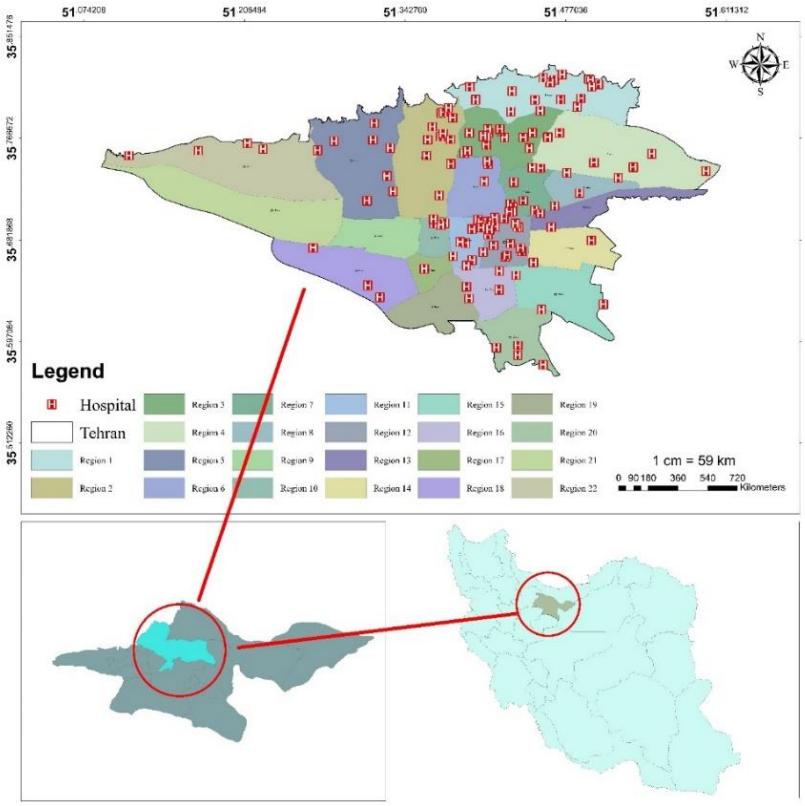

Figure 2. Tehran Province Study area

among 28 variables, such as diarrhea, vomiting, cough, headache, cancer, diabetes, and asthma, significantly impacting the outcomes. Furthermore, the study considered various climatic factors of the region, including temperature, precipitation, humidity, wind speed, and wind direction, as well as demographic attributes and the digital elevation map of the area. The random forest algorithm, a machine learning technique, was employed in the analysis.(https://behdasht.gov.ir)

3.4. Machine learning algorithms for prediction (ML)

Machine learning, a branch of artificial intelligence, enables computers to learn and improve performance through experience without explicit programming. This technology allows systems to enhance based on data and past experiences autonomously, utilizing various algorithms such as supervised learning (with labeled data), unsupervised learning (identifying patterns in unlabeled data), semi-supervised learning (using both labeled and unlabeled data), and reinforcement learning (learning through trial-and-error interactions with the environment). Extensively applied in fields like predictive analytics, cybersecurity, the Internet of Things, healthcare, e-commerce, natural language processing, and computer vision, the efficacy of machine learning primarily depends on the quality of training data and the selection of appropriate algorithms. Key tasks and algorithms in machine learning include classification, regression, clustering, dimensionality reduction, association rule learning, reinforcement learning, and deep learning with neural networks. With the surge in data availability and computational power, machine learning has significantly advanced, enabling intelligent systems to understand data patterns, make accurate predictions, and facilitate data-driven decision-making across various sectors. As data volumes grow, machine learning is set to play an increasingly crucial role in extracting insights and automating intelligent behaviors. (Sarker, 2021)

3.4.1. Naive Bayes algorithm (NB)

The Naive Bayes algorithm is a simple but powerful probabilistic classification technique based on Bayes' theorem. It determines the probability that a given data event belongs to each possible class and determines the class with the highest probability. The term "naïve" reflects the algorithm's assumption of conditional independence between the properties or attributes of the data given the class label. Despite this oversimplified assumption, Naive Bayes classifiers perform very well in various practical scenarios. The algorithm models the conditional probability of each class by calculating the product of the conditional probabilities of the features in that class by the probabilities learned from the training data. During forecasting, it estimates the conditional probability of the properties in each class. It applies the Bayes rule to determine each class's probability based on the properties, predicting the class with the highest posterior probability. Despite its simplicity, Naive Bayes is computationally efficient, requires relatively small training data, and often achieves classification accuracy comparable to more complex algorithms for many tasks. (Li et al., n.d.)

3.4.2. Random Forest Algorithm

Random Forest is an ensemble learning algorithm used for classification, regression, and other tasks that works by building multiple decision trees during training and outputs a class that is the shape of the individual trees' classification results and the regression's average prediction. The basic idea is to introduce randomness in tree growth to increase generality and avoid overfitting. This is achieved by randomly selecting a subset of

features for each distribution and using Bootstrap Aggregation, where each tree is trained on a different random sample of the training data with replacement. This randomness helps reduce variance while maintaining low bias. Random forests are easy to use, efficiently process high-dimensional data, tolerate outliers and noise, and provide priority. They have been successfully applied in many fields and are the most commonly used machine learning algorithms. (Breiman, 2001)

3.4.3. Decision tree algorithm

Decision trees are a supervised machine learning algorithm used to solve classification and regression problems that recursively divide the input data space into smaller subsets based on data properties or attributes. Starting from the root node, the data is divided into child nodes based on the most discriminating feature at each level, continuing until a stopping criterion is met, resulting in leaf nodes representing the final predictions or outcome classes. Popular decision trees handle numerical and categorical data for ease of interpretation and visualization and can automatically learn non-linear relationships in the data. Prominent decision tree algorithms include ID3, C4.5, and CART, which use metrics such as information gain or the Gini index to determine the optimal distribution of features. Despite their efficiency and intuitive modeling of complex decisions in processes, decision trees can overestimate training data, so pruning techniques and combining methods, such as random forests, which combine multiple decision trees, are used to improve accuracy and avoid overfitting. Decision trees provide a robust and comprehensive approach to understanding the underlying data models.(Lu et al., 2022; Patel & Prajapati, 2018)

3.4.4. K-Nearest Neighbor (KNN) algorithm

The k-nearest Neighbor (kNN) algorithm is a simple yet powerful machine learning algorithm for classification and regression tasks. It identifies the k-closest data points in the training set and moves them to a new data point that needs to be classified or its value predicted. The 'closeness' is typically determined by calculating a distance metric, such as Euclidean distance, between the new and all training points. For classification, the new point is assigned the class label most frequent among its k nearest neighbors, while for regression, the output value is the average of the values of its k nearest neighbors. One of kNN's key advantages is its lack of assumptions about the underlying data distribution, making it effective for non-linear decision boundaries. It is simple to understand and implement and can naturally handle multi-class classification problems. However, it is computationally expensive for large training sets due to the need to calculate distances to all training points for each new point, and choosing an appropriate value of k is crucial, as a small k can lead to overfitting noise in the training data. In contrast, a large k may miss local patterns. Despite these limitations, KNN remains popular and widely used, especially for smaller datasets. (Kashvi Taunk et al., n.d.)

3.4.4. The chi-square method for feature selection

The Chi-square method is a valuable tool for feature selection in machine learning, particularly for medical diagnostics like heart disease prediction. Evaluating feature significance relative to the target class enhances model accuracy, as demonstrated in our study, where combining Chi-square selection with the BayesNet classifier achieved an 85.0% accuracy, 84.73% precision, and 85.56% recall. The effectiveness of Chi-square selection varies with different algorithms, emphasizing the need for experimentation to optimize performance. Future research should explore other feature selection techniques and classifiers, including wrapper methods, to refine predictive models and uncover nuanced data relationships. This study highlights the importance of data analysis techniques in advancing medical diagnostics and improving disease prediction models. (Spencer et al., 2020)

3.5. The correlation matrix

Image correlation is a widely used technique for measuring surface deformations of geophysical phenomena in fields such as geoscience and natural disaster studies. Digital Image Correlation (DIC) involves finding the highest similarity between a reference image pattern and a search pattern over a larger search area in another image, which allows the displacement field between the two images to be calculated. Different correlation functions and approaches have been developed for DIC. An article comparing 15 different combinations of correlation functions (e.g., normalized cross-correlation, sum of squared differences, Fourier-based methods) and image representations (intensity, gradient, for synthetic and real images) found that frequency-based methods are generally less robust to noise methods such as zero-measure normalized cross-correlation in image intensities and original "point" correlation in directional images performed well under various noise conditions analyzed the effect of different types of noise (blurring, lighting changes, shadows) and concluded that the optimal DIC method depends on the expected environmental conditions and the types of noise present in a given application.(Moulaei et al., 2022)

4. Research results

This diagram shows the correctness or incorrectness of the laboratory tests, where the number one indicates the correctness of the tests and the number zero indicates the incorrectness of the tests. (Figure 3)

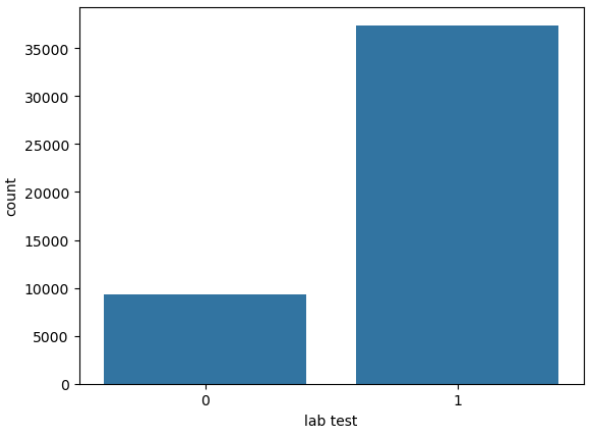

Figure 3. Laboratory tests diagram

This graph (Figure 4) shows the results of the Covid-19 test of patients. The number zero indicates that patients should follow up on their condition. The number one indicates a negative test, which includes about 15,000 people. And the number of patients whose Covid-19 test was positive is two, which is about 18,000 people. (Figure 4)

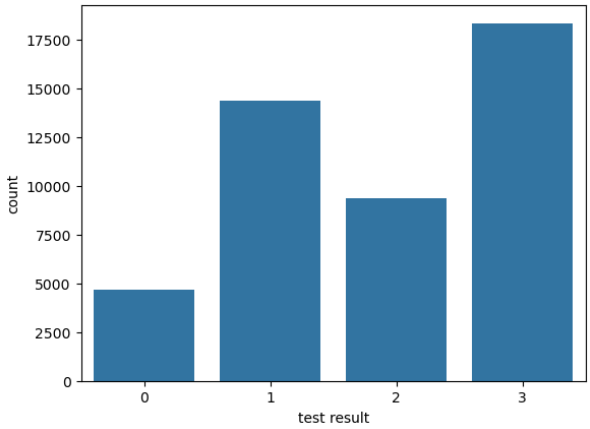

Figure 4. The results of the Covid-19 test of patients

This graph (Figure 5) shows the situation of COVID-19 patients, where the number zero indicates the patients who needed hospitalization, and there were about 5000 people. Number one is the unknown condition of the patient, and number two shows the patients who died, which were nearly 5000 people, and number three is the patients who did not need hospitalization and were discharged from the hospital. There were about 35,000 people. (Figure 5)

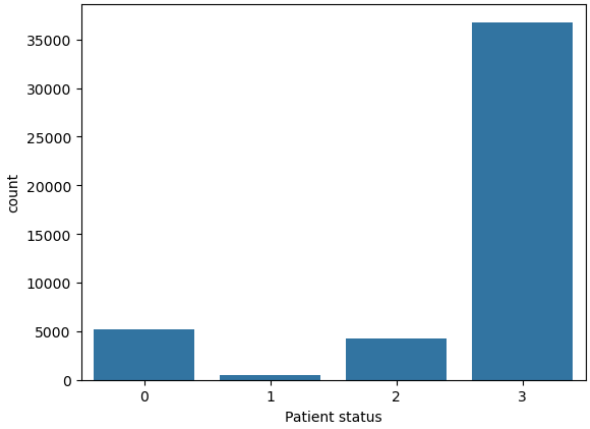

Figure 5. The situation of Covid-19 patients

This graph shows the number of Covid-19 patients according to age range. As you can see, with the increase in the age range of people, the number of COVID-19 patients also increases until the highest number of patients is related to class 6, which shows the age range between 60 and 69 years and is something over 8000 people. Class 1 is related to people aged 10 to 19, who constitute the young population of society, and class 9 is related to people over 90 years old, which includes the elderly. The elderly are included in a smaller number because they leave the house less, and their families respect health care more. (Figure 6)

The graph below shows that the highest number of COVID-19 patients is related to regions 12, 17, 18, and 1. Regions 17 and 18 have the least number of hospitals. According to the population statistics and density of these regions, it shows that the population density in these regions is due to the lack of access to hospitals and adequate health facilities and the factor of maintaining

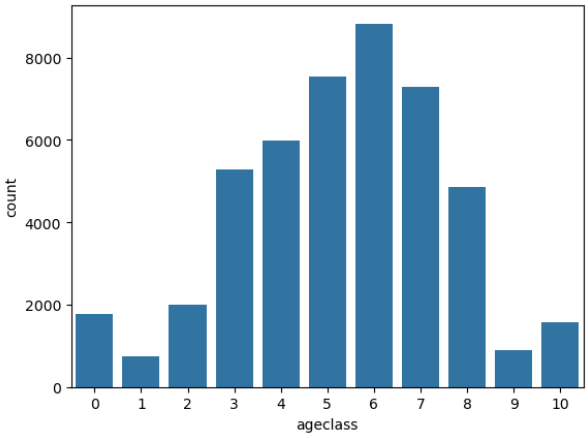

Figure 6. The number of Covid-19 patients according to an age range

distance, which is one of the essential factors of infection to COVID-19 is seen in this case. Region 12 has a significant number of hospitals. Still, the lack of hospitals in surrounding regions such as 14, 15, and 16 can increase the number of people visiting hospitals in this region, which can be justified as a result of the high number of COVID-19 patients in this region. Region 13 has the lowest number of patients in our study period, and the number of sufficient hospitals and low population density can be one of the reasons for the reduction of diseases recorded in this region. (Figure 7)

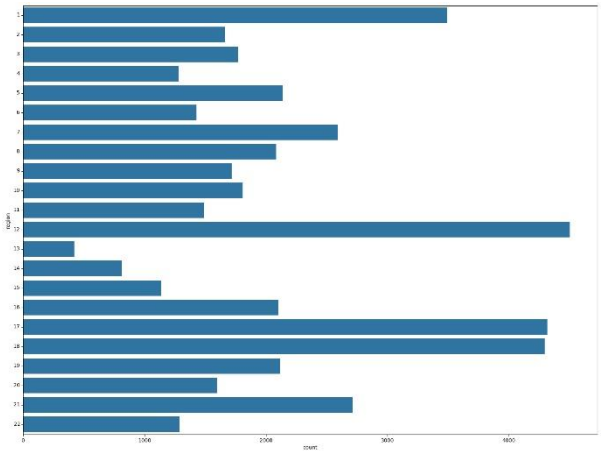

Figure 7. Number of COVID-19 patients related to regions

The graph below shows the statistics of COVID-19 every week during our study. As you can see, in the first week, there is a significant number of people with the disease, which decreases until the fourth week. After that, it remains at a fixed level until the twelfth week, and finally, it starts increasing again, which can be due to the New Year holidays, travel, shopping, and family gatherings. It is necessary to mention here that the unavailability of the COVID-19 vaccine for any reason and the resistance of the people to injecting the COVID-19 vaccine can be one of the reasons for the high number of patients in Iran compared to the global statistics. (Figure 8)

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-2024-183-2024

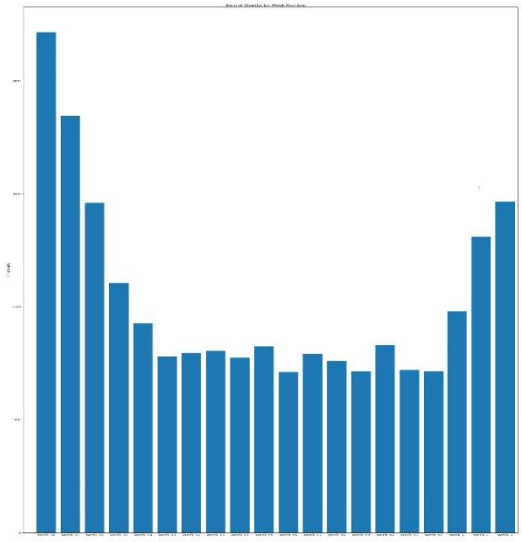

Figure 8. The statistics of the COVID-19 on a weekly

Figure 9 illustrates the prioritized characteristics of laboratory tests for diagnosing COVID-19. Age is the most critical factor, followed by muscle pain and blood oxygen levels. Individuals with a history of heart disease, diabetes, and hypertension are deemed more significant than those with symptoms such as fever, cough, or headache. The Chi-square method was employed to determine the statistical significance of these characteristics, ensuring that the most influential features are highlighted in the diagnostic process.

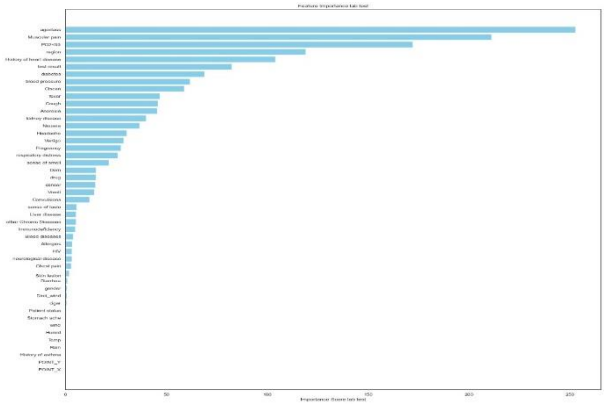

Figure 9. The characteristics of laboratory tests

4.1. The results of the algorithms

This study compared Random Forest, Decision Tree, Simple Bayes, and K-Nearest Neighbour algorithms to predict COVID-19. Results of applying machine learning algorithms to COVID-19 data to predict the spread of COVID-19 in humans: The highest accuracy is associated with the random forest algorithm with an accuracy of 0.908, and the lowest accuracy (0.50) is associated with the simple Bayesian algorithm. Therefore, according to the obtained results (Table 1), after the random forest algorithm, the highest accuracy is related to the decision tree algorithms (0.66) and the nearest Neighbour (0.61) average. (Table 1)

Algorithm	Train Score	Test Score	Precision Score	Recall Score	F1 Score	accuracy
Random Forest	1.00	0.84	0.82	0.84	0.81	0.84
Decision Trees	1.00	0.66	0.66	0.66	0.66	0.66
KNN	0.74	0.61	0.61	0.61	0.61	0.61
Naive Bayes	0.59	0.59	0.25	0.50	0.33	0.50

Table 1. compared Random Forest, Decision Tree, Simple Bayes, and K-Nearest Neighbour algorithms

To ensure the spatial accuracy of the COVID-19 prediction obtained by the model, the real information of the data extracted from the test was compared with the predicted information obtained by the model (Figure 10). In multivariate statistical analysis, there are different computational methods for measuring the dependence or relationship between two random variables. The correlation coefficient between two variables is the ability to predict the value of one relative to the other. One way to show the relationship between two variables is to calculate their covariance or correlation coefficient. For this purpose, the correlation or covariance between the two images was used, and the result was 0.999079, indicating that the two images have a high correlation.

5. Conclusion

The results of this study highlight the significant potential of machine learning algorithms, particularly the Random Forest method, to accurately predict COVID-19 infection patterns and the need for hospitalization based on geographic data. The Random Forest algorithm outperformed other models such as Decision Tree, Simple Bayes, and K-Nearest Neighbor with an accuracy of 0.84, showcasing its robustness in handling complex multivariate datasets. This high accuracy underscores the algorithm's ability to extract complex patterns from 47,029 patient records, providing valuable insights into the spatial and temporal dynamics of the spread of COVID-19 and the allocation of health services.

Central to our study is the use of a comprehensive database of 47,029 infected patient records, which improves the generalizability and reliability of the results, including various patient factors, such as age, comorbidities, and laboratory test results, allowing for more nuanced analysis and improved predictive power. In contrast, previous studies that used a limited number of patient factors may have overlooked essential variables that influence infection and hospitalization risks. Our approach emphasizes the importance of comprehensive data collection and multivariate integration to refine predictive models in epidemiological studies, ensuring that all relevant factors are considered to enhance the accuracy and reliability of predictions.

In summary, this study demonstrates the effectiveness of advanced machine learning techniques, notably the Random Forest algorithm, in predicting spatial and temporal patterns of performance of the Random Forest method, evidenced by its high accuracy, suggests that machine learning can play a crucial role in public health planning and response. Using large datasets and a wide range of variables, these models can provide healthcare COVID-19 infections and hospitalization needs. The excellent providers and decision-makers with critical information for effectively allocating resources and implementing targeted

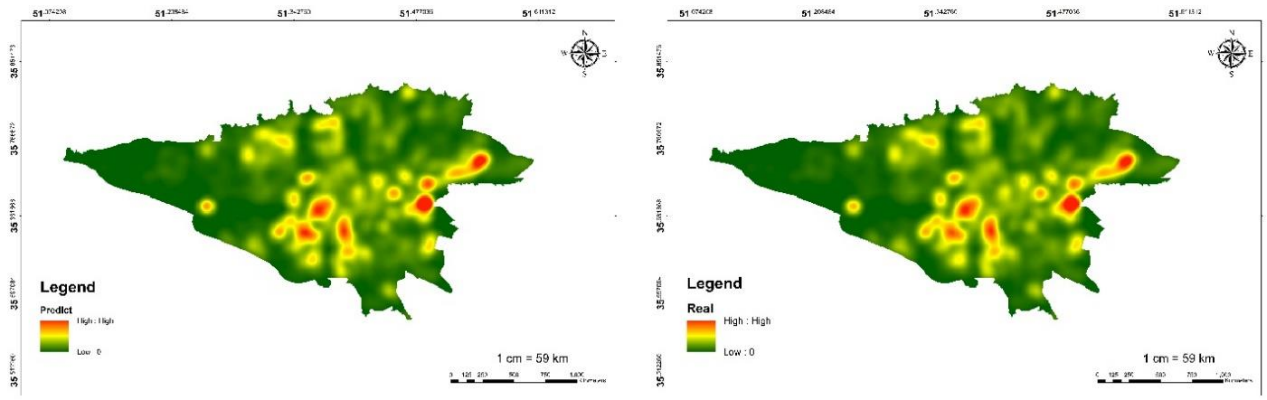

Figure 10. Predicted Covid-19 - Actual accidents

interventions. Future studies should continue to explore and improve these models by incorporating real-time data and expanding the range of variables to enhance further prediction accuracy and practical applicability in infectious disease management.

References

Breiman, L. (2001). *Random Forests* (Vol. 45).

Buvana, M., & Muthumayil, K. (2021). Prediction of covid-19 patient using supervised machine learning algorithm. *Sains Malaysiana*, *50*(8), 2479–2497. https://doi.org/10.17576/jsm-2021-5008-28

Jamshidi, E., Asgary, A., Tavakoli, N., Zali, A., Dastan, F., Daaee, A., Badakhshan, M., Esmaily, H., Jamaldini, S. H., Safari, S., Bastanhagh, E., Maher, A., Babajani, A., Mehrazi, M., Sendani Kashi, M. A., Jamshidi, M., Sendani, M. H., Rahi, S. J., & Mansouri, N. (2021). Symptom Prediction and Mortality Risk Calculation for COVID-19 Using Machine Learning. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.673527

Kashvi Taunk, Sanjukta De, Srishti Verma, & Aleena. (n.d.). *A Brief Review of Nearest Neighbor Algorithm for Learning and Classification*.

Li, D., Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers Beijing Section, IEEE International Conference on Cloud Computing and Intelligence Systems 1 2011.09.15-17 Beijing, & IEEE CCIS 1 2011.09.15-17 Beijing. (n.d.). *IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 2011 15-17 Sept. 2011, Beijing, China ; proceedings*.

Lu, Y., Ye, T., & Zheng, J. (2022). Decision Tree Algorithm in Machine Learning. *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications, AEECA 2022*, 1014–1017. https://doi.org/10.1109/AEECA55500.2022.9918857

Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-021-01742-0

Patel, H. H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. *International Journal of Computer Sciences and Engineering Open Access Research Paper*, *6*. www.ijcseonline.org

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). https://doi.org/10.3390/math8060890

Prakash, K. B. (2020). Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 2199–2204. https://doi.org/10.30534/ijeter/2020/117852020

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*, 101489–101499. https://doi.org/10.1109/ACCESS.2020.2997311

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00592-x

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. *Digital Health*, *6*. https://doi.org/10.1177/2055207620914777

Vadyala, S. R., Betgeri, S. N., Sherer, E. A., & Amritphale, A. (2021). Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM. *Array*, *11*, 100085. https://doi.org/10.1016/j.array.2021.100085

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