How to Cite:
Theresa, A. M., & Saravanan, V. (2022). Generalized feature transformative dependency magnitude decision tree classification for COVID-19 medical data analysis. *International Journal of Health Sciences*, *6*(S10), 429–445. Retrieved from https://sciencescholar.us/journal/index.php/ijhs/article/view/13493

Generalized feature transformative dependency magnitude decision tree classification for COVID-19 medical data analysis

Mrs. A. Mary Theresa
Assistant Professor, Department of Information Technology, Nirmala College for Women, Coimbatore - 641 018. Tamil Nadu, India
*Corresponding author email: mary.devanto@gmail.com

Dr. V. Saravanan
Professor & HEAD, Department of Information Technology, Hindusthan College of Arts and Science, Coimbatore - 641 028. Tamil Nadu, India
Email: vsreesaran@gmail.com

Abstract---Data mining is a method for identifying attractive patterns in an understandable format from big data. Big data mining is an integration of structured and unstructured data that are mined for extracting valuable information and used for predictive analytics. Big data analytics are used in many applications, especially in the healthcare sector. With a variety of data analytics devices and procedures, the healthcare domain uses big data to update health prevention and management. Big data analytics for healthcare create it feasible to invent smarter evaluations of patient heath conditions. One of the most current and relevant big data analytics in healthcare is the global COVID-19 prediction. Many works have been introduced for COVID-19 prediction with big data analytics. But, the precision results and time consumption was not focused on by existing methods. Therefore, a novel Log Transformative Generalized feature Mapping based Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree Classification (LTGFM-RDMDDTC) is introduced for COVID-19 prediction with higher accuracy, precision, and lesser time consumption. In medical data analytics, the performance of the model is said to be improved by including three major processes namely feature engineering, feature selection, and classification. In the LTGFM-RDMDDTC technique, first, Hosmer–Lemeshow Log Transform-based Feature Engineering model is applied to the COVID-19 medical dataset that performs mathematical transformations by handling skewed data and creating the robust features. After the transformation, the structured training dataset is obtained for further processing. Secondly, Hellinger's generalized multidimensional scaling

Manuscript submitted: 9 May 2022, Manuscript revised: 18 July 2022, Accepted for publication: 27 August 2022

feature map is applied to select the significant features from the structured training dataset for minimizing the time consumption of COVID-19 prediction. Finally, Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree Classification model is applied where the classification of medical data is performed. Here, the classification is carried out by means of the Rand Dependency Magnitude factor which in turn improves the precision accuracy for COVID-19 medical data analytics. Experimental evaluation is carried out for factors such as accuracy, precision, recall, F-measure, and prediction time with respect to a number of sample data. The quantitatively analyzed results indicate the better performance of our proposed LTGFM-RDMDDTC when compared with two state-of-the-art methods.

Keywords---big data analytics, COVID-19 prediction, hosmer–lemeshow, log transform-based feature, engineering model, hellinger generalized, multidimensional scaling, feature map, rand dependency magnitude, iterative dichotomiser-3, decision tree classification model.

Introduction

The entire world is experiencing a pandemic situation due to the wide spreading of the novel coronavirus disease namely, COVID-19 in recent days. But the detection of COVID-19 takes considerable time during this time, it may be spread among other persons. To get mitigate this unexpected condition, early identification of COVID-19 patients is required. Several machine learning-based frameworks have been developed. A novel KNN Variant (KNNV) algorithm was introduced in [1] that provides better results for accurate classification of COVID-19. But it failed to perform the accurate classification inside the medical field with minimum time. A Multi-Task Gaussian Process (MTGP) regression model was developed in [2] for improving the predictions of coronavirus COVID-19 outbreak. But it failed to improve the performance of MTGP in COVID-19 outbreak prediction. A new ensemble-based classifier of machine learning was introduced in [3] on COVID-19 dataset to predict the disease.

But it was not simple to obtain a large sample of data. The ensemble learning approach was designed in [4] for efficient recognition of COVID-19 using laboratory test results. The designed feature selection technique was not applied for improving the results of other machine learning models, thus increasing classification accuracy. A non-dominated sorting genetic algorithm (NSGA-II) was introduced in [5] to choose the interesting features and the classification phase was conducted using an AdaBoost classifier. But the designed model failed to provide satisfactory results. An integration of Harris hawks optimization (HHO) to optimize the Fuzzy K-nearest neighbor (FKNN) called HHO-FKNN was developed in [6] to differentiate the severity of COVID-19. But the performance of accurate classification was not performed. A Hybrid Model was developed in [7] for COVID-19 prediction. But it was not capable of providing an effective response to the pandemic.Adaptive Neuro-Fuzzy Inference System (ANFIS) was developed in [8] to predict the number of COVID-19 cases based on collected data. But it failed to accurately perform the feature selection to minimize the complexity.

A novel method was developed in [9] by introducing a Random Forest and Naive Bayes classifier for predicting the risk related to COVID-19 from the patient data. But it failed to handle the large volume of patient data with minimum time. A novel method was developed in [10] to optimize the hyperparameters of classifier and to balance the COVID and non-COVID classes of the dataset. But it failed to design an efficient classification technique, a namely clinically operable decision tree for data classification

Contributions

The most important contributions of the proposed LTGFM-RDMDDTC technique are discussed as given below.

- To develop the LTGFM-RDMDDTC technique that has the capability of dealing with this epidemic condition by means of big data analytics and gives the better prediction results with higher accuracy.
- To minimize the time consumption of COVID-19 medical data analysis, the LTGFM-RDMDDTC technique performs two major processes namely feature engineering and feature selection. Hosmer–Lemeshow Log Transform is applied to the COVID-19 medical dataset to create robust features and form the structured dataset. afterward, Hellinger Generalized multidimensional scaling feature map is applied to select the significant features from the structured dataset for minimizing the time consumption of COVID-19 prediction
- To increase the prediction accuracy, Finally, Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree model is applied where the classification of medical data is performed by analyzing the testing and training data samples. The Rand Dependency Magnitude factor provides the accurate classification results this in turn improves the precision as well as the accuracy of COVID-19 prediction.
- Finally, well-known experimentation is carried out to measure the performance of our LTGFM-RDMDDTC technique and other existing works. The experimental result reveals that our LTGFM-RDMDDTC technique is highly efficient than the existing classification methods.

Organization of paper

The rest of the article is arranged as follows: In Section 2, related work reviews the articles discussed followed by the detailed discussion of the proposed LTGFM-RDMDDTC technique is presented in section 3. In section 4, experimental settings with data set descriptions are presented. Section 5 discusses the performance result by making a comparison with the state-of-the-art methods. Finally, concluding remarks are provided in Section 6.

Related Works

A novel short-term forecasting approach was introduced in [11] for Novel Corona Virus (COVID -19) detection using Machine learned hybrid Gaussian method. But the time series model was not implemented for the prediction and controlling of the disease. A stacking ensemble with deep neural networks was developed in

[12] for the prediction of COVID-19. But various feature selection techniques were not implemented to handle high-dimensional data. Machine-learning applications and algorithms were developed in [13] for the detection of COVID-19 cases. But the designed algorithm failed to introduce a time-efficient model for the detection of COVID-19 cases. A deep ensemble framework of the transfer learning approach was introduced in [14] for early prediction of COVID-19. But it did not achieve maximum classification results of COVID-19 prediction. A novel hybrid model was developed in [15] to increase the prediction accuracy of collective COVID-19 confirmed data. But the performance of time consumption for COVID-19 prediction was not analyzed. A maximum likelihood estimation theory-based Kalman filter was developed in [16] for dynamic prediction of COVID-19 spread. But it failed to focus on improving the COVID-19 prediction by considering the minimum errors. Two-strain dynamic models were developed in [17] for the COVID prediction. But it failed to attain maximum classification results by using a large dataset. In a random forest, a linear model was developed in [18] for identifying COVID-19 Confirmed, Death, and Cured Cases. A long short-term memory (LSTM) algorithm of deep learning was introduced in [19] to predict the number of COVID-19 positive cases. A decision-level fusion method was developed in [20] that integrate three well-calibrated ensemble classifiers. But it was not efficient to handle the large size of the available dataset.

Methodology

The entire country faces a pandemic situation owing to the tedious virus, named COVID-19. Detection of COVID-19 at an early stage is necessary to offer sufficient treatment to the disease affected patients. Due to the large volume of patient data, detection of disease at an early stage is a challenging issue. Healthcare big data analytics are complex by applying the traditional data processing techniques. It acquires more time to detect the virus during this time, it may be spread among other people. To obtain clear of this unexpected condition, early identification of COVID-19 patients is required. In this paper, a novel technique called LTGFM-RDMDDTC is designed and optimized as a machine learning-based framework using patient information that provides an effective and time-efficient solution to this pandemic situation.

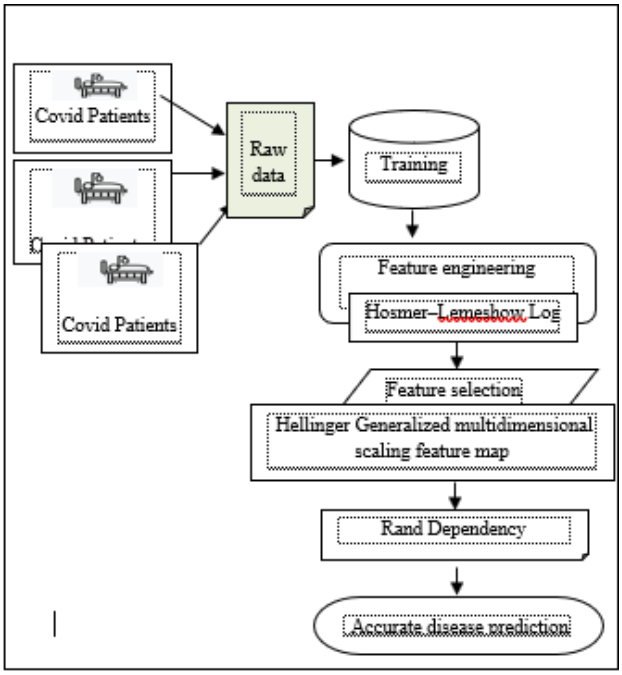

Figure 1. Architecture of the proposed LTGFM-RDMDDTC technique

Figure 1 reveals a basic architecture illustration of the proposed LTGFM-RDMDDTC technique to predict the COVID 19 disease patients with the help of different processes namely feature engineering, feature selection, and classification. Initially, the large numbers of patients' raw data are collected to form training dataset. First, the feature engineering process is carried out using Hosmer–Lemeshow Log Transform to obtain the structured dataset by creating the new features from the existing one. After that, the relevant feature selection process is performed using Hellinger Generalized multidimensional scaling feature map to select significant features for accurate Covid-19 disease prediction. Finally, the Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree Classification model is applied for disease prediction with minimum time consumption. These three different processes of the proposed LTGFM-RDMDDTC technique are explained as given below.

Hosmer–Lemeshow Log Transform-based Feature Engineering

Feature Engineering is the fundamental process of data preparation by creating the new features from the raw data. The collection of patient data from various sources is an unstructured format. The unstructured data is information that is not organized along with a pre-set data model, and hence it did not store in a relational database for further processing. Since the unstructured data format creates major impacts on obtaining accurate classification results with minimum time as well as training error. In order to solve such kinds of problems, input data must be transformed into a structured format so that the machine learning algorithm provides the accurate classification results to which prediction is said to be performed. Therefore, the proposed LTGFM-RDMDDTC technique uses the

Hosmer–Lemeshow Log Transform with the objective of providing a better representation of the patient data to the predictive learning algorithm.

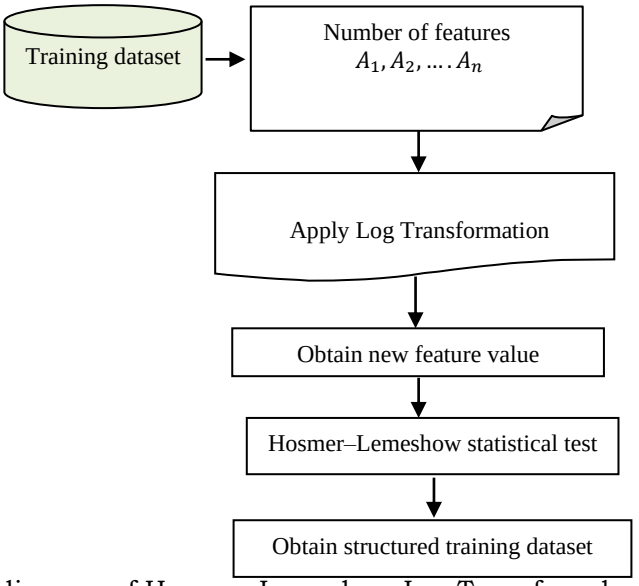

Figure 2. Block diagram of Hosmer–Lemeshow Log Transform based Feature Engineering

As given in above block diagram 2, Hosmer–Lemeshow Log Transform to obtain a new features value. Let us consider the number of features $A_1, A_2, \ldots . A_n$ represent a column in the training dataset. For each attribute, Hosmer–Lemeshow Log Transform is applied to generate new features from the existing features. In general, the transformation technique is used for creating the features that are approximately normal but have some skewness. In other words, the generated features value deviates from the normally distributed data (i.e. existing features). Therefore, the proposed transformation is employed for accurately finding the fittest feature value technique with the help of the log transformation. The new feature generation procedure is given below,

Mathematical modeling

$$A_n \leftarrow T_L = \frac{1}{m} \sum_{i=1}^{m} log\ (A_i) \qquad (1)$$

Where, A_n denotes a newly generated feature value, T_L denotes a transformation output, 'A_i' denotes a particular feature value, m denotes the number of particular column feature values. The transformation generates the new feature value from the existing single-column features. Then the Hosmer–Lemeshow test is a statistical test used for verifying the goodness of fit (i.e. whether the generated new features is fit into existing features). The statistical test for the observed feature rates matched to expected feature is given below.

$$H_L = \frac{(A_n - A_e)^2}{A_e\left(1 - \frac{A_e}{b}\right)} \qquad (2)$$

Where, H_L indicates a Hosmer–Lemeshow statistical test, A_n denotes an observed result (i.e. newly generated feature value), A_e indicates an expected feature value, b denotes the number of observations. The output of the Hosmer–Lemeshow statistical test closer to 1 indicates a good fittest feature value. In this way, the feature engineering process is carried out and obtains the structured training dataset. The algorithmic process of feature generation is given below.

Algorithm 1

Algorithm 1: Hosmer–Lemeshow Log Transformation-based feature engineering
Input: Training dataset, Number of original features $A_1, A_2, \ldots . A_n$
Output: Structured training dataset
1: Begin 2: For each feature 'A_i' 3: Apply log transformation 'T_L' 4: Obtain the new feature value 'A_n' 5: For each new feature value 'A_n' 6: For each expected feature value 'A_e' 7: Measure Hosmer–Lemeshow statistical test 'H_L' 8: $if\ (H_L = 1)$ then 9: Find best fit feature value 10: end if 11: End for 12: End for 13: Return (transformed data) 14: End for End

Algorithm 1 explains step-by-step process of feature engineering for generating the feature value from the raw dataset. It is attained through Hosmer–Lemeshow Log Transformation. As a result, the new feature values are generated. For each observed feature and expected value, Hosmer–Lemeshow statistical test is evaluated. If the statistical test value is closer to the one, then the selected feature is fittest to the original feature value. The structured training dataset is obtained for further processing.

Hellinger Generalized multidimensional scaling feature map

After the feature engineering, the second process of the proposed technique is to perform the relevant feature selection. The relevant feature selection from the structured training dataset helps to reduce the complexity of an algorithm. The proposed LTGFM-RDMDDTC technique uses the Hellinger generalized multidimensional scaling feature map to find the relevant features. Hellinger generalized multidimensional scaling feature map is a method of nonlinear dimensionality reduction that focuses on creating the mappings of relevant features based on Hellinger distance. It is a statistical metric used to quantify the similarity between two features in the given dataset. The generalized multidimensional models have the ability to adapt and respond accurately when a large number of features are taken from the same dataset.

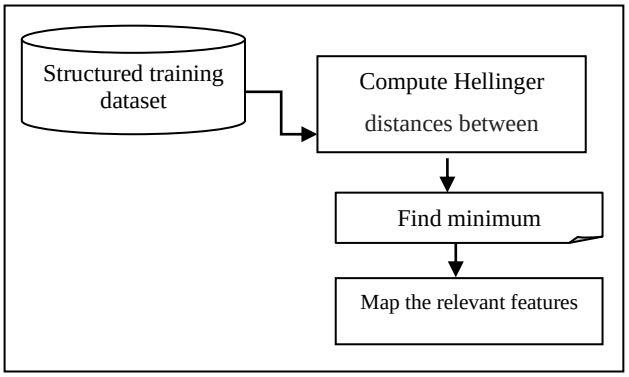

Figure 3. Block diagram of Hellinger Generalized multidimensional scaling feature map

Figure 3 illustrates the block diagram of Hellinger's generalized multidimensional scaling feature map to obtain relevant features from the dataset.

Mathematical modeling

Let us consider the number of features $A_1, A_2, \ldots . A_k$ taken from the structured training dataset after feature engineering process.

$$A_k = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1n} \\ A_{21} & A_{22} & \ldots & A_{2n} \\ \vdots & \vdots & \ldots & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mn} \end{bmatrix} \in D \quad (3)$$

From (3),A_k denotes a set of features with 'n' rows and 'm' columns in the dataset, D represents a structured training dataset. After that, the distance between the features is estimated using Hellinger distance.

$$S_{ij} = \frac{1}{\sqrt{2}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left| A_i - A_j \right| \quad (4)$$

Where, S_{ij} indicates a distance between the features, A_i and A_j , *'k'* indicates a number of features in the structured training dataset. The multidimensional scaling function finds the minimum distance between the features.

$$\varphi = \arg\min S_{ij} \quad (5)$$

Where, φ indicates a multidimensional scaling function, $\arg\min$ denotes an argument of the minimum function. As a result, minimum distances between the features are selected as relevant for accurate data classification. The algorithmic process of feature selection is given below,

Algorithm 2

// Algorithm 2: Hellinger Generalized multidimensional scaling feature map
Input: Dataset, Transformed features $A_1, A_2, \ldots . A_k$ Output: Relevant features
Begin 1. Collect the number of features $\{A_1, A_2, \ldots . A_k\}$ 2. For each feature 'A_i' and A_j 3. Measure the distance 'S_{ij}' 4. if $(arg\ \min S_{ij})$ then 5. The feature is said to be relevant 6. Select relevant features 7. else 8. The feature is said to be not a relevant 9. Remove irrelevant features 10. end if 11. end for 12. Return (relevant features) End

Algorithm 2 describes step-by-step process of the relevant feature selection based on Hellinger's generalized multidimensional scaling feature map. Initially, the number of features is taken from the structured training dataset. Then the Hellinger distance is applied for finding the similarity between the features. If the distance between the two features is minimal, then the feature is said to be relevant. Otherwise, the feature is said to be irrelevant. The relevant features are preferred for classification and other features are removed. This process of the proposed technique minimizes time complexity.

Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree Classification model

Finally, the classification process is carried out using Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree (RDMID3DT) for disease prediction based on the selected relevant features. The proposed classifier measures the Rand Dependency Magnitude between the selected features with the testing disease feature and provides accurate classification results. Rand Dependency Magnitude is a statistical factor used for comparing the similarity between the selected features and testing disease features. The Iterative Dichotomiser-3 Decision Tree consists of the root node, branch, and leaf node. The root node performs the big data analysis by means of the Rand Dependency Magnitude factor. Each branch represents the outcome of the analysis, and each leaf node represents a class label. Dichotomiser provides the two-class label such as normal and disease presence.

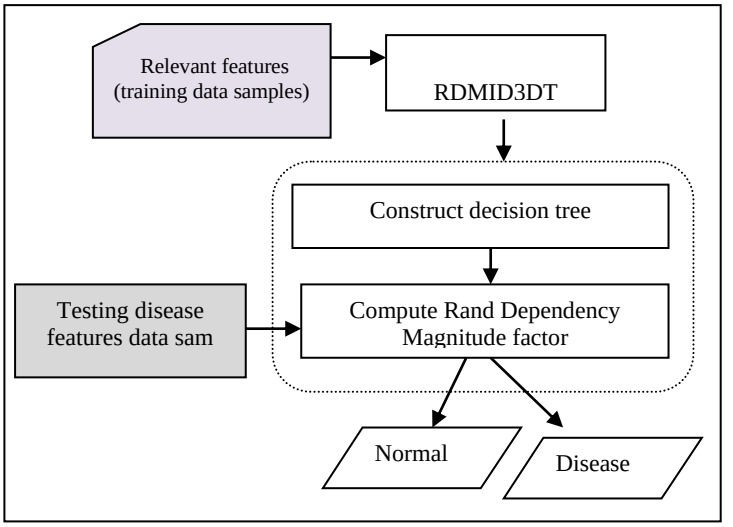

Figure 4. block diagram of RDMID3DT

Figure 4 portrays the block diagram of the RDMID3D to analyze the training feature data samples and testing disease feature data. The relationships between these two data are analyzed using the Rand Dependency Magnitude factor and obtain the final classification results as normal or disease.

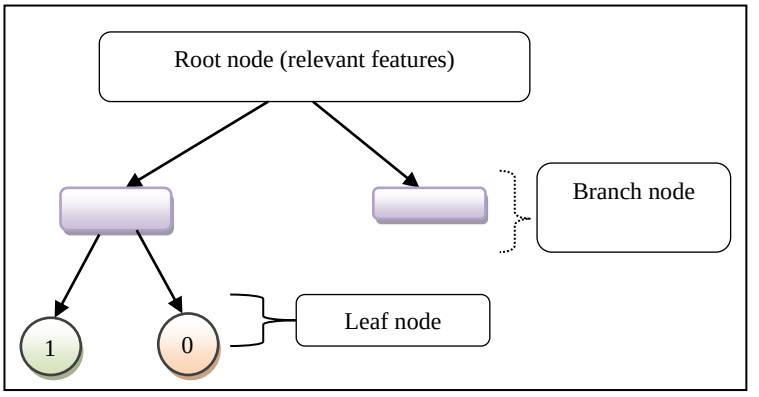

Figure 5. Iterative Dichotomiser 3-based Decision Tree

Figure 5 illustrates the Rand Iterative Dichotomiser 3-based Decision Tree which consists of one root node linked to the child node through the branch node. In the root node, the feature analysis is performed between the training and testing data using the Rand Dependency Magnitude factor.

$$\beta_{rc} = \left[\frac{A_T \cap A_R}{number\ of\ data\ samples}\right] \qquad (6)$$

Where, β_{rc} denotes a Rand Dependency Magnitude factor coefficient, A_T represents testing disease data samples, A_R indicates a training disease data samples, $A_T \cap A_R$ indicates a mutual dependence between the two sets i.e. A_T

and A_R. The coefficient (β_{rc}) returns the value between 0 and 1. Based on the similarity value, the normal and disease patterns are correctly identified.

$$Y = \begin{cases} \beta_{rc} = 1; & Disease\ presence \\ \beta_{rc} = 0; & Normal \end{cases} \quad (7)$$

Where Y denotes a classification output of the decision tree. From the analysis, the proposed classifier accurately performed the disease prediction. This process is iterated through every selected feature set to obtain the final classified results. The algorithmic step of RDMID3DT is given below,

Algorithm 3

// Algorithm 3: Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree-based classification
Input: Number of Relevant features or training data samples (A_R), testing disease data samples (A_T) Output: Disease prediction
Begin 1. For each selected feature 'A_R' and testing disease data samples 'A_T' 2. Construct the decision tree with the root node, branch node, and leaf nodes 3. Measure the Rand Dependency Magnitude factor 'β_{rc}' 4. *if* ($\beta_{rc} = +1$) then 5. Classify the data samples into a disease 6. else 7. Classify the data samples into a normal 8. end if 9. Iterate the process for all features 10. Obtain classification results 11. End for End

Algorithm 3 depicts step-by-step process for Covid 19 prediction by means of the Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree-based classification. First, the numbers of relevant significant features are given as input to the Decision Tree classifier. With the selected training feature values, the classification is performed by analyzing the testing disease features using Rand Dependency Magnitude factor. Based on the Rand Dependency Magnitude coefficient, accurate classification of normal or disease is obtained. Based on classified results, disease prediction is done with higher accuracy.

Experimental Settings

In this section, experimental evaluation of the proposed LTGFM-RDMDDTC technique and existing KNNV [1], and MTGP [2] are carried out in Java language using Novel Corona Virus 2019 Dataset taken from the Kaggle [https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset].
This dataset includes daily level information on the number of affected cases,

deaths, and recovery from the 2019 novel coronavirus. The dataset includes a individual level information file namely covid_open_line-list-data.csv file taken for conducting the experiments. The above files are used to conduct the experiments for identifying the disease at an earlier stage. First, feature engineering and feature selection is performed to select the significant features and perform classification for identifying the confirmed cases (i.e. Corona Virus 2019 disease affected). For the experimental consideration, the numbers of data samples (i.e instances) are taken in the ranges from 10000 to 10000.

Performance Results and Discussion

The results and discussion of the proposed LTGFM-RDMDDTC technique and existing KNNV [1], and MTGP [2] with the different performance metrics such as prediction accuracy, precision, recall, and F-measure and prediction time. Performance results are evaluated with the help of table values and graphical representations.

Prediction accuracy

It is used to measure the number of data samples that are correctly predicted as disease or normal from a total number of data samples. Prediction accuracy is measured as below,

$$Pr_{Acc} = \left(\frac{T_{pos}+F_{pos}}{T_{pos}+F_{pos}+T_{Neg}+F_{Neg}}\right) * 100 \qquad (8)$$

Where, Pr_{Acc} indicates a prediction accuracy, T_{pos} indicates a true positive (correctly identifies the presence of a disease or not), F_{pos}denotes a false positive (i.e. incorrectly identifies a presence of a disease), T_{Neg} indicates the true negative (correctly identifying the absence of a disease), F_{Neg} symbolizes the false negative (i.e. incorrectly identifies that an absence of disease). The accuracy is measured in terms of percentage (%).

Precision

It is measured based on true positives as well as false positives of data classification. The Precision is estimated as follows,

$$Prec = \left(\frac{T_{pos}}{T_{pos}+F_{pos}}\right) * 100 \quad (9)$$

Where, $Prec$ indicates a Precision, T_{pos} symbolizes the true positive, F_{pos} represents the false positive. The Precision is measured in percentage (%).

Recall

It is measured based on true positives and false negatives of data classification. The formula for calculating the recall rate is computed as given below,

$$Rec = \left(\frac{T_{pos}}{T_{pos}+F_{Neg}}\right) * 100 \quad (10)$$

Where 'Rec 'indicates a recall, T_{pos} represents the true positive, F_{Neg} indicates the false negative. The recall is measured in percentage (%).

F-measure

F-measure is a measure of a test's accuracy in data classification. It is calculated based on the precision as well as recall of the test. It is evaluated as below,

$$\boldsymbol{FMRE = \left[2 * \frac{Prec * Rec}{Prec + Rec}\right] * 100} \quad (11)$$

Where $FMRE$denotes an F-measure is measured based on precision $Prec$ denotes a recall 'Rec'. F-measure is measured in percentage (%).

Prediction time

It is measured as the amount of time consumed by the algorithm to predict the normal or disease sample. Therefore, time consumption is computed as given below,

$$Pr_{Time} = Number\ of\ data\ samples * T_{me}(DP)(\quad (12)$$

Where, Pr_{Time} designates a disease prediction time, $T_{me}(DP)$ denotes a time for predicting the one patient data sample$T_{me}(DP)$. Prediction time is measured in terms of milliseconds (ms).

Table 1
prediction accuracy versus number of data samples

Number of data samples	Prediction accuracy (%) LTGFM-RDMDDTC	KNNV	MTGP
10000	97.5	94.5	91
20000	98	96	91.5
30000	98.33	95.66	93
40000	97.25	96.25	94.25
50000	97.4	96	94.4
60000	97.5	95.66	94.66
70000	97.14	95.85	94.85
80000	96.87	95.25	94.12
90000	96.77	95	93.44
100000	96.9	94.6	93.2

Table 1demonstrates prediction accuracy of Covid 19 using three different classification methods namely the LTGFM-RDMDDTC technique and existing KNNV [1], MTGP [2]. The average value indicates that the overall performance of

prediction accuracy issuing LTGFM-RDMDDTC relatively increased by 2% and 4% when compared to [1], and [2] respectively.

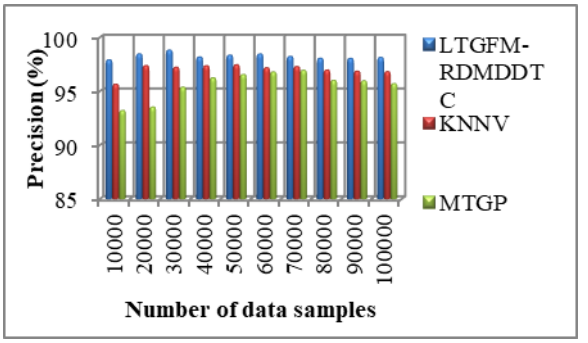

Figure 6. Performance results of precision

Figure 6 portrays the performance of precision of disease prediction with different number of data samples using three methods. LTGFM-RDMDDTC technique achieves a higher precision by 1% when compared to the C KNNV [1], and 3% when compared to MTGP [2

Table 2
Recall versus number of data samples

Number of data samples	Recall (%) LTGFM-RDMDDTC	KNNV	MTGP
10000	99.46	98.31	96.47
20000	99.47	98.39	97.19
30000	99.48	98.23	97.09
40000	98.93	98.65	97.54
50000	98.94	98.27	97.37
60000	98.94	98.20	97.46
70000	98.79	98.31	97.51
80000	98.68	97.98	97.68
90000	98.59	97.86	96.98
100000	98.63	97.41	96.93

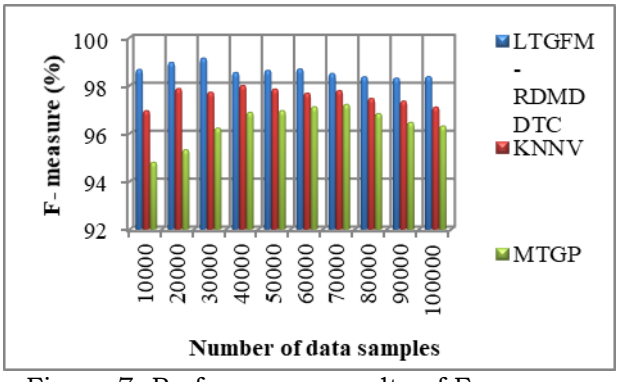

Figure 7. Performance results of F-measure

Figure 7 provide overall performance of F-measure using three methods namely LTGFM-RDMDDTC technique and existing KNNV [1], MTGP [2]. The average of ten comparison results indicates that the LTGFM-RDMDDTC technique increases overall performance of F-measure by 1% and 2% when compared to existing methods KNNV [1] and MTGP [2] respectively.

Table 3
Prediction time versus number of data samples

Number of data samples	Prediction time (ms) LTGFM-RDMDDTC	KNNV	MTGP
10000	43	47	49
20000	48	52	56
30000	54	57	60
40000	60	64	68
50000	66	70	75
60000	75	78	81
70000	81.2	85.4	87.5
80000	84	87.2	90.4
90000	90.9	92.7	95.4
100000	94	96	98

Table 3 reposts the performance of prediction time by means of three techniques namely the LTGFM-RDMDDTC technique, KNNV [1], and MTGP [2]. The overall performance results of prediction time are decreased using the LTGFM-RDMDDTC by 5% compared to [1] and 9% compared to [2] respectively.

Conclusion

An early and effective technique for the detection of COVID-19 is extremely important in this period of a global epidemic. Predicting the spread of COVID-19 is a complex task, owing to the Big Data samples stored in healthcare industry. Due to the large volume of data samples, accurate prediction of disease and assisting the physician in the decision-making are not on time and it causes the major issues. In this paper, a novel LTGFM-RDMDDTC technique is introduced for accurate disease prediction with minimum time by handling a large volume of data samples. First, the feature engineering process is carried out to make the simplicity of the algorithms to detect disease patterns and obtain a structured dataset with help of Hosmer–Lemeshow Log Transform. Then the Hellinger Generalized multidimensional scaling feature map is applied for selecting the more relevant features from the structured dataset in order to minimize disease prediction time. Finally, the Rand Dependency Magnitude Iterative Dichotomiser-3 Decision Tree identifies the normal of disease samples by measuring the similarity between training data and testing disease data. Based on the similarity estimation, final classification results are observed at the output layer. The performance of the LTGFM-RDMDDTC technique is evaluated through in-depth experimentation using COVID 19 dataset and compares the results with two conventional classification algorithms. Results have proved that the proposed LTGFM-RDMDDTC technique increases the overall performance of prediction

accuracy, precision, recall, and F-measure, as well as minimizes the prediction time.

References

1. Abdu Gumaei, Walaa N. Ismail, Md. Rafiul Hassan, Mohammad Mehedi Hassan, Ebtsam Mohamed, Abdullah Alelaiwi, Giancarlo Fortino, "A Decision-Level Fusion Method for COVID-19 Patient Health Prediction", Big Data Research, Elsevier, Volume 27, 2022, Pages 1-11. https://doi.org/10.1016/j.bdr.2021.100287
2. Aditya Gupta, Vibha Jain & Amritpal Singh, "Stacking Ensemble-Based Intelligent Machine Learning Model for Predicting Post-COVID-19 Complications", New Generation Computing, 2021, Pages 1-21. https://doi.org/10.1007/s00354-021-00144-0
3. Ahmed Hamed, Ahmed Sobhy, Hamed Nassar, "Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm", Arabian Journal for Science and Engineering, Springer, Volume 46, 2021, Pages 8261–8272. https://doi.org/10.1007/s13369-020-05212-z
4. Ameer Sardar Kwekha-Rashid, Heamn N. Abduljabbar & Bilal Alhayani, "Coronavirus disease (COVID-19) cases analysis using machine-learning applications", Applied Nanoscience, 2021, Pages 1-13. https://doi.org/10.1007/s13204-021-01868-7
5. Guohui Li, Kang Chen, Hong Yang, "A new hybrid prediction model of cumulative COVID-19 confirmed data", Process Safety and Environmental Protection, Elsevier, Volume 157, January 2022, Pages1-19.https://doi.org/10.1016/j.psep.2021.10.047
6. Hua Ye, Peiliang Wu, Tianru Zhu, Zhongxiang Xiao, Xie Zhang, Long Zheng, Rongwei Zheng, Yangjie Sun, Weilong Zhou, Qinlei Fu, Xinxin Ye, Ali Chen, Shuang Zheng, Ali Asghar Heidari Mingjing Wang, Jiandong Zhu, Huiling Chen, and Jifa L, "Diagnosing Coronavirus Disease 2019 (COVID-19): Efficient Harris Hawks-Inspired Fuzzy K-Nearest Neighbor Prediction Methods", IEEE Access, Volume 9, 2021, Pages 17787 – 17802. DOI: 10.1109/ACCESS.2021.3052835
7. Informatics in Medicine Unlocked, Elsevier, Volume 31,2022,Pages1-7. https://doi.org/10.1016/j.imu.2022.100990
8. Jialu Song, Hujin Xie, Bingbing Gao, Yongmin Zhong, Chengfan Gu, Kup-Sze Choi, "Maximum likelihood-based extended Kalman filter for COVID-19 prediction", Chaos, Solitons and Fractals, Elsevier, Volume 146, 2021, Pages 1-9. https://doi.org/10.1016/j.chaos.2021.110922
9. Kim Tien Ly, "A COVID-19 forecasting system using adaptive neuro-fuzzy inference", Finance Research Letters, Elsevier, Volume 41, 2021, Pages 1-10. https://doi.org/10.1016/j.frl.2020.101844
10. Luis Fernando Castillo Ossa, Pablo Chamoso, Jeferson Arango-López, Francisco Pinto-Santos,Gustavo Adolfo Isaza, Cristina Santa-Cruz-González, Alejandro Ceballos-Marquez, Guillermo Hernández and Juan M. Corchado, "A Hybrid Model for COVID-19 Monitoring and Prediction", Electronics, Volume 10, Issue 7, 2021, Pages 1-13. https://doi.org/10.3390/electronics10070799
11. Makram Soui, Nesrine Mansouri, Raed Alhamad, Marouane Kessentini, Khaled Ghedira, "NSGA-II as feature selection technique and AdaBoost

classifier for COVID-19 prediction using patient's symptoms", Nonlinear Dynamics, Springer, Volume 106, 2021, Pages 1453–1475 . https://doi.org/10.1007/s11071-021-06504-1

12. Md. Abdul Awal, Mehedi Masud, Md. Shahadat Hossain, Abdullah Al-Mamun Bulbul, S. M. Hasan Mahmud, Anupam Kumar Bairagi, "A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection From Inpatient Facility Data", IEEE Access, Jan 2021, Pages 10263 – 10281. DOI: 10.1109/ACCESS.2021.3050852
13. N. Deepa a,⇑, J. Sathya Priya b, Devi T, "Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy", Materials Today: Proceedings, Elsevier, 2022, Pages 1-15. https://doi.org/10.1016/j.matpr.2022.03.345
14. Olusola O. Abayomi-Alli, Robertas Damaševi˘cius, Rytis Maskeliunas and Sanjay Misra, "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples", Sensors, Volume 22, Issue 6, 2022, Pages 1-18. https://doi.org/10.3390/s22062224
15. Prabh Deep Singh, Rajbir Kaur, Kiran Deep Singh, Gaurav Dhiman, "A Novel Ensemble-based Classifier for Detecting the COVID-19 Disease for Infected Patients", Information Systems Frontiers, Springer, Volume 23, 2021, Pages 1385-1401. https://doi.org/10.1007/s10796-021-10132-w
16. Pradeep Kumar Roy and Abhinav Kumar, "Early prediction of COVID-19 using ensemble of transfer learning" Computers and Electrical Engineering, Elsevier, Volume 101, 2022, Pages 1-15. https://doi.org/10.1016/j.compeleceng.2022.108018
17. Saratu Yusuf Ilu, Prasad Rajesh, Hassan Mohammed, "Prediction of COVID-19 using long short-term memory by integrating principal component analysis and clustering techniques",
18. Shivam Bhardwaj, Sunil Sharma, Rashmi Bhardwaj, "Machine Learned Hybrid Gaussian Analysis of COVID-19 Pandemic in India", Results in Physics, Elsevier, Volume 30, November 2021, Pages1-22.https://doi.org/10.1016/j.rinp.2021.104630
19. Shwet Ketu and Pramod Kumar Mishra, "Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection", Applied Intelligence, Springer, Volume 51, 2021, Pages 1492–1512. https://doi.org/10.1007/s10489-020-01889-9
20. Vasyl Martsenyuk, Marcin Bernas, Aleksandra Klos-Witkowska, "Two-Strain COVID-19 Model Using Delayed Dynamic System and Big Data", IEEE Access, Volume 9, 2021, Pages 113866 – 113878. DOI: 10.1109/ACCESS.2021.3104519
21. Vishan Kumar Gupta, Avdhesh Gupta, Dinesh Kumar, Anjali Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model", Big Data Mining and Analytics , Volume 4, Issue 2, 2021, Pages 116 – 123. DOI: 10.26599/BDMA.2020.9020016