Prediction System for Covid-19 Upcoming Cases Using Ensemble Classification

R.Sivarasan[a], Dr. M.S. Mythili[b]
[a] Research Scholar, Department of Computer Science, Bishop Heber College (Autonomous),
Affiliated to Bharathidasan University, Tiruchirappalli-620017, Tamil Nadu, India.
[b] Associate Professor, Department of Computer Applications, Bishop Heber College (Autonomous),
Affiliated to Bharathidasan University, Tiruchirappalli-620017, Tamil Nadu, India.

Abstract

An epidemic of the novel destructive Coronavirus has been spreading rapidly around the world since 2019 and has caused a great number of deaths. Providing patients with appropriate and most timely care is crucial to combating COVID-19 spread. Testing for the disease must be done quickly and accurately. Therefore, this paper developed an ensemble classification-based country-wise COVID-19 upcoming cases prediction model. This ensemble classification and prediction model shows the upcoming month's Corona virus cases, including newly confirmed cases, recovered cases, and deaths. This analysis is carried out based on these three cases occurring in different countries on sequential dates. The proposed model uses three famous classifiers, namely ANN, Gaussian Process and SVM which have different learning characteristics and architectures at the first stage. In the second stage, they combine their predictions with average calculations. Training and assessment of the proposed model were conducted using 75065 observations comprised of 61 features from John Hopkins University in Maryland. For data preparation, the envisioned work clusters the dataset based on world countries affected by COVID-19 separately. As a result, this set of clusters fetched data once again based on death, newly confirmed, and recovered cases. The experimental result shows the proposed ensemble model provides better performance when compared with previous classification algorithms.

Keywords: COVID-19, Clustering, Classification, Upcoming Prediction, Support Vector Machine, Gaussian Process Artificial Neural Network.

*Corresponding author Email : mythili.ca@bhc.edu.in

Introduction

A previously undiscovered disease was identified in Wuhan, China, and it is currently spreading throughout the world. It was confirmed in January 2020 that SARS-CoV-2 [1] caused the disease. The coronaviruses, a large family of positive-stranded RNA viruses, are the source of the virus. Coronaviruses are well-known for bringing about a variety of diseases, including the common cold and numerous pandemics like respiratory illness in the Middle East and severe acute respiratory syndrome [2]. Globally, the World Health Organization (WHO) classified COVID-19 as an epidemic in March 2020.

The transmission of the virus from animals to humans is the main source of its spread. However, the upcoming COVID-19 cases are not associated with the subjection process. As a result, it is concluded that the virus spreads among people, and people who are infected with the virus are the main people who transmit COVID-19 [3]. Prior to the onset of symptoms, it appears that there is very little risk of COVID-19 transmission. However, there is no way to stop this virus from spreading. Besides these, everyone's advice is that asymptomatic people can overcome the virus and be protected from it by social exclusion.

A person's sneezing and coughing is believed to be the main cause of the spread of viruses, such as rhinoviruses and influenza, along with wheezing bacteria. According to data analysis [4], close contact between two people was found to be the most significant factor in the SARS-CoV-2 spread in China [4].The majority of people who can spread the virus are members of a person's family, medical professionals, and others who have been in close proximity to them.

The appropriate ways to deal with the source of the disease are early diagnosis, isolation, reporting, supportive care, and prompt preparation of outbreak information to avoid unnecessary stress. The COVID-19 virus can be prevented by each individual through special hygiene, a shaped or appropriate mask, ventilation, and avoiding crowded areas [5].

Predictive models for early detection of the COVID-19 pandemic are required to assist epidemiologists massive global. As a result, this paper proposes an ensemble classification based on the COVID-19 upcoming case prediction model. Using the data that is available, It predicts how many people will become diagnosed with COVID-19, how many will die, and how many will recover. This ensemble model can easily predict the number of newly infected cases, the number of deaths patients, and the number of recovered case. The proposed model uses three famous classifiers: Gaussian Processes, ANN, and SVM. Initially, their learning characteristics and architectures differ; however, they eventually converge on an average calculation.

The remainder of the paper is divided into five sections. Section 2 is Related Works. An explanation of the proposed methodology is provided in Section 3. Section 4 presents the simulation results, and section 5 presents some observations.

Related Works

This section provides an overview of different machine learning strategies for early COVID-19 identification.

Banerjee et al. [6] analyzed four blood test-based machine learning methods for early COVID-19 patient screening. The Random Forest (RF), ANN, Lasso-elastic-net regularized generalized linear network (GLMNET), and Logistic Regression (LR) were the methods used. The methods were evaluated with the help of a dataset from Brazil's Albert Einstein Hospital, which

included 598 blood samples from 81 guaranteed COVID- 19 cases. Only 14 blood attributes were chosen to train and test the method out of the 108 attributes of dataset samples.

Bao et al. [7] SVM and RF techniques were looked at for the early identification of COVID-19 using standard blood tests. This dataset is made up of 294 blood samples that were collected from Kunshan People's Hospital and Wuhan Union Hospital in China.Viral vs the moderate, viral vs. severe, and mild vs. severe were the three categories of classification.There were a maximum of 15 blood attributes chosen for training.

Barbosa et al. [8] generated an inexpensive COVID-19 discovery scheme from usual blood samples using multiple ML multiple ML classifiers, such as neural networks (NB), Bayesian networks (BN), RF, RT, SVM, and multilayer perceptron. The data was gathered by the authors from Albert Einstein Hospital in Brazil, which has 5644 data samples and 559 guaranteed COVID-19 cases. From the dataset with 108 attributes, Evolutionary Search (ES) optimization and Particle Swarm Optimization (PSO) algorithms were utilized to decrease the attributes to 63 and 62, respectively. The features were reduced to 24 by hand in order to train and test their methods and cut costs and time associated with blood tests. Outcomes attained tremendous classification effectiveness. The experiment showed that BN had significantly higher efficacy compared with this section presents a review of some machine learning techniques for the early detection of COVID-19.

Banerjee et al. [6] analyzed four blood test-based machine learning methods for early COVID-19 patient screening. The Random Forest (RF), ANN, Lasso-elastic-net regularized generalized linear network (GLMNET), and Logistic Regression (LR) were the methods used. The methods were evaluated with the help of a dataset from Brazil's Albert Einstein Hospital, which included 580 blood samples from 81 guaranteed COVID- 19 cases. Only 14 blood attributes were chosen to train and test the method out of the 108 attributes of dataset samples.

Bao et al. [7] examined SVM and RF methods for the premature discovery of COVID-19 using usual blood tests. This dataset includes a collection of 294 blood samples taken from Wuhan Union Hospital and Kunshan People's Hospital in China for this dataset. The three types of classification were viral vs. moderate, viral vs severe and mild vs severe. A maximum of 15 blood attributes were selected for training.

Barbosa et al. [8] generated an inexpensive COVID-19 discovery scheme from usual blood samples using multiple ML classifiers, such as neural networks (NB), Bayesian networks (BN), RF, RT, SVM, and multilayer perceptron. The data was gathered by the authors from Albert Einstein Hospital in Brazil, which has 5644 data samples and 559 guaranteed COVID-19 cases. From the dataset with 108 attributes, Evolutionary Search (ES) optimization and Particle Swarm Optimization (PSO) algorithms were utilized to decrease the attributes to 63 and 62, respectively. The features were reduced to 24 by hand in order to train and test their methods and cut costs and time associated with blood tests. Outcomes attained tremendous classification effectiveness. The experiment showed that BN had significantly higher efficacy compared with other methods.

Batista et al. [9] suggested using machine learning algorithms and blood samples from emergency care to learn to forecast COVID-19 analysis. Five famous machine learning techniques (SVM, LR, gradient boosting trees (GBT), neural networks (NN), and RF) are used for classification. The data came from Brazil's Albert Einstein Hospital, which had 235 blood samples take and 102 confirmed cases of COVID. Only 15 characteristics were selected from the blood

samples to train and test the techniques. The most accurate prediction effectiveness was achieved through the SVM algorithm.

Bayat et al. [10] utilized a combination of standard laboratory tests and vital signs to implement a COVID-19 foretelling method that was based on RF. The dataset, including 1079 ensured COVID-19 5002 rows with 68 attributes, was gathered from sites all over the United States run by the Veterans Health Administration. The authors selected the 54 most significant features by pair wise correlation. The authors identified nine key attributes that could create a satisfactory accuracy level for the method. The foretelling method includes the capability of distinguishing COVID-19 patients against other respiratory virus infections, including influenza, seasonal human coronaviruses, and respiratory syncytial viruses. The fact that the method only trained older, mostly male patients was one of its limitations.

Brinati et al. [11] looked at a variety of ML classifier classes to find COVID-19 in typical blood samples. The authors looked at these approaches: K-nearest neighbours (KNN), highly randomized trees (ET), SVM, RF, Naive Bayes (NB), LR, and DT are all examples. In addition, to improve accuracy, the authors modified the RF algorithm to be a three-way RF classifier, using a collection of 279 representative blood samples taken from Italy's San Raffaele Hospital. There are 177 confirmed COVID-19 samples in the dataset; additionally, only 15 characteristics of blood samples were considered. The RF method was the most efficient classifier. Their approach has some drawbacks, including a small number of blood sampling, a single source of information with a high percentage of positive results, and a small amount of data.

Feng et al. [12] implemented An early detection method for the detection of COVID-19 on admission. We selected four types of classifier for the technique: decision trees (DT), Ridge with LR, and LASSO with LR regularization algorithms. The method's natural potency lies in the chosen contestant characteristics, including two infection-associated biomarkers and one admission-linked variable, 17 clinical signs and symptoms variables, 20 blood value variables, four vital sign variables, and two demographic data variables. The necessary attributes were gathered from 132 patients (26 positives) at the General Hospital of the People's Liberation Army in China for the dataset. Of the 46 features chosen to train the method, only 18 were used with LASSO. Based on the results, LASSO with LR achieved the highest level of efficiency. A combination of the most commonly available data attributes on admission can precisely identify COVID-19 as this learning strategy. External validation is necessary for this method to be successful because the information from a single center is limited.

Joshi et al. [13] implemented an LR method to forecast COVID-19 from patient sex and three blood count mechanisms. A dataset of 390 patient samples from Stanford Health Care served as the basis for the method's training. With datasets from a variety of locations, including South Korea, Chicago, Northern California, and Seattle, Washington, the training process was authentic.

Kukar et al. [14] Extreme gradient boosting (XGBoost), a machine learning (ML) technique, was chosen by Kukar et al. [14] over deep neural networks (DNN) and RF because of its superior efficiency, reduced computational requirements, and built-in ability to handle lost data. The dataset included 160 assured COVID-19 people involved from the University Medical Center Ljubljana in Slovenia, along with 5333 blood samples from patients with a range of viral and bacterial infections. Out of the 117 attributes in the dataset, only the 35 the well features were chosen for the method. The dataset of the process has a very low positive ratio (2.91%), making it challenging to evaluate the accuracy of the outcomes.

Langer et al. A number of ML methods were evaluated in [15] with the use of LR, RF, DT, and ANNs to anticipate COVID-19 patients in crisis departments utilizing essential radiological, and routine laboratory data and clinical. The data comes from one of the major hospitals in Milan, Italy, and includes 74 attributes for 199 people, 127 of whom have been confirmed to have COVID-19. In order to train the machine learning algorithms, the authors employed a dimensionality reduction algorithm to select 42 crucial dataset characteristics from 74. The high-quality selected clinical information, which is typically available in emergency departments, is what makes the research useful for making a quick decision to stop the disease from spreading. The research has a few limitations; for example, it requires a few epidemiological and clinical data, a small sample size and Single-Centre analysis, which might be useful for enhancing the method's accuracy other methods.

Batista et al. [9] suggested using machine learning algorithms and blood samples from emergency care to learn to forecast COVID-19 analysis. Five famous machine learning techniques (SVM, LR, gradient boosting trees (GBT), neural networks (NN), and RF) are used for classification. The information comes from the Albert Einstein Hospital in Brazil, which has had 235 samples taken and 102 confirmed instances of COVID-19. Only 15 characteristics were selected from the blood samples to train and test the techniques. The most accurate prediction effectiveness was achieved through the SVM algorithm.

Bayat et al. [10] utilized a combination of standard laboratory tests and vital signs to implement a COVID-19 foretelling method that was based on RF. The dataset, including 1079 ensured COVID-19 5002 rows with 68 attributes, was gathered from sites all over the United States run by the Veterans Health Administration. The authors selected the 54 most significant features by pairwise correlation. The authors identified nine key attributes that could create a satisfactory accuracy level for the method. The foretelling method includes the capability of distinguishing COVID-19 patients against other respiratory virus infections, including influenza, seasonal human coronaviruses, and respiratory syncytial viruses. The fact that the method only trained older, mostly male patients was one of its limitations.

Brinati et al. [11] looked at a variety of ML classifier classes to find COVID-19 in typical blood samples. The authors looked at these approaches: K-nearest neighbours (KNN), highly randomized trees (ET), SVM, RF, Naive Bayes (NB), LR, and DT are all examples. In addition, to improve accuracy, the authors modified the RF algorithm to be a three-way RF classifier, using a collection of 279 representative blood samples taken from Italy's San Raffaele Hospital. There are 177 confirmed COVID-19 samples in the dataset; additionally, only 15 characteristics of blood samples were considered. The RF method was the most efficient classifier. Their method has some limitations, such as a limited number of blood samples characteristics, a single information source with a high positive ratio, and the relatively small size of the data.

Feng et al. [12] implemented a foretelling method for early recognition of COVID-19 on admission. Four different types of classifiers were chosen for the technique. decision trees (DT), Ridge with LR, and LASSO with LR regularization algorithms. The method's natural potency lies in the chosen contestant characteristics, including two infection-associated biomarkers and one admission-linked variable, 17 clinical signs and symptoms variables, 20 blood value variables, four vital sign variables, and two demographic data variables. The necessary attributes were gathered from 132 patients (26 positives) at The dataset was provided by the People's Liberation Army General Hospital in China. Of the 46 features chosen to train the method, only 18 were used

with LASSO. Based on the results, LASSO with LR achieved the highest level of efficiency. A combination of the most commonly available data attributes on admission can precisely identify COVID-19 as this learning strategy. External validation is required for the success of this strategy due to the little amount of data gathered from a single centre.

Joshi et al. [13] implemented an LR method to forecast COVID-19 from patient sex and three blood count mechanisms. A dataset of 390 patient samples from Stanford Health Care served as the basis for the method's training. With datasets from a variety of locations, including South Korea, Chicago, Northern California, and Seattle, Washington, the training process was authentic.

Kukar et al. [14] decided to use the ML technique of Extreme Gradient Boosting (XGBoost) rather than Deep neural networks (DNN) or Random Forest due to its superior efficiency, fewer computational necessities, and its inherent capability to manage lost data. The dataset consisted of 5333 blood samples from patients with a variety of viral and bacterial infections, including 160 guaranteed COVID-19 participants from the University Medical Center Ljubljana, Slovenia. Only the 35 most well-known features were selected from the 117 attributes in the dataset for the method. The process's dataset has a very low positive ratio (2.91 percent), making it difficult to assess the quality of the results.

Langer et al. A number of ML methods were evaluated in [15] with the use of LR, RF, DT, and ANNs to anticipate COVID-19 patients in crisis departments utilizing essential radiological, clinical, and routine laboratory data. The data comes from one of the major hospitals in Milan, Italy, and includes 74 attributes for 199 people, 127 of whom have been confirmed to have COVID-19. In order to train the machine learning algorithms, the authors employed a dimensionality reduction algorithm to select 42 crucial dataset characteristics from 74. The high-quality selected clinical information, which is typically available in emergency departments, is what makes the research useful for making a quick decision to stop the disease from spreading.

The research has a few limitations; for example, it requires a few epidemiological and clinical data, a small sample size and single-centre analysis, which might be useful for enhancing the method’s accuracy.

Materials and Methods

The method used to implement the ensemble classification method is described in detail in this section. The dataset's description is explained in the first subsection. The data clustering procedure is explained below. The final subsection describes the proposed ensemble method's execution details.

A. Dataset

The dataset utilized in this investigation was received from Johns Hopkins University in Maryland from https://ourworldindata.org/coronavirus-source-data. Only 75065 samples of COVID-19 cases with 61 features from 13 February 2020 to 15 March 2021 are included in the dataset, which demonstrates the features utilized in this study. The dataset utilized in this study contains three target classes, namely:

Recovered cases: the number of patients recovered on a specific date. It might be decreased or increased depending on the location and date.

Death cases: The number of deaths on any given date. It could be decreased or increased with the addition of a location and date.

Newly confirmed cases: The number of newly confirmed.

B. **Data Clustering**

Dataset clustering is the method of separating the rows of a dataset into several collections. Rows in similar groups are more distinct from rows in other collections than they are from rows in the same collection. This proposed work applies to two types of clustering. They are,

(1) Case-wise clustering
(2) Country-wise clustering

1. Country-Wise Clustering

Country-wise clustering separates the rows of a dataset into many groups based on the country name. Because the dataset contains many rows, each row has a record of cases on any specific date. COVID-19 cases in numerous countries may decrease or increase as a result of this location and date. Therefore, country-wise clustering is necessary for predicting COVID-19 upcoming cases by country. Figure 1 shows country-wise clustering is necessary for each country's COVID-19 upcoming case prediction. Figure 1 shows country-wise clustering.

2. Cases-Wise Clustering:

After country-wise clustering comes case-wise clustering. Case-wise clustering separates data based on three target classes (newly confirmed cases, death cases, and recovered cases). Figure 2 shows case-wise clustering.

C. **Ensemble Classification and Prediction**

Ensemble classification plans use classifiers as part of an ensemble of classifiers. These classifiers combine their predictions for the classification of new records through a variety of averaging or voting methods. It is shown in Figure 3.

Figure 1: Country- wise clustering

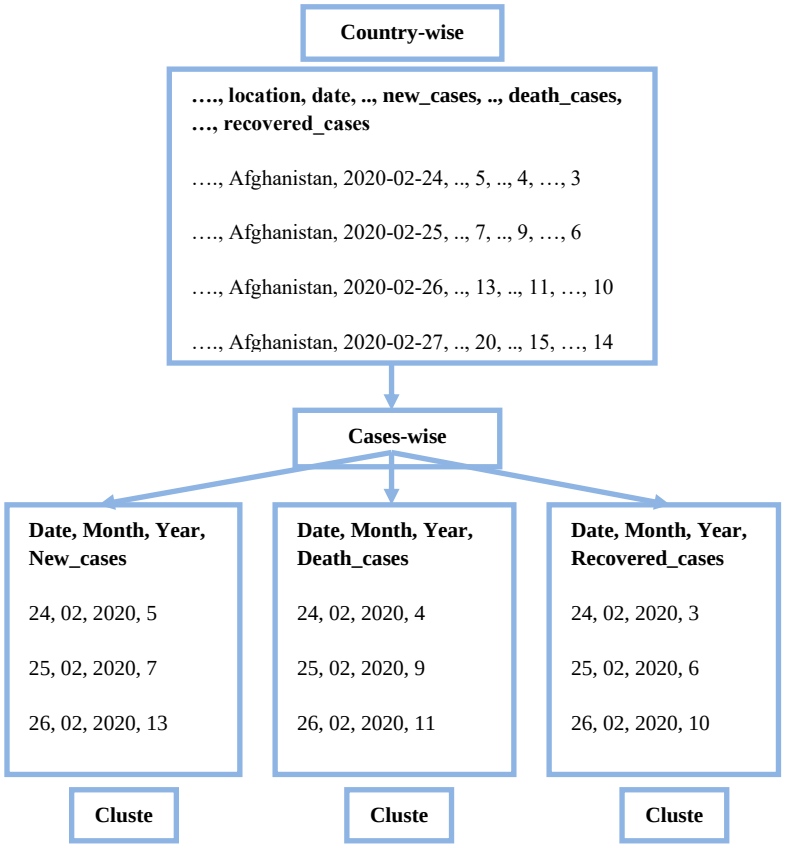


Figure2: Case-wise clustering

It was believed that the ensemble would contain a level of predictive accuracy that was superior to that of any individual classifier. The term "ensemble learning" is frequently used in place of "ensemble classification" to refer to the same thing The first, on the other hand, is a more widely utilized approach in which several techniques are combined to address any kind of problem—not simply a classification challenge. Base classifiers are the individual classifiers in an ensemble. The ensemble is called homogeneous if the base classifiers are similar (for example, decision trees). Otherwise, it is referred to as heterogeneous.

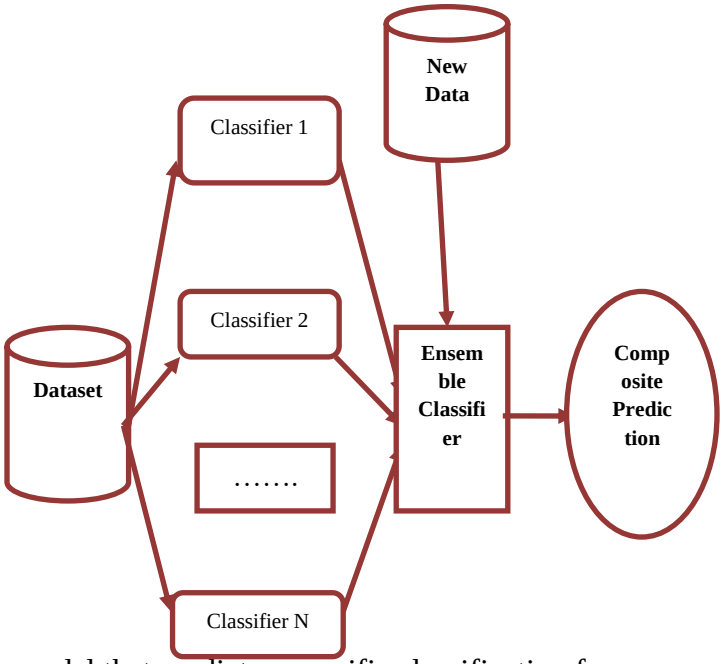
An ensemble classification algorithm's simple form is

1) Create N classifiers for a given dataset.
2) New record X
    1. 1.Determine the predicted class values that X will have for each of the N classifiers
    2. 2.If categorical is the predicted class value,

        (a) Choose the predicted class value that is most often predicted (Voting method)

1.Else if the predicted class value is numerical,
    (a) Calculate the average for all predicted class values (averaging method)
4. End for

Figure 3: Create N classifiers Model



The classification model that predicts a specific classification for a new record is counted as a single vote, and the classification with the most votes wins, i.e., it is considered the ensemble's prediction that the classification is correct.

1. Testing Dataset

The testing dataset is called new data. For prediction, it should use an ensemble classifier. The case- wise clustered datasets contain records for COVID-19 cases from February 13, 2020, to March 15, 2021. To generate the testing dataset, it should be generated between March 16, 2021 and the next 30 days (within 14 April 2021). Figure 4 shows a glimpse of the testing dataset. The dataset includes date, Total_cases, location,
New cases, new_ deaths, total_deaths, test_unit,etc.,

Figure 4: Testing Dataset

Date, Month, Year, New_cases	**Date, Month, Year, Death_cases**	**Date, Month, Year, Recovered_cases**
16, 03, 2021, ?	16, 03, 2021, ?	16, 03, 2021, ?
17, 03, 2021, ?	17, 03, 2021, ?	17, 03, 2021, ?
……..	……..	……..
14, 04, 2021, ?	14, 04, 2021, ?	14, 04, 2021, ?

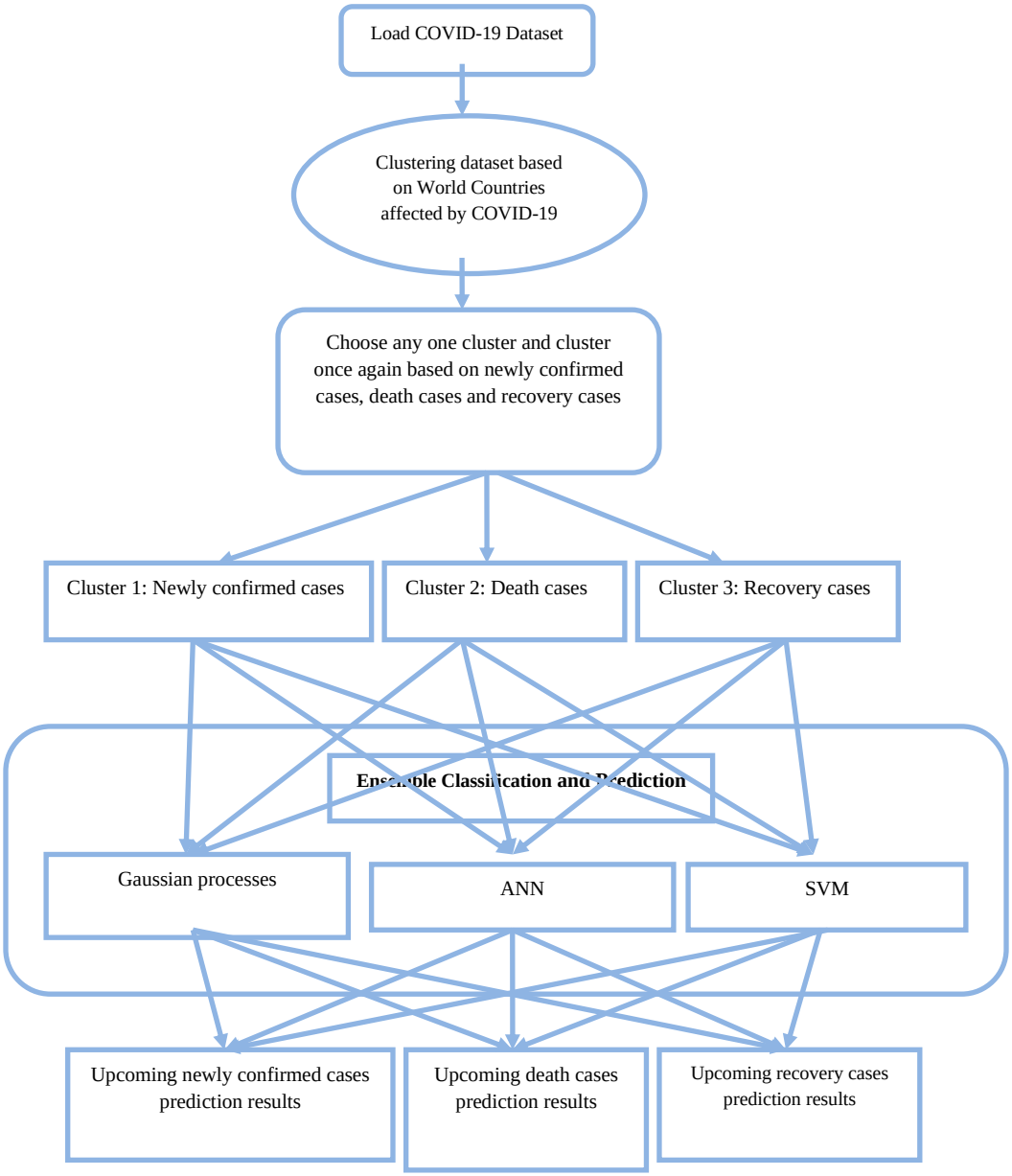
2. An Ensemble Classification-Based Country-Wise Covid-19 Upcoming Cases Prediction

After applying country- and case-based clustering, three clusters have been identified. Apply these three clusters to the proposed ensemble classification. The Weka [16] tool is utilized by the ensemble model to construct a two-level model in this setting. Figure4 illustrates the proposed ensemble workflow for predicting upcoming cases for COVID-19. The first level identifies three classifications, including the SVM, Gaussian Process, Artificial Neural Network, as shown in Figure 5. At the second level, calculate the average for all predicted cases for forthcoming dates.

Gaussian Processes

The Gaussian process is a method of supervised learning. The Gaussian probability distribution can be simplified into a process known as a Gaussian process. Stochastic processes also govern function properties in the same way that Gaussian distributions are described by their means and covariance matrices. A Gaussian process is, in any given data model, difficulty a single sample of a Gaussian distribution, whereas probability distributions describe random variables, which are vectors or scalars (for multivariate distributions). The WEKA tool was used to put the Gaussian process into action, signified by a mean and a covariance method. The mean is a method x, and the covariance is a method C (x, x") that describes the predictability of the values of the method y in the points x.

Figure 5: An ensemble classification-based country-wise covid-19 upcoming cases prediction workflow



ANN:

Artificial Neural Network is only used for classification and regression. The WEKA tool is used to implement this algorithm. A subclass of feed-forward ANN is the multilayer perceptron (MLP). MLP uses back - propagation algorithm, a supervised learning technique, for training. MLP stands out from linear perceptrons due to its nonlinear activation and numerous layers. It could differentiate data that is not linearly divisible. The MLP has one hidden layer with five nodes. The nodes of the MLP all employ the conventional sigmoid (f(x) = (1+e-x)-1). The MLP was trained to utilize backpropagation through a momentum of 0.1, a learning rate of 0.1, and a practice time of 1000 are all possible.

SVM:

SVM could be utilized for Classification or Regression. It identifies the vector-like input features that were projected onto higher-dimensional space. The most effective hyperplane was then developed to divide the various newly confirmed, deceased, or recovered cases. The WEKA tool is used to execute the SVM. It used to be used to predict a continuous variable. SVM tries to fit the most accurate line, The difference between predicted and actual values is minimized in some models, while in others it is maximized

IV Results and Discussions

The section will analyze. It compares the proposed ensemble method with other previous classification methods. For a method to be justified, its accuracy must be quantified. One straightforward way to assess a method's accuracy is to use the gap between the real value and the predicted value. Using this error could provide numerous different metrics that could provide more insight. The metrics are:

1) Mean Absolute Error (MAE)
2) Mean Absolute Percentage Error (MAPE)
3) Root Mean Squared Error (RMSE)
4) R-Squared Score (RSS)

1. MAE: It ranges from 0 to infinity, and a lesser value means a superior model. According to Eq (1) is recognize regression error metric and is exactly explained.

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n} |yi - xi|}{n} \quad (1)$$

Here, n = total number of data points, xi = actual value and yi = predicted value,

2. MAPE: Eq (3) computes accuracy by dividing the actual values actual values exceeding predictions by the average absolute percent error between the actual and predicted values

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(xi-yi)^2}{n}} \quad (2)$$

Here, n = total number of data points,
xi = actual value and yi = predicted value,

3. RMSE:

The In RMSE, the square root of the squared errors is calculated without taking into account the error direction. This is a scale from 0 to infinity, LMSM, and it is forever better at magnitude than MAE.Among the regression error metrics discussed in Eq (2), it is the most commonly used.

The RMSE error is calculated by averaging the squares across each data point, then taking the square root of it

$$\mathbf{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{xi-yi}{xi}\right| \quad (3)$$

Here, n = total number of data points, xi = actual value and yi = predicted value.

4. R-Squared/Adjusted R-Squared Score: According to Eq. (4) The variation of dependent variables caused by independent variables is expressed as a percentage.

$$\mathbf{R^2 = 1 - \frac{\sum_i(xi - yi)^2}{\sum_i(xi - x')^2}} \quad (4)$$

Here, x' = mean of actual value, yi = predicted value and xi = actual value
Table 1 shows death cases classification and prediction error rates including MAE, RMSE, MAPE and R-square error of Gaussian processes ANN, SVM and ensemble algorithms.
Table 1: Metrics based classification

Table 1 shows death cases classification and prediction error rates including MAE, RMSE, MAPE and R-square error of Gaussian processes ANN, SVM and ensemble algorithms.

Table 1: Metrics based classification

Algorithm	MAE	RMSE	MAPE	R Squared Error
Gaussian Processes	482.8176	605.6863	730.5803	71.0799
ANN	21.4829	25.2415	730.5803	2.9622
SVM	132.4631	161.1589	730.5803	18.9127
Ensemble	17.8052	22.7581	722.003	2.3905

Algorithms comparison for Death cases

Figure 6: Metrics based classification Algorithms comparison for Death cases

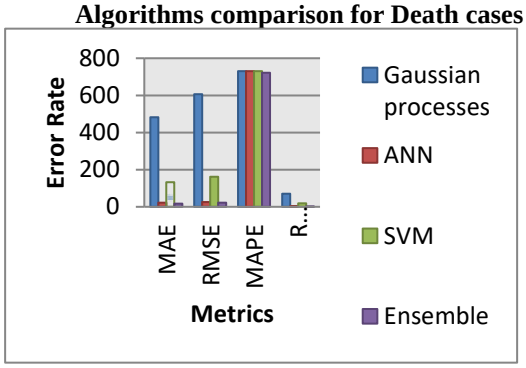


Figure 6 shows death case classification and prediction error rates for Gaussian processes, ANN, SVM, and ensemble algorithms In comparison with other algorithms, the proposed ensemble algorithm has a lower error rate for death case prediction. It concludes that the accuracy of the ensemble algorithm is high. Furthermore, Table 2 shows the recovered cases' classification and prediction error rates, including MAE, RMSE, MAPE, and R-square error of Gaussian processes, ANN, SVM, and ensemble algorithms

Table 2: Metrics based classification algorithms comparison for Recovered cases

Algorithm	MAE	RMSE	MAPE	R Squared Error
Gaussian processes	11428.05	12888.17	12209.1	92.3188
ANN	7787.59	10522.23	12209.1	75.3714
SVM	8625.228	11151.04	12209.1	79.8756
Ensemble	7780.219	10507.76	12202.27	73.7731

Figure 8: Metrics based classification algorithms comparison for Recovered cases

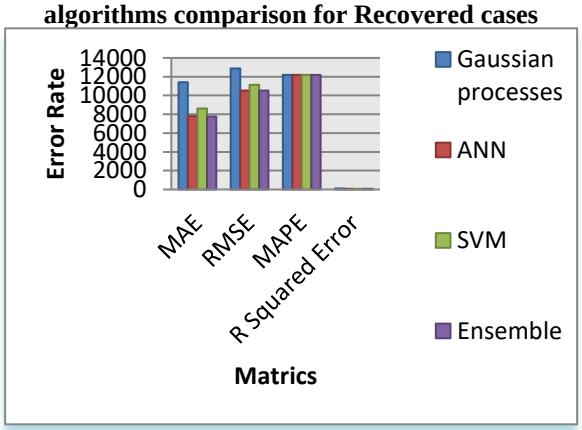


Figure 7 shows the classification and prediction error rates of Gaussian processes, ANN, SVM, and ensemble algorithms. Compared to other algorithms, the proposed ensemble algorithms have a lower error rate for predicting recovered cases. It concludes that the overall accuracy of the

ensemble algorithm is high. Furthermore, Table 3 shows newly confirmed case classification and prediction error rates, including MAE, RMSE, MAPE, and R-square error of Gaussian processes, ANN, SVM, and ensemble algorithms.*Table 3:* Table

3: Metrics based classification algorithms Comparison for newly confirmed cases

Algorithm	MAE	RMSE	MAPE	R Squared Error
Gaussian Processes	124.5358	190.9157	134.313	97.9506
ANN	124.2056	188.9923	134.213	96.9483
SVM	111.2402	207.6716	134.313	106.5354
Ensemble	109.8480	181.7537	131.5357	89.9411

Figure 8: Metrics based classification algorithms comparison for Newly Confirmed cases

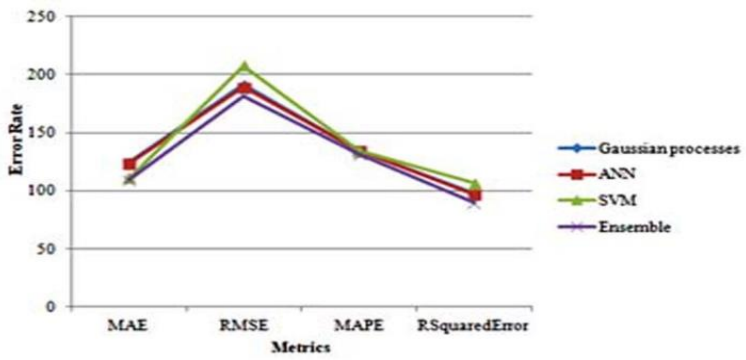


Figure 8 shows newly confirmed case classification and prediction error rates of Gaussian processes, ANN, SVM, and ensemble algorithms. Compared with other algorithms, the proposed ensemble algorithms have a lower error rate for newly confirmed cases. It concludes that the ensemble algorithm's accuracy is high

```
**********
COVID-19 upcoming New_Confirmed forecasting Results
**********

@relation NewConfirmedTestingDataset

@attribute Date numeric
@attribute Month numeric
@attribute Year numeric
@attribute NewConfirmedCount numeric

@data
16,3,2021,30
17,3,2021,30
18,3,2021,29
19,3,2021,29
20,3,2021,28
21,3,2021,28
```

V Conclusion

This paper proposed an ensemble classification-based COVID-19 upcoming cases prediction model for each country. The ensemble classification and prediction model detect novel coronavirus cases, such as newly confirmed, recovered cases separately. The proposed model uses three famous classifiers, SVM, Gaussian Processes, and ANN, which have different learning characteristics and architectures at the first stage. Combining their predictions at a second stage by averaging them out results in superior performance. The experimental results showed the proposed ensemble model provides better performance when compared with previous classification algorithms.

References

[1]. Lu X, Zhang L, Du H, Zhang J, Li YY, Qu J, Zhang W, Wang Y, Bao S, Li Y, Wu C. SARS-CoV-2 infection children. New England Journal of Medicine. (2020) Apr 23;382(17):1663-5. https://doi.org/10.1056/NEJMc2005073

[2]. Dyall J, Gross R, Kindrachuk J, Johnson RF, Olinger GG, Hensley LE, Frieman MB, Jahrling PB. Middle East respiratory syndrome and severe acute respiratory syndrome: current therapeutic options and potential targets for novel therapies. Drugs. (2017) Dec 1;77(18):1935-66. https://doi.org/10.1007/s40265-017-0830-1

[3]. Hâncean MG, Perc M, Lerner J. Early spread of COVID-19 in Romania: imported cases from Italy and human-to-human transmission networks. Royal Society open science. (2020) 22 July;7(7):200780. https://doi.org/10.1098/rsos.200780

[4]. Zhou Y, Xu R, Hu D, Yue Y, Li Q, Xia J. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data. The Lancet Digital Health. Elsevier. (2020) Aug 1;2(8):e417-24. https://doi.org/10.1016/S2589-7500(20)30165-5

[5]. Thomas-Rüddel D, Winning J, Dickmann P, Ouart D, Kortgen A, Janssens U, Bauer M. Coronavirus disease 2019 (COVID-19): update for anesthesiologists and intensivists March 2020. Der Anaesthesist. (2020) Mar 24:1-0. https://doi.org/10.1007/s00101-020-00760-3

[6]. Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, Baker M, Mackenzie LS. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. International immunopharmacology. (2020) Sep 1;86:106705. https://doi.org/10.1016/j.intimp.2020.106705

[7]. Bao FS, He Y, Liu J, Chen Y, Li Q, Zhang CR, Han L, Zhu B, Ge Y, Chen S, Xu M. Triaging moderate covid-19 and other viral pneumonia from routine blood tests. arXiv preprint arXiv:2005.06546. (2020) 13 May.

[8]. de Freitas Barbosa VA, Gomes JC, de Santana MA, Jeniffer ED, de Souza RG, de Souza RE, dos Santos WP. Heg. IA: An intelligent system to support the diagnosis of Covid-19 based on blood tests. Research on Biomedical Engineering. (2020) Oct 26:1-8. https://doi.org/10.1101/2020.05.14.20102533

[9]. de Moraes Batista AF, Miraglia JL, Donato TH, Chiavegatto Filho AD. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. medRxiv. (2020) 1 January. https://doi.org/10.1101/2020.04.04.20052092

[10]. Bayat V, Phelps S, Ryono R, Lee C, Parekh H, Mewton J, Sedghi F, Etminani P, Holodniy M. A SARS-CoV-2 Prediction Model from Standard Laboratory Tests. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America. (2020) 12 August. https://doi.org/10.2139/ssrn.3594614

[11]. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems. (2020) Aug;44(8):1-2. https://doi.org/10.1007/s10916-020-01597-4 Feng C, Huang Z, Wang L, Chen X, Zhai Y, Zhu F, Chen H, Wang Y, Su X, Huang S, Tian L. A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. MedRxiv. (2020) 1 January. https://doi.org/10.1101/2020.03.19.20039099

[12]. Joshi RP, Pejaver V, Hammarlund NE, Sung H, Lee SK, Furmanchuk AO, Lee HY, Scott G, Gombar S, Shah N, Shen S. A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. Journal of Clinical Virology.(2020) 1 August;129:104502. https://doi.org/10.1016/j.jcv.2020.104502

[13]. Kukar M, Gunčar G, Vovko T, Podnar S, Černelč P, Brvar M, Zalaznik M, Notar M, Moškon S, Notar M. COVID-19 diagnosis by routine blood tests using machine learning. arXiv preprint arXiv:2006.03476. (2020) Jun 4. https://doi.org/10.1038/s41598-021-90265-9

[14]. Langer T, Favarato M, Giudici R, Bassi G, Garberi R, Villa F, Gay H, Zeduri A, Bragagnolo S, Molteni A, Beretta A. Use of Machine Learning to Rapidly Predict Positivity to Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) Using Basic Clinical Data.

[15]. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. (2009) Nov 16;11(1):10-8. https://doi.org/10.1145/1656274.1656278