

A REVIEW ON EXTENSIVELY USED MACHINE LEARNING TECHNIQUES FOR THE PREDICTION OF COVID-19

Hafiza Zoya Mojahid^{1,2,*}, Jasni Mohamad Zain^{1,2}, Abdul Basit^{1,2}, Marina Yusoff^{1,2}, and Mushtaq Ali³

Received: May 29, 2023; Revised: December 27, 2023; Accepted: January 18, 2024

Abstract

Forecasting with a precise evaluation of new cases and the rate of occurrence is essential for the effective implementation of governmental initiatives and early prevention of any infectious illness. Despite the extensive research done on the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus since the outbreak, not enough knowledge could be gained about the virus in terms of immune function, virus-host interactions, pathogenesis, propagation, and mutations. In this paper, various statistical models, namely Supervised Machine Learning techniques (ML), are being discussed for previous diseases and for the recent COVID-19 pandemic. Namely, the use of the Support Vector Machine (SVM) model and a variety of time series regression models is demonstrated for several infectious diseases, including COVID-19. As infectious diseases evolve throughout time, they provide data on a single variable, that is, “the figure of contaminations that occurred over time”; thus, researchers tend to use time series models to fit the data and make predictions using different evaluation metrics to find the best-fitting model. This review developed ideas about how to enhance the current modeling techniques. Furthermore, findings of current Machine Learning Techniques are being evaluated, which attempts to estimate the COVID-19 spread. Researchers looking for approaches to advance SARS-CoV-2 research as well as individuals curious about the field’s current condition will find this review to be helpful

Keywords: COVID-19; Machine Learning (ML); Supervised Machine Learning; Support Vector Machine (SVM); Time Series Regression Modeling

¹ College of Computing, Informatics and Media, Universiti Teknologi Mara, Malaysia. E-mail: 2022547787@student.uitm.edu.my; jasni67@uitm.edu.my; 2021691374@student.uitm.edu.my; marina998@uitm.edu.my

² Institute of Big Data Analytics and Artificial Intelligence (IBDAAI), Komplek Al-Khawarizmi, University Teknologi Mara, Malaysia.

³ Riphah International University, Malakand, Pakistan. E-mail: mushtaq.ali@riphah.edu.pk

* Corresponding author

DOI: <https://doi.org/10.55766/sujst-2024-01-e01334>

Suranaree J. Sci. Technol. 31(1):030167(1-13)

Introduction

In December 2019, the Chinese city of Wuhan experienced the emergence of a new (“novel”) coronavirus in people (He *et al.*, 2020). Coronavirus disease, which is also known as COVID-19. Since the pandemic of COVID-19, the disease has spread vastly among humans, and non-pharmaceutical safety measures were the initial step to reduce the spread till the vaccinations were released. In March, the World Health Organization called it a pandemic after diagnosing more than 730,000 confirmed cases and about 36,000 fatalities (Cucinotta and Vanelli, 2020). There have been 6,960,038 fatalities and 700,261,068 verified cases of infection worldwide to date (Worldometer COVID-19 Coronavirus Pandemic). In Malaysia, around 5.1 million of these confirmed cases have been diagnosed, and the nation has so far reported about 37,246 fatalities (Worldometer covid-19 Malaysia). This epidemic has also caused a severe economic downturn worldwide. To combat the COVID-19 pandemic, many regulations have been put in place by governments all over the globe, (Mofijur *et al.*, 2021).

In order to comprehend underlying reasons as well as how viruses emerge and how they can be stopped or limited, statistical modeling and prediction are used in epidemiology. Particularly, when a novel epidemic (A disease that severely affects a large number of individuals in a society, country, or region) appears or a recognized infectious disease outbreak (refers to the situation where an epidemic spreads to a greater degree as compared to the projected specific cases) occurs, epidemiologists are the experts that gather, analyze, and evaluate data to determine measures to prevent future spread. Since, various infectious diseases have no regard for geographic boundaries, race, religion, tribe, or other characteristics Hence, after initially impacting just one area of the world, they spread quickly to other areas and may eventually become pandemics like COVID-19 (Yadav and Akhter, 2021). It is well-recognized that creating a vaccine for any contagious disease takes time and that producing the medication to administer the vaccine takes even longer. Several SARS-CoV-2 vaccinations and monoclonal antibodies (mAbs) have now received usage permission globally (The New York Times, 2022). There may be side effects from these vaccinations, but it is unclear what those side effects will be in aggregate and how long they will last. Compounding the problem, Mutation Tracker’s data reveals that SARS-CoV-2 has roughly 29,321 distinct mutations (GISAID, 2023). Previously, the COVID-19 virus incubated in the

human body for two to fourteen days. According to research, mild symptoms of the virus appear within 5 days and intensify later. It has been exhibited that patients with a strong immune system have tested positive for the virus while having no symptoms. Given that persons with a strong immune system do not exhibit the symptoms of this virus, but he/she can be a carrier who is undoubtedly capable of spreading the disease to others (Bugos, 2022; Centers for (Disease Control and Prevention, 2022; Hodcroft, 2022). Therefore, in order to prevent the surge of this fatal virus, it is crucial to understand the rate with which it is spreading when we should be expecting the next strain of this virus? and the number of people that will be affected by each individual mutation. Making the optimal decisions for disease control requires the use of this information in a significant way. Machine learning is incredibly useful in learning all of these different kinds of information. The performance of some machine learning approaches, namely Support Vector Machine (SVM) and common time series regression models, will be discussed in detail in this paper.

Contribution of Machine Learning Techniques in COVID-19 Pandemic

Machine learning (ML), a ground-breaking Artificial Intelligence (AI) technique, aims to accomplish four basic goals: Regression, Classification, Clustering, and Dimensionality Reduction. While Clustering and Dimensionality Reduction rely on Un-supervised Learning which requires unlabeled data, the first two, Regression and Classification, use Supervised Learning with labelled data. Supervised machine learning methods are favored for improving prediction accuracy and analyzing huge data sets. Because these techniques use algorithms to map the disease and its course, hence the problem can be solved (Khan *et al.*, 2022).

In the field of Machine Learning (ML), Support Vector Machines (SVM) have become one of the most popular classifiers compared to other supervised classifiers. It can be employed to address problems with classification or regression. For which it offers an appropriate response and great performance for multiple issues. It was first developed by Vapnik and his associates for the purpose of regression and classification (Cortes and Vapnik, 2004). The fundamental type of SVM algorithm is the SVM classifier. The SVM classifier

is designed to create a hyperplane, also known as a decision boundary, in a Q-dimensional space that splits data points into different classes. The equation of a hyperplane in r-dimensional space is:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_r x_r = 0 \quad (1)$$

If $x \in (x_1, x_2, \dots, x_r)$ such that $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r > 0$ or < 0 , then x lies on either side of the hyperplane. Additionally, if there are just two features, the hyperplane will be a straight line since the hyperplane's dimensions rely on the attributes that exist in the dataset. For other cases like if there are three features, the hyperplane will only have two dimensions. To select the best hyperplane, the maximum margin or the greatest possible separation between the sample points, is used. This best hyperplane is referred to as Optimal Hyperplane.

For datasets that cannot be separated linearly because of their high dimensions, the Non-linear classification method of SVM is carried out using the Kernel Trick. It is the function that changes data into an appropriate form by converting a low-dimensional in-put space into a higher-dimensional space. The SVM method uses a variety of kernel functions, including polynomial, linear, non-linear, radial basis functions, etc. In the non-linear SVM it plays an imperative role in regulating the classification limits used as a model.

$$f(x) = \beta_0 + \sum \alpha_i (x, x_i) \quad (2)$$

α_i are n parameters and are non-zero only for support vectors.

Any two observations can be written as a dot product using a linear kernel.

$$k(x_i, x_j) = \sum (x_{ik} \times x_{jk}) \quad (3)$$

It multiplies each pair of input values after taking the inner product between two vectors, such as x and x_i .

This kernel is used when data cannot be separated linearly. Hence, the linear kernel can be replaced by the polynomial kernel as follows:

Equations are shown centered and put on their own line.

$$k(x_i, x_j) = 1 + \sum (x_{ik} \times x_{jk})^d \quad (4)$$

The polynomial kernel has a degree parameter (d), which functions to find the optimal value in each dataset, it requires to specify by human in the learning algorithm.

The input space is mapped into an infinite dimensional space via the Radial Basis Function (RBF) kernel, which is mostly utilized in SVM

classification. It is mathematically explained in the following formula:

$$k(x, x_i) = \exp(-\text{gamma} * \text{sum}(x - x_i)^2) \quad (5)$$

The gamma lies between 0 and 1, and is the same as the polynomial kernel's degree parameter. Its value should be added by a humans in the learning algorithm. gamma = 0.1 is good value by default.

During the past years, Support Vector Machine (SVM) approach has been extensively used in the fields of bioinformatics, face identification, text classification, etc. Several researchers around the globe have utilized this technique in order to predict and forecast different diseases. This paper reviewed the use of the SVM technique on different diseases along with the vast usage for the ongoing epidemic of COVID-19. A research group combined two support vector machine (SVM) models to create an expert system that accurately predicts heart failure. The initial SVM model is linear and L1 regularized. It could eliminate extraneous attributes by setting their weights to zero. The second SVM model is a prediction model which uses L2 regularization. In order to dynamically improve the two models, they propose the hybrid grid search algorithm (HGSA). The Matthews correlation coefficient (MCC), ROC charts, accuracy, sensitivity, specificity, and area under the curve are six assessment metrics that are used to assess the success of the suggested approach. Their experimental findings showed that the proposed strategy enhances a traditional SVM model's performance by 3.3%.

Moreover, they demonstrated that their suggested strategy outperforms the 10 previously suggested methods, which had accuracy ranges between 57.85% and 91.83%. Finally, they concluded that their approach outperforms existing cutting-edge machine learning ensemble models in terms of performance (Ali *et al.*, 2019).

For diabetes detection, a new, improved SVM-based ensemble learning model is proposed. Since, the typical SVM technique starts by using the specified strategy for decreasing restrictions so, this was the first phase of the experiment. In the following phase, the original problem is transformed into an unconstrained optimization problem by using the penalty technique. Next, by incorporating the method of gradient descent the upgraded SVM outputs are displayed. Finally, a framework for collaborative learning is developed to diagnose diabetes. It incorporates CART, logistic regression, and improved SVM. The results of the experiment, which were generated using real-world datasets, showed that the proposed model outperforms alternative strategies (Yang *et al.*, 2019).

Since the efficacy of this strategy is not concealed from everyone, it has been widely employed since the outbreak of COVID-19. A group of researchers conducted a retrospective investigation to see if machine learning might offer a useful classification model for predicting ICU admissions. Data from March 2020 to January 2021 were used on the clinical features and lab outcomes of 100 individuals with laboratory-confirmed COVID-19 testing (Alshanbari *et al.*, 2022). It was crucial to foresee daily actual cases and mortalities in order to prevent the COVID-19 virus and decrease the strain on healthcare. Because there are different types of lockdowns used around the world, including partial, herd, and complete lockdowns, Support Vector Machine (SVM) and other machine learning models like random forests, K-nearest neighbors (K-NN), decision trees (DTs), polynomial regression, Holt winter, ARIMA, and SARIMA have been used to forecast the best lockdown method by a group of researchers. To forecast daily confirmed COVID-19 cases and casualties in three nations that were each subject to a different kind of lockdown, these techniques were investigated. Utilizing three performance criteria, the models' predictability and effectiveness were evaluated. This article presents the best lockdown approach and forecasting models to reduce COVID-19's causalities (Saba *et al.*, 2021).

The importance of vaccinations over COVID-19 confirm cases and fatalities was investigated in a study using linear regression analysis, followed by polynomial and OLS regression models, and SVM models. This information was useful in determining the efficacy of immunization. Since the data must be cleansed and pre-processed before any analysis can begin, as this ensures the accuracy of results. Hence, data from online databases was used for this study since they provide organized data on a daily basis. To begin with, active and new COVID-19 cases, fatalities, and those vaccinated for the first and second doses were transformed into monthly data. Logarithmic transformations were used after data standardization and normalization to make the data more dependable for analysis. After cleansing the data, statistical computations were carried out to ensure the model's correctness. According to the findings of this study, vaccination is vital to lower the risk of mortality and illness. Thus, more vaccination facilities need to be developed for individuals in order to receive vaccinations (Rustagi *et al.*, 2022).

Support Vector Machines and many pre-trained deep learning Convolutional Neural Network (CNN) models were combined and proposed by Al-Jumaili *et al.*, 2021 in their work. Lung ultrasound (LUS) image dataset which

consisted on three classes such as: Pneumonia, Regular and COVID-19 was used in this investigation. Four different types of CNN models were employed to draw characteristics from images, including (Resnet18, Resnet50, GoogleNet, and NASNet-Mobile). They concluded that their suggested technique outperformed the most recent publications of the time with the accuracy achieved more than 99% (Al-Jumaili *et al.*, 2021). While another team of researchers employed a unique dataset made up of 200 MRI images of 50 fit people and 150 individuals who had pulmonary disorders, including COVID-19. They created a hybrid AI technique based on support vector machines and particle swarm optimization to autonomously evaluate computed tomography images and provide a high likelihood of detecting the existence of pneumonia caused by COVID19. The major objective was to use this approach to enable fast detection of COVID-19. Their experimental findings demonstrated that the suggested framework performs with an accuracy of 95.78% (Sheela and Arun, 2022).

In another study, supervised machine learning models for COVID-19 infection were developed, utilizing learning algorithms like logistic regression, decision tree, support vector machine, naive Bayes, and artificial neural network. The models were trained on an epidemiology-labeled dataset containing positive and negative COVID-19 cases in Mexico. Prior to model development, a correlation coefficient analysis was conducted to assess the strength of relationships between dependent and independent features in the dataset. The training dataset constituted 80% of the data, while the remaining 20% was reserved for testing. The performance evaluation of the models revealed that the decision tree model achieved the highest accuracy at 94.99%, the Support Vector Machine Model exhibited the highest sensitivity at 93.34%, and the Naïve Bayes Model demonstrated the highest specificity at 94.30%. Despite the absence of a definitive cure, these non-clinical techniques offer promising solutions to address the challenges posed by the ongoing COVID-19 pandemic (Ebrahim *et al.*, 2021).

A team of researchers conducted a comprehensive study to assess the relevance of various factors in determining the severity of COVID-19 and predicting its severity upon admission, employing diverse machine learning methodologies. The retrospective investigation, spanning from January 26 to March 28, 2020, focused on 287 patients' records at JinYinTan Hospital. The data encompassed demographic, clinical, laboratory, C-CT features, and treatment outcomes. Real-time RT-PCR assays in the

laboratory confirmed all cases. Initially, LassoCV was employed to screen the original 86 features, resulting in the retention of 30 features. To further enhance clarity, Spearman's rank correlation was applied, leading to the exclusion of features with correlation values exceeding 0.4. The subsequent analysis involved processing the remaining 23 features using random forest, support vector machine, and logistic regression to evaluate the models' performance in predicting COVID-19 severity upon admission. The dataset exhibited a higher proportion of males at 56.4%, with severe cases accounting for 36.6%, and non-severe cases comprising 63.4%. The most frequently reported symptoms were cough at 61.3%, followed by fever at 30.7%. The performance metrics of the three models—accuracy, sensitivity, and specificity—were as follows: SVM achieved an accuracy of 88.5%, LR attained 85.2%, and RF scored 84.5%. Sensitivity results were RF at 96.7%, SVM at 93.9%, and LR at 92.3%. Specificity values were SVM at 79.0%, LR at 72.3%, and RF at 69.5%. According to the outcomes, SVM demonstrated the highest accuracy and specificity, while LR secured the second-highest accuracy and specificity. RF excelled in achieving the highest sensitivity (Xiong *et al.*, 2022).

This study introduces a refined Ensemble Classification strategy designed for predicting infection-related death and cure rates across various Indian states. The analysis utilizes data spanning January to May 2021, encompassing 5004 occurrences with 9 distinct characteristics. Employing the MATLAB environment, Neighbourhood Component Analysis (NCA) is applied for feature selection in classification, focusing on continuous features. The primary objective is predicting the virus's spread across states, with the dependent variable being "state/union territory" and predictors including "cured," "deaths," and "confirmed cases," alongside other attributes of mixed nominal and categorical data types. The dataset is split into an 80% training set and a 20% test set for model assessment. The classification model scrutinizes state-wise projected outcomes concerning "death rate" vs. "cured rate" and "cured" vs. "confirmed" incidences of COVID-19. Comparative analysis involves benchmarking the developed technique against state-of-the-art methods such as ensemble methods (AdaBoost, Random Forest, and Bagging), Decision Trees, Gaussian Naïve Bayes, and Support Vector Machine (SVM). Experimental results reveal that, in accurately predicting COVID-19 outcomes across multiple regions, the proposed Fine-tuned Ensemble model exhibits the highest accuracy, with Support Vector Machines closely following (Guleria *et al.*, 2022).

In an extensive investigation centered on SARS-CoV-2-related fatalities, a group of Chinese researchers conceived the idea of developing a predictive model employing deep learning techniques to assess the likelihood of survival or death for individuals infected with COVID-19 based on their medical history, encompassing factors like age, gender, and pre-existing conditions. Utilizing the anomaly detection capability of autoencoders, their investigation centered on crafting a model for predicting COVID-19 fatality upon diagnosis. In their dataset, death from COVID-19 was deemed an anomaly due to its infrequent occurrence. The autoencoder underwent a transformation, transitioning from a data compression algorithm to a specialized model for predicting rare events, specifically distinguishing between life and death. Essentially, the initial classification problem in their study underwent a shift to become an anomaly detection challenge. The objective was to alleviate the strain on hospital resources by identifying which infected individuals necessitated hospitalization and which could recover without such intervention. The researchers collected an actual dataset of personal epidemiological data in China. Every instance in this dataset corresponded to a person who had a COVID-19 positive test which were 2,310,111 cases. Then, a shorter database was employed by the researchers, and they differentiated both of the datasets by their size value as a larger dataset and a shorter dataset. After dataset preprocessing and training, they utilized machine learning algorithms such as Logistic regression, SVM, and random forest on both a larger and a shorter dataset for comparative analysis. The results demonstrated accuracy levels exceeding 0.9, while sensitivity fell below the 0.4 threshold. The autoencoder method, with the highest area under the curve for sensitivity and precision, exhibited superior performance, successfully predicting 36 out of 37 fatalities in the testing dataset, yielding a sensitivity ratio of 97% when employing a researcher-specified threshold (Li *et al.*, 2020).

The goal of this research is to use SVM classification to create a model that can determine if an individual is impacted by COVID-19. After carefully examining every potential infection scenario based on the symptoms, the dataset was constructed appropriately. It is challenging to make an accurate prognosis because those who do not exhibit any COVID-19 symptoms or indications are also impacted. The virus's nature is mostly unknown, and its symptoms are constantly changing and mutating. Given the severeness of this infection, only the most prevalent and serious symptoms have been included in the dataset's structuring. Since the patient's symptoms have been leaning toward

COVID-19 infection, the class of minimally infected people has been maintained under observation. Although the non-infected class indicates that the patient is temporarily safe from harm, it does not mean that they are immune to the COVID-19 chains. To assess the efficacy of supervised machine learning models, kNN, Nave Bayes, RF, AdaBoost, Binary Tree, and SVM were trained, tested, and compared. The model evaluation findings showed that SVM performs better than any other tested model. They have made predictions using metrics such as precision, recall, F1Score, area under the ROC, and CA. Additionally, the confusion matrix compiled for every model demonstrated SVM's advantage in accurately predicting COVID-19 (Guhathakurata *et al.*, 2021).

In this study, the researchers introduce an intelligent computer-aided model designed to automatically identify positive cases of COVID-19 for daily clinical applications. The model, based on a convolutional neural network (CNN) architecture, autonomously extracts discriminative features from chest X-ray images through convolution with diverse filter families, abstraction, and weight-sharing characteristics. Unlike the commonly employed transfer learning approach, this deep CNN model is trained from scratch, featuring a unique serial network with five convolution layers serving as a deep feature extractor. The extracted features are then inputted into machine learning algorithms, including k-nearest neighbor, support vector machine (SVM), and decision tree. Bayesian optimization algorithm is applied to optimize the hyperparameters of these machine learning models. The experiments are conducted on a public COVID-19 radiology database, divided into 70% training and 30% test sets. The SVM classifier emerges as the most effective, achieving an accuracy of 98.97%, sensitivity of 89.39%, specificity of 99.75%, and an F-score of 96.72%. Ultimately, this study presents an economical, rapid, and dependable intelligence tool for detecting COVID-19 infections, offering support to field specialists, physicians, and radiologists in decision-making processes (Nour *et al.*, 2020).

This paper proposes an automatic and unsupervised approach based on the hidden Markov model and Viterbi Algorithm for pleural line detection and localization in LUS data. Following pleural line localization, a supervised classification method employing the support vector machine (SVM) is carried out. The classifier evaluates the patient's overall health and, if appropriate, the severity of the pathology, assigning a score to each image in a particular LUS observation. Experiments were carried out on various LUS data from Italian hospitals, utilizing both linear and convex probes, to

demonstrate the efficacy of the suggested approach. The average overall accuracy in finding the pleura is 84% for convex probes and 94% for linear probes. SVM classification accuracy in assessing the severity of COVID-19-related pleural line changes is approximately 88% for convex probes and 94% for linear probes (Carrer *et al.*, 2020).

In this research, the researchers used a deep learning model, a subset of artificial intelligence, to detect Coronavirus. The dataset was divided into three categories: coronavirus, pneumonia, and standard X-ray images. The Fuzzy Color approach was used as an initial processing phase to rearrange the data classes, and the generated images were layered with the originals. Following that, the stacked dataset was trained with deep learning models, specifically MobileNetV2 and SqueezeNet, and the feature sets acquired from these models were processed using the social-mimic optimization approach. Following that, efficient characteristics were integrated and classified using Support Vector Machines (SVM). The proposed method had a classification rate of 99.27% overall. The study's findings imply that the model can successfully aid in the identification of COVID-19 disease (Toğaçar *et al.*, 2020).

This paper describes adaptable and unsupervised data-driven methods for detecting COVID-19 infection in blood test results. The authors approach the detection task as an anomaly detection problem with an unsupervised deep hybrid model. The model combines the feature extraction capacity of the variational autoencoder (VAE) with the detection sensitivity of the one-class support vector machine (1SVM) approach. The study used two sets of routine blood test samples from Albert Einstein Hospital in So Paulo, Brazil, and San Raffaele Hospital in Milan, Italy, to assess performance. A random forest regressor is used to impute missing values. The proposed VAE-based 1SVM detector outperforms existing models such as generative adversarial networks (GANs), deep belief networks (DBN), restricted Boltzmann machine (RBM)-based 1SVM, classic VAE, GAN, DBN, and RBM with SoftMax layer as discriminator layer, and solo 1SVM. The results show that the VAE-based 1SVM detector outperforms other standard deep learning models in terms of discriminating possible COVID-19 infections (Dairi *et al.*, 2022).

The study's goal was to create a machine-learning model that could predict crucial COVID-19 based on immune-inflammatory markers at the time of admission. A retrospective analysis was performed on 2076 consecutive COVID-19 patients from two Chinese hospitals who had significant outcomes (discharge or death) between January 27,

2020 and March 30, 2020. Admission to the intensive care unit, invasive ventilation, or death were all considered critical illnesses. The Least Absolute Shrinkage and Selection Operator (LASSO) was used to pick features. Using training, internal validation, and external validation datasets, five machine learning algorithms, Logistic Regression, Support Vector Machine, Gradient Boosted Decision Tree, K-Nearest Neighbor, and Neural Network were developed and assessed. To build the model, six features were chosen. With an area under the curve (AUC) of 0.991 (95% CI 0.979-1.000) in the internal validation cohort and 0.999 (95% CI 0.998-1.000) in the external validation cohort, the ensemble model, SPMCIIP (severity prediction model for COVID-19 by immune-inflammatory parameters), which combines SVM, GBDT, and NN, showed the best performance, forecasting critical COVID-19 occurrences about 20 days in advance (Gao *et al.*, 2021).

The study presents a deep learning-based approach to disease early detection. A dataset with three classes of COVID-19, normal, and pneumonia lung X-ray pictures was created, with each class containing 364 images. The image contrast enhancement technique was applied to the dataset during pre-processing, resulting in a new dataset. Deep learning models such as AlexNet, VGG19, GoogleNet, and ResNet were used for feature extraction. Two metaheuristic methods were used to choose the best features : binary particle swarm optimization and binary gray wolf optimization. The data was classified using Support Vector Machines (SVM) after the selected features were combined. The proposed method had an overall accuracy of 99.38% (Canayaz, 2021).

Another study utilizes the imagery dataset, suggests a classification system that uses radiomics features from CT chest images to distinguish patients with coronavirus disease 2019 (COVID-19) from those with other pneumonias. The RT-PCR nucleic acid test verified the presence of COVID-19 in 90 patients, while the other 90 patients had pneumonias of other kinds. The chest CT images were taken from both groups. To determine the region of interest (ROI) for pneumonias, data from both groups were manually delineated. The radionics approach extracted texturally, and histogram features from the ROI and produced a radionics features vector for each sample. After that, the data was divided into independent radiomic cohorts for training with 70 patients of each group, COVID-19 and other pneumonias and validation with 20 patients each of COVID-19 and other pneumonias using support vector machine (SVM). During training, the model was subjected to 20 rounds of tenfold cross-validation, and final testing

was performed on the independent validation cohort. A correlation study (Bonferroni correction, $P = 0.05/7$) investigated relationships between textural/histogram features and blood laboratory test indices in COVID-19 patients. The final model performed well on the independent validation cohort, with 89.83% accuracy, 94.22% sensitivity, 85.44% specificity, and an AUC of 0.940. The findings show that radionics properties are very distinct, and the SVM model successfully identifies and classifies COVID-19 and other pneumonia. A correlation study demonstrated that some textural aspects were positively connected to WBC and NE, while others were negatively related to SPO2H and NE, demonstrating radiomic features' potential to categorize patients with different pneumonias, including COVID-19 (Yang *et al.*, 2021).

Moreover, SVM applications on SARS-CoV-2 are numerous, and some analysts use the technique while focusing on the bio-molecular level (Kowalewski and Ray, 2020; Dutta *et al.*, 2021; Mekni *et al.*, 2021).

Time series is a type of machine learning approach that predicts target values based on a prior history of target values. It is a particular kind of regression where a list of target values serves as the input to time series analysis. In order to understand how infectious disease spreads and to provide better solutions to the problem, time series regression modeling and forecasting are important. The basic objective of time-series regression modeling is to meticulously and completely gather historical data on a time scale in order to construct the best model that can precisely explain the natural structure of the series. The future values of the series are then predicted using the established model. As a result, forecasting using the time series model that was created can be seen as a process where the past predicts the future (Raicharoen *et al.*, 2003).

To increase the accuracy of the disease prediction, many researchers have created time series models over the years. Various investigators have recommended a number of vital models for improving the precision and forecasting of time series models. Therefore, the primary time series models used for forecasting infectious disease are AR (Auto regressive time series models), MA (Moving Average), ARMA (Auto Regressive Moving Average), ARIMA (Auto Regressive Integrated Moving Average), and SARIMA (Seasonal Auto-Regressive Integrated Moving Average).

The Yule-developed autoregressive time series regression model of order p is de-noted by $AR(p)$ (Yule, 1927). It is derived as,

$$Y_t = \sum_{r=1}^p \theta_r Y_{t-r} + \varepsilon_t \quad (6)$$

where the values of $\theta_1, \theta_2, \dots, \theta_n$ are fixed. The error involved in the model is denoted by ε_t associated with the observations for Y_t the time point t and which are independently distributed with mean = 0 and rigid variance σ^2 .

Whereas, the first order autoregressive model AR (1) is given as,

$$Y_t = \theta_1 Y_{t-1} + \varepsilon_t \quad (7)$$

Hence, based on prior values and current residuals, the AR model provides linearly interpreted present values.

The moving Average (MA) model was developed by Eugen in 1937 and named as the moving average time series regression model (Eugen, 1937). When Y_t depends on the random error term ε_t and its lagged values, which is assumed to be white noise as ε_t is independently and identically distributed with mean = 0 and fixed variance σ_2 . So, a moving average model with q -lagged values of error terms is denoted by MA(q) and it can be written as,

$$Y_t = \sum_{s=1}^q \phi_s \varepsilon_{t-s} \quad (8)$$

where, $\phi_0 = 1$ and values of $\phi_1, \phi_2, \dots, \phi_q$ are fixed. Whereas, the error terms ε_t are supposed to be white noise.

Thus, the MA(q) model, known as the moving average model of order q , linearly interprets the time series current value using prior data and time series residuals.

The AR (p) model and the MA(q) model are combined to produce the ARMA model of order (p, q), known as the auto-regressive moving average model. In the AR component, the variable is regressed for its own lagged values. The MA component accounts for error components that occur simultaneously at various historical times. It is because of the MA, or moving average, and its behavior that ARMA is ideal for unobserved disasters, such as in the case of a pandemic. Below is a presentation of the World's ARMA (p, q) model,

$$Y_t = \sum_{r=1}^p \theta_r Y_{t-r} + \sum_{s=1}^q \phi_s \varepsilon_{t-s} \quad (9)$$

Where, $\theta_1, \theta_2, \dots, \theta_p$ and $\phi_0, \phi_1, \phi_2, \dots, \phi_q = 1$ and the error terms ε_t is white noise.

A generalization of the ARMA model is the ARIMA model, which was proposed by Box *et al.* in 1994 and is known as the auto-regressive integrated moving average model (Box *et al.*, 1994). In certain situations when data indicate traces of mean non-stationarity, ARIMA models are being used, and an initial differencing step (according to the “integrated” component of the model) can be

applied one or more times to eradicate the non-stationarity of the mean function.

The letter I for “integrated” denotes that the differences between the new values and the old values have been used in place of the data values and this differencing process may have been performed more than once.

First order difference Z_t of the Y_t series will be,

$$Z_t = Y_t - Y_{t-1} \quad (10)$$

Where $t = 2, 3, \dots, n$. Next, the second-order difference Z_t of the Y_t series will be,

$$Z_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \quad (11)$$

Where $t = 3, 4, \dots, n$.

ARIMA model with order p, d, q is written as ARIMA (p, d, q) where d denotes the degree of differencing.

SARIMA model is a special form of ARIMA model with an additional component that is ‘seasonality’. It is appropriate for time series with seasonality. The SARIMA model is denoted by ARIMA (P, D, Q)_m, where ‘m’ represents observations per year.

The Box Jenkins technique, which consists of Identification, Estimation, and Diagnostic Checking, enables SARIMA model to identify patterns through seasonality differencing, auto-regression (AR), and moving average (MA) for forecasting.

Several studies have employed these time series regression models in conjunction with cutting-edge computational methods to anticipate various viral illnesses and to get a better knowledge of these diseases. A few of which are discussed below.

Since the Governments and health policy makers can benefit with the aid of estimating the number of potential patients and death cases by making essential policies and imposing limitations to minimize frequency of the disease.

A group of researchers from Iran conducted a comparison of Time Series forecasting methods using nine models such as Auto-regressive Integrated Moving Average (ARIMA), Hybrid, Holt-Winter (HW), Bayesian Structural Time Series (BSTS), Neural Network Auto-Regression (NNETAR), Trigonometric Box-Cox ARMA errors Trends and Seasonal components (TBATS), Multilayer Perceptron (MLP), Extreme Learning Machine (ELM) network, and Prophet models to obtain the optimal model for estimating the number of confirmed infected cases and death cases for the following 30 days. Their study concluded that the MLP network, who showed to be the best model

with reduced forecasting error on unseen data of confirmed cases, can more reliably anticipate confirmed cases in the future than other models. Whereas, The Holt-Winter model, which has a smaller forecasting error and can be used to anticipate death cases, may also be used to forecast death cases (Talkhi *et al.*, 2021). Contrarily, Purwandari, along with fellow researchers, conducted a study in Malaysia for the application of the Time Series approach using different Neural Network Techniques to find the best model for the number of confirmed cases, recoveries, and deaths. These researchers precisely used Neural Network Auto-Regressive (NNETAR), Multi-Layer Perceptron (MLP), and Extreme Learning Machine (ELM). This study identified that ELM model is better than MLP for the testing data (Purwandari *et al.*, 2022).

The forecasting techniques that use monthly typhoid fever incidence were thoroughly analyzed by some researchers. The seasonal auto-regressive integrated moving average (SARIMA) was compared to back propagation neural networks (BPNN), radial basis function neural networks (RBFNN), and Elman recurrent neural networks model (ERNN). The advantages and disadvantages of the SARIMA model and neural networks were discussed, along with their differences. Data for modeling and forecasting purposes were given by the Chinese Center for Disease Control and Prevention for the years 2005 through 2009 and 2010. MSE, MAE, and MAPE were used to assess the performances. Their analysis revealed that for both modeling and forecasting, RBFNN had the lowest MAE, MAPE, and MSE (Zhang *et al.*, 2013). Using ARIMA and PROPHET, the confirmed, dead, and recovered COVID-19 cases in Indonesia over the course of a 30-day period were forecasted. For this research, projecting began in April 2020, and it was completed in May 2020. For PROPHET, they used the pre-tuned default settings, however for ARIMA, they just adjusted the order of p , d , and q . Utilizing ARIMA was challenging since getting the best fit required adjusting the p , d , and q orders. Although they were able to make the dataset steady enough that order d could be considered as 1, ARIMA consistently performed poorer than PROPHET in terms of predicting accuracy for this research (Aditya Satrio *et al.*, 2021).

Using daily basis data of the confirmed cases globally from the World Health Organization (WHO) website, the Autoregressive Integrated Moving Average (ARIMA) model was used to evaluate the dynamic behavior of COVID-19's global spread across the time period from January 22, 2020, to April 7, 2020. Over the course of three months, a total of 9,189,262 to 14,906,483 validated

COVID-19 patients were forecasted worldwide (Chyon *et al.*, 2022). The same machine learning model (ARIMA model) has been built using time series analysis with the goal of predicting new instances in India in the next days. The data collection for this study has covered the period from January 1 to July 31, 2020. While the analysis was done using the Python. Hence, the study showed when the results were compared to the AR model for predicting new cases in India, the ARIMA model was demonstrated to be the best-fit model (Kulshreshtha and Garg, 2021).

A team of researchers conducted a comparative analysis between the autoregressive integrated moving average (ARIMA) model and the eXtreme Gradient Boosting (XGBoost) model to gauge their accuracy in predicting the incidence of COVID-19 in the United States.

The daily COVID-19 data for the United States spanning from December 13, 2020, to June 30, 2021, underwent division into training data from December 13, 2020, to June 16, 2021 and validation sets from June 17, 2021, to June 30, 2021. The models were developed using the training dataset and subsequently assessed on the validation set for performance evaluation employing three accuracy metrics-mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The study revealed that the XGBoost model consistently outperformed the ARIMA model for both the training and validation sets. Demonstrated by lower MAE, RMSE, and MAPE values, the XGBoost model exhibited enhanced accuracy in predicting the number of COVID-19 cases in the USA. In conclusion, the study affirmed that the XGBoost model surpassed the ARIMA model, thereby contributing to an improved prediction of COVID-19 cases in the United States. Additionally, the research incorporated vaccination data from the USA into the XGBoost model and took into account the seasonality of data in both models (Fang *et al.*, 2022).

This article tackles challenges encountered in the analysis of infectious diseases and weather patterns using time series regression. The issues discussed include variations in immune population, strong auto-correlations, diverse lag structures, association patterns, seasonality adjustments, and over-dispersion. The proposed solutions involve adapting traditional regression methods with modifications like using past cases as proxies for the immune population and employing logarithms of lagged disease counts to control auto-correlation. Additionally, alternative approaches such as mathematical modeling, wavelet analysis, and ARIMA models are introduced and illustrated using datasets from cholera cases and rainfall in

Bangladesh, as well as influenza and temperature in Tokyo. The article underscores the importance of considering biological mechanisms, exploring lag structures with distributed lag non-linear models, and using alternative distribution models for over-dispersed data (Imai *et al.*, 2015).

In an attempt to predict the disease's trend in Indonesia and estimate the approximate time when normalcy will return, a team of researchers utilized machine learning models. Utilizing a dataset downloaded from the Kaggle website and recorded from January 20, 2020 to May 21, 2020, this study compares the accuracy and performance of Facebook's Prophet Forecasting Model and ARIMA Forecasting Model to show the trends of the disease's time series four weeks in advance. The dataset contains 27618 rows and 8 columns. This data's attributes include: Serial Number, Observation Date, Province/State, Country/Region, Last Update, Verified, Deaths, and Recuperated. The forecast models were subsequently compared to the most recent two weeks of actual data to see how they stacked up against one another. They utilized the pre-set, automatically tuned settings for PROPHET. The order of p , d , and q are only the three hyperparameters for ARIMA that have been adjusted. Early on, the PROPHET-predicted data exhibits remarkable accuracy with very little deviation from the actual data. But as time goes on, the disparities tend to widen and become more apparent. However, both the actual and PROPHET results show an upward tendency; however, PROPHET displays a more linear increase in the recovered instances. The most noticeable difference between the actual data and PROPHET's prediction is in the recovered cases section. ARIMA was a bit difficult to utilize. Selecting the optimal fit while adjusting the p , d , and q order is crucial. They did, however, manage to make the dataset stationary so that order d can be regarded 1, although ARIMA continues to have a lower accuracy than PROPHET even since the start of forecasting. Therefore, the outcome demonstrates that Prophet generally performs better than ARIMA, even if the more days it anticipates, more substantially it deviates from the real data (Aditya Satrio *et al.*, 2021). As shown in Table 1.

Conclusion and Future Perspectives

Since the size of the dataset has a significant impact on the effectiveness of any ML algorithm. One of the most important challenges with ML and COVID-19 systems is the lack of consistent datasets. To ensure that ML produces correct results, particularly in terms of classification

accuracy, larger databases are required as data grows in volume. One of the reasons for having an insufficient dataset is because the data includes information on an individual's personal activities like images from MRI, X-ray, or CT scans, travel history, and daily activities during the COVID-19 pandemic. These algorithms can then be trained using this data for virus prediction, diagnostic management, and vaccine development. However, due to privacy and security concerns, people hesitate to provide their information to others until it is explicitly announced or requested.

Further, it is necessary to advance ML techniques to boost the precision and reliability of data analytics in order to enable better COVID-19 identification and treatment. As ML, in combination with other innovative strategies, could offer effective COVID-19 substitutes. Several ML-based systems do not incorporate cross-validation methods, which are necessary to ensure that the findings can be applied to other databases.

In order to control the pandemic, a lot of coordination and collaboration between industry and academics is required in terms of potential discoveries to use ML. Since ML has been acknowledged as a fantastic technique for creating smart solutions to battle the epidemic. Thus, it is applied by many researchers to reduce pandemics in a number of areas, such as patient monitoring, imaging techniques, diagnosis and treatment, medication development, and viral spread tracking etc. The authors of these researches hope that the results will help advance machine learning (ML) strategies in real world settings. But the problem is the bulk of data sources are available online and most of them were generated by academics and researchers through fusing data from many sources and systems and yet no standard database has been compiled for COVID-19. As a results, a variety of COVID-19 datasets have been employed by the researchers for various applications of ML that are reviewed in the literature. Thus, it is hard to decide which model is better for COVID-19 prediction, prevention as well as detection because different researchers have used datasets from different resources (Qin *et al.*, 2011; Ismail *et al.*, 2017; Azemin *et al.*, 2022; Buttia *et al.*, 2023).

Acknowledgement

We would like to articulate our whole heartedly appreciation to the Institute of Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi and College of Computing, Informatics and Media Universiti Teknologi MARA (UiTM) for their high gratihitude support and assistance.

Table 1. Compares past research on the extensive use of SVM and Time Series Machine Learning models as well as enhanced algorithms of these techniques to estimate COVID-19 instances from actual data

References	Data Source	Methods	Limitations
Alshanbari <i>et al.</i> , 2022	100 patients' data from Prince Mohammed Bin Abdulaziz Hospital, Saudi Arabia, from May 2020 to January 2021 data	LDA and SVM (linear, polynomial, radial, and weighted radial kernels)	The sample size was limited. Ignored various factors such as location and culture etc. data 1
Saba <i>et al.</i> , 2021	Dataset from January 22, 2020 to September 30, 2020 of nine different countries from: https://github.com/CSSEGISandData	Random forests, K-nearest neighbors, SVM, Decision trees, polynomial regression, Holt winter, ARIMA, SARIMA	All models performed differently for selected countries. There is no single model that outperformed other models in all datasets. Only Asian countries were selected.
Rustagi <i>et al.</i> , 2022	COVID-19 database 'Our World in Data' from February 24, 2020, to September 26, 2021	Linear Regression, Polynomial Regression, and SVM	Dataset contained null values. Analysis might get better predictions with person-level data or by trying different datasets available online
Al-Jumaili, <i>et al.</i> , 2021	Covid19 ultrasound original dataset from: https://www.kaggle.com/bachaboos/covid19-ultrasound-original-dataset	CNN and SVM	Genetic data missing
Sheela and Arun, 2022	Used a custom dataset of 200 MRI images	hybrid PSO-SVM algorithm	Genetic data missing results are limited to Iran only
Ebrahim <i>et al.</i> , 2021	https://www.kaggle.com/marianarfranklin/mexico-covid-19-clinical-data/metadata	naive Bayes, logistic regression, decision tree, SVM, and ANN	No information provided regarding on what basis the features were being selected and data was reduced from 41 columns to 10 columns
Xiong <i>et al.</i> , 2022	the JinYinTan Hospital from January 26 to March 28, 2020	RF, SVM, and LR	Small dataset, with incomplete or missing information, such as smoking and drinking, and educational level),
Guleria <i>et al.</i> , 2022	January-May 2021 dataset from: https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-india	ensemble methods (AdaBoost, Random Forest, Bagging), Decision Trees, Gaussian Naïve Bayes, and Support Vector Machine (SVM).	Small dataset, no deep learning-based models has been employed or compared to verify results
Li <i>et al.</i> , 2020	two datasets obtained from: (https://github.com/beoutbreakprepared/nCoV2019), and (https://datarepositary.wolframcloud.com/resources/Patient-Medical-Data-for-Noel-	Auto-encoder model, Logistic regression, SVM, and Random forest	Insufficient information about the patients, importantly the data didn't include if patient received medical care or not
Guhathakurata <i>et al.</i> , 2021	A multicriteria dataset based on the most common COVID-19 symptoms has been generated with only 200 records and 8 attributes	kNN, Naïve Bayes, RF, AdaBoost, BinaryTree, and SVM	Very small study design. Lacks the applicability on bigger appropriate dataset.
Nour <i>et al.</i> , 2020	Dataset was combined from 4 different sources: https://arxiv.org/abs/2003.13145 . https://www.sirm.org/category/senza-categoria/covid-19/ . https://arxiv.org/pdf/2003.11597.pdf . http://dx.doi.org/10.1016/j.cell.2018.02.010 .	CNN, k-NN, SVM, and decision tree	Dataset was small, no information was provided that how different samples were joined together to create one dataset
Carrer <i>et al.</i> , 2020	Italian COVID-19 Lung Ultrasound Data Base (ICLUSDB). Available: https://covid19.disi.unitn.it/iclusdb/login	hidden Markov model and the Viterbi Algorithm, SVM	learning phase of the algorithm takes hours even while working offline. real-time implementation is required
Gao <i>et al.</i> , 2021	January 27, 2020 and March 30, 2020. Data source not provided	Logistic Regression, SVM, Gradient Boosted Decision Tree (GBDT), KNN, and Neural Network (NN)	limited diversity in patient populations
Yang <i>et al.</i> , 2021	Data source not provided	SVM utilizing radiomics features	lack of a broader and more diverse dataset limits the generalizability of proposed technique
Talkhi <i>et al.</i> , 2021	COVID-19 dataset for IRAN from February 20 to August 15, 2020. https://www.worldometers.info/coronavirus/	ARIMA, ELM, MLP, NN, ETAR, , Hybrid, Holt-Winter, BSTS, TBATS, and Prophet model	Models performed poorly in trained data. Conclusion was based on test data without the inclusion of other fac-tors such as age, gender, clinical history etc.
Purwandari <i>et al.</i> , 2022	COVID-19 Cases Data by Humanitarian Data Exchange from January 22, 2020, to June 13, 2020.	Neural Network Auto Regressive Multi-Layer Perceptron, , and Extreme Learning Machine	Factors which influenced the forecast value for the number of cases have not included like government interventions, lockdown, community's co-operations.
Aditya Satrio <i>et al.</i> , 2021	COVID-19 dataset from Kaggle website starting from January 20, 2020 to May 21, 2020 the World Health Organization (WHO) website (Novel Corona-virus (COVID-19) Cases Data - Hu-manitarian Data Exchange, 2020). Data from January 22, 2020 to April 7, 2020.	ARIMA and PROPHET	Both models did not perform well.
Chyon <i>et al.</i> , 2022	(COVID-19) Cases Data - Hu-manitarian Data Exchange, 2020). Data from January 22, 2020 to April 7, 2020.	ARIMA	Limited data was taken and Prediction was made on the as-sumption that data remain relatively sta-ble
Kulshreshtha and Garg, 2021	Dataset from January 1, 2020 to July 31, 2020	AR and ARIMA	Predictions are not ac-curate as the mean is not equal to zero for residuals.
Fang <i>et al.</i> , 2022	USA COVID-19 dataset from December 13, 2020, to June 30, 2021, using this website: https://COVID-19.cdc.gov	XGBoost, and ARIMA	Proposed technique was not validated through applying on other datasets
Imai <i>et al.</i> , 2015	Not provided	Timeseries regression, mathematical modeling, wavelet analysis, and ARIMA models	further development and evaluation of their technique's validity is needed
Aditya Satrio <i>et al.</i> , 2021	From January 20, 2020 until May 21 2020, https://www.kaggle.com/datasets/	ARIMA and PROPHET,	ignored the crucial factors in determining how the disease develops

Funding

This study was supported by the Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi and Universiti Teknologi MARA (UiTM), Malaysia with grant number 600-RMC/GSS 5/3 (056/2022).

References

- Aditya Satrio, C.B., Darmawan, W., Nadia, B.U., and Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, 179:524-532. <https://doi.org/10.1016/j.procs.2021.01.036>
- Ali, L., Niamat, A., Khan, J.A., Golilarz, N.A., Xingzhong, X., Noor, A., Nour, R., and Bukhari, S.A.C. (2019). An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. *IEEE Access*, 7:54,007-54,014. <https://doi.org/10.1109/ACCESS.2019.2909969>
- Al-Jumaili, S., Duru, A.D., and Ucan, O.N. (2021). COVID-19 Ultrasound image classification using SVM based on kernels deduced from Convolutional neural network. *ISMSIT 2021 - 5th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, p. 429-433. <https://doi.org/10.1109/ISMSIT52890.2021.9604551>
- Alshanbari, H.M., Mehmood, T., Sami, W., Alturaiki, W., Hamza, M.A., and Alosaimi, B. (2022). Prediction and Classification of COVID-19 Admissions to Intensive Care Units (ICU) Using Weighted Radial Kernel SVM Coupled with Recursive Feature Elimination (RFE). *Life*, 12(7):1,100. <https://doi.org/10.3390/life12071100>
- Azemin, M.Z.C., Tamrin, M.I.M., Ali, M.A.M., and Jamaludin, I. (2022). Unsupervised Chest X-ray Opacity Classification using Minimal Deep Features. *International Journal of Advanced Computer Science and Applications*, 13(3):259-262. <https://doi.org/10.14569/IJACSA.2022.0130332>
- Box, G., Jenkins, G., and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*. New Jersey: Prentice Hall.
- Buttia, C., Llanaj, E., Raeisi-Dehkordi, H., Kastrati, L., Amiri, M., Meçani, R., Taneri, P. E., Ochoa, S. A. G., Raguindin, P. F., Wehrli, F., Khatami, F., Espínola, O. P., Rojas, L. Z., de Mortanges, A. P., Macharia-Nimietz, E. F., Alijla, F., Minder, B., Leichtle, A. B., Lüthi, N., Ehrhard, S., Que, Y.-A., Fernandes, L.K., Hautz, W., and Muka, T. (2023). Prognostic models in COVID-19 infection that predict severity: a systematic review. *European Journal of Epidemiology*, 38(4):355-372. <https://doi.org/10.1007/s10654-023-00973-x>
- Canayaz, M. (2021). MH-COVIDNet: Diagnosis of COVID-19 using deep neural networks and meta-heuristic-based feature selection on X-ray images. *Biomedical Signal Processing and Control*, 64:102257. <https://doi.org/10.1016/j.bspc.2020.102257>
- Carrer, L., Donini, E., Marinelli, D., Zanetti, M., Mento, F., Torri, E., Smargiassi, A., Inchingolo, R., Soldati, G., Demi, L., Bovolo, F., and Bruzzone, L. (2020). Automatic Pleural Line Extraction and COVID-19 Scoring from Lung Ultrasound Data. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(11):2,207-2,217. <https://doi.org/10.1109/TUFFC.2020.3005512>
- Centers for Disease Control and Prevention. (2022). SARS-CoV-2 Variant Classifications and Definitions. Available from: <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html>. Accessed date: September 20, 2022.
- Chyon, F.A., Suman, M.N.H., Fahim, M.R.I., and Ahmmed, M.S. (2022). Time series analysis and predicting COVID-19 affected patients by ARIMA model using machine learning. *Journal of Virological Methods*, 301:114433. <https://doi.org/10.1016/j.jviromet.2021.114433>
- Cortes, C. and Vapnik, V.N. (2004). Support-Vector Networks. *Machine Learning*, 20(3):273-297. <https://doi.org/10.1007/BF00994018>
- Cucinotta, D. and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Biomedica*, 91(1):157-160. <https://doi.org/10.23750/abm.v91i1.9397>
- Cucinotta, D. and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Biomedica*, 91(1):157-160. <https://doi.org/10.23750/abm.v91i1.9397>
- Dairi, A., Harrou, F., and Sun, Y. (2022). Deep Generative Learning-Based l-SVM Detectors for Unsupervised COVID-19 Infection Detection Using Blood Tests. *IEEE Transactions on Instrumentation and Measurement*, 71:1-11. <https://doi.org/10.1109/TIM.2021.3130675>
- Dutta, K., Elmezayen, A.D., Al-Obaidi, A., Zhu, W., Morozova, O.V., Shityakov, S., and Khalifa, I. (2021). Seq12, Seq12m, and Seq13m, peptide analogues of the spike glycoprotein shows antiviral properties against SARS-CoV-2: An in silico study through molecular docking, molecular dynamics simulation, and MM-PB/GBSA calculations. *Journal of Molecular Structure*, 1246:131113. <https://doi.org/10.1016/j.molstruc.2021.131113>
- Ebrahim, L.J.M., Sani, A.A., Usman, S., Ahmad, A., Chakraborty, C., and Mohammed, I.A. (2021). Supervised Machine Learning Models for Prediction of COVID - 19 Infection using Epidemiology Dataset. *SN Computer Science*, 2(1):1-13. <https://doi.org/10.1007/s42979-020-00394-7>
- Eugen, S. (1937). The Summation of Random Causes as the Source of Cyclic Processes. *Econometrica*, 5(2):105-146. <https://doi.org/10.2307/1907241>
- Fang, Z.G., Yang, S.Q., Lv, C.X., An, S.Y., and Wu, W. (2022). Application of a data-driven XGBoost model for the prediction of COVID-19 in the USA: a time-series study. *BMJ Open*, 12(7):1-8. <https://doi.org/10.1136/bmjopen-2021-056685>
- Gao, Y., Chen, L., Chi, J., Zeng, S., Feng, X., Li, H., Liu, D., Feng, X., Wang, S., Wang, Y., Yu, R., Yuan, Y., Xu, S., Li, C., Zhang, W., Li, S., and Gao, Q. (2021). Development and validation of an online model to predict critical COVID-19 with immune-inflammatory parameters. *Journal of Intensive Care*, 9(1):1-12. <https://doi.org/10.1186/s40560-021-00531-1>
- GISAID. (2023). Mutation Tracker for genome-wide analysis, Available from : https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html. Accessed date : December 24, 2023.
- Guhathakurata, S., Kundu, S., Chakraborty, A., and Banerjee, J.S. (2021). A novel approach to predict COVID-19 using support vector machine. *ELSEVIER*, 68(1):1-12. <https://doi.org/https://doi.org/10.1016/B978-0-12-824536-1.00014-9>
- Guleria, P., Ahmed, S., Alhumam, A., and Srinivasu, P.N. (2022). Empirical Study on Classifiers for Earlier Prediction of COVID-19 Infection Cure and Death Rate in the Indian States. *Healthcare (Switzerland)*, 10(1):85. <https://doi.org/10.3390/healthcare10010085>
- He, F. J., Tan, M., Song, J., and MacGregor, G.A. (2020). Salt substitution to lower population blood pressure. *Nature Medicine*, 26(3):313-314. <https://doi.org/10.1038/s41591-020-0784-9>

- He, F.J., Tan, M., Song, J., and MacGregor, G.A. (2020). Salt substitution to lower population blood pressure. *Nature Medicine*, 26(3):313-314. <https://doi.org/10.1038/s41591-020-0784-9>
- Hodcroft, E. (2022). CoVariants, Available from: <https://covariants.org/>. Accessed date: September 19, 2022.
- Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., and Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, 142:319-327. <https://doi.org/10.1016/j.envres.2015.06.040>
- Ismail, K.A., Abdul Majid, M., Mohamed Zain, J., and Abu Bakar, N.A. (2017). Big Data prediction framework for weather Temperature based on MapReduce algorithm. *ICOS 2016 - 2016 IEEE Conference on Open Systems*, p. 13-17. <https://doi.org/10.1109/ICOS.2016.7881981>
- Khan, M.A., Khan, R., Algarni, F., Kumar, I., Choudhary, A., and Srivastava, A. (2022). Performance evaluation of regression models for COVID-19: A statistical and predictive perspective. *Ain Shams Engineering Journal*, 13(2):101574. <https://doi.org/10.1016/j.asej.2021.08.016>
- Kowalewski, J. and Ray, A. (2020). Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space. *Heliyon*, 6(8):e04639. <https://doi.org/10.1016/j.heliyon.2020.e04639>
- Kulshreshtha, V. and Garg, N.K. (2021). Predicting the New Cases of Coronavirus [COVID-19] in India by Using Time Series Analysis as Machine Learning Model in Python. *Journal of The Institution of Engineers (India): Series B*, 102(6):1303-1309. <https://doi.org/10.1007/s40031-021-00546-0>
- Li, Y., Horowitz, M.A., Liu, J., Chew, A., Lan, H., and Liu, Q. (2020). Individual-Level Fatality Prediction of COVID-19 Patients Using AI Methods. 8(September):1-12. <https://doi.org/10.3389/fpubh.2020.587937>
- Mekni, N., Coronello, C., Langer, T., Rosa, M., and Perricone, U. (2021). Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors. *International Journal of Molecular Sciences*, 22(14):7714. <https://doi.org/10.3390/ijms22147714>
- Mofijur, M., Fattah, I.M.R., Alam, M.A., Islam, A.B.M.S., Ong, H.C., Rahman, S.M.A., Najafi, G., Ahmed, S.F., Uddin, M.A., and Mahlia, T.M.I. (2021). Impact of COVID-19 on the social, economic, environmental and energy domains: Lessons learnt from a global pandemic. *Sustainable Production and Consumption*, 26(January):343-359. <https://doi.org/10.1016/j.spc.2020.10.016>
- Nour, M., Cömert, Z., and Polat, K. (2020). A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization. *Applied Soft Computing*, 97(Part A):106580. <https://doi.org/10.1016/j.asoc.2020.106580>
- Purwandari, T., Zahroh, S., Hidayat, Y., Sukono, Mamat, M., and Saputra, J. (2022). Forecasting model of COVID-19 pandemic in Malaysia: An application of time series approach using neural network. *Decision Science Letters*, 11(1):35-42. <https://doi.org/10.5267/j.dsl.2021.10.001>
- Qin, H., Ma, X., Herawan, T., and Zain, J.M. (2011). An Adjustable Approach to Interval Valued Intuitionistic. *Asian Conference on Intelligent Information and Database Systems*, p. 80-89. https://doi.org/10.1007/978-3-642-20042-7_9
- Raicharoen, T., Lursinsap, C., and Sanguanbhokai, P. (2003). Application of critical support vector machine to time series prediction. *Proceedings of the 2003 International Symposium on Circuits and Systems*, p. V-V. <https://doi.org/10.1109/ISCAS.2003.1206419>
- Rustagi, V., Bajaj, M., Tanvi, Singh, P., Aggarwal, R., AlAjmi, M.F., Hussain, A., Hassan, M.I., Singh, A., and Singh, I.K. (2022). Analyzing the Effect of Vaccination Over COVID Cases and Deaths in Asian Countries Using Machine Learning Models. *Frontiers in Cellular and Infection Microbiology*, 11(February):1-13. <https://doi.org/10.3389/fcimb.2021.806265>
- Saba, T., Abunadi, I., Shahzad, M.N., and Khan, A.R. (2021). Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types. *Microscopy Research and Technique*, 84(7):1462-1474. <https://doi.org/10.1002/jemt.23702>
- Sheela, M.S. and Arun, C.A. (2022). Hybrid PSO-SVM algorithm for Covid-19 screening and quantification. *International Journal of Information Technology (Singapore)*, 14(4):2049-2056. <https://doi.org/10.1007/s41870-021-00856-y>
- Talkhi, N., Akhavan Fatemi, N., Ataei, Z., and Jabbari Nooghabi, M. (2021). Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. *Biomedical Signal Pro-processing and Control*, 66(February):102494. <https://doi.org/10.1016/j.bspc.2021.102494>
- The New York Times. (2022). Covid-19 Vaccine Tracker, Available from: <https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html>. Accessed date : September 19, 2022.
- Toğaçar, M., Ergen, B., and Cömert, Z. (2020). COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Computers in Biology and Medicine*, 121:103805. <https://doi.org/10.1016/j.compbiomed.2020.103805>
- Worldometer covid-19 Coronavirus Pandemic, Available online <https://www.worldometers.info/coronavirus/>, Accessed on. 24/12/2023
- Worldometer covid-19 Malaysia, Available online <https://www.worldometers.info/coronavirus/country/malaysia/>, Accessed on. 24/12/2023
- Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., and Huang, L. (2022). Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious Diseases of Poverty*, 11(1):1-9. <https://doi.org/10.1186/s40249-022-00946-4>
- Yadav, S.K. and Akhter, Y. (2021). Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread. *Frontiers in Public Health*, 9(June):1-27. <https://doi.org/10.3389/fpubh.2021.645405>
- Yang, N., Liu, F., Li, C., Xiao, W., Xie, S., Yuan, S., Zuo, W., Ma, X., and Jiang, G. (2021). Diagnostic classification of coronavirus disease 2019 (COVID-19) and other pneumonias using radiomics features in CT chest images. *Scientific Reports*, 11(1):1-7. <https://doi.org/10.1038/s41598-021-97497-9>
- Yang, Z., Zhou, Y., and Gong, C. (2019). Diagnosis of diabetes based on improved support vector machine and ensemble learning. *ACM International Conference Proceeding Series*, p. 177-181. <https://doi.org/10.1145/3319921.3319954>
- Yule, G.U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A*, 226(636-646):267-298. <https://doi.org/10.1098/rsta.1927.0007>
- Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A.A., and Li, X. (2013). Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China. *PLoS ONE*, 8(5):e63116. <https://doi.org/10.1371/journal.pone.0063116>