




Data Analysis of COVID-19 Hospital Records Using Contextual Patient Classification System

Vrushabh Gada¹ · Madhura Shegaonkar¹ · Madhura Inamdar¹ · Sharath Dinesh¹ · Darshan Sapariya¹ · Vedant Konde¹ · Mahesh Warang¹ · Ninad Mehendale¹ 

Received: 14 April 2021 / Revised: 1 November 2021 / Accepted: 19 February 2022 /
Published online: 22 March 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Humanity today is suffering from one of the most dangerous pandemics in history, the Coronavirus Disease of 2019 (COVID-19). Although today there is immense advancement in the medical field with the latest technology, the COVID-19 pandemic has affected us severely. The virus is spreading rapidly, resulting in an escalation in the number of patients admitted. We propose a contextual patient classification system for better analysis of the data from the discharge summary available from the research hospital. The classification was done using the Knuth–Morris–Pratt algorithm. We have also analyzed the data of COVID-19 and non-COVID-19 patients. During the analysis, studies on the medicines, medical services and tests, pulse count, body temperature, and the overall effect of age and gender was done. The death versus survival ratio for the COVID-19 positive patients has also been studied. The classification accuracy of the contextual patient classification system achieved was 97.4%. The combination of data analysis and contextual patient classification will be helpful to all the sectors to be better prepared for any future waves of the COVID-19 pandemic.

Keywords Data analysis · Patient classification system · Contextual search

1 Introduction

A catastrophic virus originated in early December 2019, in the Wuhan province of China. Later on, it became a worldwide crisis termed Coronavirus Disease of 2019 (COVID-19) by the World Health Organization (WHO) which is still affecting the world [1]. COVID-19 is still a serious challenge for doctors and hospitals. Even though

✉ Ninad Mehendale
ninad@somaiya.edu

¹ K. J. Somaiya College of Engineering, Mumbai, Maharashtra 400077, India

hospitals are trying their best to overcome this difficult situation, this pandemic is becoming more severe day by day as the number of variants is increasing.

The COVID-19 pandemic has resulted in uncontrollable havoc in India. Since this was an unexpected circumstance, many local hospitals were not prepared to handle this crisis. The number of patients getting admitted because of COVID-19 is still increasing rapidly and this has caused a strain on hospital resources like ventilators, beds, medication (drugs), ICU beds, oxygen supply, etc. [2]. It makes the situation even more difficult for doctors and related staff such as nurses, ward boys, etc. This chaotic situation has majorly affected the patients as well. The proper allocation of resources has become a tough challenge for hospitals. Because of this, there is a possibility that many patients may not get proper treatment. If the trends in the current situation of the COVID-19 pandemic in terms of patient condition and availability of hospital resources are studied and analyzed correctly, it can help in the organized planning of any future waves of the COVID-19 pandemic [3].

This will eventually help in quick decision-making and proper allocation of the hospital resources. Data science is one of the tools to get the trends from a large dataset. Data science uses scientific methods and algorithms on unstructured data to extract useful insights, which help different businesses, health care, and other organization to improve their goods and services [4].

In India, different hospitals have different ways and software for maintaining their patient records [5]. A centralized system of maintaining the records is required. For proper resource management, we need the history of a patient to be presented in a well-organized manner. There are a good deal of software already available that can be used for hospital resource management if organized data is present. The patient summary written by doctors varies from doctor to doctor [6]. Hence, we need a context-based patient classification system that can give segregated data which can be useful for hospital resource allocation.

In our proposed method, data analysis is done on the anonymous data provided by a local hospital. This data was present in an unorganized form. The received raw data from the hospital contained eight different databases as excel sheets. Out of the eight databases provided by the hospital, seven were used. They named the seven sheets as patient list, registration list, ward list, medicine list, service list, test list, discharge summary list. We then organized this data and passed it as an input to the contextual patient classification system. The organized data was given to the contextual patient classification system to classify COVID-19 and non-COVID-19 patients. The classification was done using the KMP algorithm [7]. The classification that was done helped us in performing a comparative analysis between the COVID-19 and non-COVID-19 patient characteristics. We could compare the COVID-19 and non-COVID-19 patients based on the effect of gender, age, and services provided to them by the hospitals in terms of treatment. The death versus survival ratio of COVID-19 patients was obtained based on differences in gender and differences in age. This classification and comparison will help in the early prediction for the resource allocation and treatment process of COVID-19 patients using the data present in the discharge summary section of the organized data.

Figure 1a shows the conceptual diagram of the process of data analysis and contextual patient classification system. The data is filtered and arranged in an organized

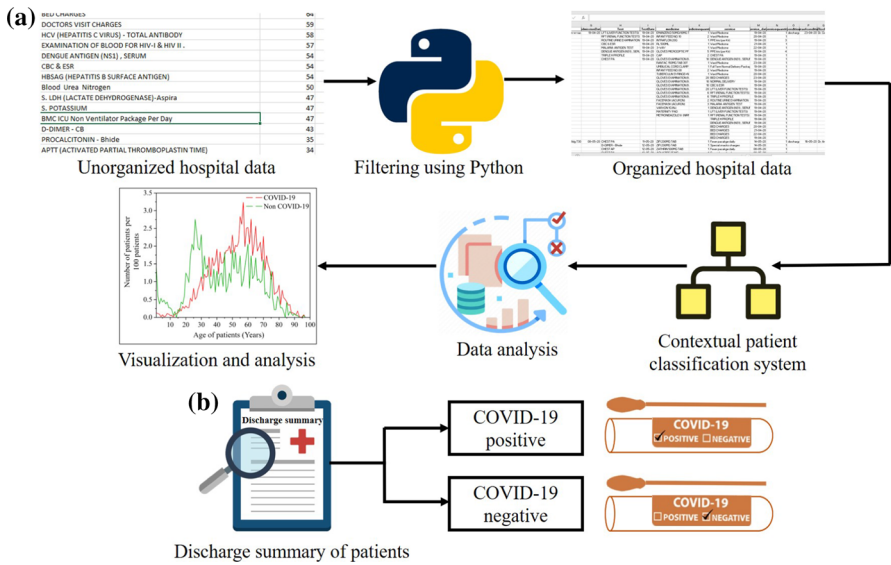


Fig. 1 **a** Concept diagram of the proposed method for data analysis. The unorganized data obtained from the hospital was filtered with the help of python programming. The filtration resulted in an organized representation of the raw data given by the hospital. In the organized data, the data for attributes related to hospital services, medicines, and discharge details were mapped against the unique MR numbers of each patient. The organized dataset was given to a contextual patient classification system. The discharge summary list from the organized dataset was then used to classify COVID-19 and non-COVID-19 patients. This classification further helped in data analyses and visualizing the differences in various aspects between COVID-19 and non-COVID-19 patients. **b** Based on the discharge summary of the patients obtained from the records provided by the hospital, the patients were segregated into non-COVID-19 and COVID-19. This was done using a contextual patient classification system that used the KMP algorithm for pattern mapping

manner using python programming. After that, the classification system is implemented for the analysis of the organized data. Figure 1b shows a conceptual diagram of a contextual classification system based on the discharge summary section present in the organized data. Depending on the records, the patients were grouped into non-COVID-19 and COVID-19.

2 Literature Review

A lot of studies have been reported in the literature for contextual classification systems and the analysis of the trends for hospital data. An early study by Wharton et al. [8] performed a contextual classification method for recognizing land-use patterns in randomly generated synthetic data developed to simulate four mixture classes, which differ only in terms of their frequency distribution of components. A subsequent study was done by Jhung et al. [9] which used modified M-estimates and Markov random fields on various classifiers and their individual and overall accuracy with different tests were determined. The experimental results show that the suggested scheme outperforms conventional non-contextual classifiers and contextual classifiers.

Tian et al. [10] analyzed 262 patients and the cases were divided into severe, mild, common cases, asymptomatic, etc. They identified the common symptoms along with the age group to predict who was most likely to get infected with the coronavirus. Hospitalized patients with COVID-19 were characterized by substantial in-hospital mortality and a high rate of thromboembolic complications by Lodigiani et al. [11]. The high rate of positive Venous Thromboembolism (VTE) imaging tests among the few COVID-19 patients suggested an urgent need to improve specific VTE diagnostic strategies and investigate the efficacy and safety of thromboprophylaxis in ambulatory COVID-19 patients. Interim measures were adopted by the hospitals that included online consultation, region separation, epidemic priority, etc.

The hospital emergency management plan as designed by Cao et al. [12] could ease the workload, protect health care personnel, and control the cross-infection during the COVID-19 epidemic. Cai et al. [13] studied and analyzed all the confirmed COVID-19 cases treated in the Third People's Hospital of Shenzhen, from January 11 to February 6, 2020. The epidemiological and clinical features were analyzed of these cases to better inform patient management in normal hospital settings. COVID-19 patients were mainly characterized by mild symptoms and could be effectively managed using the existing hospital system.

A simple patient simulation model (called ICU-covid-sim) was developed by Alban et al. [14]. The ICU-covid-sim tool uses queueing theory and patient flow simulations. It describes the maximum rate of COVID-19 patients which can be handled for a given number of ICU beds dedicated to COVID-19 patients.

Sun et al. [15] discussed the characteristics of COVID-19 for providing a reference for future studies and help for the prevention and control of the COVID-19 epidemic. It includes the epidemiological characteristics of COVID-19, mechanism, symptoms, and diagnosis of COVID-19. Also, effective ways for the prevention and treatment of COVID-19 were discussed.

Predictors [16] of a fatal outcome in COVID-19 cases included age, the presence of underlying diseases, the presence of secondary infection, and elevated inflammatory indicators in the blood. The results obtained from this study by Ruan et al. [16] also suggest that COVID-19 mortality might be due to virus-activated "cytokine storm syndrome" or fulminant myocarditis.

A study of 1438 patients hospitalized in metropolitan New York to determine the association between the use of hydroxychloroquine, with or without azithromycin, and clinical outcomes among hospitals in patients diagnosed with COVID-19 was done by Rosenberg et al. [17]. Analysis of COVID-19 patient's clinical data from December 2019 to February 2020 was done using a single-arm meta-analysis by Li et al. [18]. The results showed that the major symptoms experienced by the patients were fever and that the males took a larger distribution in the gender distribution of COVID-19 patients along with other significant results.

Association of COVID-19 with obesity was analyzed by the data provided by the Third People's Hospital of Shenzhen by Cai et al. [19]. Obese people showed the symptoms of fever and cough as compared to non-obese people. In their study, they observed that obese patients had increased odds of progressing to severe COVID-19. According to the analysis of Guan et al. [20], they claimed that patients with any co-morbidity yielded poor clinical outcomes for COVID-19 as compared to those that

did not. The data was obtained from 1590 laboratory-confirmed hospitalized patients from 575 different hospitals. The mean age of patients was 48.9 years and 42.7% of patients were female.

In the study by Han et al. [21], they detected and analyzed the main laboratory indicators related to heart injury in 273 patients with COVID-19 and investigated the correlation between heart injury and severity of the disease. 671 hospitals from six continents took part in the research by Mehra et al. [22] on patients hospitalized between Dec 20, 2019, and April 14, 2020, with a positive laboratory finding for SARS-CoV-2. All, 96032 patients were divided into 5 groups depending on which medicine is given to them. Each of these drug regimens was associated with decreased in-hospital survival and an increased frequency of ventricular arrhythmias when used for the treatment of COVID-19.

Balli [23] in his study used time series machine learning algorithms to analyze COVID-19 data and performed short-term prediction of cases. His dataset consisted of data collected over 35 weeks from World Health Organization (WHO). Machine learning models such as linear regression, multi-layer perceptron, random forest, and Support Vector Machines (SVM) were used to predict when the peak of the pandemic will be reached. After comparing the models with different performance metrics, the SVM classifier showed the best results.

Muhammad et al. [24] in their study proposed various machine learning models to perform data analysis to be better prepared to deal with the COVID-19 pandemic. They applied supervised learning methods such as logistic regression, decision tree, support vector machine, naïve Bayes, and artificial neural network on a dataset with two labels namely, positive COVID-19 and negative COVID-19. Before training, the correlation between different dependent and independent features of the two classes was obtained. After training, the decision tree showed the highest accuracy of 94.99% in predicting if a particular case was COVID positive or negative depending on the data.

Chao et al. [25] used image and non-image data to predict the progression of COVID-19 in patients with the aim to mitigate the adverse progression of the disease in high-risk patients. The percentage of the lung abnormalities such as opacity, etc., and other key features using image segmentation based on deep learning and non-image data such as vitals and other findings were used to predict whether a patient would require ICU support or not. Their dataset included data from different countries and upon training, they concluded that adding contextual data to a predictive algorithm can significantly improve the performance.

Wang et al. [26] have proposed a method for automatically classifying clinical text data using machine learning algorithms. Supervised learning models require labeled data which requires human effort, hence they propose a method to automatically generate labels in the dataset by using the Natural Language Processing (NLP) technology. The model is trained using a dataset with labels generated using NLP, this is known as weak supervision. SVM, Random Forest (RF), Multilayer Perceptron Neural Networks, and Convolutional Neural Networks (CNN) were used for training three different text datasets. The CNN algorithm showed the highest accuracy but the algorithm is seen to be susceptible to the size of the dataset and the authors conclude that their proposed method may not be efficient on complex datasets.

Hughes et al. [27] have proposed a system to automatically classify clinical text using deep-CNN models. Their approach was able to outperform several other NLP approaches by 15%. Their dataset contained a vast amount of health-related data. Their method used CNN to perform text classification at a sentence level by facilitating semantic classification. Nguyen et al. [28] developed a model that could classify technical publications on the basis of the research topics. For the classification, the text present in the title, abstract, introduction, and conclusion is mainly used. Title Bi-Gram and Title SigNoun, which are new features, are also used. They also developed a back-off model to classify the type of paper based on the text present in it. This model was able to get an accuracy of 60.45%. Their model showed better results as compared to a few other existing models.

In a study by Kumar et al. [29], cluster analysis, one of the data mining techniques is used to classify real groups of infectious disease “novel coronavirus disease (COVID-19)” data set of different states and union territories (UTs) in India according to their high similarity to each other. The results obtained displayed a sense of clusters of affected Indian states and UTs. The main objective of clustering in this study is to optimize monitoring techniques such as screening, closedown, curfews, lockdown, evacuations, legal actions, etc. in affected states and UTs in India which will be very valuable to the government, doctors, police, and others involved in understanding seriousness of the spread of novel coronavirus (COVID-19) to improve government policies, decisions, medical facilities, treatment, etc. to reduce the number of infected and deceased persons. Hierarchical cluster analysis was performed to determine relationships depending upon the observations obtained from the three types of cases of COVID-19 of Indian states and UTs. Here, cluster analysis grouped 27 states and 5 UTs into six clusters (I–VI), and further conclusions were drawn based on these clusters.

In a research by Li et al. [30], several ways to effectively combat COVID-19 through global collaboration have been discussed. The authors present suggestions for this cross-culture collaboration, especially among scientific and technological communities. They believe sharing data and information about the pandemic could help effectively track and trace the virus. Along with the data, every country should adopt from the experiences learned in other countries. Based on this, the government can evaluate its current public health systems and can improve wherever needed. Finally, they suggest countries identify the systems that may destroy the environment and work on them to keep the environment clean, which is essential to life on the planet as coronavirus shutdowns have yielded unintentional climate and environmental benefits.

A study was performed regarding interactions among people by Liu et al. [31] which is an essential factor that characterizes the disease transmission patterns. A computational model was created to reveal the interactions between the population of different age groups in terms of social contact patterns. The retrospective and prospective situations of the disease outbreak, including the past and future transmission risks, the effectiveness of different interventions, and the disease transmission risks of restoring normal social activities, are computationally analyzed and reasonably explained with an in-depth characterization of age-specific social contact-based transmission. The study’s findings not only provide a comprehensive explanation of the underlying COVID-19 transmission patterns in China, but they also provide social contact-based

risk analysis methods that can be easily applied to guide intervention planning and operational responses in other countries, reducing the impact of the COVID-19 pandemic.

A study was performed involving Joint Modeling of Longitudinal CD4 Count and Time-to-Death of HIV/TB Co-infected Patients by Temesgen et al. [32]. The study followed 254 HIV/TB co-infected individuals who were 18 years old or older and receiving antiretroviral treatment in Jimma University Specialized Hospital in West Ethiopia from February 2009 to July 2014. Since the development of AIDS, tuberculosis (TB), and HIV have been tightly related; TB promotes HIV replication by speeding up the natural evolution of HIV infection, which is the major cause of disease and death among HIV/AIDS patients. The study discovers factors that influence the mean change in square root CD4 measurement over time as well as risk factors for HIV/TB co-infected patients' survival time.

Shi et al. [33] has presented the theory and the applications with up-to-date progress in Multiple Criteria Programming and support vector machines from their research and application activities. In the very first chapter, they have presented the C-SVM for classification problems and extended it to problems and nominal attributes. The next chapter introduces LOO Bounds, which can speed up the process of searching for appropriate parameters for Support Vector Machines' several algorithms. SVMs for multi-class, unsupervised, and semi-supervised problems are discussed in Chapters 3 and 4 using various mathematical programming models. The fifth chapter discusses robust optimization models for a variety of uncertain problems. Chapter 6 employs p-norm minimization to combine standard SVMs with feature selection strategies at the same time. Part two focuses on MCP for data mining. Chapter 7 covers fundamental MCP ideas and models before constructing penalized Multiple Criteria Linear Programming (MCLP) and regularized MCLP. Chapters 8, 9, and 11 present several modifications of MCLP and Multiple Criteria Quadratic Programming (MCQP) to create distinct models with varied aims and restrictions.

When interactions between characteristics are permitted for classification, Chapter 10 offers non-additive measured MCLP. Part three discusses several real-world applications of MCP and SVM models. Finance applications are covered in Chapters 12, 13, and 14, which include firm financial analysis, personal credit management, and health insurance fraud detection. Web services are covered in Chapters 15 and 16, which include network intrusion detection and a study of the pattern of deleted VIP email client accounts. Chapter 17 is concerned with HIV-1 informatics for the development of particular treatments, whereas Chapter 18 is concerned with antigen and anti-body informatics. Geochemical analyses are covered in Chapter 19. Each chapter of applications is self-contained and self-explained for the reader's convenience. Finally, Chap. 20 presents intelligent knowledge management for the first time and discusses the theoretical foundation of intelligent knowledge in depth. This chapter's contents expand beyond the usual realm of data mining to investigate ways to provide knowledgeable assistance to end-users by combining hidden patterns from data mining with human expertise.

A study involving the Internet of Things has been performed by Tien et al. [34], Real-Time Decision Making, and Artificial Intelligence shows the way these three technologies are interrelated, complementary, and very much mutually supportive.

Servgoods are suitably defined as ‘things in the Internet of Things (IoT) in this study. It is also described as a physical good or product that is encased in a services-oriented layer that makes it smarter, more adaptable, and customizable for a specific function. It considers the wireless infrastructure that underpins the Internet of Things, as well as the IoT power infrastructure.

It also emphasizes the need for real-time decision making (RTDM), which is based on decision informatics and encompasses enhanced sensing, processing, reacting, and learning technologies. It also covers the growing importance and impact of artificial intelligence (AI) on IoT and RTDM, as well as the underlying machine learning technology, the expanding breadth of AI successes, and the unique possibility of autonomous vehicles or servgoods. It finishes with various observations on how the Internet of Things, Real-Time Data Management, and Artificial Intelligence have impacted and will continue to impact the twenty-first century.

3 Methodology

3.1 Dataset

A database comprising different data of patients that have undergone treatment at a local hospital of Mumbai, Maharashtra, India was obtained from the hospital authorities. Data provided was from March 2020 to February 2021. The database comprised seven different excel sheets which were patient list, registration list, ward list, medicine list, service list, test list, and discharge summary list. Registration and patient lists comprised details of patients like visit date, admission date, gender, and other details. Medicine list also showed the types and quantity of medicines provided to individual patients. Service list depicted the types of services provided to the patients like the ward, machines, and the date on which the service was provided. Test list comprised different tests performed on patients with the associated dates. The sheet of bed details comprised the data regarding types of wards allotted to patients and admission date and duration. The discharge summary comprised details such as time and date of discharge with recovered patients and time and date of death with non-recovered patients. The discharge summary also consisted data of what disease the patient had contracted.

The entire database with seven sheets was converted into one organized database. Personal details of the patients like their names, contact details were pre-removed by the hospital to maintain patients’ privacy (none of our studies needed any personal information). Each patient was allotted a unique Medical Record (MR) number. The total number of MR numbers was extracted using a python script. All the unorganized hospital data was organized into one dataset using python.

The attributes which were included in the organized single sheet dataset were MR number, gender, age, admission date, tests, test date, medicine, the quantity of medicine, service, service date, service quantity, condition of the patient, ward details with the duration of allotment, field name, discharge summary, and specialist doctors.

Using python, the data from each excel sheet was first imported into the python list of dictionaries. An empty dictionary was created which contained empty columns named after each of the attributes. Firstly, all the MR numbers were inserted into the

dataset and their index was stored in a list. A dictionary was created for every unique MR number. Then, from the first excel sheet, the attribute data was inserted for the respective MR numbers. If the attribute had multiple values for the same MR number, it created an empty dictionary into the dataset for the MR number and incremented the index, and inserted the attribute. If a new MR number is encountered, the system creates a new dictionary for the new MR number. The whole process was repeated for every MR number in the excel sheet.

The dataset contains the data from March 2020 to February 2021. In June 2020, the hospital was converted into a COVID-19 hospital, so all non-COVID-19 patients were transferred to another hospital. Hence, the non-COVID-19 patient data is from March 2020 to June 2020. Thus the classification model is used on the data from March 2020 to June 2020, as the data after this day contains information of only COVID-19 patients. The data storing format of the hospital was updated once they converted the hospital to a COVID-19 hospital i.e. after June 2020. Because of this, the data acquired after June 2020 consisted on only the discharge summary list, medicine list which contained the amount of times the medicine was consumed, registration list with initial patient body temperature and pulse count and the patient list. The test details, service details and ward details were not provided after June 2020. The dataset consisted of details of 3409 COVID-19 positive patients and 1850 non-COVID-19 patients.

Figure 2 shows that the unorganized raw data received from the hospital was organized into individual databases namely, patient list, registration list, ward list, medicine list, service list, test list, and discharge summary list. Using filtration and contextual patient classification system the databases are classified into two parts- COVID-19 patients and non-COVID-19 patients. Further analysis on the data of COVID-19 patients was done to find out the effect of COVID-19 based on patient age and gender. The death to discharge ratio based on the age and gender of COVID-19 patients was also studied using the available data. The time required for the treatment of COVID-19 based on age was also analyzed. Further analyses were done to find out the top ten services and medicines used by COVID-19 patients as compared to non-COVID-19 patients. Plots were obtained for each of the analyses.

3.2 Data Filtration and Rearrangement

We defined four functions for inserting all the attributes. First function ‘singleN’ for those attributes which were to be added one at a time. These included gender, age, admission date, discharge summary from a doctor, and hospital Ward. Then a function ‘doubleN’ inserted two attributes simultaneously. These included medicine and medicine quantity, test and test date, etc. These attributes were treated as a group and inserted one at a time for each MR number. Then a function ‘tripleN’ inserted three attributes simultaneously. These included service, service quantity, and service date, and ward, from the admission date and to the discharge/death date. These attributes were treated as a group and inserted one at a time for each MR number. Then the last function ‘condition’ was to insert the condition of the patient. If the patient had a death date then we set the condition as ‘Death’ or if the patient data only included discharge date, we set the condition as ‘Discharge’. We set the date of discharge as the date of

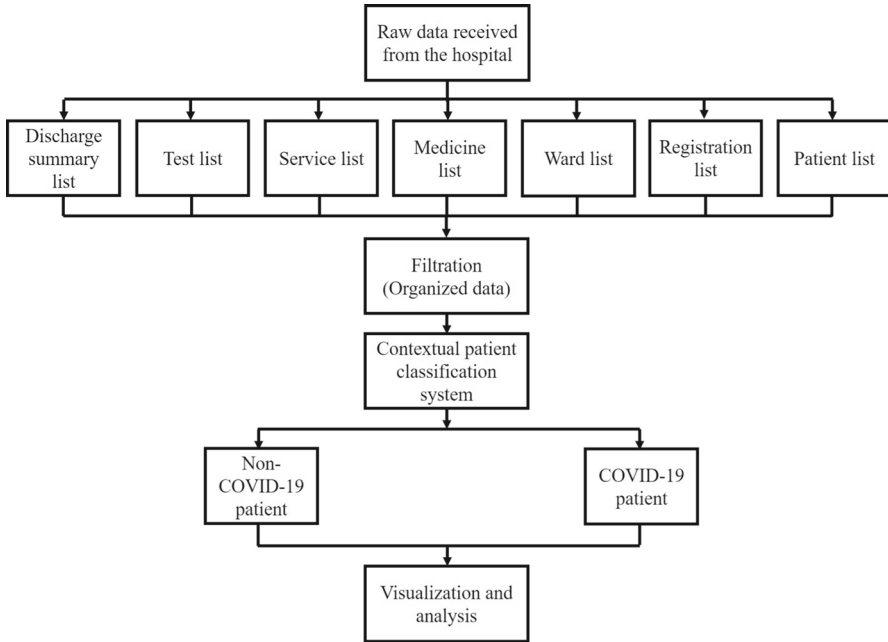


Fig. 2 Flow diagram of the proposed method. The raw data acquired from the local hospital was in an unorganized form. This data contained lists which were named as patient list, registration list, ward list, medicine list, service list, test list, and discharge summary list. This data was given to the proposed contextual patient classification system which classified COVID-19 and non-COVID-19 patients. This classification was used to further analyze data and find out the differences in various aspects of the COVID-19 and non-COVID-19 patients

condition. After the insertion of every attribute, the dataset was exported as an excel sheet using the Pandas library. The dataset included data of non-COVID-19 patients too.

3.3 Contextual Patient Classification

After organizing the raw dataset received from the hospital, we used a contextual patient classification system to classify COVID-19 patients and non-COVID-19 patients. The classification was done based on the data present in the discharge summary of the patient. We designed the patient classification system using the contextual search method. We performed data filtering on the discharge summary to extract the required features for classification. The contextual classification system made use of the Knuth–Morris–Pratt (KMP) algorithm [7] for pattern matching. The KMP algorithm is a linear time algorithm and hence, backtracking of the string is avoided. Then the keywords were figured out by inspecting the discharge summary dataset manually, for example, ‘COVID’, ‘COVID-19’, and ‘CORONA’. After determining the list of keywords, the string search operation was done on the dataset for the keywords using python, and results were re-verified using the in-built ‘find’ function. The MR num-

Table 1 The top 10 medicines used for the treatment of COVID-19 patients

10 most used medicines for the treatment of COVID-19 patients	Frequency of use (%)
Vitamins	100
Paracetamol	98.56039326
Antibiotic	88.86938202
Ivermectin	80.33707865
LMWH	70.92696629
Mehtylprednisolone	63.83426966
Remdesivir	55.75842697
O2 therapy o/a	22.0505618
HCQS	13.02668539
IV-fluids	11.16573034

It shows that vitamins and paracetamols are used in the treatment of almost all the patients while IV fluids are the least used for the treatment

bers of COVID-19 patients and non-COVID-19 patients were separated. For the MR numbers of COVID-19 patients, we imported the data from the discharge summary sheet from the organized dataset to a new sheet. Then, a manual search was done on the imported data to re-verify the results of the classification system to obtain accuracy.

3.4 Data Visualization

Using all the data available, various graphs were plotted on Origin 2020 (demo license) for pictorial representation, comparison, and understanding. The individual data for each plot was segregated by python code (Supplementary material). The dataset was imported into a python dictionary, then those columns were extracted for which the analysis was to be done.

For Fig. 3a, b, the data was made by extracting the service name and its quantity for every patient. Then the total count of each service was calculated using the ‘count’ function. The services were further divided into hospital services and medical services. And the graph of the top 5 most used hospital services and medical services by COVID-19 and non-COVID-19 patients were plotted. The same procedure was repeated for Tables 1, 2, and Fig. 3c where medicines and tests were extracted respectively.

For Fig. 4a, the data was extracted using the ‘count’ function of the list, the number of patients aged between 0 to 100 years was calculated (As we knew no patient aged more than 100 has arrived at the hospital). Using the count, the data were extracted for both COVID-19 and non-COVID-19 patients. For Fig. 4b, patients were classified based on age and the number of days patients were admitted to the hospital was calculated.

Figure 4c the patient’s death and discharged count was extracted from the data set and later on it was classified along with the gender also. For Fig. 4d, the count of COVID-19 patients who got discharged or died was calculated and later categorized by their age. Using the count of each age year the graph was plotted. For Fig. 5a, in

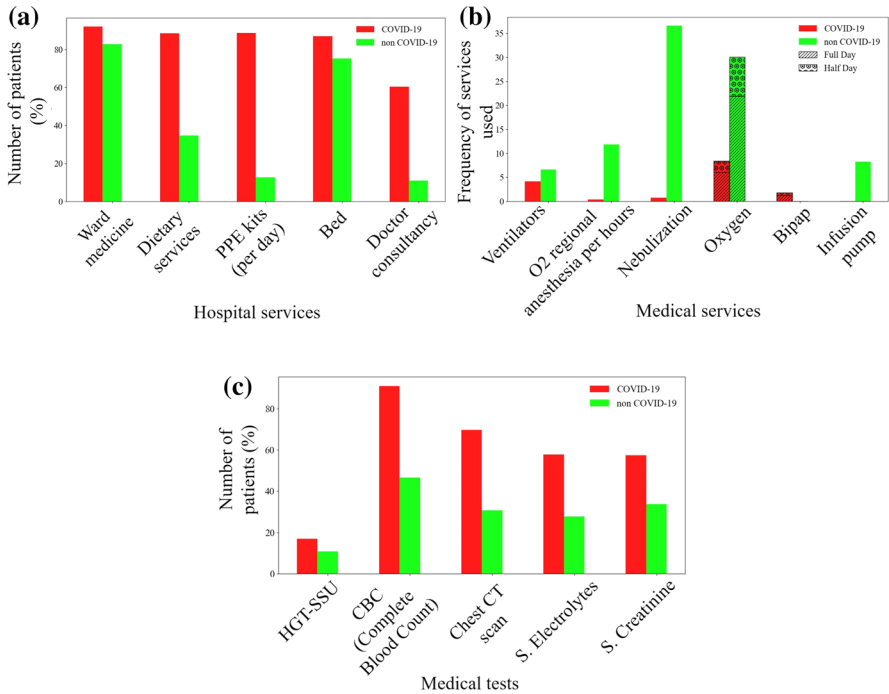


Fig. 3 a Histogram of the number of patients serviced versus the top five used hospital services for COVID-19 and non-COVID-19 patients. Ward medicine is the most used hospital service for both types of patients. The demand for PPE kits and doctor consultancy is much higher for COVID-19 patients than for non-COVID-19 patients. The demand for hospital beds also increased during COVID-19. b Histogram of the frequency of usage versus medical service used by COVID-19 and non-COVID-19 patients. During the early months of COVID-19, the need for ventilators, medical oxygen, and Bipap had increased. c Histogram of the number of patients tested versus the top five most used tests. The demand for all tests has increased for COVID-19 patients. CBC tests were required by a larger number of COVID-19 patients as compared to non-COVID-19 patients

Table 2 The top 10 medicines used for the treatment of non-COVID-19 patients

10 most used medicines for the treatment of non-COVID-19 patients	Frequency of use (%)
Pentovar-40 injection	28.16216
Frusemide 20 mg injection 2 ml	7.675676
Hydroxychloroquine 200 mg tab	10.75676
Duolin respules 2.5 ml	10.81081
Varxon 1 g injection	20.75676
RL 500 ml	26.59459
Frusemide (Rasix) 2 ml injection	3.243243
Panam-40 injection	15.67568
Budecort 0.5 mg respules	8.486486
Metronidazole IV (nirmet)	11.08108

Pentovar-40 injection and RL 500 ml were used in the highest proportion while Frusemide was used in the least proportion

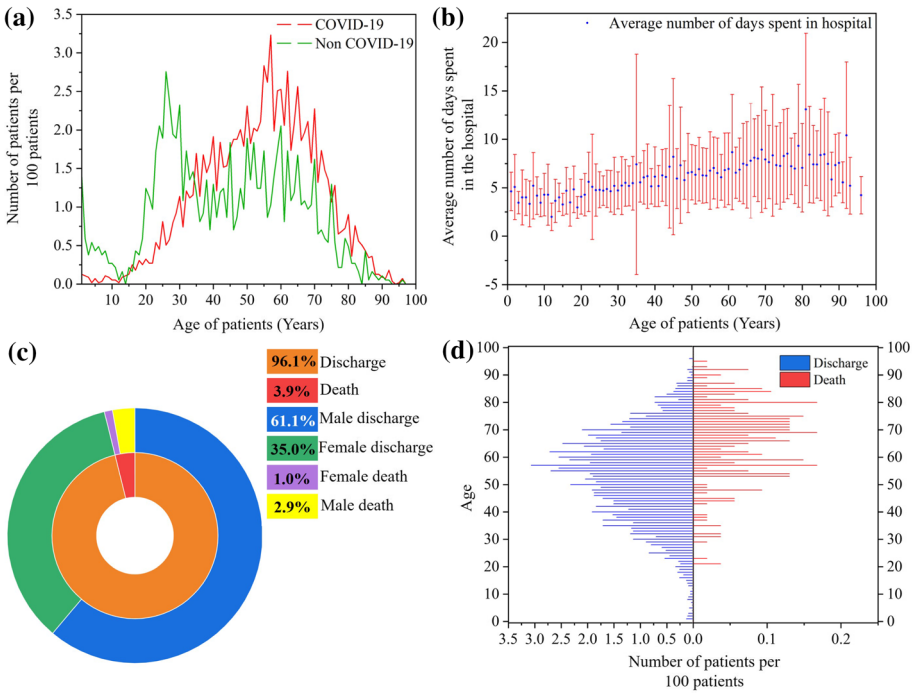


Fig. 4 **a** The plot shows the number of COVID-19 and non-COVID-19 patients per 100 patients versus age (in years). The curve is bell-shaped. The highest number of patients suffering from COVID-19 belong to the age group of 55–60 years. Most non-COVID-19 patients were aged around either 25 years or 60 years. In non-COVID-19 patients, there was a peak of 3.6% at 0 due to newborn babies. **b** The plot shows the average number of days spent in the hospital versus the age of patients in years. The average time spent in the hospital is 7 days. **c** Pie chart of overall study regarding the percentage of patients who passed away and who were discharged. Further gender-based analysis of the two types of patients is carried out. The male population is more likely to get affected as compared to the female population. The overall discharge rate is greater than the death rate. **d** The plot of the discharge versus death is based on the age of COVID-19 patients. The death percentage is around 3.9%

the data, the pulse value of the patients at the time of admission was also available. For Fig. 5b, in the data on which day which medicine was given to patients was also available, so we calculated the average number of days the medicine was given to patients, and out of all the medicine 10 medicine which were most used the data is plotted.

The following things were analyzed

- The effect of COVID-19 on the age distribution of patients
- Death vs survival rate of COVID-19 patients based on gender
- Death vs survival rate of COVID-19 patients based on age
- Duration of treatment in the hospital of COVID-19 patients based on age
- A difference of services used by COVID-19 and non-COVID-19 patients.
- The difference between tests used by COVID-19 and non-COVID-19 patients.
- The difference between medicines used by COVID-19 and non-COVID-19 patients.

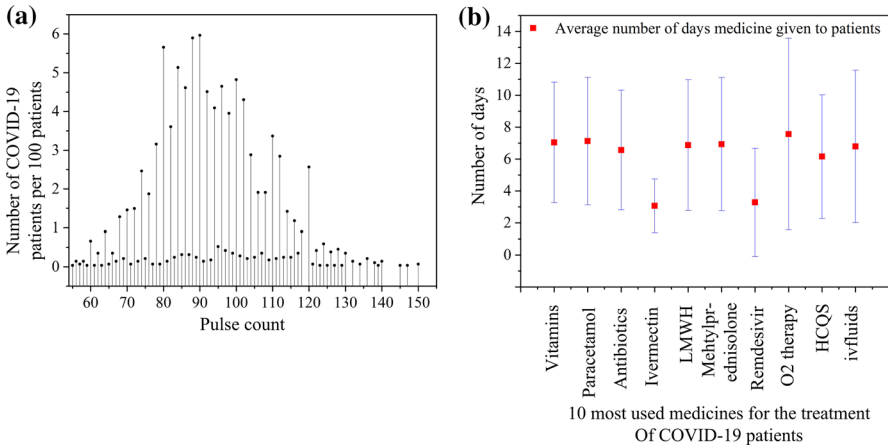


Fig. 5 **a** Plot of pulse count versus the number of COVID-19 positive patients. From our analyses, it was evident that the maximum number of patients showed a pulse count in the range of 80–110 Beats Per Minute (BPM). This indicates that most patients infected with COVID-19 show a high resting heart rate. **b** A plot of the top 10 most used medicines for the treatment of COVID-19 patients and the number of days they are consumed by the patients. It was seen that patients receiving Oxygen (O2) were required to take it for a longer number of days. Whereas Ivermectin was given for the least number of days

- Gender biasing of COVID-19 patients
- Pulse count at the time of admission of patients
- Body temperature of the patients at the time of admission

4 Results and Discussions

The Patient classification system was able to identify 576 COVID-19 patients. After verifying the list manually, we counted 561 COVID-19 patients. This shows that our patient classification system worked with an accuracy of 97.4%. The number of non-COVID-19 was 1850 patients (130 of them were tested as negative for COVID-19).

Figure 3 was plotted based on the patient dataset, which was last updated on 12th June 2020. Hence, the figures and the analysis are not based on the most recent data. Figure 3a shows a plot of different hospital services. Ward medicine was the most used hospital service. Use of dietary service was roughly same as of ward medicine and its demand had also increased a lot during COVID-19. The use of PPE kits had increased almost 8 times in post COVID era. As the era of COVID-19 begins, the requirement of patients observation increased to result in an increase in requirement for doctor consultancy. Figure 3b shows the medical services most used by COVID-19 and non COVID-19 patients. The most used service by COVID-19 patients was medical oxygen, which was divided based on the time it was used by the patient, i.e., full-day or half a day. The ventilator was the second most used service and the use of the infusion pump had reduced. Also, the use of nebulization had decreased from 36 to 1 per 100 patients post COVID-19 outbreak. Figure 3c shows the most common tests done for COVID-19 and non COVID-19 patients. It can be seen that in

the era of COVID-19, the number of tests performed on patients has increased. CBC test was performed on roughly 90% of the COVID-19 patients and roughly 46% of non COVID-19 patients. Figure 4a shows the plot of the number of COVID-19 and non-COVID-19 patients versus age of the patients in years. It can be observed that for COVID-19 patients, the curve appears to be of bell-shaped nature, from which we can observe that the age distribution has a peak at around 57 years. For non-COVID-19 patients, the peaks were around the age of 25 years and 60 years with a very large peak at the age 0 because the age of new born babies was considered as 0.

Figure 4b shows a plot of average number of days spent in the hospital versus age of patients in years. It can be inferred from the plot that the average time taken by patients for treatment was around 7 days. It should be also noted that maximum time spent in the hospital was of 66 days, and minimum time spent was 3 days. For the age group of 35, there were 176 patients. Out of which 6 patients had a treatment duration of 66 days. The rest of the patients had a treatment duration of fewer than 13 days. Due to which the average days of treatment for the age group of 35 came to be 7.412429 and the error came to be 11.36127. Since the error is greater than the average the error bar goes below zero.

Figure 4c shows that males were more affected than females and because of this, the number of patients discharged and number of deaths in the male population is also more than that of the female population. Figure 4d shows a graph of the number of patients discharged versus the number of patients who died based on their ages. From the total number of admitted patients, 96.1% of patients were eventually discharged and the discharge curve had the highest peak at the age of 45. While 3.9% of people died, the graph of dead patients peaks between the age of 50 to 80 years and the highest peak around the age of 50 years.

Figure 5a shows the distribution of pulse count of patients at the time of admission. Maximum number of patients had a pulse count in range of 80 to 110 beats per minute. The normal pulse for healthy adults ranges from 60 to 100 beats per minute. The increased pulse rate in infected patients can be seen as a sign of increasing infection. Thus, increased pulse rate in a person can be used as an early sign of COVID-19 infection. Figure 5b shows the average number of days for which top 10 medicines were given to the patients. O₂ therapy is clearly given to the patients for most number of days which signifies the need of oxygen for COVID-19 infected patients. Vitamins and paracetamol were also given for a long duration which shows the deficiency that the virus creates in the patients.

Table 1 shows the 10 most used medicines for treatment of COVID-19 patients with the number of patients given that medicines in percentage. Vitamins and paracetamol being used by maximum patients whereas HCQs and IV fluids by very few patients. O₂ therapy is used by 22% of COVID-19 patients. This percentage can be brought down further by proper management and picking up early signs of COVID-19 so that the severity of the disease does not increase. As vitamins and paracetamol are given to almost all patients, in the future it may happen that the demand for the vitamins may increase drastically.

Table 2 shows the 10 most used medicines for treatment of non-COVID-19 patients. Table 2 also includes how frequently the said medicines were given to the patients. Around 28% and 26 % of patients were provided Pentovar and RL 500ML and these

Table 3 Comparison with the reported literature states that the minimum and maximum number of COVID-19 patients used for analysis were 150 and 96032 patients respectively

Method	Number of COVID-19 patients analysed	Mortality rate (%)	Age (in years)	Gender (male) (%)
Tian et al. [10]	262	0.90	47.5(median)	48.50
Ruan et al. [16]	150	45	51 (mean)	–
Guan et al. [20]	1590	–	48.9 (mean)	57.30
Han et al. [21]	273	8.79	58.39	35.53
Lodigiani et al. [11]	388	–	66 (median)	68
Cai et al. [13]	298	1.00	47.5 (median)	48.66
Rosenburg et al. [17]	1438	20.30	63 (median)	59.70
Cai et al. [19]	383	2.40	48 (median)	17.50
Mehra [22]	96,032	11.10	53.8 (mean)	53.70
CPCS (proposed system)	3409	3.9	48 (median)	63.90

Rosenburg et al. recorded the highest mortality rate of 20.30% the least mortality rate noted was 0.90% by Tian et al. The mean and median values of age of patients (in years) ranged from 47.5 to 63. It was observed that a maximum of 63.90% of male patients and a minimum of 17.50% were diagnosed with COVID-19

constitute the top two medicines used. On the other hand, Frusemide was delivered to only 3 percent of patients.

From the analysis, we also found that 78.16% of the patients at the time of admission had a body temperature of 98 degrees Fahrenheit, 13.6% of patients had body temperature of less than 98 degrees Fahrenheit and 8.24% of patients had more than 98 degrees. As the majority of patients have body temperature in the normal range, the temperature cannot alone determine whether the person have COVID-19 or not.

As shown in Table 3, we have analyzed data of 3409 patients which is higher than most of the reported literature with the exception of Mehra et al. [22]. The average mortality rate of the reported literature was 12% and our value was 3.9%. The average age of patients affected by COVID-19 is 58 years for rest of the world, but for India, it is 48 years.

In Table 4 the proposed system is compared to similar system reported in the literature. Convolutional Neural Networks (CNN) are a common choice for contextual patient classification. Rule-based Natural Language Processing (NLP) and CNN, along with Word2vec, are the other choices used in making such a system. The proposed system uses KMP for contextual patient classification. The proposed system has the highest accuracy of 97.4% from among the reported literature. Among the reported literature the highest accuracy values were of 97% and 92%. Various types of datasets have been used by the reported literature like data on smokers from the Mayo Clinic, proximal femur fracture patient data, etc. We found a strong gender bias on COVID-19 patients. We averaged the values in the reported literature along with our value, the average male percentage affected is 50.91%. In India, the number of men traveling to and from their workplaces was more than the number of women traveling to and from

their workplaces. Hence higher number of men contracted the virus than the number of women.

A contextual patient classification system on the discharge summary is helpful in effective data filtration and hence, a better data analysis system becomes available. With the assistance of the type of analysis of data and the study done in this paper, hospitals would be able to generate their own structure and plans for effective handling of the ongoing situation. We analyzed data obtained from only one hospital. If data from more hospitals is studied simultaneously then it will be strongly supportive for the final conclusions made in this paper. Making predictions using the trends in patient data is a comparatively simple thing compared to actual practical implementation. There can be some cases where the resources available are inadequate. So even if the analysis of data is done and the allocation plan is generated, it becomes difficult to deal with limited number of resources available at that time. Therefore, this study can be somewhat restricted to the condition of the availability of adequate resources, services, and doctors in the hospital. As we have used this data for the patient's classification, similarly any type of user data can be analyzed to achieve business goals for companies. Pharmaceutical companies can analyze customer feedback and sales data to improve their products. On similar lines, Olson et al. [35] suggests various ways of analyzing and using customer data to increase revenue for companies. It is also proposed that methods like machine learning and deep learning can help in efficiently understanding sales data to conduct deeper market research.

5 Future Prospects

In our proposed system we have analyzed the data available from one local hospital to map out the different requirements of COVID-19 and non-COVID-19 patients. This mapping of data will prove to be extremely useful when planning and preparing for any possible future waves of the pandemic. By doing analyses on the current requirement and current availability of resources the country can get an estimation of future requirements and stock up on resources. This will avoid the risk of incomplete treatment due to a lack of resources. Our proposed method can be used to classify several other diseases based on the available data at hospitals. By acquiring the diagnostic data we can build a classification model that will help doctors in faster diagnosis in the future. Therefore, acquiring more robust data from various parts of the country will also help us increase the geographical area of prediction of the effects of the COVID-19 pandemic.

6 Conclusions

Any conclusive result is not easily extractable from the raw data. The raw data analysis was simplified with the help of a contextual patient classification system. The proposed method demonstrates contextual patient data classification on the raw data obtained from the hospital. Around 5200 patients' data was studied, manual reading of these would have been time-consuming. An automated system helped in achieving

Table 4 A comparison of the existing studies present in the literature that are based on text dataset classification

Author	Dataset	Algorithm/Model	Accuracy (in %)
Wang et al. [26]	Mayo Clinic smoking	CNN	92
Wang et al. [26]	Proximal femur (hip) fracture	CNN	97
Wang et al. [26]	i2b2 2006 smoking	Rule-based NLP	89
Nguyen et al. [28]	Language Resources and Evaluation Conference (LREC)	Back off	66.39
Hughes et al. [27]	Merck Manual	CNN + Word2vec	68
Proposed method	Hospital records	KMP	97.4

Wang et al. stated that CNN gave the best accuracy for text classification. Deep learning models are seen to be commonly used for this purpose. It is observed that the proposed method gave the highest accuracy

this difficult task easily. With a contextual patient data classification system, medical healthcare workers can arrange data in an easily readable manner. The accuracy achieved was 97.4%. The algorithm designed will help society in general for analyzing raw data and can be applied to all fields. Data analysis done on the patients can be used for large-scale implementation of resource allocation systems in hospitals.

Acknowledgements The authors would like to thank the local hospital of Thane near Mumbai, Maharashtra, India for all the anonymous patient data provided for our contextual patient classification study. Finally, we would like to thank all the colleagues at Somaiya Vidyavihar University for providing the facilities to carry out and complete our research. We would like to thank Vruddhi Shah, Vrushali Sule and Rutwik Patel for their continuous feedback and help in completing the work.

Author Contributions Conceptualization was done by V. Gada (VG), M. Warang (MW) and N. Mehendale (NM). All the literature reading and data gathering were performed by VG, M. Shegaonkar (MS) and V. Konde (VK). All the experiments and coding were performed by VG, S. Dinesh (SD) and D. Sapariya (DS). The formal analysis was performed by VG, MS. Manuscript writing- original draft preparation was done by VG, MS and VK. Review and editings were done by MW and NM. Visualization work was carried out by M. Inamdar (MI), VG, and NM.

Funding No funding was involved in the present work.

Code availability The codes will be made available upon reasonable request to the authors.

Declarations

Conflict of interest Authors V. Gada, M. Shegaonkar, M. Inamdar, S. Dinesh, D. Sapariya, V. Konde, M. Warang, and N. Mehendale, declare that he has no conflict of interest.

Ethical standards All authors consciously assure that the manuscript fulfills the following statements: (1) This material is the authors' own original work, which has not been previously published elsewhere. (2) The paper is not currently being considered for publication elsewhere. (3) The paper reflects the authors' own research and analysis in a truthful and complete manner. (4) The paper properly credits the meaningful contributions of co-authors and co-researchers. (5) The results are appropriately placed in the context of prior and existing research.

Human and animal rights All the necessary permissions were obtained from the Institute Ethical Committee and concerned authorities to run our algorithms on patient data.

Consent for publication Authors have taken all the necessary consents for publication wherever required.

Informed consent Informed consent was obtained from participants whose data was used to do analysis.

Data availability The data will be made available upon reasonable request to the authors.

References

1. Chauhan S (2020) Comprehensive review of coronavirus disease 2019 (COVID-19). *Biomed J* 43(4):334
2. Saha A, Ahsan MM, Quader TU, Shohan MUS, Naher S, Dutta P, Akash AS, Mehedi HH, Chowdhury AAU, Karim H et al (2021) Characteristics, management and outcomes of critically ill COVID-19 patients admitted to ICU in hospitals in Bangladesh: a retrospective study. *J Prev Med Hyg* 62(1):E33
3. Malki Z, Atlam ES, Ewis A, Dagnew G, Alzighaibi AR, ELmarhomy G, Elhosseini MA, Hassanien AE, Gad I (2021) ARIMA models for predicting the end of COVID-19 pandemic and the risk of second rebound. *Neural Comput Appl* 33(7):2929

4. Sajid S, Haleem A, Bahl S, Javaid M, Goyal T, Mittal M (2021) Materials today: proceedings
5. Aanestad M, Jolliffe B, Mukherjee A, Sahay S (2014) Infrastructuring work: building a state-wide hospital information infrastructure in India. *Inf Syst Res* 25(4):834
6. Kripalani S, Jackson AT, Schnipper JL, Coleman EA (2007) Promoting effective transitions of care at hospital discharge: a review of key issues for hospitalists. *J Hosp Med* 2(5):314
7. Régnier M (1989) International symposium on mathematical foundations of computer science. Springer, pp 431–444
8. Wharton SW (1982) A contextual classification method for recognizing land use patterns in high resolution remotely sensed data. *Pattern Recogn* 15(4):317
9. Jhung Y, Swain PH (1996) Bayesian contextual classification based on modified M-estimates and Markov random fields. *IEEE Trans Geosci Remote Sens* 34(1):67
10. Tian S, Hu N, Lou J, Chen K, Kang X, Xiang Z, Chen H et al (2020) Characteristics of COVID-19 infection in Beijing. *J Infect* 80:401–406
11. Lodigiani C, Iapichino G, Carenzo L, Cecconi M, Ferrazzi P, Sebastian T, Kucher N et al (2020) Venous and arterial thromboembolic complications in COVID-19 patients admitted to an academic hospital in Milan, Italy. *Thromb Res* 191:9–14
12. Cao Y, Li Q, Chen J, Guo X, Miao C, Yang H, Chen Z, Li C, Li L (2020) Hospital emergency management plan during the COVID-19 epidemic. *Acad Emerg Med* 27(4):309
13. Cai Q, Huang D, Ou P, Yu H, Zhu Z, Xia Z, Su Y et al (2020) COVID-19 in a designated infectious diseases hospital outside Hubei Province, China. *Allergy* 75(7):1742–1752
14. Alban A, Chick SE, Dongelmans DA, Vlaar APJ, Sent D (2020) ICU capacity management during the COVID-19 pandemic using a process simulation. *Intensiv Care Med* 46(8):1624–1626
15. Sun P, Lu X, Xu C, Sun W, Pan B (2020) Understanding of COVID-19 based on current evidence. *J Med Virol* 92(6):548
16. Ruan Q, Yang K, Wang W, Jiang L, Song J (2020) Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med* 46(5):846
17. Rosenberg ES, Dufort EM, Udo T, Wilberschied LA, Kumar J, Tesoriero J, Weinberg P et al (2020) Association of treatment with hydroxychloroquine or azithromycin with in-hospital mortality in patients with COVID-19 in New York State. *Jama* 323(24):2493–2502
18. Li LQ, Huang T, Wang YQ, Wang ZP, Liang Y, Huang TB, Zhang HY, Sun W, Wang Y (2020) COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis. *J Med Virol* 92(6):577
19. Cai Q, Chen F, Wang T, Luo F, Liu X, Wu Q, He Q, Wang Z, Liu Y, Liu L et al (2020) COVID-19 severity in a designated hospital in Shenzhen, China. *Diabetes Care* 43(7):1392–1398
20. Guan W, Liang W, Zhao Y, Liang H, Chen Z, Li Y, Liu X et al (2020) Comorbidity and its impact on 1590 patients with Covid-19 in China: a nationwide analysis. *Eur Respir J* 55(5):2000547
21. Han H, Xie L, Liu R, Yang J, Liu F, Wu K, Chen L, Hou W, Feng Y, Zhu C (2020) Analysis of heart injury laboratory parameters in 273 COVID-19 patients in one hospital in Wuhan, China. *J Med Virol* 92:819–823
22. Mehra MR, Desai SS, Ruschitzka F, Patel AN (2020) RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. *Lancet*
23. Balli S (2021) Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. *Chaos Solitons Fractals* 142:110512
24. Muhammad L, Algehyne EA, Usman SS, Ahmad A, Chakraborty C, Mohammed IA (2021) Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput Sci* 2(1):1
25. Chao H, Fang X, Zhang J, Homayounieh F, Arru CD, Digumarthy SR, Babaei R, Mobin HK, Mohseni I, Saba L et al (2021) Integrative analysis for COVID-19 patient outcome prediction. *Med Image Anal* 67:101844
26. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H (2019) A clinical text classification paradigm using weak supervision and deep representation. *BMC Med Inform Decis Mak* 19(1):1
27. Hughes M, Li I, Kotoulas S, Suzumura T (2017) Informatics for health: connected citizen-led wellness and population health. IOS Press, Amsterdam, pp 246–250
28. Nguyen TH, Shirai K (2013) International conference on application of natural language to information systems. Springer, pp 278–284
29. Kumar S (2020) Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Ann Data Sci* 7:417

30. Li J, Guo K, Viedma EH, Lee H, Liu J, Zhong N, Gomes LFAM, Filip FG, Fang SC, Özdemir MS et al (2020) Culture versus policy: more global collaboration to effectively combat COVID-19. *Innovation* 1(2):10003
31. Liu Y, Gu Z, Xia S, Shi B, Zhou XN, Shi Y, Liu J (2020) What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. *EClinicalMedicine* 22:100354
32. Temesgen A, Gurmesa A, Getchew Y (2018) Joint modeling of longitudinal CD4 count and time-to-death of HIV/TB co-infected patients: a case of Jimma University Specialized Hospital. *Ann Data Sci* 5(4):659
33. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) *Optimization based data mining: theory and applications*. Springer, Berlin
34. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. *Ann Data Sci* 4(2):149
35. Olson DL, Shi Y, Shi Y (2007) *Introduction to business data mining*, vol 10. McGraw-Hill, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.