Computers in Biology and Medicine 182 (2024) 109211

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

COVID-19 from symptoms to prediction: A statistical and machine learning approach

Bahjat Fakieh [a], Farrukh Saleem [b,*]

[a] Department of Information System, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
[b] School of Built Environment, Engineering, and Computing, Leeds Beckett University, Leeds, LS6 3QR, UK

ARTICLE INFO

Keywords:
Machine learning
Statistical analysis
Ensemble algorithms
COVID-19
Public health informatics
Predictive models

ABSTRACT

During the COVID-19 pandemic, the analysis of patient data has become a cornerstone for developing effective public health strategies. This study leverages a dataset comprising over 10,000 anonymized patient records from various leading medical institutions to predict COVID-19 patient age groups using a suite of statistical and machine learning techniques. Initially, extensive statistical tests including ANOVA and t-tests were utilized to assess relationships among demographic and symptomatic variables. The study then employed machine learning models such as Decision Tree, Naïve Bayes, KNN, Gradient Boosted Trees, Support Vector Machine, and Random Forest, with rigorous data preprocessing to enhance model accuracy. Further improvements were sought through ensemble methods; bagging, boosting, and stacking. Our findings indicate strong associations between key symptoms and patient age groups, with ensemble methods significantly enhancing model accuracy. Specifically, stacking applied with random forest as a meta leaner exhibited the highest accuracy (0.7054). In addition, the implementation of stacking techniques notably improved the performance of K-Nearest Neighbors (from 0.529 to 0.63) and Naïve Bayes (from 0.554 to 0.622) and demonstrated the most successful prediction method. The study aimed to understand the number of symptoms identified in COVID-19 patients and their association with different age groups. The results can assist doctors and higher authorities in improving treatment strategies. Additionally, several decision-making techniques can be applied during pandemic, tailored to specific age groups, such as resource allocation, medicine availability, vaccine development, and treatment strategies. The integration of these predictive models into clinical settings could support real-time public health responses and targeted intervention strategies.

1. Introduction

The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first identified in Wuhan, China, in December 2019 and later declared a global pandemic by the World Health Organization (WHO) [1]. As of February 17, 2023, it has infected approximately 756 million individuals and resulted in nearly 6.85 million deaths worldwide [2]. The rapid development and distribution of vaccines have significantly decreased the number of new cases, with over 13 billion doses administered globally. This advancement is gradually restoring economic activities to pre-pandemic levels [3].

Throughout the pandemic, the collaborative efforts of the scientific community, including paramedics, have been pivotal. Researchers have contributed through diverse means, from enhancing educational tools [4] and supporting national economies [5]to boosting corporate and healthcare responses [6]. Significant advancements have also been made by the artificial intelligence (AI) and machine learning (ML) communities. They have developed digital tools for reducing the pandemic's impact, such as smart applications for future case prediction [7], contact tracing systems [8], enhanced testing capabilities [9], and advanced diagnostic systems [10]. These communities are a helpful source for researchers to analyze, evaluate, and generate new patterns and recommendations to authorities to respond efficiently during this crisis.

This study aims to develop a statistical and ML model to assist healthcare professionals in managing COVID-19 across different patient age groups. The primary research question addresses predicting COVID-19 patient age groups using various attributes through statistical and ML techniques. Our approach involves analyzing two datasets: one previously published [11] and another newly collected from several Pakistani

* Corresponding author. School of Built Environment, Engineering, and Computing, Leeds Beckett University, Leeds, LS6 3QR, UK.
E-mail address: f.saleem@leedsbeckett.ac.uk (F. Saleem).

https://doi.org/10.1016/j.compbiomed.2024.109211
Received 23 May 2024; Received in revised form 2 September 2024; Accepted 23 September 2024
Available online 28 September 2024

hospitals post-identification of the Delta variant [12]. This comparative analysis across different datasets aims to identify the best dataset for ML implementation and understand the impact of COVID-19 mutations on symptom presentation. The specific symptoms analyzed include cough, fever, sore throat, shortness of breath, and headache [13,14].

Contributions of this research include.

- Curating a dataset from Pakistani medical institutions for current and future research.
- Publishing this dataset on prominent research platforms for global accessibility.
- Employing statistical and ML methods to delineate the relationship between COVID-19 symptoms and patient age groups.
- Analyzing the significance of each symptom in the infected population through ANOVA, Chi-Square Test of Independence, and t-tests.
- Implementing and comparing various ML algorithms to determine the most effective method for age group prediction.
- The use of ensemble approaches and cross validation techniques to avoid overfitting issues.
- To provide a better understanding of the different symptoms and their association with age group.

Machine learning has been widely used for predicting disease outcomes or diagnosis, while applying it to predict a patient's age group based on symptoms is a novel approach. The idea presented in this study can lead to customized and age-specific treatment strategies, which is critical in managing during a pandemic like COVID-19. The findings of this study will help decision-makers develop specific strategies tailored to particular age groups, including resources allocation, medicine availability, vaccine development, and treatment strategies. The methodology applied in this study focused on symptoms-based modeling for predicting age group, which is not commonly explored in previous COVID-19 epidemiological studies. Overall, age-specific impacts of COVID-19 explored in this research, which can be useful for adopting preventive measures in future.

The remainder of the paper is organized as follows: Section 2 outlines the research context, Section 3 details the methodological approach, Section 4 presents the statistical analysis, Section 5 discusses the ML implementations, and the final section concludes the study.

2. Research background

The academic response to the COVID-19 pandemic has been extensive, exploring a wide array of topics that range from clinical to societal impacts. Key areas of study include symptoms [15], testing methodologies [16], diagnostic tools [17], treatment protocols [18,19], preventative measures [20], virus transmission [21,22], vaccine development [23,24], and forecasting future trends [25,26]. Each of these topics is crucial for comprehensive pandemic management and has been extensively documented in the literature. The pandemic's reach has extended beyond healthcare, affecting various sectors such as education [27,28], medicine more broadly [29,30], corporate operations [31,32], social interactions [33,34], and even human psychology [35,36].

Previous research has also identified several other elements that have significant impact on the virus's transmission. For example, the regression method was used in the Australian context to estimate virus propagation over the road network system. In addition, socio-demographic factors used to enhance the effectiveness of the proposed model [37]. Furthermore, suburban road network investigated in research and identified as one of the common vulnerabilities in spreading the COVID-19 virus. For the investigation, authors used a dataset collected from Greater Sydney of New South Wales [38,39]. This breadth of research demonstrates the virus's pervasive impact and the multidisciplinary approach required to address it.

Given the scope of these studies, significant attention has been paid to developing and refining diagnostic tools, and predicting the relationship between common symptoms, age groups, and prevention measures. Decision making techniques are further enhanced the idea of proposing treatment strategies during this pandemic situation [40]. These tools have been pivotal in detecting and monitoring the disease, offering critical insights into its progression and treatment efficacy. As such, this research contributes to the ongoing discussion by focusing on advanced machine learning techniques to enhance diagnostic accuracy, and patient management system. Specifically, this research used the data for analysis and ML implementation, collected while performing laboratory test in diagnosing COVID-19.

2.1. COVID-19 symptoms

This study explores common and clinical symptoms of COVID-19, emphasizing their correlation with findings of the test report, which are crucial for the identification of suspected cases. According to the World Health Organization, key symptoms such as fever, cough, and shortness of breath may appear singly or in combination and are often the first indicators of the disease [41]. Using similar findings, a study proposed a model to understand the prominent and expected cases of COVID-19 using five basic symptoms. The idea presented in that research is to provide a clear understanding of infected persons using basic clinical findings. The goal is to develop a predictive framework to identify potential COVID-19 cases using clinical symptom [11].

Another research highlighted the common signs and symptoms of COVID-19 for predicting infected people in Jordan. The dataset used in that study, originating from Jordan, enables the validation of these models across diverse populations, ensuring robustness and applicability [42]. Furthermore, the study in Geneva that focused on tracking the evolution of COVID-19 symptoms and to monitor disease progression, ensuring continuous patient care [43]. This integrated approach highlights the value of continuous data collection in managing patient outcomes. Similarly, several machine learning models were used to identify early-stage symptoms such as fever, cough, lung infection, and runny nose in order to predict COVID-19 infection in different age groups, with XGBoost recorded the highest accuracy [44].

Additionally, Wang, H. Y. et al. (2020) highlighted the occurrence of neurological symptoms such as instability in walking, headaches, and dizziness in COVID-19 patients [45]. These findings suggest the importance of including neuroimaging studies to identify early markers of COVID-related neurological complications, further complicating the diagnostic process. The absence of traditional respiratory symptoms in cases presenting with primary neurological signs underscores the need for a multi-modal diagnostic approach that combines symptom assessments using different laboratory tests. This section underscores the integration of clinical findings collected through laboratory tests, utilizing advanced machine learning techniques to enhance the predictive accuracy of COVID-19 diagnostics. This approach is vital for analyzing, identifying, and predicting COVID-19 positive cases, where detailed laboratory data can significantly contribute to comprehensive patient management and cure strategies.

2.2. Statistical analysis

During the COVID-19 pandemic, the integration of statistical analysis on different datasets has played a crucial role in supporting the prevention and control strategies implemented by government and health agencies. Therefore, this study also provided a detailed statistical analysis of real-world data collected from hospitals and medical agencies. Statistical tools such as ANOVA, T-Test, and Gaussian modeling have been employed to analyze recovery and mortality rates across various European and Asian countries, revealing that infection rates correlate strongly with population density and testing rates [46]. The main purpose of the research was to determine recovery and mortality rates. The results of this study suggested that the COVID-19 infection rate largely depends on the population of the country and

the number of tests conducted. Therefore, it is evident from the study that identifying valuable predictive results using statistical analysis can provide potential evidence for government agencies to develop future control and prevention strategies.

Furthermore, another research leverages the Auto-Regressive Integrated Moving Average Model (ARIMA) applied to Pakistani health data [47]. ARIMA's capability in time series analysis allowing for a more nuanced analysis of how symptoms and disease markers evolve over time. Despite challenges such as limited data samples, the use of statistical analysis on medical dataset offers comprehensive insights, leading to robust public health strategies. The interdisciplinary approach includes collaboration between statisticians, medical experts, and machine learning specialist to develop models that interpret complex data sets effectively. This collaborative effort is essential for advancing diagnostic and prognostic tools that are capable of handling diverse and large-scale data sets.

Ethical considerations, particularly regarding data privacy, are paramount. The study adheres to strict data protection measures to ensure confidentiality and integrity in the handling of sensitive patient data. Additionally, it addresses potential biases by ensuring demographic diversity within the datasets, thus enhancing the fairness and applicability of the models across different populations.

In conclusion, the integration of statistical models with medical data not only provides a deeper understanding of the COVID-19 pandemic but also significantly impacts the development of data-driven public health policies. The subsequent section will detail the methodologies used in this study, emphasizing how these integrated models are implemented to refine management strategies for COVID-19.

2.3. ML approaches for COVID-19: enhancing predictive models

The integration of machine learning, with healthcare data has transformed how medical authorities manage and respond to COVID-19. ML techniques have been pivotal in developing predictive models that not only forecast the disease progression but also assist in analyzing medical data to provide deeper insights into the infection mechanisms. The implementation of ML algorithm using supervised and ensemble approaches has achieved remarkable success in the medical industry [48,49]. These traditional and ensemble methods offer a comprehensive understating of problems, optimize model performance, and provide reliable solutions [50].

Studies using ML to manage COVID-19 data have shown significant advancements. For example, optimized regression models were applied to predict mortality rates in countries like France, Spain, Turkey, Sweden, and Pakistan. Notably, even without a lockdown, Sweden's mortality rates were lower than those in the studied countries [51]. These models, when integrated with clinical dataset, can enhance the understanding of underlying factors that influence mortality rates. Additionally, studies have highlighted other clinical findings in data sets using ML algorithms. By analyzing historical numbers of positive cases, researchers proposed an improved model to predict the growth rate and future trends of COVID-19.

ML algorithms have also improved the accuracy of predicting COVID-19 trends. The Weibull model, known for its robustness, has outperformed baseline models in forecasting the virus's growth rates [52]. Ultimately, sharing of these results with medical agencies, could provide a more comprehensive approach in creating future strategies. Moreover, epidemiological models have identified the exponential nature of the virus spread [53] and demonstrated that multilayer perceptron have better predictive accuracy than traditional linear regression [54]. These findings could be further enriched by correlating them with dynamic changes observed in the new collected data, providing a multidimensional view of the pandemic's progression.

The implementation of machine learning algorithms on medical data has demonstrated a wide range of predictive models, including supervised models and ensemble approaches. While supervised learning approaches are fundamental to many ML tasks and can offer strong performance [55,56], this study further explores the use of ensemble algorithms to reduce bias and variance by combining multiple predictive models. Ensemble algorithms typically provide better generalization of the data and help mitigate overfitting issues [57]. By incorporating both supervised and ensemble approaches, the diversity of the models in this study is enhanced. This comparative study is beneficial for understanding the strength and weaknesses of each technique when applied to similar dataset.

The wide range of ML algorithms applied to different applications highlights the variety of optimization methods discussed in previous studies. A review analysis on the applicability of ML algorithms for disease prediction found that Support Vector Machine (SVM) is the most frequently used, while Random Forest (RF) showed the highest accuracy comparatively [55]. Additionally, ensemble approaches such as stacking have the potential to combine multiple models, known as base learners, to create a final prediction model (i.e. meta model). Stacking has proven to be a successful ensemble method, demonstrating the highest prediction accuracy compared to other ensemble methods, particularly when applied to medical datasets [58].

The main goal of ML in this context is to equip computers with the ability to simulate human decision-making processes based on training data, which then can be applied to new datasets for predictive purposes. For instance, an XGBoost model used to predict respiratory failure in COVID-19 patients showed a 91 % accuracy rate [59]. Recently, ML and statistical analysis techniques have been widely applied to data from various countries to predict infection rates, mortality rates, and the potential spread of the COVID-19 virus. One analysis aimed to forecast the death ratio trend in India over the coming weeks [60]. Additionally, using historical data, a model was proposed to estimate future trends and the potential number of infections in China [61].

3. Methodology

This section delineates the comprehensive methodology employed to address the research problem stated earlier. The methodology is structured into four distinct phases, each integral to the research process, as illustrated in Fig. 1. Detailed descriptions of each phase are provided in the following subsections to elucidate the steps and techniques utilized. This structure not only facilitates a clear understanding of the research approach but also aligns with the rigorous analytical frameworks typical in medical studies.

3.1. Phase 1: data collection

In this study, our data collection strategy was meticulously designed to source datasets pertinent to our research objectives. We explored a variety of online resources such as Kaggle, GitHub, HDX, and data.world to secure appropriate data. An initial dataset was chosen based on its mention in a relevant academic publication and available on GitHub [62], as referenced in an influential study [11]. However, subsequent analysis indicated that this dataset was insufficient for our research needs, leading us to augment our data collection efforts.

Given the necessity for more specific data, particularly in the context of the recent COVID-19 variants, we opted to collect additional data through a quantitative survey. Initial plans to gather data directly from medical institutions were adjusted due to stringent requirements for permissions and ethical consent. The Delta variant, first identified in India on June 21, 2021, significantly influenced our data collection strategy. Although we initially planned to collaborate with local researchers in India, logistical challenges, including a critical shortage of specific protective equipment, necessitated a shift to Pakistan. From August 1 to August 15, 2021, we successfully conducted the survey in Pakistan, where it was feasible to deploy temporary staff for this purpose.

The datasets collected—both the initial and the newly acquired—are

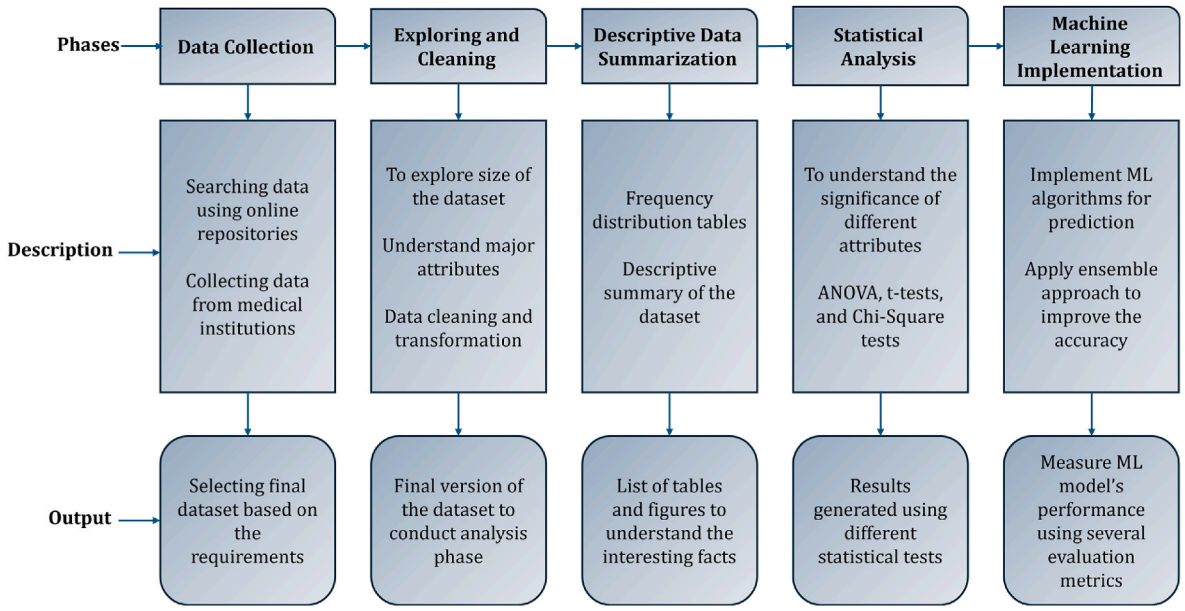

Fig. 1. Stepwise implementation of methodology.

extensively analyzed in later phases of this research. Preliminary findings and a comprehensive statistical evaluation of these datasets are presented in Section 4, whereas the main attributes of collected dataset and their description is shown in Table 1. We collected our own dataset using various attributes from previously published datasets. The main purpose of collecting a new dataset is to introduce novelty in the findings applied to different geographical regions and groups of people (i.e. Pakistan). Table 1 highlights the main factors considered for data collection. In addition to demographic data, the major attributes define the association between several symptoms and a COVID-19 positive result.

3.2. Phase 2: data exploration and cleaning

The objective of this phase is to thoroughly explore the dataset to ascertain its essential attributes, setting the stage for advanced statistical analysis and machine learning implementation. The initial step involves a detailed assessment of data quality and adherence to ethical standards as outlined in the Declaration of Helsinki, which emphasizes the minimization of harm and protection of participant privacy. To this end, all identifiable participant information was meticulously removed.

Table 1
Dataset attributes.

Attribute	Description	Categories
Corona Result	The test findings obtained to identify the patient were either positive or negative.	0 = non-infected 1 = infected
Age	Describe the age of the patients.	Numerical values denoting age of the patient
Gender	Explain the gender of the patients.	0 = Female 1 = Male
Sore Throat	The first symptom to identify its relationship with COVID-19.	0 = No 1 = Yes
Shortness of Breath	To analyze whether a patient suffered shortness of breath.	0 = No 1 = Yes
Headache	Third symptoms to know the condition of the patients.	0 = No 1 = Yes
Cough	To determine whether a patient experienced cough or not.	0 = No 1 = Yes
Fever	Last investigated symptom before COVID-19 test.	0 = No 1 = Yes

Data cleaning was particularly focused on ensuring the integrity and uniformity of the dataset. This included employing advanced techniques to manage missing values and normalize data columns, which are critical for maintaining the reliability of subsequent analyses. The 'Age' attribute was discretized into categorical groups to aid in nuanced analysis, where age-related variations can significantly influence diagnostic outcomes.

The number of preprocessing steps used before ML model development. For example, to convert all data denoting the presence of symptoms (yes/no) to binary values, "1" and "0". Similarly, the "infected" column values were transformed to binary values. After carefully examining the obtained data, several transactions were eliminated due to missing information, such as presence of infection, age, and unexplained symptoms.

Data values from various sources were standardized to create a consistent dataset, essential for accurate statistical analysis and modeling. This standardization process included the use of specialized software tools designed for data discretization and feature extraction, ensuring that the dataset is suitable for rigorous analysis.

The comprehensive methods and results of this data preparation phase are elaborated upon in the results section, detailing how these processes have influenced the study's findings and the implications for statistical and machine learning implementation.

3.3. Phase 3: descriptive data summarization

This crucial phase addresses the primary research questions through comprehensive descriptive statistical analysis. Each dataset attribute was meticulously analyzed for value counts, category diversity, and frequency distribution. Findings from this analysis are presented in Section 4 through detailed tables and figures, facilitating an understanding of the associations between different attributes. This phase is especially important for analyzing a complex medical data, as statistical insights can inform the development of algorithms that enhance diagnostics prediction, such as extracting positive cases based on detected patterns linked to clinical outcomes.

3.4. Phase 4: statistical analysis

Statistical analysis was conducted to uncover the relationships and features within the datasets. This involved the use of multiple ANOVA

tests, which helped determine the significance and correlations between variables, focusing on those attributes most relevant to medical data, such as features correlated with disease markers. The F-values, p-values, and t-tests offered detailed insights into the reliability and relevance of these attributes, which are essential for predictive modeling. Results are thoroughly detailed in Section 4, underscoring their implications for predicting disease progression and treatment efficacy using symptoms data.

3.5. Phase 5: machine learning implementation

The final phase involved selecting the optimal dataset for implementing various machine learning algorithms, including advanced ensemble methods to improve prediction accuracy. This stage was crucial for evaluating the applicability of machine learning models for predicting positive cases using multiple attributes, where precision, recall, F-measure, and accuracy are paramount. These metrics, discussed in detail in Section 4, validate the efficacy of the models in clinical settings, particularly in their ability to enhance diagnostic accuracy and reliability.

4. Descriptive and statistical analysis – results

4.1. First dataset - statistical analysis

The objective of this phase was to evaluate the first dataset obtained from an open data repository [62] for its potential to develop an intelligent computing model for predicting COVID-19 infection based on five symptoms. The dataset, encompassing 278,848 records from March 11, 2020, to April 30, 2020, underwent rigorous statistical analysis, organized into three stages: data exploration and cleaning, descriptive statistical analysis, and significance testing using ANOVA to determine the predictive power of the symptoms for COVID-19.

4.1.1. Exploring and cleaning

In this dataset, critical attributes include test date, symptoms (cough, fever, sore throat, shortness of breath, headache), corona test result, age (60 and above), gender, and test indication. The initial cleaning targeted the 'gender' and 'corona result' attributes, where inconsistencies and missing values were prevalent. Specifically, the dataset contained 19,563 records with unspecified or blank gender values and 3892 records with indeterminate corona test results, as detailed in Table 2. These records were excluded to ensure the integrity of the data used in the analysis, which is crucial for maintaining the accuracy of statistical and, subsequently, symptoms-based predictive models.

The cleaned dataset ensures reliable input for further statistical evaluation, which is particularly important when these methodologies are paralleled with machine learning implementation. For example, the application of ANOVA in this context helps identify significant features or patterns that might correlate with major findings in medical studies, where similar attributes can influence diagnostic outcomes.

4.1.1.1. Attributes and values summary. During the data exploration and cleaning phase, a significant issue was identified with the 'age 60 and above' attribute, where approximately half of the records were unspecified. These records were retained in the analysis due to the binary nature of the available age categorization (under or over 60). This simplistic age grouping limits the depth of analysis that can be conducted. For instance, younger age groups may show different physiological responses compared to older groups.

4.1.1.2. Enhanced data collection recommendations. To align better with the rigorous standards expected in related research, future data collection should consider more detailed age stratifications. Suggested age categories could include under 18, 18–25, 26–30, and so forth, which would allow for more precise correlation between age-related changes in clinical outcomes. This nuanced approach is essential for developing predictive models and diagnostic tools, as it enhances the capability to tailor interventions and understand disease impacts across different life stages.

4.1.2. Descriptive data summarization

Following the data cleaning process, our analysis was refined to 255,911 valid entries after removing 22,937 records due to unspecified information regarding gender and corona test results. The refined dataset indicates a balanced gender distribution with 127,370 males and 128,541 females as shown in Table 3. Of these, 13,560 entries, representing 5.3 % of the total, were confirmed as COVID-19 infections, with a higher incidence in males (7,519) compared to females (6,041).

Cough was identified as the most prevalent symptom, reported by 15.3 % of the participants, followed by fever at 7.8 %. This data is crucial for subsequent analyses that explore correlations between these symptoms and diagnostic findings.

4.1.3. Statistical analysis

This section evaluates the significance of specific symptoms and gender in predicting COVID-19 infections using an ANOVA test as illustrated in Table 4. The analysis focuses on the relationship between these variables and the likelihood of infection.

Despite the significant associations indicated by the ANOVA results for all symptoms, a deeper examination reveals challenges due to insufficient sample sizes for certain symptoms like shortness of breath, which appeared in only 0.4 % of cases. This rarity impacts the robustness of statistical conclusions, as indicated by the near-zero P values. Further analysis shows that a significant portion of the infected cohort (36.3 % or 4929 out of 13,560) exhibited none of the primary symptoms. This finding suggests the presence of asymptomatic carriers and highlights the potential for other undetected symptoms or markers of infection.

This result suggests that researchers may encounter challenges in developing a machine learning model with the current data, due to the lack of significant findings among the explored factors. The next phase of this study will focus on analyzing a different dataset.**4.2.**

Table 2
Data exploration – an overview.

Attributes/Values				
Gender				
Value	Male	Female	Unspecified	Total
Total	129127	130158	19563	278848
Corona Result				
Value	Infected	Not infected	Other	Total
Total	14729	260227	3892	278848

Table 3
Descriptive statistics overview (N = 255,911).

Attribute	Entry	Frequency	n (%)
Gender	Male	127370	49.8 %
	Female	128541	50.2 %
Infections	Infected	13560	5.3 %
	Not Infected	242351	94.7 %
Gender-based Infections	Infected Males	7519	5.9 %
	Infected Females	6041	4.7 %
Age	Under 60	112788	44.1 %
	60 and above	23749	9.3 %
	None	119374	46.6 %
Symptoms	Cough	39054	15.3 %
	Fever	19855	7.8 %
	Sore throat	1498	0.6 %
	Shortness of breath	1080	0.4 %
	Headache	2108	0.8 %

Table 4
ANOVA test of significance based on gender.

	Males & Females	Males	Females
F value (df)	14778.89 (1,255908)	7293.83 (1,127368)	7560.72 (1,128539)
P value of all symptoms	0	0	0
P value of Cough	1.913 x 10-246	5.543 x 10-117	3.681 x 10-132
P value of Fever	0	0	0
P value of Sore throat	0	0	0
P value of Shortness of breath	0	0	0
P value of Headache	0	0	0

4.2. Second dataset - statistical analysis

Upon encountering validation challenges with the first dataset, we turned our focus to a second dataset specifically collected for this study from Pakistan, comprising 516 participants primarily infected with COVID-19.

4.2.1. Exploring and cleaning

Initial preprocessing involved the removal of six non-infected entries to ensure the dataset focused exclusively on infected cases. Additional cleaning addressed inconsistencies in the age attribute, standardizing various representations into a numerical format to maintain data consistency and reliability, essential for subsequent analyses, including potential imaging studies.

4.2.2. Descriptive data summarization

After cleaning, the dataset's gender distribution was balanced, with 58 % males (297) and 42 % females (219). Detailed age stratification facilitated the categorization of participants into several age groups, enhancing the granularity of the analysis. Since the data collection process was planned and monitored by the authors of this study, classifying participants into various age groups was more straightforward. As shown in Table 5 and Fig. 2, 48 % of participants were in the 18–30 age group. This group was followed by three relatively balanced groups: 51–60 (16 %, 84 participants), 31–40 (14 %, 72 participants), and 41–50 (11 %, 56 participants). The prevalence of symptoms and demographic distributions are summarized as follows:

The descriptive analysis provides a foundation for further investigation into the correlations between reported symptoms and infections, if available. For instance, the high incidence of cough and headache might be linked with specific critical features observed in clinical findings.

Table 5
Descriptive statistics (N = 516).

Attribute	Entry	Frequency	Percentage (%)
Gender	Male	297	58 %
	Female	219	42 %
Infections	Infected	510	99 %
	Not Infected	6	1 %
Gender-based Infections	Infected Males	295	58 %
	Infected Females	215	42 %
Age Groups	Less than 18	12	2 %
	18–30	244	48 %
	31–40	72	14 %
	41–50	56	11 %
	51–60	84	16 %
	61 and above	42	8 %
Symptoms	Cough	391	77 %
	Fever	388	76 %
	Sore throat	224	44 %
	Shortness of breath	109	21 %
	Headache	418	82 %
Asymptomatic Infections	Without Symptoms	6	1.18 %

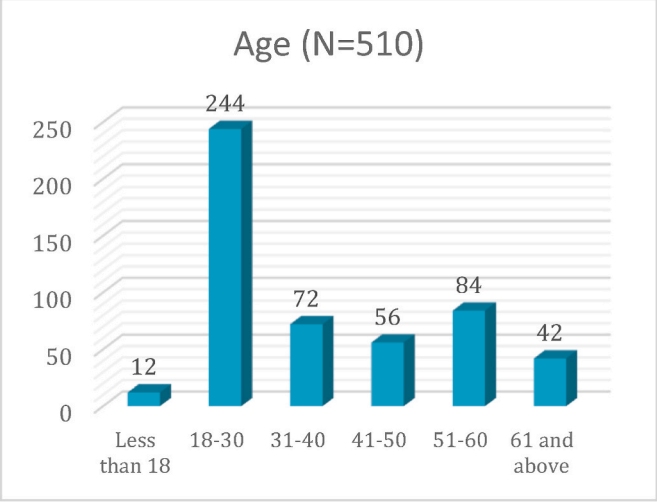

Fig. 2. Frequency distribution based on the age groups.

As detailed in Table 5 and Fig. 3, a significant majority of the infected participants reported experiencing symptoms, with headache being the most common at 82 % (418 individuals). Close behind were cough and fever, reported by 77 % (391) and 76 % (388) of participants, respectively. Sore throat and shortness of breath were also reported, albeit at lower frequencies of 44 % (224) and 21 % (109), respectively. Unlike the findings from the first dataset, this dataset indicates that nearly all infected individuals experienced at least one of the five investigated symptoms, with only 6 cases presenting as asymptomatic.

Given the high prevalence of these symptoms, it is crucial to delve deeper into understanding their interrelationships and potential diagnostic significance. This insight could be particularly useful for enhancing predictive models, where symptoms might correlate with specific features such as age group and gender, indicative of COVID-19 infection.

4.2.3. Statistical analysis

The aim of this analysis was to assess the relationship between specific symptoms of COVID-19 and the likelihood of infection. For this purpose, the predominance of infected individuals in the dataset necessitated a shift in focus to ANOVA tests to determine the influence of demographic factors (age and gender) on symptom prevalence. Prior to conducting the ANOVA tests, age was categorized into six distinct groups based on the descriptive statistics outlined in Table 5. This

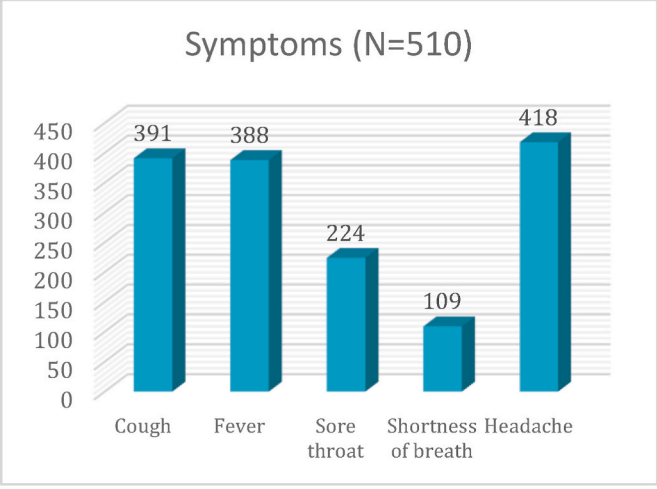

Fig. 3. Frequency distribution based on the symptoms.

structured approach facilitated a nuanced examination of symptom distribution across various age demographics, enhancing the specificity of our findings. Therefore, six age groups were included in this test. The ANOVA test presented in Table 6 examines the relationship between age and gender groups, with a significance level of $\alpha = 0.05$. Consequently, any P values lower than 0.05 would indicate statistical significance for the investigation group.

The ANOVA test result in Table 6 shows a significant relationship between the combination of the five symptoms and the age groups, with a P value of 6.151 x 10^-9, which is less than the selected α value. However, these symptoms do not reliably indicate patient gender, as the P value is 0.106, which is greater than α.

Moreover, to analyze frequency-based data, contingency table showing the relationship between gender and age group as depicted in Table 7. As per the collected data, age group (2) has the highest number of individuals for both genders, with 118 females and 131 males. Age group (6) has relatively fewer individuals, particularly among females. The table is showing an understanding of frequency distribution of individuals in different age groups categorized by gender.

To statistically test the relationship between gender and age group, a Chi-Square Test of Independence was conducted. The results are illustrated in Table 8. This test helps identify significant association between the variables. The table shows the expected frequencies for each category in the contingency table to determine if gender and age group are independent. according to the p-value (0.0358), which is less than 0.05, there is a statistically significant association between gender and age group in this dataset. This indicates that the distribution of age group is not independent of gender.

The next step in th7e analysis is using a *t*-test to examine the significance of each symptom, as shown in Table 9.

Table 9 shows that the most significant symptoms indicating age groups were sore throat and trouble breathing, with P values of 0.000240693 and 1.93033 x 10^-5, respectively. Conversely, dry cough, fever or chills, and headache showed low significance in the investigated sample based on the resulting P values. The ANOVA test supported our research problem, allowing us to predict age groups based on different symptoms.

4.3. Discussion – descriptive and statistical analysis

This study conducted a comparative analysis across two diverse datasets to investigate the manifestation of COVID-19 symptoms, collected from different geographic regions. This approach not only broadens the empirical base of the research but also enriches the understanding of symptomatic variations across populations. This research contributed in the following ways:

Dual-Dataset Analysis: By evaluating two distinct datasets, this research underscores the variability and commonality in COVID-19 symptom presentation across international cohorts. Such comprehensive data analysis enhances the reliability of the findings, providing a robust basis for subsequent predictive modeling.

Statistical Methodologies: The application of advanced statistical techniques allowed for the detailed exploration of relationships between COVID-19 symptoms and other patient characteristics, such as age and gender. This methodological rigor helps in identifying statistically significant symptom patterns that could be vital for early disease detection and management.

Significance of Symptoms in Demographic Segmentation: The study identified "Sore throat" and "Trouble breathing" as particularly significant symptoms for determining age-related differences in COVID-19 infection rates. These symptoms' prominence suggests potential diagnostic markers that could be targeted in imaging studies. The above conclusion was based on the data samples presented in the second dataset. A deeper examination of the first dataset reveals challenges due to insufficient sample sizes for certain symptoms, such as shortness of breath. Further analysis shows that a significant portion of the infected cohort (36.3 %) exhibited none of the primary symptoms. The reason could be the null values, or presence of asymptomatic carriers. Since the first dataset was not used in the development phase of the predictive models, no claims are made regarding the importance of symptoms in this dataset. *Foundation for Machine Learning Models:* The second dataset's analysis demonstrated potential for constructing machine learning models capable of predicting patient demographics based on symptom presentation. This indicates a promising direction for integrating statistical findings with machine learning to enhance diagnostic accuracy.

Table 6
Second dataset ANOVA test of significance based on age and gender.

Variable	F value (df)	P value
Age Group	9.7798 (5, 504)	6.151 x 10^-09
Gender	1.8244 (1, 508)	0.10

Table 7
Contingency Table showing the relationship between Gender and Age Group.

Age Group	1	2	3	4	5	6
Female	7	118	30	25	29	10
Male	5	131	42	32	55	32

Table 8
Chi-Square test of independence.

Age Group	1	2	3	4	5	6
Female	5.09	105.68	30.56	24.19	35.65	17.83
Male	6.91	143.32	41.44	32.81	48.35	24.17

p-value: 0.0358.
Degree of freedom: 5.

Table 9
T-tests on the five investigated symptoms by age group.

Symptom	*t*-Test	P value
A dry cough	1.560683073	0.11
Fever or chills	0.015488516	0.98
Sore throat	3.698201119	0.0002
Trouble breathing	4.313471746	1.93033 x 10^-05
Headache	−0.245418305	0.80

5. Machine learning implementation and results

Following the insights from previous statistical analyses, particularly the ANOVA test indicating 'Age Group' as a significant factor, this study embarked on developing a machine learning-based model to predict an individual's age group using various COVID-19 symptoms. This model, aimed at integrating with an intelligent system, could enhance our understanding of COVID-19 outcomes across different age demographics. Therefore, a prediction models developed using the most significant factor (age group) as a class variable. This model will be useful for predicting a person's age group based on the symptoms described in Table 9. Additionally, the prediction model can be integrated with an intelligent system to help identify COVID-19 outcomes based on different age groups.

The study conducted several machine learning (ML) experiments to find the optimal solution. Initially, the authors evaluated different age groups as highlighted in Fig. 2. However, due to the dataset's bias towards the 18–30 age group, the results were unsatisfactory. To improve prediction accuracy and avoid dataset bias, the class variable divided into two classes: Class_1 (<30) and Class_2 (≥30). This division balanced the number of transactions in each class and resolved the bias issue.

For the prediction model, the following classification ML algorithms

and approaches were selected, which frequently implemented and provide a better understanding applied on COVID-19 datasets: Decision Tree (DT) [63], Naïve Bayes (NB) [64], K-Nearest Neighbors (KNN) [65], Gradient Boosted Trees (GBT) [63], Random Forest (RF) [66], SVM [55] Bagging [67], Bagging [68], Stacking [69]. The predictive models developed offer promising avenues for integration with medical data. By correlating predicted age groups and symptom data, these models could potentially enhance diagnostic accuracy and provide deeper insights into the pathophysiological variations of COVID-19 across different patient demographics. This could lead to improved patient management strategies, tailored treatment plans, and ultimately, better clinical outcomes.

This study used several machine learning models for prediction. The number of preprocessing steps used before ML model development. For example, to convert all data denoting the presence of symptoms (yes/no) to binary values, "1" and "0". Similarly, the "infected" column values were transformed to binary values. After carefully examining the obtained data, several transactions were eliminated due to missing information, such as presence of infection, age, and unexplained symptoms. Furthermore, throughout the ML implementation, different features were chosen to improve the model's performance. For example, the number of trees (100), maximal depth (10), and number of bins (10) were identified as optimal features for applying the GBT algorithm. To enhance overall performance of the model, different parameters were tested, and Table 10 highlights the optimal features selected for each algorithm.

5.1. Experiment 1 [all selected algorithms]

5.1.1. Model implementation and setup

The initial phase of our machine learning exploration involved deploying various algorithms using Rapid Miner, an open-source tool favored by the research community for its robust ML capabilities [70–72]. The setup avoided ensemble techniques initially to isolate the performance of each algorithm. Although RF and GBT inherently employ ensemble strategies, their integration was managed distinctly to evaluate their standalone capabilities.

Table 10
The Optimal Features selected for different ML Model Implementation.

Decision Tree		Random Forest		Gradient Boosted Tree	
Criterion	Gain Ratio	No. of trees	100	No. of trees	100
Maximal depth	10	Criterion	Gain Ratio	Maximal depth	10
Apply pre-pruning	Yes	Apply pre-pruning	Yes	Min rows	10
Number of pre-pruning alternatives	3	Random Splits	Yes	No. of bins	10
Confidence Level	0.1	Voting strategy	Majority vote	Early Stopping	Yes
Minimal size of split	3	Parallel execution	Yes	Distribution	Auto
Naïve Bayes		**KNN**		**SVM**	
		Number of K	5	Kernel Type	Radial
		Weighted Vote	Yes	Kernel Gamma	1.0
		Measure Type	Mixed Measures		
Bagging		**Boosting**		**Stacking**	
Sample Ratio	0.9	Cross Validation	10	Cross Validation	10
Iterations	10	Iterations	10	Attributes	All
Average Confidence	Yes				

For comprehensive evaluation, the "Cross Validation" operator was used, which is standard for assessing model accuracy through repetitive testing and training cycles. This dual-phase process allows each ML model to be assessed independently, ensuring the integrity and reliability of the performance metrics. All algorithms were executed in a single process described in Fig. 4. In this figure, the operator representing the ML classifier is known as 'Cross Validation,' which is primarily used to estimate the accuracy of the learning model through multiple iterations (i.e., 10). This operator, also called a nested operator, is divided into two sub-processes: training and testing. During the training phase, each ML classifier was used separately, where the model was validated and various evaluation metrics were measured, as discussed below.

5.1.2. Performance metrics and analysis

The experiment's outcomes, detailed in Table 11, showcase the comparative effectiveness of each classifier based on precision, recall, F-measure, and accuracy, across the two predefined age groups: Class 1 (age_group <30) and Class 2 (age_group ≥ 30). These results highlight potential relationships between varying COVID-19 symptoms and age demographics, which are critical for refining public health strategies tailored to specific age groups. The performance was measured using the prediction values for the two classes. The experiment suggested a relationship between COVID-19 symptoms and different age groups. The model can be used to predict age groups and to design new strategies, such as vaccination and testing, for people in each age group.

The analysis confirmed that tree-based classifier such as GBT (0.6607) typically outperform others due to their robust nature in handling complex datasets, which is also identified in the previous work [73], followed by the SVM (0.6531) The other two tree-based models like DT, and RF showed better performance with respective accuracies of 0.6332, and 0.6357 respectively. Conversely, NB and KNN exhibited limitations, possibly due to inadequate transaction volumes, which negatively impacted their predictive accuracy. Specifically, KNN struggled with false positives, leading to lower recall and F-measure scores. The primary reason for the low accuracy of the KNN model is the high number of false positive transactions in Class_2. Additionally, the recall and f-measure for this classifier are notably low. It appears that the learning process requires further parameter adjustments or the integration of classifiers with other techniques. Therefore, to explore potential enhancements in model performance, this study conducted additional experiments using various ensemble approaches.

5.2. Experiment 2 [with ensemble approach – bagging]

5.2.1. Implementation details

The bagging technique, known for stabilizing predictions and mitigating overfitting, was applied using the "Bagging Operator" in Rapid Miner. This operator allows for any classifier to be used as a sub-process, thereby creating a robust ensemble by aggregating predictions through a voting mechanism. For this experiment, ten iterations were conducted to establish a series of models, enhancing the reliability of the predictions. Although bagging is typically more advantageous for tree-based algorithms, it was effectively adapted for both tree-based and non-tree-based algorithms in this study.

The models, integrated with the bagging approach, were executed as detailed in Fig. 5. This figure illustrates the setup within Rapid Miner, where each classifier operates as a sub-process under the "Bagging Operator".

5.2.2. Performance evaluation

The outcomes of this experiment are summarized in Table 12, which displays the evaluation metrics for the machine learning algorithms post-bagging implementation. The results indicated varying degrees of performance enhancements across the models. Notably, while KNN

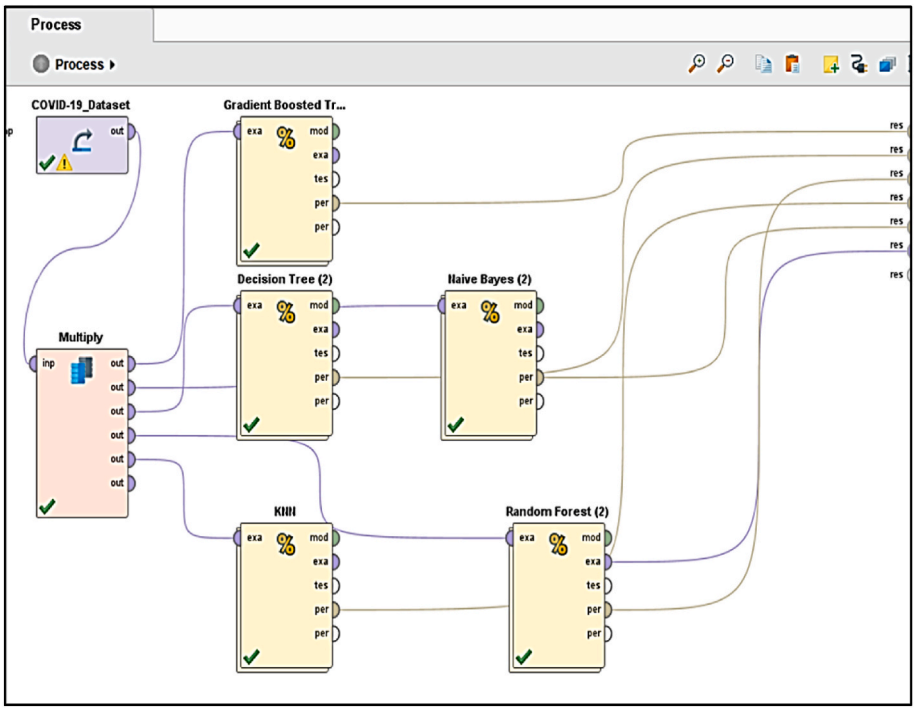

Fig. 4. ML Models Implementation using Rapid Miner.

Table 11
Evaluation metrics for all selected ML Models.

Decision Tree		Actual		Precision	Recall	F-measure	Accuracy
Predicted	Class	1	2	0.6433	0.5878	0.6106	0.6332
	1	177	105				
	2	84	150				
Naïve Bayes							
Predicted	Class	1	2	0.5367	0.7732	0.6299	0.5541
	1	89	58				
	2	172	197				
KNN							
Predicted	Class	1	2	0.6729	0.106	0.1808	0.529
	1	246	228				
	2	15	27				
GBT							
Predicted	Class	1	2	0.6399	0.7371	0.6802	0.6607
	1	153	67				
	2	108	188				
Random Forest							
Predicted	Class	1	2	0.6433	0.6037	0.6168	0.6357
	1	174	101				
	2	87	154				
SVM							
Predicted	Class	1	2	0.683	0.5862	0.6309	0.6531
	1	153	71				
	2	108	184				

maintained a consistent performance (0.529), Decision Tree, SVM and Gradient Boosted Trees showed improvements in accuracy to 0.6377, 0.6783, and 0.6628, respectively. In contrast, Naïve Bayes and Random Forest displayed slight declines in performance compared to Experiment 1. The observed improvements in recall for Decision Tree and Naïve Bayes underscore the efficacy of the bagging approach in enhancing model accuracy.

The experiment aimed to reassess model learning performance using a bagging approach. The results indicated that bagging can effectively enhance learning performance and improve model accuracy.

5.3. Experiment 3 [with ensemble approach – boosting]

5.3.1. Overview of boosting implementation

In this phase of the study, the AdaBoost algorithm was utilized to enhance the ensemble's predictive performance. Employed over 10 iterations, AdaBoost focuses on misclassified instances from previous iterations, adjusting the model progressively to improve accuracy. AdaBoost is a common boosting algorithm, also known as a meta-algorithm in RapidMiner. The implementation scenario is as follows: Cross Validation (10-fold) → AdaBoost → ML Algorithm (each) → Apply Model → Measure Performance. This scenario allowed researchers to measure the performance of each model by integrating boosting with all classifiers. Fig. 6 depicts the implemented model.

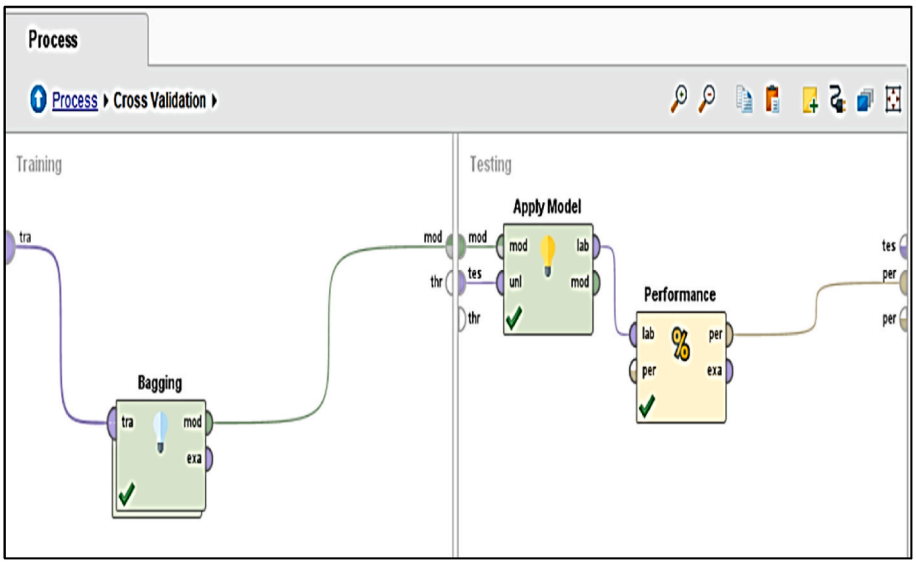

Fig. 5. ML Models Implementation with Bagging using Rapid Miner.

Table 12
Evaluation metrics for all selected ML models with Bagging.

Decision Tree		Actual		Precision	Recall	F-measure	Accuracy
Predicted	Class	1	2	0.6351	0.6466	0.6388	0.6377
	1	164	90				
	2	97	165				
Naïve Bayes							
Predicted	Class	1	2	0.5309	0.8009	0.6354	0.5503
	1	80	51				
	2	181	204				
KNN							
Predicted	Class	1	2	0.6729	0.106	0.1808	0.529
	1	246	228				
	2	15	27				
GBT							
Predicted	Class	1	2	0.6515	0.69	0.6693	0.6628
	1	166	79				
	2	95	176				
Random Forest							
Predicted	Class	1	2	0.619	0.6274	0.6214	0.6222
	1	161	95				
	2	100	160				
SVM							
Predicted	Class	1	2	0.6951	0.6126	0.6513	0.6783
	1	155	68				
	2	98	195				

5.3.2. *Performance analysis using boosting*

Boosting was shown to be particularly effective for tree-based models such as Decision Tree (DT), Gradient Boosting Trees (GBT), and Random Forest (RF). Table 13 displays the improved accuracy metrics obtained through this approach, underscoring the potential of boosting to enhance weaker models significantly, as evidenced by the performance gains in Naïve Bayes (NB). In this study, the boosting experiment proved to be a powerful technique for generating stronger ensembles compared to bagging, as shown in the results in Table 13. Previous studies have also noted the superior performance of boosting over bagging [74]. When compared to single model performance, creating ensembles using boosting resulted in better prediction performance for DT (0.6377), GBT (0.661), and RF (0.6376). In contrast, SVM displayed slight declines in performance compared to bagging ensemble method.

In this experiment, 10 iterations of AdaBoost were used, with the learning process based on the misclassified transactions identified in the previous model/iteration. This approach significantly improved accuracy by reassessing the errors found in previous models. Additionally, boosting increased the number of correct predictions for each class. For example, as shown in Table 13, NB, which had the second lowest prediction ratio in the first experiment, improved its accuracy by 6 % with boosting.

5.4. *Experiment 4 [with ensemble approach – stacking]*

5.4.1. *Stacking Implementation details*

In the culmination of our model development series, the stacking technique was applied to generate ensembles that potentially offer superior predictive performance. Stacking, distinct from bagging and boosting, combines various model types through a meta-learner that learns how to best integrate the predictions of base models. This method is particularly advantageous in settings where diverse approaches need to be synthesized to enhance prediction accuracy.

The implementation involved two primary phases visualized in Figs. 7 and 8. Initially, multiple base learners were trained; their predictions were then used as inputs for a second-level model, the stacking

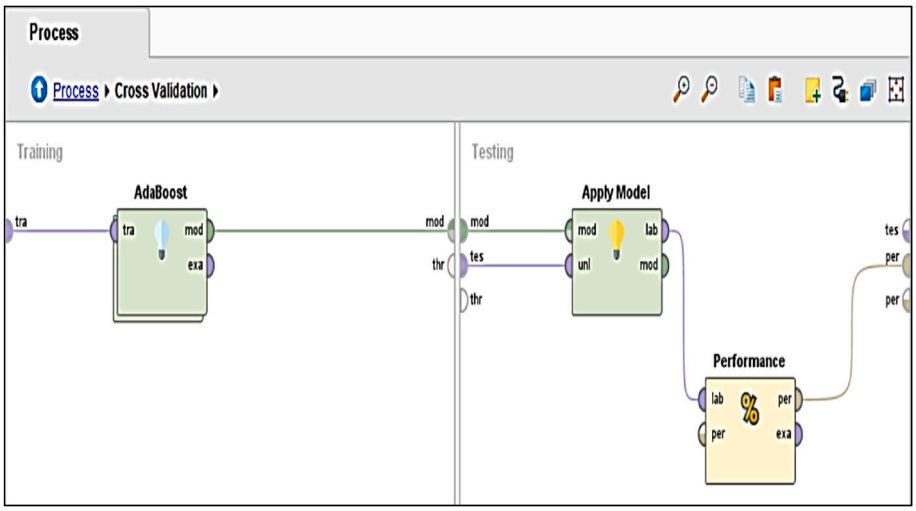

Fig. 6. ML Models Implementation with Boosting using Rapid Miner.

Table 13
Evaluation metrics for all selected ML Models with Boosting.

Decision Tree		Actual		Precision	Recall	F-measure	Accuracy
Predicted	Class	1	2	0.6347	0.6505	0.6399	0.6377
	1	163	89				
	2	98	166				
Naïve Bayes							
Predicted	Class	1	2	0.5898	0.7223	0.6446	0.6105
	1	131	71				
	2	130	184				
KNN							
Predicted	Class	1	2	0.6729	0.106	0.1808	0.529
	1	246	228				
	2	15	27				
GBT							
Predicted	Class	1	2	0.6368	0.7495	0.6871	0.661
	1	150	64				
	2	111	191				
Random Forest							
Predicted	Class	1	2	0.6335	0.6428	0.6369	0.6376
	1	165	91				
	2	96	164				
SVM							
Predicted	Class	1	2	0.6939	0.6415	0.6667	0.6705
	1	170	75				
	2	95	176				

model, to finalize predictions.

This figure displays the stacking operator configured within a cross-validation framework in Rapid Miner, ensuring robust evaluation by simulating multiple training and testing scenarios. The subsequent figure details how the base learners' outputs are combined and processed by the stacking model. This dual-phase approach enhances the robustness of the final predictions by refining them through an additional layer of learning.

5.4.2. Performance outcomes and analysis

The stacking method was evaluated across various configurations of base and stacking learners, as shown in Table 14. This experimentation aimed to identify the most effective combinations for maximizing prediction accuracy, especially in complex diagnostic settings. Table 14 identifies the stacking model learners used in this experiment known as GBT, RF, NB, and KNN. The results show significantly better performance for two algorithms (NB and KNN), but the prediction ratio for GBT decreased compared to bagging and boosting. Finally, multiple base learners (the combination of SVM, DT, and GBT), with RF as a meta learner demonstrated greater accuracy of 70 %.

Overall, stacking ensemble method has improved the performance of different classifiers as identified in previous work [58]. The performance of NB and KNN was outstanding, showing an improvement from their low performance in previous experiments. Specifically, for KNN as a stacking model learner, the recall (0.106), F-measure (0.1808), and accuracy (0.529) improved remarkably by 0.6115, 0.6214, and 0.63, respectively. A possible reason behind this performance is the stacking approach, which combines the performance of base learner models, reducing the bias and variation. Additionally, as a stacking learner model, NB also improved prediction accuracy, achieved 0.622, which is much better than with bagging (0.5503) and as a individual performance of NB (0.5541).

This emphasizes the importance and predictive capabilities of the stacked generalization method, which can facilitate multiple kinds of ML algorithms and provide a better learning process, ultimately enhancing accuracy [57]. Finally, the subsequent section emphasizes the major findings and the overall comparison of all ML experiments conducted in this study with previous work.

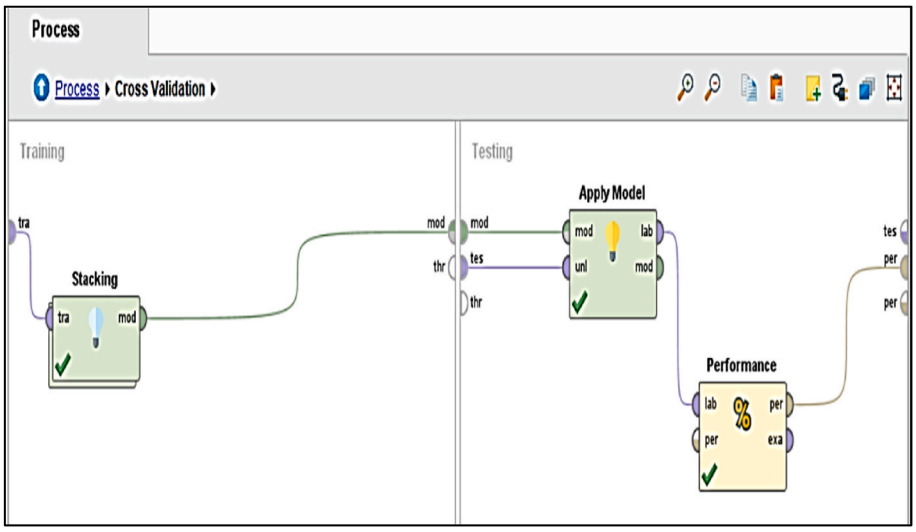

Fig. 7. Initial Setup for Stacking Implementation using Rapid Miner.

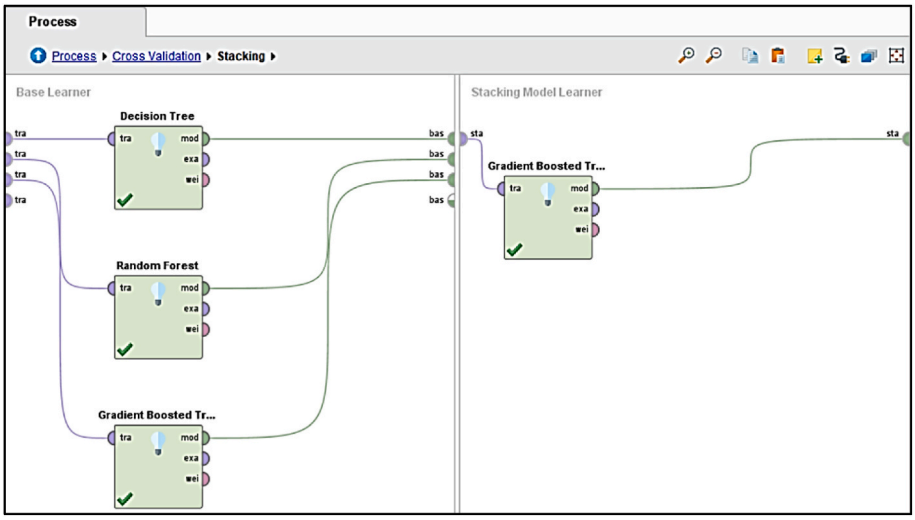

Fig. 8. Sub-processes of stacking [base and stacking model learners].

Table 14
Evaluation metrics for all selected ML Models with Stacking.

Base Learners	Stacking Model		Actual			Precision	Recall	F-measure	Accuracy
(i) DT	GBT	Predicted	Class	1	2	0.6314	0.7335	0.6792	0.6552
(ii) RF			1	151	68				
(iii) GBT			2	110	187				
(i) KNN	NB		Class	1	2	0.6092	0.7185	0.6541	0.622
(ii) RF			1	138	72				
(iii) GBT			2	123	183				
(i) SVM	RF		Class	1	2	0.7126	0.6848	0.6984	0.7054
(ii) DT			1	176	71				
(iii) GBT			2	81	188				
(i) DT	KNN		Class	1	2	0.6372	0.6115	0.6214	0.63
(ii) RF			1	169	99				
(iii) GBT			2	92	156				

5.5. Comparison of machine learning approaches

This study systematically applied a comprehensive machine learning (ML) strategy to develop an intelligent system for predicting the age groups of COVID-19-infected patients. Across four main experimental phases, various ML models and ensemble techniques were evaluated to determine the most effective configurations for accurate predictions.

5.5.1. Overview of experiments

The research was structured into four distinct phases, each utilizing different combinations of ML models and techniques. These phases were designed to explore the potential of ensemble methods in enhancing

prediction accuracy over single-model approaches. Fig. 9 provides a visual comparison of the performance outcomes across these phases, illustrating the advantages of ensemble approaches.

5.5.2. Detailed performance analysis

- Single Model Implementation: Gradient Boosted Trees (GBT) exhibited superior performance among the individual algorithms, achieving an accuracy of 66 %, followed by the SVM (65 %). This demonstrates the applicability of GBT, and SVM in handling complex datasets and predictive tasks.
- Ensemble Techniques - Boosting and Bagging: The boosting technique significantly improved the accuracies of Decision Tree (DT), Naive Bayes (NB), SVM, and Random Forest (RF), with respective accuracies of 0.6377, 0.6105, 0.0.6705, and 0.6376. Bagging also showed strong performance, particularly enhancing DT, SVM and GBT, with SVM reaching an accuracy of 0.6783. These results highlight the effectiveness of ensemble methods in stabilizing and enhancing model predictions.
- Stacking Approach: Stacking provided a notable improvement, and showed the superior overall performance, especially for K-Nearest Neighbors (KNN), which saw the highest accuracy increase of 10 % across all experiments. Naive Bayes, used as a stacking model, also improved substantially, achieving an accuracy of 62 %. Specifically, the implementation of multiple base learners (the combination of SVM, DT, and GBT), with RF as a meta learner demonstrated greater accuracy of 70 %. These outcomes emphasize the stacking approach's capability to leverage multiple model strengths, thereby optimizing the predictive performance.

Comparatively, several models' performances highlighted and confirmed findings from previous studies, which discussed the superior performance of SVM and RF as supervised models [55], and stacking as an ensemble method when applied on medical data [58]. Ensemble methods helped improve disease prediction accuracy by reducing bias and variance. This study utilized multiple variations of ensemble approaches, such as bagging, boosting, and stacking, which demonstrated improvements in prediction accuracy [57]. During the experiments, different statistical features for each class were identified and presented in the dataset section. The aim was to understand how statistical analysis can help elucidate the effect of meta-level ratio features on different ML algorithms [75]. Overall, tree-based machine learning algorithms have demonstrated superior performance both as individual models and when applied with ensemble approaches, as identified in the previous work [73].

- Finally, the study proposed an intelligent system using ML algorithms to predict the age groups of people infected by COVID-19. The implementation of ML algorithms with ensemble approaches has yielded valuable results and can be considered a vital component of the intelligent system. It can offer several benefits to the healthcare industry as follows:

- By finding the association between multiple symptoms with different age groups, healthcare providers can identify treatment and medication effectively based on different age groups.
- Proactive measures can be taken using multiple demographic features, enabling rapid and effective response system.
- Proper resource allocation strategies can be applied in the hospitals and other caring facilities according to the predicted needs of multiple age groups.

Fig. 9. Comparative analysis of ML experiments.

- Further preventive measure can be implemented utilizing trend analysis to identify which types of symptoms can affect which age group.
- Age-specific symptoms presentation is a novel approach presented in this study for developing targeted strategies in future.

6. Conclusion

This study demonstrates that statistical methods and machine learning (ML) algorithms can effectively predict and prioritize COVID-19 symptoms across different age groups. We analyzed two distinct datasets sourced from varied locations to detect common patterns and assess data utility for addressing specific research questions. The analysis revealed that the first dataset was unsuitable for our objectives due to inadequate age categorization. In contrast, the second dataset proved valuable, enabling effective prediction and stratification of COVID-19 symptoms by age. Statistical tests identified key symptoms indicative of different age groups, bolstering our research approach. Subsequently, we applied ML algorithms with an emphasis on ensemble methods to enhance prediction accuracy. While stacking applied with random forest as a meta leaner exhibited the highest accuracy (0.7054), other techniques such as Gradient Boosted Trees with a bagging approach showed notable performance (accuracy of 0.6628) as well. Overall, ensemble approaches helped to improve the model's performance effectively. Notably, K-Nearest Neighbors (KNN) and Naive Bayes (NB) initially recorded lower accuracies of 0.529 and 0.554, respectively. However, these algorithms significantly improved under the stacking model, achieving accuracies of 0.63 and 0.622. This study explored the association of COVID-19 symptoms across different age groups, highlighting the novelty of the proposed methodology. This unique approach can assist healthcare providers in preparing future strategies, such as customized age-specific treatments, vaccine development, and understanding age-dependent symptom profiles. Future research should extend to analyzing multiple datasets including medical images of COVID-19 patients. This will further refine and validate the findings and applicability of proposed idea-. The insights gained from this study could assist medical and public health authorities in developing targeted COVID-19 strategies based on age-specific symptom presentation.

Funding statement

This research work was funded by Institutional Fund Projects under Grant No. (IFPHI-035-611-2020). Therefore, the authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Data availability

All data generated or analyzed during this study are included in this published article.

Ethics approval

This research was reviewed by the ethics committee at King Abdulaziz University, which granted a waiver as the study involved no personal data exposure or high-risk activities. Participants in the online survey provided informed consent, affirming their voluntary participation and the confidentiality of their responses.

CRediT authorship contribution statement

Bahjat Fakieh: Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Farrukh Saleem:** Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization.

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.

References

[1] A. Du Toit, Outbreak of a novel coronavirus, Nat. Rev. Microbiol. 18 (2020) 123.
[2] WHO, Dashboard, Coronavirus disease (COVID-19). https://covid19.who.int/, 2024. (Accessed 31 August 2024).
[3] E. de Lara-Tuprio, C.D.S. Estadilla, J.M.R. Macalalag, T.R. Teng, J. Uyheng, K. E. Espina, C.E. Pulmano, M.R.J.E. Estuar, R.F.R. Sarmiento, Policy-driven Mathematical Modelling for COVID-19 Pandemic Response in the Philippines, Epidemics, 2022 100599.
[4] K. Alshaikh, S. Maasher, A. Bayazed, F. Saleem, S. Badri, B. Fakieh, Impact of COVID-19 on the educational process in Saudi Arabia: a technology–organization–environment framework, Sustainability 13 (2021) 7103.
[5] W. McKibbin, R. Fernando, The Economic Impact of COVID-19, Economics in the Time of COVID-19 45, 2020.
[6] S. Gautam, L. Hens, COVID-19: Impact by and on the Environment, Health and Economy, 2020.
[7] B. McCall, COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread, Lancet Digit Health 2 (2020) e166–e167.
[8] S. Wang, S. Ding, L. Xiong, A new system for surveillance and digital contact tracing for COVID-19: spatiotemporal reporting over network and GPS, JMIR Mhealth Uhealth 8 (2020) e19457.
[9] J. Laguarta, F. Hueto, B. Subirana, COVID-19 artificial intelligence diagnosis using only cough recordings, IEEE Open J Eng Med Biol 1 (2020) 275–281.
[10] F.M. Salman, S.S. Abu-Naser, E. Alajrami, B.S. Abu-Nasser, B.A.M. Alashqar, Covid-19 Detection Using Artificial Intelligence, 2020.
[11] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, NPJ Digit Med 4 (2021) 1–5.
[12] A. Saeed, S. Riaz, Delta Variant Accounts for over 70 Percent Coronavirus Cases in Pakistan — NCOC, Arab News (n.d.).
[13] M.A. Alzubaidi, M. Otoom, N. Otoum, Y. Etoom, R. Banihani, A novel computational method for assigning weights of importance to symptoms of COVID-19 patients, Artif. Intell. Med. 112 (2021) 102018.
[14] M. Otoom, N. Otoum, M.A. Alzubaidi, Y. Etoom, R. Banihani, An IoT-based framework for early identification and monitoring of COVID-19 cases, Biomed. Signal Process Control 62 (2020) 102149.
[15] T. Struyf, J.J. Deeks, J. Dinnes, Y. Takwoingi, C. Davenport, M.M.G. Leeflang, R. Spijker, L. Hooft, D. Emperador, J. Domen, A. Tans, S. Janssens, D. andsymptoms, V. Lannoy, SR A. Horn, A. Van den Bruel, Cochrane COVID-19 Diagnostic Test Accuracy Group. Signs andsymptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19, Cochrane Database Syst. Rev. 5 (2022) CD013665.
[16] A.S. Bhagavathula, P.M. Massey, J. Khubchandani, COVID-19 testing demand amidst Omicron variant surge: mass hysteria or population health need? Brain Behav. Immun. 101 (2022) 394.
[17] O. Attallah, An intelligent ECG-based tool for diagnosing COVID-19 via ensemble deep learning techniques, Biosensors 12 (2022) 299.
[18] M.T. Rahman, S.Z. Idid, Can Zn be a critical element in COVID-19 treatment? Biol. Trace Elem. Res. 199 (2021) 550–558.
[19] M. Makhoul, F. Abu-Hijleh, H.H. Ayoub, S. Seedat, H. Chemaitelly, L.J. Abu-Raddad, Modeling the population-level impact of treatment on COVID-19 disease and SARS-CoV-2 transmission, Epidemics 39 (2022) 100567.
[20] A.L. Skinner-Dorkenoo, A. Sarmal, K.G. Rogbeer, C.J. André, B. Patel, L. Cha, Highlighting COVID-19 racial disparities can reduce support for safety precautions among White US residents, Soc. Sci. Med. 301 (2022) 114951.
[21] J.D. Forrester, A.K. Nassar, P.M. Maggio, M.T. Hawn, Precautions for operating room team members during the COVID-19 pandemic, J. Am. Coll. Surg. 230 (2020) 1098–1101.
[22] F. Trentini, A. Manna, N. Balbo, V. Marziano, G. Guzzetta, S. O'Dell, A.G. Kummer, M. Litvinova, S. Merler, M. Ajelli, Investigating the relationship between interventions, contact patterns, and SARS-CoV-2 transmissibility, Epidemics (2022) 100601.
[23] R. Al-Amer, D. Maneze, B. Everett, J. Montayre, A.R. Villarosa, E. Dwekat, Y. Salamonson, COVID-19 vaccination intention in the first year of the pandemic: a systematic review, J. Clin. Nurs. 31 (2022) 62–86.
[24] F.T. Goh, Y.Z. Chew, C.C. Tam, C.F. Yung, H. Clapham, A country-specific model of COVID-19 vaccination coverage needed for herd immunity in adult only or population wide vaccination programme, Epidemics (2022) 100581.
[25] S.F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A.R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, P.M. Atkinson, Covid-19 Outbreak Prediction with Machine Learning, Available at: SSRN 3580188, 2020.
[26] M.L. Daza-Torres, M.A. Capistrán, A. Capella, J.A. Christen, Bayesian sequential data assimilation for COVID-19 forecasting, Epidemics 39 (2022) 100564.
[27] J. Crawford, J. Cifuentes-Faura, Sustainability in higher education during the COVID-19 pandemic: a systematic review, Sustainability 14 (2022) 1879.
[28] S.Z. Salas-Pilco, Y. Yang, Z. Zhang, Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: a systematic review, Br. J. Educ. Technol. 53 (2022) 593–619.

[29] L. Alschuler, A.M. Chiasson, R. Horwitz, E. Sternberg, R. Crocker, A. Weil, V. Maizes, Integrative medicine considerations for convalescence from mild-to-moderate COVID-19 disease, Explore 18 (2022) 140–148.
[30] D. Paez, M. Mikhail-Lette, G. Gnanasegaran, M. Dondi, E. Estrada-Lobato, J. Bomanji, S. Vinjamuri, N. El-Haj, O. Morozova, O. Alonso, Nuclear medicine departments in the era of COVID-19, in: Semin Nucl Med, Elsevier, 2022, pp. 41–47.
[31] S.X. Zhang, J. Chen, A. Afshar Jahanshahi, A. Alvarez-Risco, H. Dai, J. Li, R. M. Patty-Tito, Succumbing to the COVID-19 pandemic—healthcare workers not satisfied and intend to leave their jobs, Int. J. Ment. Health Addiction 20 (2022) 956–965.
[32] C. Costa, M. Teodoro, C. Mento, F. Giambò, C. Vitello, S. Italia, C. Fenga, Work performance, mood and sleep alterations in home office workers during the COVID-19 pandemic, Int. J. Environ. Res. Publ. Health 19 (2022) 1990.
[33] A.M. Almars, I. Gad, E.-S. Atlam, Applications of AI and IoT in COVID-19 vaccine and its impact on social life, in: Medical Informatics and Bioimaging Using Artificial Intelligence, Springer, 2022, pp. 115–127.
[34] L.O. Gostin, Life after the COVID-19 pandemic, in: JAMA Health Forum, American Medical Association, 2022 e220323.
[35] Y. Miyah, M. Benjelloun, S. Lairini, A. Lahrichi, COVID-19 impact on public health, environment, human psychology, global socioeconomy, and education, Sci. World J. 2022 (2022).
[36] R.I. Sifat, F. Ahmed, M.R.A. Miah, M. Khisa, Effects of COVID-19 on livelihood, health, and psychology of hijra population: insights from Dhaka, Bangladesh, J. Homosex. (2022) 1–17.
[37] S. Uddin, A. Khan, H. Lu, F. Zhou, S. Karim, F. Hajati, M.A. Moni, Road networks and socio-demographic factors to explore COVID-19 infection during its different waves, Sci. Rep. 14 (2024) 1551.
[38] S. Uddin, A. Khan, H. Lu, F. Zhou, S. Karim, Suburban road networks to explore COVID-19 vulnerability and severity, Int. J. Environ. Res. Publ. Health 19 (2022) 2039.
[39] S. Uddin, H. Lu, A. Khan, S. Karim, F. Zhou, Comparing the impact of road networks on COVID-19 severity between delta and omicron variants: a study based on greater Sydney (Australia) suburbs, Int. J. Environ. Res. Publ. Health 19 (2022) 6551.
[40] F. Alsolami, A.S. Al-Malaise Alghamdi, A.I. Khan, Y.B. Abushark, A. Almalawi, F. Saleem, A. Agrawal, R. Kumar, R.A. Khan, A unified decision-making technique for analysing treatments in pandemic context, Comput. Mater. Continua (CMC) (2022) 2591–2618.
[41] WHO, World Health Organization, Coronavirus Disease 2019 (COVID-19): Situation Report, 67, 2020 n.d.
[42] E. Fayyoumi, S. Idwan, H. AboShindi, Machine learning and statistical modelling for prediction of novel covid-19 patients case study: Jordan, Mach. Learn. 11 (2020) 3–11.
[43] M. Nehme, O. Braillard, G. Alcoba, S. Aebischer Perone, D. Courvoisier, F. Chappuis, I. Guessous, COVID-19 symptoms: longitudinal evolution and persistence in outpatient settings, Ann. Intern. Med. 174 (2021) 723–725.
[44] M.M. Ahamad, S. Aktar, M. Rashed-Al-Mahfuz, S. Uddin, P. Liò, H. Xu, M. A. Summers, J.M.W. Quinn, M.A. Moni, A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients, Expert Syst. Appl. 160 (2020) 113661.
[45] H.-Y. Wang, X.-L. Li, Z.-R. Yan, X.-P. Sun, J. Han, B.-W. Zhang, Potential neurological symptoms of COVID-19, Ther Adv Neurol Disord 13 (2020) 1756286420917830.
[46] S.R. Nayak, V. Arora, U. Sinha, R.C. Poonia, A statistical analysis of COVID-19 using Gaussian and probabilistic model, J. Interdiscipl. Math. 24 (2021) 19–32.
[47] M. Yousaf, S. Zahir, M. Riaz, S.M. Hussain, K. Shah, Statistical analysis of forecasting COVID-19 for upcoming month in Pakistan, Chaos, Solit. Fractals 138 (2020) 109926.
[48] Z. Ullah, M. Jamjoom, M. Thirumalaisamy, S.H. Alajmani, F. Saleem, A. Sheikh-Akbari, U.A. Khan, A deep learning based intelligent decision support system for automatic detection of brain tumor, Biomed. Eng. Comput. Biol. 15 (2024) 1–13.
[49] Z. Ullah, N. Alsubaie, M. Jamjoom, S.H. Alajmani, F. Saleem, EffiMob-Net: a deep learning-based hybrid model for detection and identification of tomato diseases using leaf images, Agriculture 13 (2023) 737.
[50] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, F. Kateb, A.M. Ali, B. Shah, Detecting high-risk factors and early diagnosis of diabetes using machine learning methods, Comput. Intell. Neurosci. 2022 (2022).
[51] Y.A. Khan, S.Z. Abbas, B.-C. Truong, Machine learning-based mortality rate prediction using optimized hyper-parameter, Comput. Methods Progr. Biomed. (2020) 105704.
[52] S. Tuli, S. Tuli, R. Tuli, S.S. Gill, Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, Internet of Things (2020) 100222.
[53] B.F. Maier, D. Brockmann, Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China, Science 368 (2020) 742–746, 1979.
[54] R. Sujath, J.M. Chatterjee, A.E. Hassanien, A machine learning forecasting model for COVID-19 pandemic in India, Stoch. Environ. Res. Risk Assess. (2020) 1.
[55] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making 19 (2019) 1–16.
[56] R.H. Al-Mohammdi, Machine Learning Approach for Measuring the Impact of COVID-19 on Distance Education: an Applied Case on Saudi Arabia Universities, (n. d.).
[57] P. Mahajan, S. Uddin, F. Hajati, M.A. Moni, E. Gide, A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets, Health Technol. 14 (2024) 597–613.
[58] P. Mahajan, S. Uddin, F. Hajati, M.A. Moni, Ensemble learning for disease prediction: a review, in: Healthcare, MDPI, 2023, p. 1808.
[59] S. Bolourani, M. Brenner, P. Wang, T. McGinn, J.S. Hirsch, D. Barnaby, T.P. Zanos, N.C.-19 R. Consortium, A machine learning prediction model of respiratory failure within 48 hours of patient admission for COVID-19: model development and validation, J. Med. Internet Res. 23 (2021) e24246.
[60] S. Ghosal, S. Sengupta, M. Majumder, B. Sinha, Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020), Diabetes Metabol. Syndr.: Clin. Res. Rev. 14 (2020) 311–315.
[61] Q. Li, W. Feng, Y.-H. Quan, Trend and forecasting of the COVID-19 outbreak in China, J. Infect. 80 (2020) 469–496.
[62] Y. Zoabi, S. Deri-Rozov, N. Shomron, Corona tested individuals ver 006. https://github.com/nshomron/covidpred/blob/master/data/corona_tested_individuals_ver_006.english.csv.zip, 2021. (Accessed 1 August 2021).
[63] K.B. Prakash, S.S. Imambi, M. Ismail, T.P. Kumar, Y.N. Pawan, Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms, Int. J. 8 (2020).
[64] W.M. Shaban, A.H. Rabie, A.I. Saleh, M.A. Abo-Elsoud, Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy, Pattern Recogn. (2021) 108110.
[65] H. Arslan, H. Arslan, A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier, Engineering Science and Technology, an International Journal 24 (2021) 839–847.
[66] V.A. de F. Barbosa, J.C. Gomes, M.A. de Santana, C.L. de Lima, R.B. Calado, C. R. Bertoldo Júnior, J.E. de A. Albuquerque, R.G. de Souza, R.J.E. de Araújo, L.A. R. Mattos Júnior, Covid-19 rapid test by combining a Random Forest-based web system and blood tests, J. Biomol. Struct. Dyn. (2021) 1–20.
[67] Z. Ullah, F. Saleem, M. Jamjoom, B. Fakieh, Reliable prediction models based on enriched data for identifying the mode of childbirth by using machine learning methods: development study, J. Med. Internet Res. 23 (2021) e28856.
[68] E.A. Amrieh, T. Hamtini, I. Aljarah, Mining educational data to predict student's academic performance using ensemble methods, International Journal of Database Theory and Application 9 (2016) 119–136.
[69] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman and Hall/CRC, 2019.
[70] S. Angra, S. Ahuja, Implementation of data mining algorithms on student's data using rapid miner, in: International Conference on Big Data Analytics and Computational Intelligence (ICBDACI), 2017, pp. 387–391.
[71] P. Tripathi, S.K. Vishwakarma, A. Lala, Sentiment analysis of English tweets using rapid miner, in: Computational Intelligence and Communication Networks (CICN), 2015 International Conference on, IEEE, 2015, pp. 668–672.
[72] F. Saleem, Z. Ullah, B. Fakieh, F. Kateb, Intelligent decision support system for predicting student's E-learning performance using ensemble machine learning, Mathematics 9 (2021) 2078.
[73] S. Uddin, H. Lu, Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data, PLoS One 19 (2024) e0301541.
[74] R. Maclin, D. Opitz, An Empirical Evaluation of Bagging and Boosting, 1997, pp. 546–551. AAAI/IAAI 1997.
[75] S. Uddin, H. Lu, Dataset meta-level and statistical features affect machine learning performance, Sci. Rep. 14 (2024) 1670.

Bahjat Fakieh is working as an associate professor in the information systems department at King Abdulaziz University in Jeddah - Saudi Arabia. Previously, Bahjat was working as a lecturer in Federation University in association with ATMC college campus in Sydney for 1.5 years to teach cloud computing for a postgraduate program. In addition, he was working as an Information Systems tutor in Macquarie University for 4.5 years.

His research interest is data analysis and digital transformation for business and healthcare. In addition, Bahjat is holding several globally recognized professional certificates in IT service management, as well as working as a coach for business development from the technology perspective.

Farrukh Saleem received his Ph.D. Degree in Computer Science from the University of Technology Malaysia in 2017. He is currently working as a Senior Lecturer, School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, UK. Previously, he worked as an Assistant Professor, Department of Information System, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. He has overall 20 years of experience in the education field, specially in teaching higher educational institutes. He has published several journals and conference papers, mainly in ICT evaluation, data mining, Machine Learning, ERP, and IT with business alignment.