Egyptian Informatics Journal 27 (2024) 100508

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

A hybrid contextual framework to predict severity of infectious disease: COVID-19 case study

M. Mehran Bin Azam [a], Fahad Anwaar [b], Adil Mehmood Khan [b], Muhammad Anwar [a,c], Hadhrami Bin Ab Ghani [c,*], Taiseer Abdalla Elfadil Eisa [d], Abdelzahir Abdelmaboud [e]

[a] Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, Pakistan
[b] School of Computer Science, University of Hull, HU6 7RX, United Kingdom
[c] Faculty of Data Science and Computing, Universiti Malaysia Kelantan, 16100, Kota Bharu, Kelantan, Malaysia
[d] Department of Information Systems-Girls Section, King Khalid University, Mahayil, 62529, Saudi Arabia
[e] Humanities Research Center, Sultan Qaboos University, Muscat, Oman

ARTICLE INFO

Keywords:
Artificial intelligence
Natural language processing
Severity rating
Profile learner
COVID-19

ABSTRACT

Infectious disease is a particular type of disorder triggered by organisms and transmitted directly or indirectly from an infected one like COVID-19. The global economy and public health are immensely affected by COVID-19, a recently emerging infectious disease. Artificial Intelligence can be helpful to predict the severity rating of COVID-19 which assists authorities to take appropriate measures to mitigate its spread in different regions, hence it results in economic reopening and reduces the degree of mortality. In this paper, a hybrid contextual framework is proposed which incorporates content embedding of Standard Operating Procedure's (SOPs) auxiliary description along with COVID-19 temporal features of the respective region as side information. The word embedding techniques are incorporated to generate distributed representation of SOPs auxiliary description. The higher representation of auxiliary description is obtained by utilizing content embedding and then combined with temporal features to build counties profiles. These county profiles are fed into a profile learner based on an ensemble algorithm to predict the severity level of COVID-19 in different regions. The proposed contextual framework is evaluated on public datasets provided by healthdata.gov and the National Centers for Environmental Information. A comparison of the proposed contextual framework with other state-of-the-art approaches has demonstrated its ability to accurately predict the severity level of COVID-19 in different regions.

1. Introduction

In December 2019 the sudden outbreak of a novel Coronavirus in Wuhan city of China rapidly spread across the country during the spring festival. This is a family of viruses including SARS, and ARDS, surprisingly fast and propagated throughout the world. COVID-19 has been declared a Public Health Emergency by the World Health Organization (WHO) as it spreads through respiratory droplets transmitted through coughing, sneezing, or when people meet with each other in close proximity [1–3]. Then these droplets can be inhaled or land on a surface that may come in contact with a healthy person through the eyes, mouth, and nose. For example, it can live on plastic and stainless steel for a few days and on cardboard and copper for a few hours. Its symptoms usually appear within 1 to 14 days after infection. This virus may have a severe effect on patients suffering from diabetes, asthma, heart, and lung problems. Such patients require major attention and care [4–9].

The countries with high populations, elder aging, and wealthier health systems are currently in more critical situations as compared to the countries with weaker health systems, lower income, and younger age groups [10–12]. The relationship between weather variables like humidity, temperature, and the spread of COVID-19 in the respective region has earned the special attention of the researchers. It has been observed that there is a notable correlation between weather variables and the spread of COVID-19 in locations with major outbreaks and in the same temperature zone. Northeastern US, Hubei, Iran, Spain, South Korea, Italy, Japan, Germany, and England share the same average temperature of 5C–11 C, and humidity of 47%–79% are the outbreak epicentres in January and February 2020. Regions having temperatures

* Corresponding author.
E-mail addresses: bsf1702281@ue.edu.pk (M.M.B. Azam), f.anwaar-2022@hull.ac.uk (F. Anwaar), a.m.khan@hull.ac.uk (A.M. Khan), anwar.muhammad@ue.edu.pk (M. Anwar), hadhrami.ag@umk.edu.my (H.B.A. Ghani), teisa@kku.edu.sa (T.A.E. Eisa), a.elnour@squ.edu.om (A. Abdelmaboud).

https://doi.org/10.1016/j.eij.2024.100508
Received 10 November 2023; Received in revised form 1 July 2024; Accepted 17 July 2024

up to 15C and humidity 75% have less spread of the outbreak [13–15]. In the current situation smart lockdown, wearing a face mask, working from home, and hand sanitizers are only the solution. In addition to medical procedures, Artificial Intelligence (AI) has significant potential in the diagnosis of diseases through computer vision and natural language processing (NLP). Machine learning can be used to identify this novel coronavirus and help us to forecast nature and spread around the globe [16–18]. It is not necessary to give thorough care to every Coronavirus-confirmed patient. Identifying people who are more likely to suffer from acute illness will assist in providing assistance, planning, allocating, and utilizing medical resources effectively. Many countries entering into a new phase of fighting against this pandemic and at the same time, all types of industries are opening under the strict Standard Operating Procedures (SOPs) guidelines. Recommender Systems can help us in defining the new set of SOPs that have an optimal effect against the pandemic and at the same time helps to revive the industries. Artificial intelligence can effectively help smart cities cope with this COVID-19 pandemic by collecting data every minute from various embedded devices. The use of AI-powered chatbots can assist patients in self-awareness and answer questions they may not be able to get answers to during a pandemic [19–22].

Artificial Intelligence helps us in predicting the number of infected patients, identifying positively diagnosed patients who need to be hospitalized or not, and predicting whether they are likely to be successfully cured [23–26]. The use of AI can also be used to forecast future fatalities from the COVID-19 pandemic as well as calculate the mortality rate for the COVID-19 pandemic based on the number of newly infected patients [27–29]. As a large number of patients must be treated during a pandemic, it becomes very difficult to allocate resources effectively. As a result, artificial intelligence is a highly efficient tool for allocating resources during pandemics [30–32]. The process of discovering a new vaccine based on pandemic epidemiological features may take a longer time. The use of Artificial Intelligence now makes it possible to generate new studies that might lead to a better treatment for COVID-19 by learning from drugs and protein structures [33–38].

In terms of predicting disease spread and tracking, artificial intelligence has not been very successful so far. Furthermore, AI lacks historical training data for COVID-19, as well as problems with big data such as data derived from social media and noise within it, so that real trends must be filtered before they can be discerned [39]. Previous research on COVID-19 severity rating prediction techniques has primarily focused on statistical and weather features without considering how standard operating procedures (SOPs) may affect the outcome. Research using natural language processing techniques in the COVID-19 domain primarily involved sentiment analysis. However, methods used for extracting deep semantics were unable to capture the deep features in auxiliary descriptions. Our perspective motivates us to come up with an ensemble algorithm based on word embeddings that combines both statistical and weather factors to determine COVID-19's severity rating. The proposed framework can also help the organization in resource allocation based on the severity ratings within specific regions. Our research has made the following major contributions:

- A hybrid framework, HCF-SR is proposed which incorporates content embeddings of SOPs auxiliary description being followed by different regions in the USA. The contextual features are integrated with statistical and weather features to predict severity levels.
- Additionally, HCF-SR uses temporal information (active cases, recovered cases, temperature, humidity, etc.) of the region to create counties' profiles as side information; thus, extracting the deep semantics of SOP auxiliary descriptions along with temporal features that result in better county profiles and hence better prediction.
- The county profiles are incorporated into a profile learner which predicts the severity rating of COVID-19. This severity rating will assist the authorities to make appropriate plans to mitigate the spread of the pandemic, consequently reducing the mortality rate and economic reopenings.

In the remaining paper, the following sections are included: Sections Section 2 contain related works, Sections Section 3 provide a description of the proposed hybrid contextual framework for predicting severity, Sections Section 4 provide performance evaluation and analysis of experimental results. As a final section, Section 5 concludes the paper.

2. Related work

Artificial intelligence has produced encouraging outcomes in health care through its ability to make accurate decisions based on epidemiological features. AI (Artificial Intelligence) can play a key role in combating COVID-19; however, finding the appropriate solution is currently the biggest challenge facing the healthcare system.

A hybrid model (ARIMA-WBF) is proposed in [40] (a) to generate a short-term prediction of the daily number of the confirmed case in the UK, Canada, and South Korea based on statistical features, (b) An optimal autoregressive tree algorithm is used to identify important causal variables that affect mortality. However, this study is based on some assumptions (a) virus mutation rates are comparable, (b) the recovered person achieved lifelong immunity, and (c) climate changes are also ignored. A classical machine learning algorithm [41] is proposed using climate, census, and health center data to investigate temperature and humidity's effects on the spread of COVID-19. Based on the results, the decision tree algorithm yielded the best results. However, deep neural networks could be utilized for analyzing the better relationship between COVID-19 statistical and weather features. Artificial intelligence based hybrid framework (ISI + NLP + LSTM) [42] is proposed to analyze the development trends and transmission laws by estimating the variety of infection rates. The effect of raising public awareness of prevention along with the use of different control measures extracted from different news websites is considered. However, as the infection spreads and the social media traffic around it accumulates, data becomes more cluttered and noisy. Traditional machine learning models and feature extraction techniques [43] (TF-IDF, bag of words) are used to classify clinical reports into four groups (ARDS, COVID, SARS, both ARDS and COVID). However, traditional embedding techniques, including Bag-of-Words and TF-IDF, cannot capture the semantic similarity between words. A hybrid framework [44] is proposed to extract multi-time scale features from the convolution layers of a convolution neural network. Then these multi-time scale features are combined and input into the second layer of LSTM to predict hotspot location with high precision. However, a number of predictors can be added to the framework to increase its performance, such as Meta-IDVP, Meta-ID6MA, and HLPpred-Fuse. Bayesian optimization guided shallow LSTM [45] is used to predict long-term country-specific risk based on trend data (number of confirmed patients, number of deaths, number of recoveries) including weather features. However, the proposed model was trained on a small and incomplete dataset. Training of AI models required large and consistent datasets but this study is based on small, fusion and uncertain datasets.

Classical machine learning [46] algorithms are utilized for offline sentiment analysis to train the models and used for online sentiment prediction pipeline in real-time sentiment analysis based on tweets collected with the hashtags (#COVID-19, #Coronavirus). Twitter streaming APIs, Apache Spark and Kafka were used to develop an online prediction pipeline. Results shows that the random forest model based on unigram feature extraction produces the most accurate results. However, feature extraction techniques unigram and TF-IDF lack carrying the deep semantics from textual description beside these as social media traffic accumulates the noise around it also accumulates.

The MRPMC model [47] is proposed to predict the mortality risk of COVID-19 patients by utilizing clinical data. The proposed framework allows the prediction of physiological deterioration and death at least twenty days in advance. This study, however, has not evaluated MRPMC's prognostic implications in prospective cohorts due to its retrospective nature. A global deep learning-based framework has been trained by utilizing blockchain-based federated learning on a small amount of data from different healthcare institutions. Federated learning techniques and blockchain technology were used to authenticate and train the models globally while maintaining the privacy of the organizations. A hybrid framework [48] for has been proposed that consists of (a) methods for normalizing heterogeneous data (b) Capsule Networks for segmenting and categorizing COVID-19 patients (c) And a novel framework that uses blockchain technology with federated learning to train global models. The following challenges arise from federated learning (a) Each node might have some bias based on the general population, and dataset sizes may differ greatly. In addition, local dataset distributions may change with time, depending on their temporal heterogeneity.

A COUNTERACT framework [49] proposed that uses backpropagation through the subject's location data to calculate an incubation phase for the infected subject. During the incubation period, contextualize each location as a contagious location and notify exposed suspects who are moving toward the contagious location in real time. It is possible that the projected location for a specific day differs from where the individual has been during the incubation period. This is because position variation fails to track with the GPS or if the mobile phone is off. A hybrid model [50] is proposed to predict the number of confirmed patients in the USA, Germany, and Turkey for a short time frame. To improve forecasting accuracy, the proposed framework used rolling mechanism to update data in equal dimensions. For optimizing grey modeling parameters with a minimum error rate, particle swarm optimization algorithm (PSO) is employed. However, genetic algorithms or gray wolf algorithms can be used to optimize gray prediction model parameters.

An epidemic model [51] that classifies infection rates into three categories is proposed to predict COVID-19 spread by incorporating transmission rates and social distance conditions. However, the SIPHERD model was unable to consider the impact of public prevention awareness and weather variables. A hybrid model [52] using data-driven estimation techniques (Curve fitting and LSTM) is developed to predict the total number of cases in India 30 days ahead of time and how social isolation can be used to prevent the spread of the pandemic. In spite of this, the proposed model has some limitations. (i) The model was evaluated using limited data, and (ii) the initial data was almost flat.

Artificial Intelligence based speech processing framework [63] is proposed to identify COVID-19 by using biomarker extractors to extract acoustic biomarkers from cough recordings. After pre-screening cough recordings, the Mel Frequency Cepstral Coefficient is applied to them and a Convolutional Neural Network is used to diagnose COVID-19. The suggested model, however, does not account for regional, age, and cultural variations in coughs, as well as other sources of sound or input modalities, such vision and descriptions of symptoms in natural language. Logistic algorithm [64] for fitting epidemic trend caps and feeding them into the FbProphet model to estimate epidemic trends. In contrast, the proposed model assumes a maximum outbreak and models the epidemic curve using a logarithmic fit. In reality, some small peaks may occur during the pandemic because the government intervenes differently and cooperates differently with the public. A deep convolution neural network [65] was developed to extract deep semantics from chest X-ray images to discriminate between infected and non-infected COVID-19 patients. The region of interest (ROI) is extracted using multiple image processing techniques and image data generator classes are used to overcome the problem of a small image dataset. However, the proposed framework ignored other segmentation techniques like threshold-based segmentation which can accurately identify ROI. A deep learning-based mechanism INASNET [66] was developed to identify the infected patients by analyzing the chest X-ray images. INASNET framework is composed of InceptionNet and neural network search architecture to create more accurate and faster predictions. As a result, the proposed framework was trained using biased datasets since there are very few COVID-19 images compared to other classes. Various machine learning [53] methods including Decision Trees (DTs), SVM, Multilayer Perceptrons (MLPs), and K-Nearest Neighbors were used to analyze the data of 1225 COVID-19 patients between February 9, 2020, and July 20, 2021. A Horse Herd Optimization algorithm was first used to identify the most significant predictors (HOA). The retrospective study reported here has some limitations (i) Poor data quality (noisy, imbalanced, and meaningless values) and low data quantity (duplicate and missing values). (ii) Limited sample size and a single-center dataset limit the generalizability of the proposed framework. Abul Hasan et al. [54] used conditional random fields to extract from patient social media posts symptoms associated with COVID-19, such as severity, duration, negations, and parts of the body. The pipeline was then further refined by applying an unsupervised rule-based algorithm to establish relationships between concepts. A vector representation of each post was created using the extracted concepts and relations. The support vector machines were applied separately to classify patients into three categories and diagnose them. However, the proposed framework work on social media data not on doctor's consultations. A dual convolutional neural network (CNN) [55] framework was proposed using the COVID-19 PHM dataset, which contains upto 11,000 annotated tweets. Dual convolutional neural network provide additional information to the primary convolutional neural network in order to detect PHMs more effectively from tweets. However, the proposed method extracts only semantic information from tweets and ignores other information such as the time and location of tweet posts. An innovative web-based medical decision support system [56] combines near-field communication tags with a cloud-based structured data platform to facilitate remote diagnosis of COVID-19 in quarantine wards. The proposed framework works in four stages: (i) extraction of contactless health information from patients using NFC tags (ii) conversion of medical reports into auxiliary information by optical character recognition (iii) medical parameters are extracted by using natural language processing techniques (iv) in the fourth step, key parameters are visualized. The early COVID-19 warning system [57] was developed using Twitter data and machine learning algorithms. Pretrained NLP models are used to fine-tune BERT's Twitter classification. Additionally, the proposed system includes a linear regression model based on Twitter for forecasting COVID-19 outbreaks. However, the annotated dataset is too small for text classification. This study [58] leverages artificial intelligence to predict COVID-19 severity using various machine learning and deep learning algorithms, focusing on clinical markers and vital signs. It evaluates multiple pipelines combining five data-balancing techniques and twelve classifiers, identifying Random Forest with Borderline SMOTE as the best-performing model with an 83% recall. The research aims to develop an explainable decision-support system for healthcare professionals in regions with limited medical resources. However, the reliance on retrospective data may introduce biases, and the performance of the models could vary with different datasets or populations. The study does not address the potential for overfitting with complex models and multiple classifiers. A machine learning module [59] trained on data from 930 COVID-19 patients hospitalized in Italy during the first wave. The dataset includes 25 biomarkers and Chest X-ray (CXR) images. The system diagnoses low- or high-risk patients with an accuracy, sensitivity, and F1-score of 89.03%, 90.44%, and 89.03%, respectively, outperforming systems that use only CXR images or biomarker data by 6%. Additionally, it employs a multivariate logistic regression-based nomogram to calculate mortality risk. However, the study has limitations. The reliance on data from a single country and the initial pandemic wave may limit the generalizability of the results. Additionally, the performance might vary with

Table 1
This table compares HCF-SR with other leading algorithms.

Model	Input	Contextual embeddings	Retrospective data	Single-center data	Reliance on noisy social media data	Generalizability	Open source	Metric
Varzaneh et al. [53]	Clinical Data	No	Yes	Yes	No	No	No	F1-Score, Recall, Precision
Hasan et al. [54]	Social Media	Yes	Yes	Yes	Yes	No	Yes	F1-Score, Recall, Precision
Luo et al. [55]	Tweets Data	No	Yes	Yes	Yes	No	Yes	F1-Score, Recall, Precision
Balasubramanian et al. [56]	Clinical Data	No	Yes	Yes	No	No	No	F1-Score, Recall, Precision
Yiming et al. [57]	Tweets Data	Yes	Yes	Yes	Yes	No	Yes	F1-Score, Recall, Precision
Varada et al. [58]	Demographic Data, Clinical Data	No	Yes	No	No	Yes	Yes	F1-Score, Recall, Precision, Accuracy, AUC Score
Tawsifur et al. [59]	X-ray Images, Clinical Data	No	Yes	No	No	Yes	No	F1-Score, Recall, Precision
Sanzida et al. [60]	Clinical Data	No	Yes	Yes	No	No	Yes	F1-Score, Recall, Precision, Accuracy
Krishnaraj et al. [61]	Clinical Data, Demographic and Blood markers	No	Yes	No	No	Yes	No	F1-Score, Recall, Precision, Accuracy, AUC Score
M.T. Huyut [62]	Demographic Data, Clinical Data	No	Yes	No	No	Yes	No	F1-Score, Recall, Precision, Mean Absolute Error, Root Mean Square Error
HCF-SR	**Statistical Data, Weather Data, SOPs Auxiliary Description**	**Yes**	**No**	**No**	**No**	**Yes**	**Yes**	**F1-Score, Recall, Precision, Mean Absolute Error, Root Mean Square Error**

Table 2
COVID-19 severity rating.

Level of spread	Severity rating	Severity rating class
ratio > 75%	1	High Risk
50% < ratio < 75%	2	High - Medium Risk
25% <ratio < 50%	3	Low - Moderate Risk
ratio < 25%	4	Low Risk

new variants and different demographic groups. A web based machine learning framework [60] is developed to detect COVID-19. The dataset is preprocessed using techniques like feature engineering and SMOTE, was used to train various classifiers, including logistic regression, random forest, SVM, and deep learning models. The hybrid CNN-LSTM model with SMOTE showed the best performance, achieving 96.34% accuracy and a 0.98 F1 score. Explainable AI with the LIME framework was used to interpret results. This model was deployed to a website for users to obtain instant COVID-19 prognoses based on their symptoms. However, the single, open-source dataset may limit the generalizability of the model to other populations. The preprocessing techniques, such as SMOTE, could introduce biases and potential overfitting. A decision support system [61] is proposed to enhance COVID-19 diagnosis using machine learning and deep learning techniques, complementing the RT-PCR test. Patient data from two Manipal hospitals in India were used to train a stacked multi-level ensemble classifier, along with deep neural networks (DNN) and one-dimensional convolutional networks (1D-CNN). Explainable AI techniques like SHAP, ELI5, LIME, and QLattice were applied for model interpretability. This system aims to support initial screening and alleviate the burden on healthcare infrastructure. However, The dataset's regional specificity may limit the model's generalizability to other populations. The use of complex ensemble methods and deep learning techniques might result in overfitting, especially with smaller datasets. A recent study [62] aimed to ease healthcare pressures by identifying severe and mild COVID-19 cases at admission using routine blood values (RBV) and demographic data. Analyzing data from 4,202 patients hospitalized between March and September 2021, various machine learning models were applied to predict disease severity. Feature selection used MRMR, PCA, and logistic regression, identifying 28 RBV parameters and age as significant predictors. However, the study's reliance on a specific patient cohort limits generalizability. The use of complex feature selection and multiple models could introduce biases and overfitting. Our proposed model incorporates SOPs auxiliary descriptions alongside statistical and weather features to capture deep semantics and the context of SOPs, thereby building comprehensive county profiles. By leveraging contextual embeddings and county-specific data, HCF-SR minimizes bias associated with retrospective data and enhances generalizability across diverse populations and regions. Additionally, the use of ensemble algorithms and feature selection methods effectively mitigates the risk of overfitting which is a common issue with complex models. Table 1 presents a comparative analysis of the proposed model against state-of-the-art models, highlighting each model's strengths and weaknesses.

3. Hybrid contextual framework for predicting the severity rating using content embeddings

A hybrid contextual framework HCF-SR is proposed for predicting COVID-19 severity ratings for United States counties. The proposed model incorporates COVID-19 statistical and weather features with word embedding-based content extraction techniques to extract the deep semantics SOPs auxiliary descriptions. The overall computational methodology is presented in Fig. 1. Experiments are carried out on preprocessed datasets prepared from healthdata.gov and the National Centers for Environmental Information to demonstrate the model's efficacy. An example that elaborates the functionality of HCF-SR for predicting the severity rating of USA counties is provided in Table 2, Table 3, and Table 5.

3.1. System description

Our proposed hybrid contextual framework HCF-SR utilizes the SOPs description (Auxiliary information) obtained from the raw content of the healthdata.gov SOPs dataset. Auxiliary information, such as SOPs descriptions, is required for each county in the USA. After the information is obtained, natural language processing techniques such as stopword removal and tokenization are used to preprocess it. In the second step, preprocessed SOP auxiliary descriptions are used to generate vector representations of SOP auxiliary descriptions using the word embedding techniques TF-IDF and Word2Vec [67]. Word2Vec is trained along with Hierarchical Softmax. We initialized $window$= 5, min_count=1, and $size$=100 where size is used to represent then N-dimensional feature vector and min_count specifies the minimal frequency of words considered in the context. The creation process of the

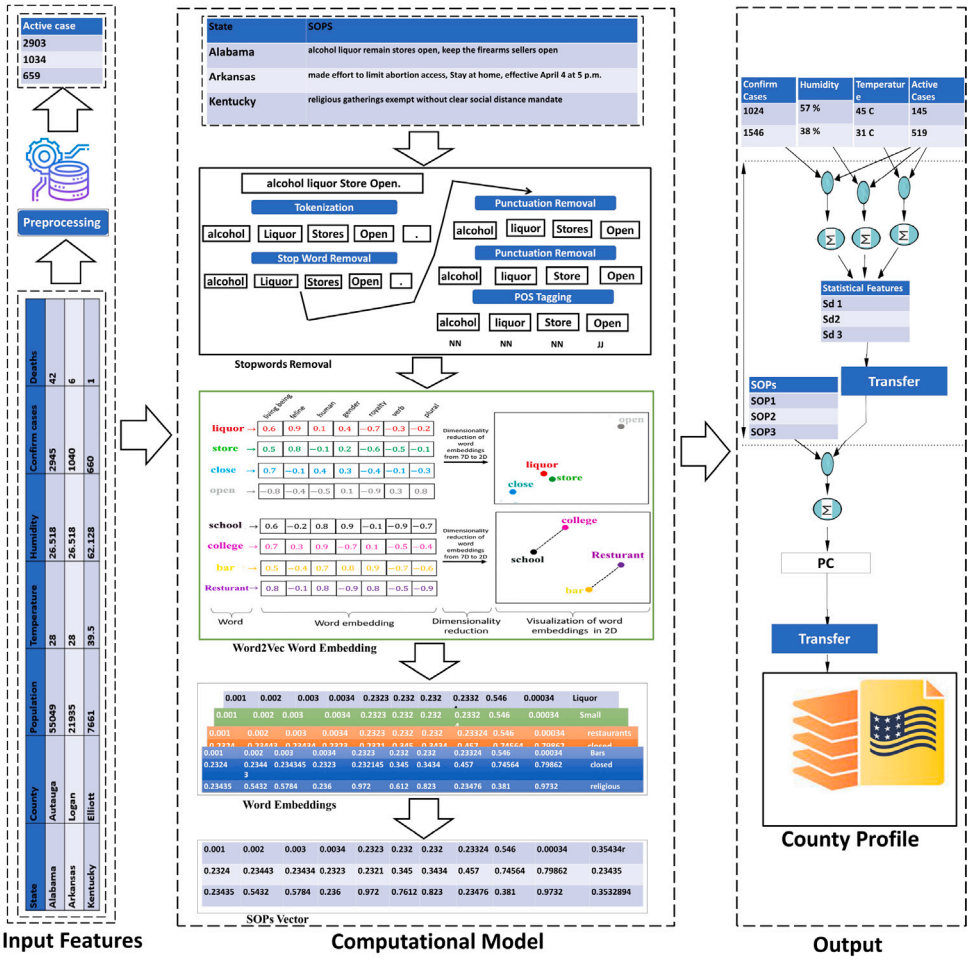


Fig. 1. Counties Profile generation using SOPs word embeddings and Statistical Features.

SOPs vector is shown in Fig. 1. The feature vectors of SOPs description along with the COVID-19 statistical and weather features (Eq. (8)) are incorporated to generate counties profiles. The proposed hybrid contextual framework HCF-SR is then utilized to predict the severity rating based on county profiles.

3.2. Mathematical problem formulation

Let $Conf_i = [Conf_1, Conf_2, \ldots., Conf_\gamma]$ represent the of number of confirmed case, $Death_i = [Death_1, Death_2, \ldots, Death_n]$ represents the number of deaths and $Rec_i = [Rec_1, Rec_2, \ldots, Rec_n]$ represents number of recovered patients based on these statistical features (Stats) we have calculated the number of active case $Active_i = [Active_1, Active_2, \ldots, Active_\alpha]$ using Eq. (1) for counties $Coun_i \in [Coun_1, Coun_2, \ldots, Coun_n]$ located within USA states $S \in [S_1, S_2, \ldots, S_\beta]$, having temperature $Temp_i \in [Temp_1, Temp_2, \ldots, Temp_n]$, Humidity $humi_i \in [humi_1, humi_2, \ldots., humi_n]$ and set of imposed SOPs description $SOP_i \in [SOP_1, SOP_2, \ldots, SOP_s]$.

3.3. Severity of the COVID-19 pandemic

COVID-19 statistical and weather features along with the SOPs description are utilized to predict the severity rating for USA counties. The number of active cases $Active_i$ of the current situation for each county is calculated using Eq. (1).

$$Active_i = Conf_i - Death_i - Rec_i \tag{1}$$

According to Eq. (1), the total number of active cases in a county ($Active_i$) is equal to the total number of confirmed cases ($Conf_i$) minus total number of recovered cases (Rec_i) and the total number of deaths. The COVID-19 statistical and SOPs auxiliary description is extracted from healthdata.gov [68]. Weather features are extracted from National Centers for Environmental Information dataset [69].

$$ratio_i = \frac{Active_i}{\sum_{J=1}^{n} Active_j} \tag{2}$$

As $ratio_i$ increases the spread of COVID-19 in a county becomes more severe as shown in Table 2.

3.4. Input feature

In the proposed hybrid contextual framework HCF-SR the combination of input features i.e. statistical features (Stats), preprocessed SOP description extracted from healthdata.gov and weather features gathered from the National Centers for Environmental Information (NCEI) dataset are incorporated to predict the severity rating of COVID-19. Let $SOP = \{SOP_1, SOP_2, SOP_3, \ldots, SOP_n\}$ be the set of imposed SOPs in different counties $Coun_i \in [Coun_1, Coun_2, \ldots., Coun_n]$ of USA along with the combination of statistical and weather features represented as S_d shown in Tables 3 and 4. SOPs input feature composed of words $w_1, w_2, w_3, w_4, \ldots, w_l$. The SOPs description D for a county of USA State is obtained as:

$$D \xleftarrow[\text{SOPs Description}]{} \{w_1, w_2, w_3, w_4, \ldots, w_n\}$$

Table 3
SOPs description and word embeddings.

County	SOPs description (Pre-processed)	Vector embeddings
Calcasieu Parish	D_1 = residents shelter home essential tasks effective April 12	V_1 =Word2Vec{D_1}
Montgomery	D_2 = alcohol liquor stores open	V_2 =Word2Vec{D_2}
Anchorage Municipality	D_3 = The order allows dine-in restaurants to resume serving customers and allows retail shops to reopen	V_3 =Word2Vec{D_3}

Table 4
Example: Statistical features (Stats).

State	SOPs	Confirm case	Deaths	Recoveries	Active cases	County	Temperature	Humidity
LA	SOP_1	1000	117	80	803	Calcasieu Parish	28 C	72%
AL	SOP_2	2000	400	600	1000	Montgomery	29 C	52%
Ak	SOP_3	3000	500	1000	1500	Anchorage Municipality	29 C	51%

Table 5
Example: Counties profiles.

County profile	Mean of vector embeddings, statistical data
$P_{CalcasieuParish}$	$P = [(V_1 + V_n)/n, S_d]$
$P_{Montgomery}$	$P = [(V_1 + V_n)/n, S_d]$
$P_{AnchorageMunicipality}$	$P = [(V_1 + V_n)/n, S_d]$

The total number of words in the SOPs for a county in the USA is represented by n.

The textual description of SOPs is used to generate a corpus C, which includes all the SOPs descriptions (D) implemented in a county ($Coun$) during the pandemic and is represented as follows:

$$C \xleftarrow[\text{Corpus Generation}]{} \{D_1, D_2, D_3, D_4, \ldots, D_\beta\}$$

Where $C \in \{w_1, w_2, w_3, w_4, \ldots, w_l, w_{l+1}, \ldots, w_t\}$, t represents the total number of words in a corpus C. By utilizing the corpus C, the proposed framework builds county profiles (P_{Coun}) along with statistical and weather information as S_d.

3.5. Feature extraction from text corpus

Hybrid Contextual Framework (HCF-SR) provides distributed representations of SOP auxiliary descriptions in a vector space by using the Word2Vec [67] algorithm. It employs unsupervised learning techniques to identify word representations in large text corpora within the context of preserving regularities among vectors. Word2vec captures these regularities with two models: (1) Skip-Gram (SG) model, in which words are predicted from the target context, and (2) Continuous Bag of Words (CBOW), in which words are predicted from the target context. Using a sequence of training data words such as $w_1, w_2, w_3, w_4, \ldots, w_T$ for a given corpus C, the proposed framework maximizes log probability using the SG model.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} log\, p(w_{t+j}|w_t) \tag{3}$$

Assume that $T = |C|$, where C represents the size of the training window. Using hierarchical softmax [67] as a basis for defining the basic probability $p(w_{t+j}|w_t)$ of the output word, the following is proposed:

$$p(w_{t+j}|w_t) = \Pi_{i=1}^{L(w_{t+j})-1}\{\sigma(\mathbb{1}(n(w_{t+j}, i+1) = child(n(w_{t+j}, i))).v_{n(w_{t+j},i)}.v_{w_t})\} \tag{4}$$

Where $\sigma(x) = 1/(1+\exp(-x))$, in the binary tree, $L(w)$ represents the path length from the root to w, where $n(w,1)$ represents the root node, and $n(w,/,L(w))$ represents the path length from w to w. The Child (n) node indicates the fixed child of the node n. Vector representations of inner nodes are v_n, and vector representations of input words are v_{w_t}. If x is true, then $\mathbb{1}$ has value 1, otherwise it has value −1. Eqs. (3) and (4) are used to compute the vector representation of each word w. The preferred technique for calculating the vector notation for each county ($Coun_i$) of USA state is [67], since it is based on word embeddings, which provides the best composite representation of text and their similarity:

$$V_{C_{Coun_i}} = \frac{1}{|l|}\sum_{w=1}^{l} V_l \tag{5}$$

$V_{C_{Coun_i}}$ represents the vector representation of each word in the text corpus. where $Coun_i \in [Coun_1, Coun_2, \ldots., Coun_n]$. l represents a set of words depicting the properties of SOPs enforced within a County ($Coun_i$) of the USA state and $l \neq 0$. The complete creation process of SOPs vectors elaborated in Fig. 1 and Table 3.

3.6. County profile generation

The proposed hybrid contextual framework HCF-SR takes account into the content feature of SOPs auxiliary description vectors $\bar{V}_{C_{Coun_i}}$, statistical and weather features as S_d along with the known rating to predict the severity rating.

$$S_d \in \{Active_i, Conf_i, death_i, Rec_i, Coun_i, Temp_i, humi_i\} \tag{6}$$

$$Sta.SOPs \in \{Active_i, Conf_i, death_i, Rec_i, Coun_i, \bar{V}_{C_{Coun}}\} \tag{7}$$

County profile (P_{Coun}) for each county is generated by combining statistical and weather features along with the SOPs auxiliary description vectors($\bar{V}_{C_{Coun}}$):

$$P_{Coun} \xleftarrow[\text{County profile}]{} \{\bar{V}_{C_{Coun}}, S_d\} \tag{8}$$

$\bar{V}_{C_{Coun_i}}$ the vector of the complete SOPs description can be calculated by taking the centroid of the SOPs vector representation of a county ($Coun_i$) using Eq. (9):

$$\bar{V}_{C_{\hat{Coun}_i}} = \frac{1}{|M|}\sum_{j=1}^{M} V'_{C_j} \tag{9}$$

where $V'_{C_j} \subset \bar{V}_{C_{Coun_i}}$, $j \subset Coun_i$ and M is total number of SOPs description enforced with in a county. County profiles for all the counties of USA states are generated by utilizing the feature vectors along with the statistical features (Stats) and weather features. The structure of county profiles is shown in Fig. 1 and an example can be found in Table 5.

Algorithm 1 Counties Profile Generation

Input$_1$: SOPs auxiliary description
Input$_2$: $\leftarrow \{Active_i; Conf_i; death_i; Rec_i; Coun_i\}$
Input$_3$: $\leftarrow \{Temp_i; humi_i\}$
Output : Each SOP statement $SOP_i \in SOPs$ represented with a collection of tokens in a refined word list TK
for each SOP statement $SOP_i \in SOPs$ **do**
  $TK \leftarrow tokenize(SOP_i)$
  for each token $tk_i \in TK$ **do**
    $tk_i \leftarrow tk_i.SpellChecker()$
    $tk_i \leftarrow stem(tk_i)$
    $tk_i \leftarrow tagger(tk_i)$
    $tk_i \leftarrow lemmatize(tk_i)$
    return SOP statements SOP_i with a list of TK refined words
  end for
end for
D, mapping dictionary from n-grams to integers
$|V|$, Vocabulary Size
for $R \in tk_i$ **do**
  $n-grams \leftarrow n-gram[R]$
  for $n-gram \in n-grams$ **do**
    $D[n-gram] \leftarrow D[n-gram] + 1$
  end for
end for
$keys \leftarrow sort(D)$
$D.clear()$
$i \leftarrow 1$
while $i \leq |V|$ **do**
  $D[key[i] \leftarrow i$
  $i \leftarrow i + 1$
end while
$write(D)$
li, list of integers
vv, 2D vector of integers
for $R \in tk_i$ **do**
  v_i, a vector of integer
  $n-grams \leftarrow n-gram[R]$
  for $n-gram \in n-grams$ **do**
    $l_i.push(D[n-gram])$
    $v_i.push(D[n-gram])$
  end for
  $vv.push(v_i)$
end for
$write(l_i)$
$write(vv)$

3.7. Hybrid contextual framework HCF-SR as profile learner

A Hybrid Contextual Framework HCF-SR is utilized to predict the severity rating for USA counties by working as a profile learner. The Hybrid Contextual Framework is a supervised learning approach that trains on the preprocessed dataset to predict the severity rating. Counties profiles P_{Coun} are utilized to train the Hybrid Contextual Framework HCF-SR. HCF-SR is composed of more than one model, each model of the proposed hybrid contextual framework predicts a single class from the targeted classes. After each prediction from a specific model, the counter of the targeted class gets updated. When all the models of the proposed hybrid model predict the severity rating, the class with the highest counter is predicted as the final prediction. Fig. 2 illustrates the structure of the hybrid contextual framework proposed.

The proposed Hybrid Contextual Framework HCF-SR has the precision benefit over the classical machine learning algorithms using both word embedding techniques i.e. TF-IDF and Word2Vec. In comparison to classical machine learning algorithms, the Hybrid Contextual Framework is able to handle large dimensions of input more efficiently.

Algorithm 2 Hybrid Contextual Framework

L_i, list of files with integer list
$first\_run \leftarrow true$
for $l_i \in L_i$ **do**
  if $first\_run$ **then**
    $E \leftarrow Createmodel(l_i)$
    $first\_run = false$
  else
    $E \leftarrow Updatemodel(l_i, E)$
  end if
end for
$SOP_i\_vectors \leftarrow X.avg()$
$D \leftarrow combine(Input_2, Input_3, SOP_i\_vectors)$
$trainEnsemble \in \{model_1, model_2, ..., model_n\}$
$predictions = trainEnsemble(D)$
$severityClass = \{1 = 0, 2 = 0, 3 = 0, 4 = 0\}$
FUNCTION VoteMechanism(predictions)
if $predictions$ in $severityClass$ **then**
  $severityClass[Predictions] += 1$
end if
EndFunction
$VoteMechanism(predictions)$
$vote = max(count.values())$
$key\_list = list(count.keys())$
$val\_list = list(count.values())$
$position = val\_list.index(vote)$
$key\_list[position]$
Function voting(prediction)
if $prediction$ in $count$ **then**
  $count[prediction] += 1$
end if
EndFunction
$voting(prediction)$
$vote = max(count.values())$
$key\_list = list(count.keys())$
$val\_list = list(count.values())$
$position = val\_list.index(vote)$

Thus, proposed HCF-SR provides more compact and effective features to predict severity ratings of COVID-19 pandemic.

3.8. An illustrative example to demonstrate the working of HCF-SR

The working of the proposed Hybrid Contextual Framework HCF-SR is demonstrated using an illustrative example comprising three counties (Calcasieu Parish, Montgomery, Anchorage Municipality) are rated from 1 to 4 using a rating scale as shown in Table 2. The value of the rating scale ranges from 1 to 4. The three counties' SOPs auxiliary descriptions are extracted from the healthdata.gov dataset. Natural language processing techniques are used to remove Stop words from the SOPs auxiliary description, which is represented as D_1, D_2, D_3 for the three counties. As a result of these descriptions, Word2Vec is used to extract the content embeddings, resulting in vector space representations V_1, V_2, V_3 for the three counties. The process is shown in Table 7. These vectorial representations along with the statistical and weather features as S_d are utilized to generate the counties profiles for each county as shown in Table 5.

4. Performance evaluation

The proposed Hybrid Contextual Framework HCF-SR is evaluated in this section. A description of the dataset is presented in the subsection Dataset Preparation, followed by a description of the preparation steps. We discuss the experimental setting and results, before comparing

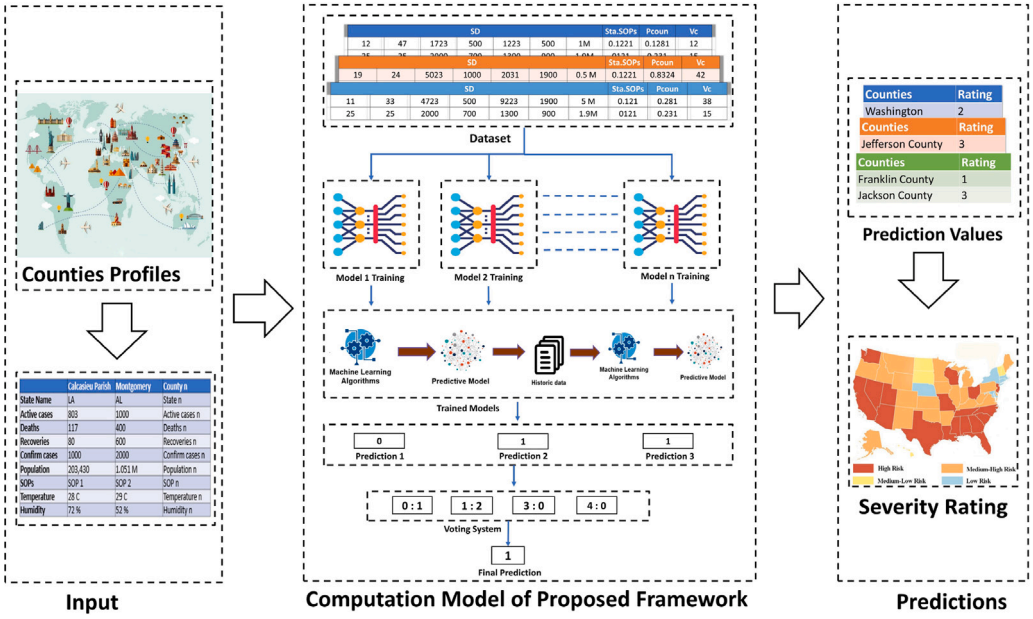


Fig. 2. Proposed Hybrid contextual framework.

Table 6
Dataset statistic.

States	Counties	Profiles	Rating Scale	Max. Avg. Temp.	Min. Avg. Temp.	Max. Humidity	Min. Humidity
55	1929	1929	[1–4]	19.07 C	5.77 C	74.0%	38.3%

Table 7
Illustrative example of HCF-SR.

County profile	Severity rating
$P_{CalcasieuParish}$	1
$P_{Montgomery}$	4
$P_{AnchorageMunicipality}$	3

it with the classical machine learning algorithms based on different combinations of features.

4.1. Dataset preparation

The proposed Hybrid Contextual Framework (HCF-SR) is evaluated on a preprocessed dataset formed by combining the features from different state-of-the-art COVID-19 statistical and weather datasets along with the SOPs auxiliary description extracted from benchmark COVID-19 datasets provided by healthdata.gov[1] [68] and National Centers for Environmental Information[2] [69]. The National Centers for Environmental Information and healthdata.gov, are the agencies of the United States government. SOPs auxiliary descriptions are collected from a dataset provided by healthdata.gov [68] implemented during the COVID-19 time period. The SOPs auxiliary information contains irrelevant words, symbols, and links that do not provide semantic meanings. In order to remove the noise from SOPs auxiliary description, some natural language processing techniques like stemming, word tokenization, stop word removal and regex scripts are utilized to remove the words, links, and symbols which do not carry semantic content, such as the, of, etc. SOPs descriptions are then converted into vector representations using Word2Ved and TF-IDF word embeddings. SOPs vector $V_{C_{Coun_i}}$ are computed for each word vector V_l using Eq. (5). Preprocessed SOPs descriptions are then incorporated along with the statistical and weather features of COVID-19 to shape counties' profiles.

The proposed dataset is collected from January 2020 to December 2020 and consists of 54 states with 1929 counties having a maximum temperature of 19.07C, maximum humidity of 74.0% and minimum temperature of 5.77C, minimum humidity of 38.3% along with SOPs Auxiliary description which is composed of 5731 unique words. The rating scale described in the proposed dataset ranges from 1 to 4 and the label for each rating class is described in Table 2. Complete statistics of the proposed dataset are elaborated in Table 6. COVID-19 severity rating prediction scenarios are stimulated by dividing the datasets into two disjoint subsets, the Training dataset and the Testing dataset. The training dataset contains 80% of the counties' profiles, whereas the testing datasets contain the remaining profiles. The training set is used to calculate the prediction using the proposed Hybrid Contextual Framework HCF-SR, while the testing set measures the prediction performance of the proposed framework.

4.2. Evaluation metric

Various assessment metrics, such as Root Mean Square Error (RMSE), and Mean Squared Error (MSE), are used to evaluate the effectiveness of the proposed hybrid contextual framework. The RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{P_{Coun}}(\tilde{x}_{P_{Coun}} - \hat{x}_{P_{Coun}})^2} \quad (10)$$

[1] https://healthdata.gov
[2] https://www.ncei.noaa.gov/access/search/dataset-search

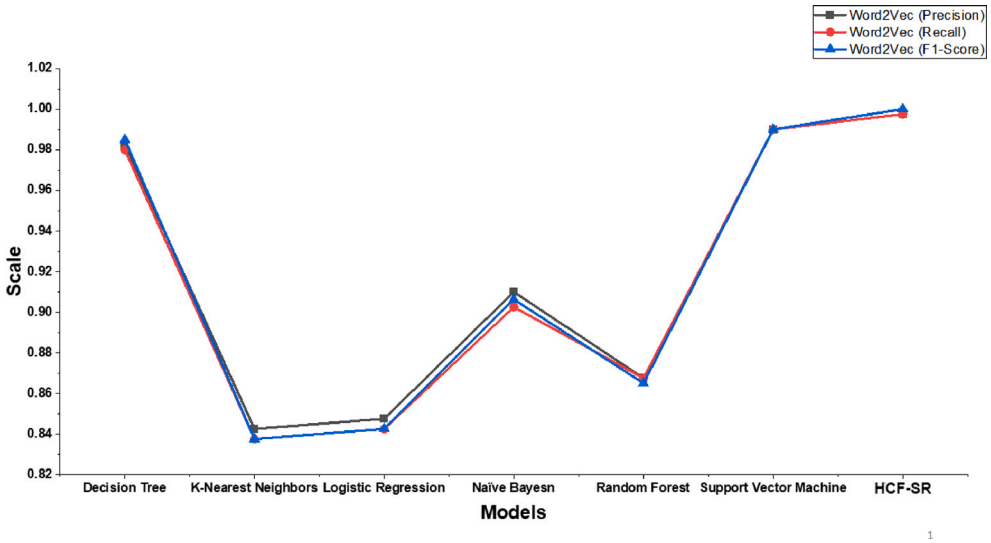


Fig. 3. Performance Comparison with Machine Learning Models.

MSE is calculated as follows:

$$MSE = \frac{1}{N}\sum_{P_{Coun}} (\bar{x}_{P_{Coun}} - \hat{x}_{P_{Coun}})^2 \tag{11}$$

where N refers to the total number of predicted ratings, $\bar{x}_{P_{Coun}}$ and $\hat{x}_{P_{Coun}}$ represent predicted and actual ratings for USA state county ($Coun_i$) based on the county profile (P_{Coun}) respectively. We have measured the quality of the predictions using precision, recall, and F1-score metrics, commonly used in information retrieval.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{14}$$

where TP denotes true positive (actual severity rating and predicted severity rating). FP is false positive (not actual severity rating and predicted severity rating). FN is a false negative (actual severity rating and not predicted severity rating).

4.3. Results and analysis

The performance of the proposed Hybrid Contextual Framework HCF-SR is evaluated on the preprocessed dataset using 5-fold cross-validation. Experiments are conducted using different combinations of features individually and in combination to examine the impact of the hybrid contextual framework HCF-SR. In each combination of features as the value of k increases, the value of MSE and RMSE decreases. The value of MSE and RMSE is lower as compared to other classical machine learning algorithms at the different values of k. Sklearn confusion matrix is utilized to compute recall, precision and f1-score shown in Table 9, Table 8 and Fig. 3. The proposed Hybrid Contextual Framework HCF-SR yields better results in terms of predicting the severity rating of COVID-19 for USA counties.

The performance of the proposed model HCF-SR using the different combination of features, only statistical features (Stats) 1.0904 RMSE, the combination of statistical and weather features as S_d 0.3491 RMSE, statistical features (Stats) along with the SOPs auxiliary description 0.2117 RMSE shown in Table 10, Table 11 and Fig. 5.

The Performance of the proposed Hybrid Contextual Framework is also compared using two different word embedding techniques i.e. Word2Vec and TF-IDF in Tables 12, 13 and Fig. 6 shows that the proposed framework outperforms using counties profiles with Word2Vec decrease the RMSE to 0.0455.

A comparison of the proposed model to other best-performing hybrid models reported in the literature is also conducted. The Table 14 and Fig. 4 summarizes the comparison between the proposed hybrid model and other Best Performing Systems, including [53–62].

The exceptional performance of the Hybrid Contextual Framework HCF-SR can be attributed to its unique approach compared to other

Table 8
Performance comparison using Word2Vec.

Model	Precision	Recall	F1-Score
Logistic regression	0.8475	0.8425	0.8425
Naïve bayes	0.91	0.9025	0.9063
Decision tree	0.9825	0.98	0.985
k-nearest neighbors	0.8425	0.8365	0.8375
Random forest	0.8675	0.8675	0.865
SVM	0.99	0.99	0.99
HCF-SR	**0.9975**	**0.9975**	**1**

Table 9
Performance comparison using TF-IDF.

Model	Precision	Recall	F1-Score
Logistic regression	0.7675	0.0.7725	0.76
Naïve bayes	0.9125	0.91	0.9075
Decision tree	0.995	0.9925	0.9925
k-nearest neighbors	0.8375	0.825	0.825
Random forest	0.865	0.8675	0.8675
SVM	0.975	0.975	0.9775
HCF-SR	**0.99**	**0.99**	**0.99**

Table 10
Performance comparison with MSE using SOPs Auxiliary Description and Temporal Features.

Model	Stats	S_d	Sta.SOPs	P_{Coun}
Logistic regression	1.4	0.4097	0.3419	0.1597
Naïve bayes	1.5612	0.5604	0.3017	0.2883
Decision tree	1.5083	0.3474	0.2517	0.0165
k-nearest neighbors	1.2505	0.5010	0.3224	0.1929
Random forest	1.3229	0.4684	0.3529	0.1701
SVM	1.4632	0.2962	0.2014	0.0103
HCF-SR	**1.0904**	**0.2334**	**0.1519**	**0.0020**

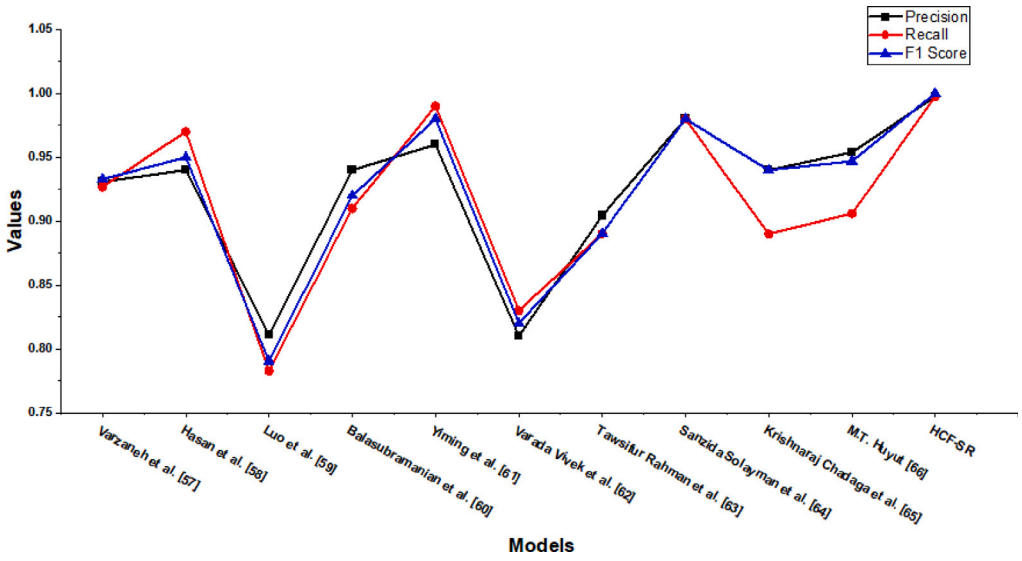


Fig. 4. Performance comparison with other best-performing state-of-the-art models.

Table 11
Performance comparison with RMSE using SOPs Auxiliary Description and Temporal Features.

Model	Stats	S_d	Sta.SOPs	P_{Coun}
Logistic regression	1.6472	0.6661	0.3515	0.2406
Naïve bayes	1.8755	0.6949	0.3274	0.2614
Decision tree	1.6912	0.4899	0.3065	0.1288
k-nearest neighbors	1.4981	0.5456	0.3819	0.1937
Random forest	1.5381	0.7521	0.3138	0.1479
SVM	1.6020	0.4245	0.2576	0.1018
HCF-SR	**1.3020**	**0.3491**	**0.2117**	**0.0455**

Table 12
Performance comparison with Counties Profiles using TF-IDF.

Model	MSE	RMSE
Logistic regression	0.2406	0.4905
Naïve bayes	0.2614	0.5112
Decision tree	0.0062	0.0788
k-nearest neighbors	0.1937	0.4401
Random forest	0.1479	0.3845
SVM	0.0125	0.1118
HCF-SR	**0.0083**	**0.0912**

Table 13
Performance comparison with Counties Profiles Word2Vec.

Model	MSE	RMSE
Logistic regression	0.1597	0.3996
Naïve bayes	0.2883	0.5370
Decision tree	0.0165	0.1288
k-nearest neighbors	0.1929	0.4392
Random forest	0.1701	0.4124
SVM	0.0103	0.1018
HCF-SR	**0.0020**	**0.0455**

best-performing algorithms for COVID-19 severity rating prediction in the literature. While previous models do not incorporate the SOPs auxiliary description, the HCF-SR framework utilizes these descriptions

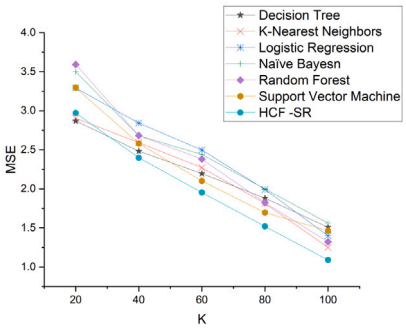
Table 14
A comparison between the proposed model and other high-performing algorithms.

Model	Precision	Recall	F1-Score
Varzaneh et al. [53]	0.931	0.927	0.933
Hasan et al. [54]	0.94	0.97	0.95
Luo et al. [55]	0.8111	0.7829	0.7907
Balasubramanian et al. [56]	0.94	0.91	0.92
Yiming et al. [57]	0.96	0.99	0.98
Varada et al. [58]	0.81	0.83	0.82
Tawsifur et al. [59]	0.9044	0.8903	0.8903
Sanzida et al. [60]	0.98	0.98	0.98
Krishnaraj et al. [61]	0.94	0.89	0.94
M.T. Huyut [62]	0.954	0.906	0.947
HCF-SR	**0.9975**	**0.9975**	**1**

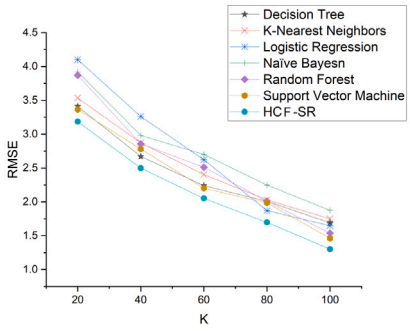
in conjunction with statistical and weather features such as active cases, deaths, temperature, and humidity. By capturing the deep semantics of SOPs descriptions along with statistical and weather data, HCF-SR creates a comprehensive county profile that significantly enhances the accuracy of severity rating predictions for COVID-19. Moreover, the proposed model addresses several critical limitations found in previous studies show in Table 1. By incorporating contextual embeddings and county-specific data, HCF-SR reduces bias associated with retrospective data and improves generalizability across different populations and regions. The use of ensemble algorithms and feature selection methods effectively mitigates the risk of overfitting, a common issue with complex models. Additionally, external validation on public datasets demonstrates the model's applicability and effectiveness beyond a single location or demographic, further underscoring its robustness and reliability for real-world implementation.

5. Conclusion

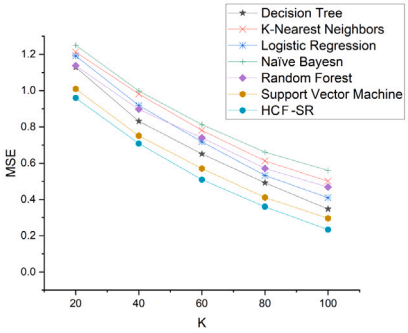
This paper proposes a Hybrid Contextual Framework HCF-SR based on content embedding. Two different word embedding techniques are being utilized to extract deep semantics of auxiliary description. The model predicts the COVID-19 severity rating by exploiting SOPs description. Distributed representation of SOPs auxiliary descriptions is produced using a word embedding model. The proposed Hybrid Contextual Framework HCF-SR incorporates the SOPs auxiliary description



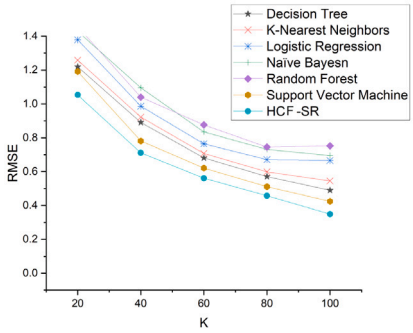
(a) MSE : Statistical Features comparison



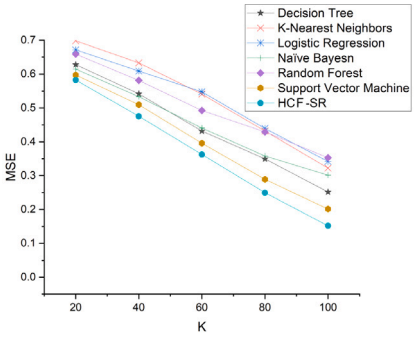
(b) RMSE : Statistical Features comparison



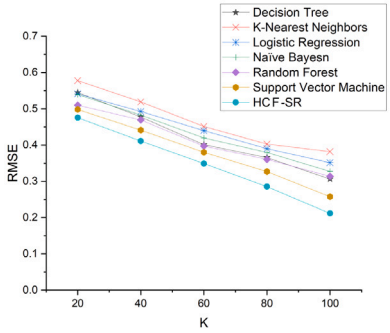
(c) MSE : Weather Features comparison



(d) RMSE : Weather Features comparison



(e) MSE : Statistical and SOPs auxiliary description comparison



(f) RMSE : Statistical and SOPs auxiliary description comparison

Fig. 5. Performance comparison using a different combination of features.

along with the statistical and weather features to predict the COVID-19 severity rating for USA counties by constructing counties profiles. Our model is evaluated on a preprocessed dataset formed by combining the features from state-of-the-art datasets obtained from healthdata.gov and National Centers for Environmental Information datasets. Comparison with classical machine learning prediction algorithms and other best-performing algorithms shows significant improvements in performance. The use of the SOPs auxiliary description along with the combination of statistical and weather features, using content embedding, captures the deep semantics of SOPs auxiliary description, leading to much better results than just using statistical information (Stats), weather data, or combined statistical information with the auxiliary description of SOPs. The proposed Hybrid Contextual Framework HCF-SR is not only limited to COVID-19 severity rating, it can also be applicable to other domains such as cardiovascular disease, lung diseases, e.t.c. to predict severe health conditions based on clinical notes. A rich set of auxiliary information about SOPs is required to predict COVID-19 severity ratings with the proposed hybrid context framework HCF-SR. In this study we only focus on the COVID-19 pandemic, in the future, we intend to extend our framework to predict the severe health conditions for other diseases based on statistical, and epidemiological features along with clinical notes auxiliary information.

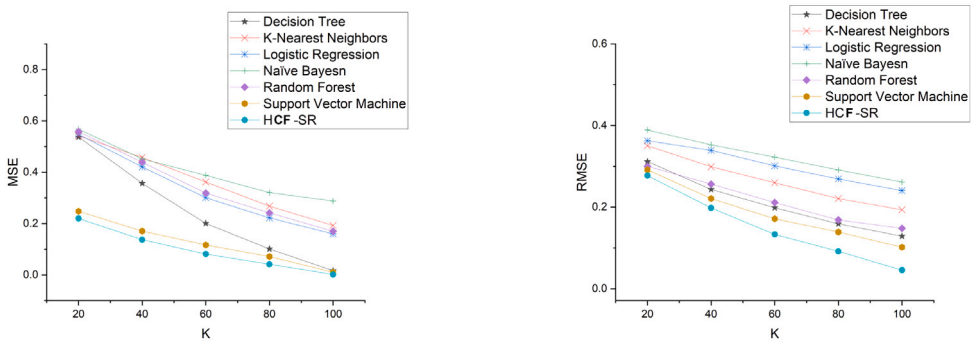


Fig. 6. Performance comparison using Counties profiles.

CRediT authorship contribution statement

M. Mehran Bin Azam: Writing – review & editing. **Fahad Anwaar:** Investigation. **Adil Mehmood Khan:** Formal analysis. **Muhammad Anwar:** Investigation, Visualization. **Hadhrami Bin Ab Ghani:** Supervision, Writing – review & editing. **Taiseer Abdalla Elfadil Eisa:** Resources. **Abdelzahir Abdelmaboud:** Data curation, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their gratitude to those who contributed to the writing of this article and make some valuable comments.

Funding statement

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/259/45.

References

[1] Sohrabi Catrin, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020;76:71–6.
[2] Zhu Na, et al. A novel coronavirus from patients with pneumonia in China, 2019. New Engl J Med 2020.
[3] Wang Chen, et al. A novel coronavirus outbreak of global health concern. Lancet 2020;395(10223):470–3.
[4] Boldog Péter, et al. Risk assessment of novel coronavirus COVID-19 outbreaks outside China. J Clin Med 2020;9(2):571.
[5] Wu Zunyou, McGoogan Jennifer M. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. jama 2020;323(13):1239–42.
[6] Sun Liping, et al. Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. J Clin Virol 2020;128:104431.
[7] Li Wendong, et al. Bats are natural reservoirs of SARS-like coronaviruses. Science 2005;310(5748):676–9.
[8] Vaishya Raju, et al. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr: Clin Res Rev 2020;14(4):337–9.
[9] Real de Asua Diego, et al. Comparison of COVID-19 and non-COVID-19 pneumonia in down syndrome. J Clin Med 2021;10(16):3748.
[10] Holshue Michelle L, et al. First case of 2019 novel coronavirus in the United States. New Engl J Med 2020.
[11] Rothe Camilla, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. New Engl J Med 2020;382(10):970–1.
[12] Razai Mohammad S, et al. Coronavirus disease 2019 (covid-19): a guide for UK GPs. Bmj 2020;368.
[13] Sajadi Mohammad M, et al. Temperature, humidity, and latitude analysis to predict potential spread and seasonality for COVID-19. Soc Sci Res Netw 2020.
[14] Dangi Ravi Rai, George Mathew. Temperature, population and longitudinal analysis to predict potential spread for COVID-19. In: Population and longitudinal analysis to predict potential spread for COVID-19 (March 24 2020). 2020.
[15] Demongeot Jacques, Flet-Berliac Yannis, Seligmann Hervé. Temperature decreases spread parameters of the new Covid-19 case dynamics. Biology 2020;9(5):94.
[16] Lalmuanawma Samuel, Hussain Jamal, Chhakchhuak Lalrinfela. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020;139:110059.
[17] Bullock Joseph, et al. Mapping the landscape of artificial intelligence applications against COVID-19. J Artificial Intelligence Res 2020;69:807–45.
[18] Yadav Milind, Perumal Murukessan, Srinivas M. Analysis on novel coronavirus (COVID-19) using machine learning methods. Chaos Solitons Fractals 2020;139:110050.
[19] Allam Zaheer, Jones David S. On the coronavirus (COVID-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. Healthcare 2020.
[20] Kunzmann Klaus R. Smart cities after COVID-19: Ten narratives. disP- Plan Rev 2020;56(2):20–31.
[21] Xu Chunwen, et al. The 2019-nCoV epidemic control strategies and future challenges of building healthy smart cities. Indoor Built Environ 2020;29(5):639–44.
[22] Ijab Mohamad Taha, Shahril Mohamad Syahmi, Hamid Suraya. Infodemiology framework for COVID-19 and future pandemics using artificial intelligence to address misinformation and disinformation. In: International visual informatics conference. Cham: Springer; 2021.
[23] Ulhaq Anwaar, et al. Computer vision for COVID-19 control: a survey. 2020, arXiv preprint arXiv:2004.09420.
[24] Khemasuwan D, Sorensen JS, Colt HG. Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19. Eur Respir Rev 2020;29(157).
[25] Ulhaq Anwaar, et al. Covid-19 control by computer vision approaches: A survey. IEEE Access 2020;8:179437-179456.
[26] Bhargava A, Bansal A. Novel coronavirus (COVID-19) diagnosis using computer vision and artificial intelligence techniques: a review. Multimedia Tools Appl 2021;80(13):19931–46.
[27] Rustam Furqan, et al. COVID-19 future forecasting using supervised machine learning models. IEEE Access 2020;8:101489–99.
[28] Pham Quoc-Viet, et al. Artificial intelligence (AI) and big data for coronavirus (COVID-19) pandemic: a survey on the state-of-the-arts. IEEE Access 2020;8:130820.
[29] Wynants Laure, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj 2020;369.
[30] van Der Schaar Mihaela, et al. How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. Mach Learn 2021;110(1):1–14.
[31] Assaf D, Gutman YA, Neuman Y, Segal G, Amit S, Gefen-Halevi S, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. Intern Emerg Med 2020;15(8):1435–43.
[32] Rela Iskandar Zainuddin, et al. COVID-19 risk management and stakeholder action strategies: conceptual frameworks for community resilience in the context of Indonesia. Int J Environ Res Public Health 2022;19(15):8908.
[33] Alimadadi Ahmad, et al. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics 2020;52(4):200–2.

[34] Lalmuanawma Samuel, Hussain Jamal, Chhakchhuak Lalrinfela. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020;139:110059.
[35] Mohanty Sweta, et al. Application of artificial intelligence in COVID-19 drug repurposing. Diabetes Metab Syndr: Clin Res Rev 2020;14(5):1027–31.
[36] Keshavarzi Arshadi Arash, et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artif Intell 2020;3.
[37] Keshavarzi Arshadi Arash, et al. Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artif Intell 2020;3.
[38] Mohapatra Sovesh, et al. Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking. PLoS One 2020;15(11):e0241543.
[39] Naudé Wim. Artificial intelligence vs COVID-19: limitations, constraints and pitfalls. AI Soc 2020;35(3):761–5.
[40] Chakraborty Tanujit, Ghosh Indrajit. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. Chaos Solitons Fractals 2020;135:109850.
[41] Malki Zohair, et al. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Chaos Solitons Fractals 2020;138:110137.
[42] Zheng Nanning, et al. Predicting COVID-19 in China using hybrid AI model. IEEE Trans Cybern 2020;50(7):2891–904.
[43] Khanday Akib Mohi Ud Din, et al. Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol 2020;12(3):731–9.
[44] Khan Sultan Daud, Alarabi Louai, Basalamah Saleh. Toward smart lockdown: a novel approach for COVID-19 hotspots prediction using a deep hybrid neural network. Computers 2020;9(4):99.
[45] Pal Ratnabali, et al. Neural network based country wise risk prediction of COVID-19. Appl Sci 2020;10(18):6448.
[46] Zhang Xiongwei, et al. Predicting coronavirus pandemic in real-time using machine learning and big data streaming system. Complexity 2020;2020.
[47] Gao Yue, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. Nat Commun 2020;11(1):1–10.
[48] Kumar Rajesh, et al. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. IEEE Sens J 2021;21(14):16301–14.
[49] Ghayvat Hemant, et al. Recognizing suspect and predicting the spread of contagion based on mobile phone location data (counteract): a system of identifying covid-19 infectious and hazardous sites, detecting disease outbreaks based on the internet of things, edge computing, and artificial intelligence. Sustainable Cities Soc 2021;69:102798.
[50] Ceylan Zeynep. Short-term prediction of COVID-19 spread using grey rolling model optimized by particle swarm optimization. Appl Soft Comput 2021;109:107592.
[51] Mahajan Ashutosh, Solanki Ravi, Sivadas Namitha. Estimation of undetected symptomatic and asymptomatic cases of COVID-19 infection and prediction of its spread in the USA. J Med Virol 2021;93(5):3202–10.
[52] Tomar Anuradha, Gupta Neeraj. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. Sci Total Environ 2020;728:138762.
[53] Varzaneh Zahra Asghari, et al. A new COVID-19 intubation prediction strategy using an intelligent feature selection and K-NN method. Inform Med Unlocked 2022;28:100825.
[54] Hasan Abul, et al. Monitoring COVID-19 on social media: development of an end-to-end natural language processing pipeline using a novel triage and diagnosis approach. J Med Internet Res 2022;24(2). e30397.
[55] Luo Linkai, Wang Yue, Liu Hai. COVID-19 personal health mention detection from tweets using dual convolutional neural network. Expert Syst Appl 2022;200:117139.
[56] Balasubramanian Vishal, Vivekanandhan Sapthagirivasan, Mahadevan Venkatesh. Pandemic tele-smart: a contactless tele-health system for efficient monitoring of remotely located COVID-19 quarantine wards in India using near-field communication and natural language processing system. Med Biol Eng Comput 2022;60(1):61–79.
[57] Zhang Yiming, et al. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. Expert Syst Appl 2022;198:116882.
[58] Khanna Varada Vivek, et al. A machine learning and explainable artificial intelligence triage-prediction system for COVID-19. Decis Anal J 2023;100246.
[59] Rahman Tawsifur, et al. BIO-CXRNET: A robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data. Neural Comput Appl 2023;35(24):17461–83.
[60] Solayman Sanzida, et al. Automatic COVID-19 prediction using explainable machine learning techniques. Int J Cognit Comput Eng 2023;4:36–46.
[61] Chadaga Krishnaraj, et al. A decision support system for diagnosis of COVID-19 from non-COVID-19 influenza-like illness using explainable artificial intelligence. Bioengineering 2023;10(4):439.
[62] Huyut MT. Automatic detection of severely and mildly infected COVID-19 patients with supervised machine learning models. IRBM 2023;44(1):100725.
[63] Laguarta Jordi, Hueto Ferran, Subirana Brian. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol 2020;1:275–81.
[64] Wang Peipei, et al. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solitons Fractals 2020;139:110058.
[65] Javaid Mohd, et al. Industry 4.0 technologies and their applications in fighting COVID-19 pandemic. Diabetes Metab Syndr: Clin Res Rev 2020;14(4):419–22.
[66] Perumal Murukessan, et al. INASNET: Automatic identification of coronavirus disease (COVID-19) based on chest X-ray using deep neural network. ISA Trans 2022;124:82–9.
[67] Mikolov Tomas, et al. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013;26.
[68] [Dataset]. Health.gov COVID-19. 2021, https://healthdata.gov/.
[69] [Dataset]. Weather dataset of USA counties. 2021, https://data.nodc.noaa.gov.