Human-Centric Intelligent Systems (2023) 3:441–460
https://doi.org/10.1007/s44230-023-00049-9

RESEARCH ARTICLE

Machine Learning-Based Screening Solution for COVID-19 Cases Investigation: Socio-Demographic and Behavioral Factors Analysis and COVID-19 Detection

K. M. Aslam Uddin[1] · Farida Siddiqi Prity[1] · Maisha Tasnim[1] · Sumiya Nur Jannat[2] · Mohammad Omar Faruk[2] · Jahirul Islam[3] · Saydul Akbar Murad[4] · Apurba Adhikary[1] · Anupam Kumar Bairagi[5]

Received: 20 July 2023 / Accepted: 10 October 2023 / Published online: 28 October 2023

Abstract

The COVID-19 pandemic has unleashed an unprecedented global crisis, releasing a wave of illness, mortality, and economic disarray of unparalleled proportions. Numerous societal and behavioral aspects have conspired to fuel the rampant spread of COVID-19 across the globe. These factors encompass densely populated areas, adherence to mask-wearing protocols, inadequate awareness levels, and various behavioral and social practices. Despite the extensive research surrounding COVID-19 detection, an unfortunate dearth of studies has emerged to meticulously evaluate the intricate interplay between socio-demographic and behavioral factors and the likelihood of COVID-19 infection. Thus, a comprehensive online-based cross-sectional survey was methodically orchestrated, amassing data from a substantial sample size of 500 respondents. The precisely designed survey questionnaire encompassed various variables encompassing socio-demographics, behaviors, and social factors. The Bivariate Pearson's Chi-square association test was deftly employed to unravel the complex associations between the explanatory variables and COVID-19 infection. The feature importance approach was also introduced to discern the utmost critical features underpinning this infectious predicament. Four distinct Machine Learning (ML) algorithms, specifically Decision Tree, Random Forest, CatBoost, and XGBoost, were employed to accurately predict COVID-19 infection based on a comprehensive analysis of socio-demographic and behavioral factors. The performance of these models was rigorously assessed using a range of evaluation metrics, including accuracy, recall, precision, ROC-AUC score, and F1 score. Pearson's Chi-square test revealed a statistically significant association between vaccination status and COVID-19 infection. The use of sanitizer and masks, the timing of infection, and the interval between the first and second vaccine doses were significantly correlated with the likelihood of contracting the COVID-19 virus. Among the ML models tested, the XGBoost classifier demonstrated the highest classification accuracy, achieving an impressive 97.6%. These findings provide valuable insights for individuals, communities, and policymakers to implement targeted strategies aimed at mitigating the impact of the COVID-19 pandemic.

Keywords COVID-19 · Socio-demographic · Behavior · Pearson Chi-square · Machine Learning

✉ Anupam Kumar Bairagi
anupam@cse.ku.ac.bd

[1] Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali 3814, Bangladesh

[2] Department of Statistics, Noakhali Science and Technology University, Noakhali 3814, Bangladesh

[3] Department of Computer Science and Engineering, New Mexico Institute of Mining and Technology, Socorro, USA

[4] School of Computing Sciences and Engineering, University of Southern Mississippi, Hattiesburg, USA

[5] Computing Sciences and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh

1 Introduction

COVID-19, a contagious ailment triggered by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) virus, originated in Wuhan, China, in 2019 and was subsequently classified as a pandemic by the World Health Organization (WHO) [1, 2]. Until April 10, 2023, the global tally of officially confirmed COVID-19 cases had surpassed 684 million, with an unfortunate death toll of 6.8 million [3]. The COVID-19 pandemic has affected many lives and spread to more than 229 countries worldwide [4]. Globally, the pandemic has significantly affected social and economic conditions. Implementing lockdowns and restrictions to mitigate the virus's spread had far-reaching consequences. These measures disrupted daily life and led to social isolation, interruptions in education, and adverse mental health effects. Economically, the pandemic resulted in substantial job losses and posed significant business challenges, leading to widespread company closures. These factors and economic recessions in various regions have underscored the pandemic's enduring and multifaceted impact on societies and economies worldwide [5].

Marginalized populations have been disproportionately impacted by the epidemic, worsening pre-existing inequities. Countries have implemented various measures in interaction with the pandemic, including lockdowns, social distance regulations, and the usage of masks. Some nations have witnessed a spike in cases because of challenges in implementing and enforcing preventive measures, limited healthcare resources, and vaccine hesitancy, while others have been able to prevent the virus's spread more effectively through rigorous public health measures, efficient vaccination campaigns, and better healthcare systems. With over 106 million confirmed cases and 1.16 million deaths, the United States is most adversely afflicted by the ongoing crisis [6]. India, Brazil, and Russia are some more nations that have significantly suffered [7]. Several nations have been endeavoring to immunize their citizens against COVID-19. The world has administered over 13 billion vaccination doses since March 20, 2023 [8]. Nevertheless, there are considerable differences in vaccination rates between various nations. Some nations, like the United States and the United Kingdom, have successfully managed to vaccinate a sizable portion of their communities due to factors such as robust healthcare infrastructure, vaccine availability, and efficient distribution networks, whereas others, especially low-income nations, have had difficulty accessing vaccines primarily because of limited resources, logistical challenges, and global vaccine supply disparities.

The severity of the outbreak is intricately tied to behavioral and social factors. The transmission dynamics of COVID-19 can be significantly influenced by many socio-economic factors, including population density, inter-provincial mobility, cultural norms, lack of awareness, poverty, limited healthcare infrastructure, and the implementation of national lockdown measures [9]. Furthermore, certain behavioral factors have been identified as potential contributors to the spread of the virus, such as fear of the virus, inadequate physical activity, smoking, poor health status, urban residency, and suboptimal health conditions [10]. The lack of understanding regarding the significance of social distancing, hand hygiene, and mask usage among individuals has resulted in an elevated risk of infection. The dissemination of misinformation and rumors about the virus has further perpetuated public ignorance and unawareness [11]. Some studies have also explored the impact of individuals' socio-economic status, family dynamics, sense of community, and perception of the stringency of lockdown measures, all of which have proven to be significant factors influencing the spread of COVID-19 [12]. A recent cohort study in the United Kingdom revealed that individuals between the ages of 40 and 69 were more susceptible to hospitalization due to COVID-19, primarily due to smoking, excessive alcohol consumption, and physical inactivity [10]. The likelihood of hospitalization escalates with the number of behaviors exhibited, as these factors have demonstrated a correlation with unfavorable COVID-19 outcomes [13]. The COVID-19 pandemic necessitates the adoption of socio-demographic and behavioral factors that can be scrutinized and analyzed to mitigate the risk of infection. Moreover, healthcare institutions face significant challenges in delivering high-quality services at affordable rates. Accurate patient diagnosis and effective disease management constitute vital elements of superior healthcare provision. Inadequate clinical judgment can lead to catastrophic consequences, an outcome deemed unacceptable. Within the medical realm, establishing a precise diagnosis proves to be a formidable and time-consuming endeavor. Relying solely on human intelligence proves inadequate for achieving accurate diagnoses. Numerous obstacles arise during the diagnostic process, including decreased accuracy of results, limited experience, performance constraints tied to time sensitivity, and difficulties keeping pace with evolving knowledge and advancements.

The healthcare industry is experiencing significant transformations due to the widespread implementation of Machine Learning [14–16]. Machine Learning, a subset of Artificial Intelligence (AI), strives to improve the effectiveness and precision of medical procedures. This technological advancement offers significant potential, especially in regions struggling with overburdened healthcare systems and a shortage of healthcare professionals. ML plays a crucial role in medicine by uncovering patterns within extensive

datasets and identifying diagnostic indicators associated with risks or diseases. ML techniques have demonstrated remarkable performance in various medical applications, including medical image analysis [17–20], language processing [21, 22], and the detection of tumors or cancer cells [23, 24], thereby supporting clinical management and assisting specialists. ML-based approaches have recently been utilized in the investigation of COVID-19 disease.

1.1 Previous Study

Several research studies have employed Machine Learning techniques to investigate demographic factors' impact on COVID-19 infection and vaccine efficacy. In one particular study [25], the authors conducted a comparative analysis of ML and soft computing models to forecast the outbreak of COVID-19. The outputs of two ML models, namely Multilayer Perceptron (MLP) neural network and Adaptive Neuro-Fuzzy Inference Systems (ANFIS), demonstrated robust generalizability in long-term forecasting. Given the complex nature of the COVID-19 outbreak and regional variations, the authors proposed ML as a valuable approach for simulating the time series of the outbreak. Another research conducted by Han, Y., Huang, et al. [26] focused on the occurrence and severity of COVID-19. They meticulously filtered out irrelevant components from a pool of 113 variables and employed four ML-based classification and regression models to predict the occurrence and intensity of COVID-19. The impact of each significant factor was carefully evaluated. The optimal regression model achieved an impressive R2 value of 0.778 for predicting intensity, and their ideal classification model exhibited an accuracy of 91.91% in predicting COVID-19 incidence. The study revealed that the likelihood of COVID-19 increased with extreme weather conditions, high minimum relative humidity, and affluent communities near the epicenter and at higher elevations. Suthar et al. [27] noted the impact of vaccination coverage on COVID-19-related mortality and morbidity in counties encompassing approximately 300 million individuals, accounting for about 80% of the total US population. An autoregressive correlation structure accommodates a substantial amount of data per county. The findings revealed that a 10% increase in vaccination coverage corresponded to an 8% (with a 95% confidence interval ranging from 8 to 9%) reduction in fatality rates and a 7% (6% to 8%) decrease in disease incidence. Notably, higher vaccination coverage levels were associated with lower death and incidence rates, particularly in the presence of the alpha and delta variants.

Deep learning has emerged as a transformative technology in COVID-19 detection. Zhang and colleagues [28] employed a SqueezeNet (SN) model equipped with a complex bypass mechanism for analyzing CT images related to COVID-19. They utilized the extreme learning machine (ELM) as the classifier. The SNELM model achieved a remarkable accuracy rate of 96.14% with a narrow margin of error at 0.96% in detecting COVID-19. In another study, Zhang et al. [29] proposed a novel innovative diagnosis model called the 7L-CNN-CD (Seven-Layer Convolutional Neural Network for COVID-19 Diagnosis). Their approach, consisting of seven layers, exhibited remarkable performance, boasting a sensitivity of $94.44\% \pm 0.73\%$, a specificity of $93.63\% \pm 1.60\%$, and an accuracy rate of $94.03\% \pm 0.80\%$. Wang et al. [30] introduced an innovative approach for COVID-19 detection, presenting the Attention-Based VGG-Style Network for COVID-19 (AVNC). This specialized network, comprised of 12 layers, demonstrated remarkable performance, with sensitivity, precision, and F1 scores for each class consistently exceeding 95%.

Another investigation conducted by researchers [31] delved into the potential relationship between SARS-CoV-2 infection/COVID-19 and blood type, utilizing Electronic Health Record (EHR) data obtained from the New York Presbyterian/Columbia University Irving Medical Center (NYP/CUIMC) hospital in New York City, USA. The study examined the severity of the initial infection and two critical outcomes of COVID-19, namely, intubation and death, while also exploring the association with ABO and Rh(D) blood types. The researchers employed Chi-squared independence tests to scrutinize the variation in blood-group frequencies between individuals who underwent SARS-CoV-2 testing and those who did not. However, the analysis yielded insufficient evidence to conclude that there were discernible differences in blood-group frequencies between the two groups. Viswanath et al. were the pioneers in identifying the intricate interplay of individual, communicative, and social factors influencing vaccination uptake, as elucidated in their seminal research [32]. The survey was conducted before the widespread availability of COVID-19 vaccines, and the research team adopted a comprehensive approach drawing upon data from a nationwide online probability-based panel comprising 1012 representative adults from diverse backgrounds across the United States. As for the independent variables, the study encompassed a multifaceted array of factors, including political party affiliation, trust in scientists, exposure to COVID-19-related news across different media platforms, risk perceptions, and social determinants of health. Various variables were underscored as potential influences on COVID-19 vaccination adoption, encompassing race and ethnicity, risk perceptions, exposure to COVID-19 news through diverse media channels, party identification, and trust in scientists. In a parallel endeavor, spanning the period between January and April 2021, a meticulously executed cross-sectional survey study was carried out to gather comprehensive insights into the impact of COVID-19 vaccination on individuals residing in the United Arab Emirates [33]. The study analyzed demographic information,

vaccination rates, and the perspectives of those who resisted COVID-19 vaccination. The analysis incorporated variables such as age, educational attainment, marital status, employment status, nationality, and region of residence, thereby capturing a holistic view of the study participants. Moreover, an in-depth exploration of vaccination hesitancy was undertaken, coupled with an assessment of participants' health status, ongoing medical concerns, past COVID-19-related history, vaccination status, previous infection, number of vaccine doses administered, and pertinent medical records associated with COVID-19. The collected data was meticulously analyzed and presented in frequency and percentage distributions, facilitating a comprehensive understanding of the intricate dynamics.

1.2 Knowledge Gap

Previous investigations [28–30] have predominantly focused on the development of classifiers capable of accurately detecting instances of COVID-19 infection. However, a considerable gap persists in the literature, as studies have yet to comprehensively explore the entirety of patient data, aiming to identify the key variables indispensable for the precise prediction of COVID-19 disease. Research conducted in related domains has demonstrated that the strategic selection of these critical features holds immense potential for empowering data mining professionals operating within the healthcare industry, enabling them to unravel the intricate interdependencies among recorded risk factors within a given dataset and discern the individual impact of each on the accuracy of COVID-19 disease prediction. Furthermore, more attention should be paid to demographic variables within the purview of previous studies [26, 27, 31–34], including the employment of face masks, utilization of hand sanitizers, or other potential confounding factors. Most investigations have primarily relied on data sourced from a solitary medical center, thereby inadvertently limiting the external validity of the prediction models. This limitation arises from the inherent variations in COVID-19 dynamics across diverse countries and populations, warranting a more expansive and diverse dataset for robust generalization and broader applicability of the developed prediction models. Therefore, future research endeavors should prioritize the comprehensive analysis of patient data, encompassing various essential variables, and extend the scope to encompass diverse populations and regions to enhance the accuracy and reliability of COVID-19 disease prediction models.

1.3 Contribution of This Study

This study aims to comprehensively explain the numerous risk variables for COVID-19 prediction by exploring the social and behavioral factors associated with the spread of COVID-19. Many risk factors in patient records are examined to determine the most crucial elements required for COVID-19 disease prediction. Different socio-demographic and behavioral factors such as area, gender, age, profession, family members, birth division, vaccination division, vaccinated status, vaccine name, vaccine dose, the mask uses, type of mask, sanitizer uses, side effects duration, the severity level of coronavirus, taste disturbance, fever, headache, fatigue, muscle pain, pain at the side of the injection, no symptoms, and time of infection have been collected from eight divisions of Bangladesh to analyze the infection of COVID-19. Statistical analyses are conducted to find the association between the target variable and other factors. This research has used different classification techniques of ML such as Decision Tree, Random Forest, CatBoost, and XGBoost to determine the prediction of COVID-19 infection in public health based on these socio-demographic variables and vaccine-related information. This study also compared the performance of different classification techniques and found the best one. Machine Learning-based feature importance has been applied to find the significant factors responsible for COVID-19 infection. The proposed system can be used to control the pandemic effectively by managing individual actions and broader social influences. Through a better understanding of socio-demographic and behavioral factors, public health interventions can be developed to reduce the spread of the virus, protecting people worldwide.

2 Material and Methods

The prediction of COVID-19 infections has been formulated using Machine Learning algorithms, harnessed to train and test datasets to envision future values effectively. The intricate processes and diverse methodologies entwined within this endeavor have been meticulously delineated in Fig. 1, encapsulating the comprehensive workflow. The inception of this investigation entails the meticulous accumulation of patient-specific information, which subsequently undergoes a scrupulous standardization process, paving the way for rigorous statistical analyses. During the training and testing phases, the normalized data is deployed, serving as the bedrock for deriving reliable and robust prediction models for disease prognosis. The final selection of the disease prediction model is predicated upon methodically comparing and examining the training and testing outputs, ensuring their alignment with pre-established satisfaction thresholds. Furthermore, an insightful exploration of feature importance has been precisely conducted, enabling the identification and elucidation of the pivotal factors that significantly influence

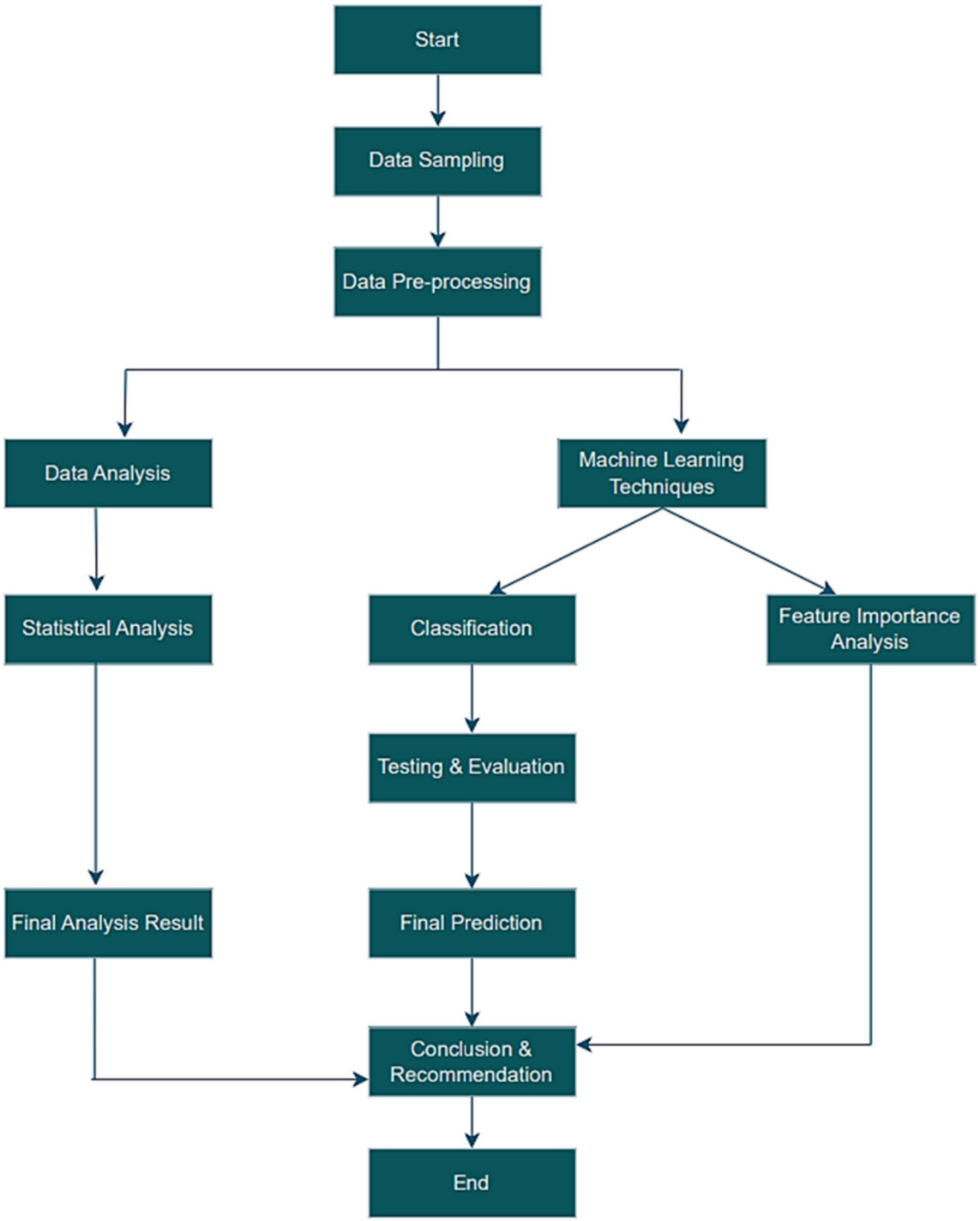

Fig. 1 Workflow diagram of the proposed model

COVID-19 infection. This facet of the research elucidates the intrinsic interplay and relative significance of various factors, casting light on their impact on COVID-19 infection dynamics.

2.1 Data Collection and Processing

Data collection and analysis is the initial step in the model development process. Below, we describe the specifics of data collection and processing.

2.1.1 Survey Design and Participants

This nationwide cross-sectional questionnaire survey was conducted among Bangladeshi citizens from October 20, 2020, to January 25, 2020. Our study included all Bangladeshi citizens aged 16 years and above, with the only exclusion criterion being individuals under 16. The study employed a meticulously designed questionnaire to collect valuable data. Before the extensive survey, a pilot examination was conducted with 25 participants to pinpoint and address any potential anomalies in the questionnaire. Due to the formidable challenges posed by the COVID-19 pandemic, particularly the constraints on in-person interactions, data collection was facilitated through the Google form distributed via popular social networking platforms such as Facebook and Messenger. This digital approach ensured the safety and convenience of participants while adhering to pandemic-related restrictions.

Five hundred observations were meticulously collected and prepared for the final analysis, ensuring a robust dataset for the study's objectives. The research protocol underwent rigorous scrutiny and gained approval from the Ethical Review Committee of Noakhali Science and Technology University. Participants' rights and privacy were upheld, with individuals providing informed consent online before participating in the survey. The consent form explicitly outlined the study's objectives, assured confidentiality, emphasized participants' rights, and reaffirmed their freedom to withdraw from the study at any point. For participants under 18, parental or guardian consent was a prerequisite. This approach allowed for an inclusive dataset, enabling a more thorough analysis of the factors related to COVID-19 knowledge, preventive behaviors, and other relevant aspects among the Bangladeshi population.

2.1.2 Questionnaire Design

A questionnaire was developed with 17 questions. It explored social-demographic characteristics (age, gender, area, education, source of income, and family size), knowledge, and preventive behaviors related to COVID-19, such as prevention and treatment measures, practices related to handwashing, wearing a face mask, using hand sanitizers, intention to receive COVID-19 vaccine. The current survey also assessed the prevalence of COVID-19 prevention practices among participants. The construction of these questions was informed by a thorough literature review encompassing various research reports and scholarly publications [34–36]. By collaboratively analyzing these papers, we successfully identified pertinent questions employed in previous studies. The rationale behind selecting these questions was their efficacy in assessing critical factors relevant to the study objectives. These questions have previously shown their utility in capturing essential data related to participants' social-demographic profiles, knowledge regarding COVID-19, and their practices and intentions concerning preventive measures. The questionnaire was designed in English and was pre-tested to ensure its quality and suitability with the statistical analysis. The questionnaire's multiple-choice and yes/no response format makes it easy to understand the question.

2.1.3 Data Pre-Processing

This research study has used county-level surveillance data. A total of 500 data was collected online, and the data details are included in Table 1. Data must be normalized because of the apparent discrepancies in the variable ranges. According to the Eq. (1), normalization was achieved as follows:

$$\text{Xnormalized} = \frac{\text{x} - \min(\text{x})}{\max(\text{x}) - \min(\text{x})} \tag{1}$$

where, Xnormalized is the updated normalized value. Each feature's lowest value is assumed to be "0," its highest value to be "1," and all other values are transformed to an integer between "0" and "1."

2.2 Statistical Analysis

The primary data collected for the study was modified for statistical analysis. Due to the qualitative and categorical nature of the data, Non-parametric tests were used throughout the study. The demographic and socio-economic factors frequency distribution table was developed to understand the respondents' backgrounds better. This study used Pearson's Chi-square association test between responses and predictor variables. The utilization of a Chi-square test to explore the relationship between socio-economic, demographic, and behavioral factors and COVID-19 infection was motivated by existing literature and research findings. Numerous prior studies [31, 37] have indicated that these factors can significantly affect an individual's susceptibility to COVID-19 and adherence to preventive measures. By employing the Chi-square test, we aimed to empirically assess the statistical associations between these variables and COVID-19 infection. This approach allows us to identify our dataset's potential patterns, correlations, or significant differences. Furthermore, the Chi-square test is well-suited for analyzing qualitative and categorical data, making it an appropriate choice for this study's research objectives. SPSS (version 25.0) was used to analyze the data.

2.3 Machine Learning Techniques

This study utilizes four distinct Machine Learning algorithms to analyze the dataset. The input parameters are preprocessed to ensure data quality and then fed into the ML

Table 1 Dataset description

Name of the variables	Categories and their assigned code
Area	0 = "Urban" 1 = "Rural"
Gender	0 = "Female" 1 = "Male"
Age	0 = "16–20" 1 = "21–25" 2 = "26–30" 3 = "Above 30 years"
Professions	0 = "Government job holder" 1 = "Student" 2 = "Others"
Family members	0 = "2–3 members" 1 = "4–6 members" 2 = "Above 6 members"
Birth division	0 = "Dhaka" 1 = "Chattagram" 8 = "Others"
Vaccination division	0 = "Dhaka" 1 = "Chattagram" 8 = "Others"
Vaccinated or not?	0 = "No" 1 = "Yes"
Vaccine name	0 = "Pfizer" 1 = "Moderna" 2 = "Covishield" 3 = "Sinopharm"
Vaccine dose	0 = One 1 = Two 3 = Three 4 = None
Mask uses	0 = "No" 1 = "Yes" 2 = "Sometimes"
Type of mask	0 = "Surgical" 1 = "Cotton" 2 = "Others"
Sanitizer uses	0 = "No" 1 = "Yes" 2 = "Sometimes"
Side effects duration	1 = "1 day" 2 = "2 days" 3 = "3–7 days"
Severity level of coronavirus	0 = "Moderate" 1 = "High" 2 = "Mild" 3 = "No symptom"
Taste disturbance	0 = "No" 1 = "Yes"
Fever	0 = "No" 1 = "Yes"
Headache	0 = "No" 1 = "Yes"
Fatigue	0 = "No" 1 = "Yes"
Muscle pain	0 = "No" 1 = "Yes"
Pain at the side of the injection	0 = "No" 1 = "Yes"
No symptoms	0 = "No" 1 = "Yes"
Time of infection	0 = "Not infected" 1 = "After vaccination" 2 = "Before Vaccination"
Target	0 = "Not infected" 1 = "Infected"

models. These models encompass a variety of computational techniques, working together to uncover meaningful patterns within the dataset. The algorithms are:

(a) Random Forest
(b) Decision Tree
(c) CatBoost
(d) XGBoost

(a) *Random Forest* The Random Forest classifier, renowned for its robust prediction capabilities in medical fields, is employed to discern the most trustworthy predictors. It entails amalgamating many tree classifiers, each constructed utilizing a distinct random vector sampled from the input vector, enabling the classification of an input vector with enhanced accuracy. The utilization of RF in this study exemplifies its efficacy in extracting valuable insights and identifying critical predictors within complex medical datasets [38, 39].

(b) *Decision Tree* As a non-parametric supervised learning method, Decision Trees offer a valuable approach for classification and regression tasks. Its main objective is to acquire intuitive decision rules from the data's features, enabling the construction of a model that accurately predicts the target variable's value. Notably, the Decision Tree approach has proven helpful in forecasting stroke outcomes, diagnosing cardiac arrhythmias, detecting cancer early, and managing chronic disease [40–45].

(c) *CatBoost* CatBoost employs a highly effective strategy that mitigates overfitting, enabling the utilization of the entire dataset for training purposes. The application of CatBoost extends notably to the health domain, where it excels in predictive modeling for various diseases, including brain, asthma, prostate, and breast cancer [46–48]. Its impressive performance and versatility in healthcare settings make CatBoost a valuable tool for accurate disease prediction and improved patient outcomes.

Table 2 Proportions of data in the dataset

Data set	Proportion
Training set	75%
Test set	25%

(d) *XGBoost* The implementation of Gradient Boosted Decision Trees is facilitated by utilizing the XGBoost framework. Notably, within XGBoost, weights play a crucial role in the modeling process. Before inputting into the Decision Tree, each independent variable is assigned a weight, which influences the prediction of outcomes. Subsequently, the second Decision Tree incorporates these variables, emphasizing increasing the weights of those variables that were inaccurately predicted by the preceding tree. This iterative approach has found widespread application in disease prediction research, yielding significant advancements in the field [49–53].

2.3.1 Training Set and Testing Set

This section discussed the percentage of training and testing data sets. Out of 500 data, there are 75% of the training data and 25% of the test data, as illustrated in Table 2.

2.3.2 Performance Evaluation

This study evaluated the efficacy and applicability of various ML classification algorithms in predicting COVID-19 infections through model development. The model's performance was evaluated using a confusion matrix and a comprehensive set of metrics, including True Positives, True Negatives, False Positives, False Negatives, precision, recall, F1-score, accuracy, and other relevant measures. The rationale behind employing this evaluation methodology is its ability to provide a detailed and rigorous assessment of the model's predictive capabilities in many literature reviews [38–53]. It offers insights into how well the model performs across different classification aspects.

Confusion matrix The assessment of the model's effectiveness and accuracy entails the utilization of a confusion matrix, a method of analysis that offers simplicity and clarity. The confusion matrix is a two-dimensional table characterized by its "Actual class" and "Predicted class" dimensions. Within this framework, the rows represent the COVID-19 illness classifications as they genuinely exist, while the columns depict the corresponding classifications projected by the model. Specifically, the dataset comprises classes denoted as Class 0 and Class 1. A confusion matrix designed for COVID-19 infection detection can be seen in Table 3 below:

Table 3 Confusion matrix

Actual	Predicted	
	Negative (Class 0)	Positive (Class 1)
Negative (Class 0)	TN	FP
Positive (Class 1)	FN	TP

True Positives (TP) These accurately predicted positive values indicate that both the actual and predicted classes are affirmative.

True Negatives (TN) These accurately predicted negative values indicate that both the actual and predicted classes are negative.

False Positives (FP) When the actual class is negative, and the predicted class is positive.

False Negatives (FN) When the actual class indicates a positive outcome, the predicted class indicates a negative outcome.

Accuracy Accuracy is determined by dividing the count of accurate predictions for COVID-19 disease by the total size of the dataset.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Precision It indicates the proportion of positive class predictions that correspond to actual COVID-19 disease positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

Recall Recall pertains to the capability of a test to identify an individual with COVID-19 disease as positive correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F1-score The harmonic mean of precision and recall is called F1-score.

$$\text{F1} - \text{score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{5}$$

The rationale behind this evaluation methodology is to ensure a thorough and transparent assessment of our model's predictive accuracy and ability to identify COVID-19 cases correctly. This rigorous evaluation approach is integral to the reliability and credibility of our study's results, enabling us to draw meaningful conclusions regarding the model's performance.

2.4 Feature Importance Analysis

The concept of feature importance encompasses techniques that assign scores to each input feature within a specific model, indicating the relative significance of each feature. A higher score indicates that a particular characteristic will influence the model's ability to predict a specific variable. In this study, Machine Learning-based feature importance analysis was conducted on the dataset to identify the most influential features to the prevalence of COVID-19 infection.

3 Result and Discussion

The COVID-19 pandemic has profoundly impacted human social and behavioral aspects, resulting in a devastating loss of life. The primary objective of this research is to identify the factors contributing to the spread of COVID-19 infection. To achieve this, various Machine Learning-based methodologies have been employed to predict and forecast COVID-19 infection rates. The effectiveness of these approaches has been thoroughly evaluated, and a comprehensive comparison between them has been conducted. Additionally, the study investigates the significance of individual features in detecting COVID-19 by assessing their respective importance. Before the performance evaluation of different algorithms, a detailed statistical analysis of the dataset was performed, yielding insightful results in the subsequent section.

3.1 Statistical Analysis

Statistical analysis has been employed on the dataset to find the association between the target variable and other features.

3.1.1 Socio-Demographic and Behavioral Characteristics

Five hundred participants completed the survey, accepting the request to participate. The demographic characteristics of all the participants are presented in Table 4. Among the respondents, a maximum was male (n = 279, 55.8%), and the rest were female (n = 221, 44.2%), respectively. More than 50% of the participants were between the ages of 21 and 25 (n = 296, 59.2%). The study discloses that more than 95% of respondents were vaccinated (n = 487, 97.4%). Since nearly 98% of respondents took the vaccine, most were not infected with the virus (n = 381, 76.2%). The survey included the highest number of students (n = 382, 76.4%), Govt. jobs (n = 27, 5.4%), and others (n = 91, 18.2%). Most respondents were from urban (n = 353, 70.6%). The majority came from a family of 4 to 6 members (n = 388, 77.6%). Most respondents (n = 308, 61.6%) live in the Chattogram division.

3.1.2 Knowledge of Transmission of COVID-19 Infection

The findings from Table 5 showed that a maximum of the respondents had correct knowledge of COVID-19 transmission since (n = 320, 64%) use sanitizer and the majority of respondents (n = 308, 61.6%) in the Chattogram division had taken the vaccine to minimize the risk of infection. The highest number of respondents (n = 320, 64% %) reported that they had taken the vaccine twice, and the maximum number of respondents took the Sinopharm vaccine (n = 255, 51%). The rest of the respondents took Phizer (n = 110, 22%), Moderna (n = 58, 11.6%), and Covishield (n = 77, 15.4%). The study found that most respondents (n = 352, 70.4%) wore surgical masks. Cotton masks are worn by (n = 106, 21.2%) of respondents, whereas other forms of masks are used by (n = 42, 8.4%) respectively. Table 5 illustrates that the highest number of respondents who took the vaccine for the COVID-19 virus had experienced adverse effects for one day (n = 157, 31.4%). 50% of the respondents waited one month (n = 256, 51.2%) between the first and second doses of the immunization. Following the study, most respondents delayed the second and third immunization doses over three months (n = 422, 84.4%).

Table 5 indicates the associations between social and behavioral factors with COVID-19 infection. Pearson Chi-square tests were used to explore the relationship between factors associated with knowledge of COVID-19. Maximum respondents aged 21 to 25 (n = 231, 78.04%) reported not being infected with the virus. Age is statistically significant at a 5% significance level since ($\chi^2 = 12.832$, $p < 0.05$) infected by COVID-19 infection in Bangladesh. A significant association ($p < 0.05$) was found between 'Vaccinated or not' and "Infected by covid-19 virus". As in this case ($\chi^2 = 4.168$, $p < 0.05$). The study also discloses that the Maximum number of respondents used sanitizer (n = 232, 72.50%) and weren't infected with the COVID-19 virus. The use of sanitizer ($p < 0.05$) was also significantly associated with those who are infected with the COVID-19 virus ($\chi^2 = 7.284$, $p < 0.05$). That means the use of sanitizer impacts the spread of COVID-19 infection. Remarkable association ($\chi^2 = 440.65$, $p < 0.05$) between "Time of infection" and "Infected by COVID-19 virus". So, there is a significant relation between the two variables. Maximum respondents who had taken two vaccine doses were infected by the virus (n = 69, 21.56%) and not infected with the virus (n = 251, 78.43%). These findings showed that most respondents who had taken the vaccine were not affected much by COVID-19 infection. So, there is a significant relationship between vaccine doses and infection with the virus ($\chi^2 = 11.845$, $p < 0.05$). Regarding the profession, most respondents were

Table 4 Frequency distributions of socio-demographic and behavioral characteristics of respondents

Variable name	Categories	Frequency	Percent (%)
Gender	Female	221	44.2
	Male	279	55.8
Age	16–20	75	15
	21–25	296	59.2
	25–30	38	7.6
	> 30	91	18.2
Infected_with_virus	No	381	76.2
	Yes	119	23.8
Vaccinated_Or_not	No	13	2.6
	Yes	487	97.4
Sanitizer_Uses	No	26	5.2
	Yes	320	64
	Sometimes	154	30.8
Time_of_Infection	Not infected	371	74.2
	After vaccination	31	6.2
	Before Vaccination	98	19.6
Area	Rural	147	29.4
	Urban	353	70.6
Vaccine_Name	Phizer	110	22
	Moderna	58	11.6
	Covishield	77	15.4
	Sinopharm	255	51
Vaccine_Doses	1st dose	31	6.2
	2nd dose	320	64
	3rd dose	149	29.8
Profession	Govt_job	27	5.4
	Students	382	76.4
	Others	91	18.2
Family	2–3 members	42	8.4
	4–6 members	388	77.6
	6+ members	70	14
Birth division	Dhaka	77	15.4
	Chattogram	308	61.6
	Others	115	23
Vaccination division	Dhaka	106	21.2
	Chattogram	308	61.6
	Others	86	17.2
Mask	Surgical	352	70.4
	Cotton	106	21.2
	Others	42	8.4
Side effect time duration	1 day	157	31.4
	2 Days	142	28.4
	3 Days	85	17
	More than 3 days	116	23.2
Time_gap_1st and_2nd dose	1 month gap	256	51.2
	2 months gap	125	25
	3 months gap	54	10.8
	More than 3 months gap	65	13

Table 4 (continued)

Variable name	Categories	Frequency	Percent (%)
2nd and 3rd dose gap	1 month gap	23	4.6
	2 months gap	25	5
	3 months gap	30	6
	More than 3 months	422	84.4

students and infected with the virus (n = 74, 19.37%) while not infected with the virus (n = 308, 80.63%). The variable "Profession" is highly correlated ($p < 0.05$) with "Infected by COVID-19 virus" ($\chi^2 = 18.977$, $p < 0.05$). Respondents who took the vaccine from Dhaka were infected with the virus (n = 32, 30.18%) and not contaminated with the virus (n = 74, 69.8%). The majority of the respondents took the vaccine from the Chattogram division. People who had taken the vaccine from Chattogram were infected with the COVID-19 virus (n = 57, 18.50%) and not infected with the virus (n = 251, 81.49%). From this relation, it is seen that there is a strong association between vaccination place and COVID-19 infection ($\chi^2 = 12.97$, $p < 0.05$). The study revealed that the people who wear Surgical masks and are infected with the COVID-19 virus (n = 94, 26.80%), Cotton masks (n = 22, 20.20%), and other masks (n = 3, 7.50%).

Table 5 shows that most people used a surgical mask and were not infected with the virus (n = 258, 73.2%). Respondents who used cotton masks were not infected with the virus (n = 84, 79.8%). There, we found a remarkable association between "Type of Mask" and "Infected by COVID-19 virus" ($\chi^2 = 8.605$, $p < 0.05$). The time interval between the first and second dose is one month for a maximum number of respondents infected with the COVID-19 virus (n = 60, 23.43%). "Time Gap between First and Second Dosage" is significantly associated with "Infected with COVID-19 virus" ($\chi^2 = 8.835$, $p < 0.05$).

3.2 COVID-19 Prediction Using Distinct ML Techniques

This section presented the performance analysis of each algorithm for COVID-19 infection prediction.

Figure 2 illustrates the formation of a confusion matrix resulting from applying Machine Learning techniques to the collected dataset for training and testing purposes. Within the confusion matrix, the intersection of purple and grey cells signifies a match between the output and the target, while a black cell indicates a mismatch. Upon examining the confusion matrix generated by the Random Forest model, it becomes evident that there are 93 True Positive values, indicating accurate predictions of the negative class, and 28 True Negative values. The Random Forest classifier correctly predicts 93 instances of the negative class while making two incorrect predictions. Furthermore, it accurately identifies 28 instances of the positive class. There are 121 instances of correct predictions and 4 instances of incorrect predictions.

In the confusion matrix generated by the Decision Tree model, it can be observed that there are 93 instances classified as True Positives and 27 instances classified as True Negatives. Figure 2 visually represents the accurate prediction of 93 samples for the negative class by the Decision Tree, while two samples are mispredicted. Additionally, the Decision Tree classifier correctly identifies 27 instances and incorrectly identifies three instances for the positive class. There are 120 instances of correct predictions and five instances of incorrect predictions.

Within the confusion matrix of the CatBoost model, the True Positive value is observed as 93, while the True Negative score is recorded as 27. Notably, the CatBoost classifier accurately predicts 93 negative class samples, yet it misclassifies two samples. This classifier successfully identifies 27 instances regarding the positive class, whereas it mistakenly labels three cases within the negative class. The total number of accurate predictions is 120 instances, while there are five instances where the predictions are erroneous.

In the confusion matrix associated with the XGBoost model, the True Positive value stands at 93, whereas the True Negative score is 29. Examining its performance, the XGBoost classifier effectively predicts 93 samples within the negative class, with only two misclassifications. Furthermore, this classifier demonstrates the ability to identify 29 instances within the positive class accurately but only correctly distinguishes one case within the negative class. Comparatively, there were three instances of incorrect predictions, while a total of 122 predictions were correct.

A comparison of performances of different classifiers—Random Forest, Decision Tree, CatBoost, and XGBoost is carried out. Table 6 compares the performance of Random Forest, Decision Tree, CatBoost, and XGBoost in terms of accuracy, precision, recall, ROC-AUC score, and F1 score.

Figure 3 represents the graphical representation of comparative analysis among all algorithms. In Table 6 and Fig. 3, XGBoost outperformed other models regarding accuracy, precision, recall, F1-score, and ROC-AUC score. The most outstanding 97.6% accuracy is attained by Random Forest, with 94% precision, 97% recall, 95% F1-score, and 97.2% ROC-AUC score. Therefore, it has been concluded

Table 5 Pearson's Chi-square association test of socio-economic, demographic, and behavioral factors with COVID-19 infection

Variable name	Categories	Infected by COVID-19 Virus		Chi-square value	P-value
		No	Yes		
Gender	Female	161	60	2.45	0.072
	Male	220	59		
Age	16–20	65	10	12.832	0.005
	21–25	231	65		
	25–30	26	12		
	> 30	59	32		
Vaccinated_Or_not	No	13	0	4.168	0.028
	Yes	368	119		
Sanitizer_Uses	No	20	6	7.284	0.026
	Yes	232	88		
	Sometimes	129	25		
Time_of_Infection	Not infected	370	1	440.65	0
	After vaccination	0	31		
	Before vaccination	11	87		
Area	Rural	119	28	2.593	0.107
	Urban	262	91		
Vaccine_Name	Phizer	86	24	5.559	0.135
	Moderna	43	15		
	Covishield	51	26		
	Sinopharm	201	54		
Vaccine_Doses	1st dose	29	2	11.845	0.003
	2nd dose	251	69		
	3rd dose	101	48		
Profession	Govt_job	14	13	19.437	0
	Students	308	74		
	Others	59	32		
Family	2–3 members	26	16	5.865	0.053
	4–6 members	298	90		
	6+ members	57	13		
Birth division	Dhaka	59	18	6.004	0.05
	Chattogram	244	64		
	Others	78	37		
Vaccination division	Dhaka	74	32	12.97	0.002
	Chattogram	251	57		
	Others	56	30		
Mask	Surgical	258	94	8.605	0.014
	Cotton	84	22		
	Others	39	3		
Time_gap_1st and_2nd Dose	1 month gap	196	60	8.835	0.032
	2 months gap	93	32		
	3 months gap	35	19		
	More than 3 months gap	57	8		
2nd and 3rd dose gap	1 month gap	16	7	1	0.801
	2 months gap	20	5		
	3 months gap	24	6		
	More than 3 months	321	101		

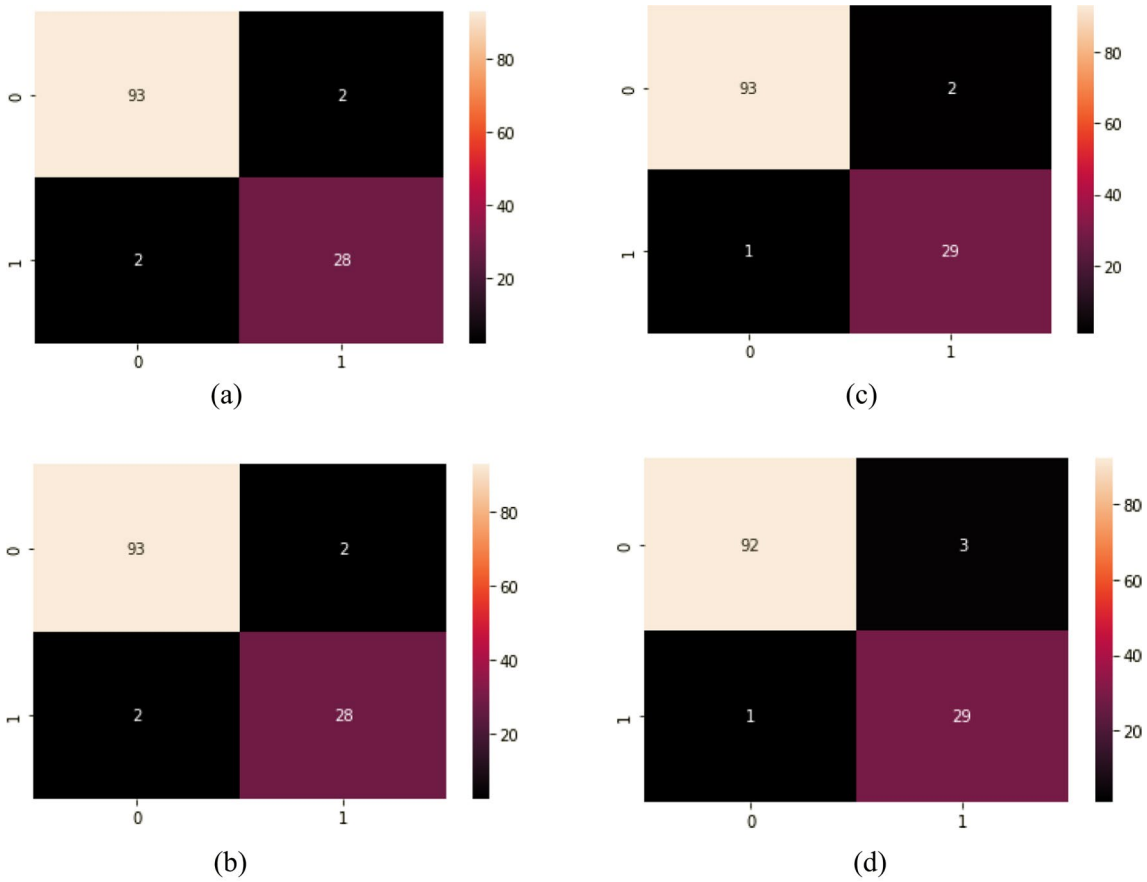

Fig. 2 Confusion matrix for random forest (**a**), decision tree (**b**), XGBoost (**c**), and CatBoost (**d**)

Table 6 Comparative analysis of all algorithms

	Accuracy (%)	Precision (%)	Recall (%)	ROC-AUC score (%)	F1-score (%)
Random forest	96.8	93	93	95	93
Decision tree	96	93	90	93	92
CatBoost	96.8	91	97	96	94
XGBoost	97.6	94	97	97.2	95

that XGBoost is superior to the other three classification algorithms for COVID-19 detection. Random Forest has achieved 96.8% accuracy, 93% precision, 93% recall, 93% F1 score, and 95% ROC-AUC score. CatBoost had 96.8% accuracy with 91% precision, 97% recall, 94% F1 score, and 96% ROC-AUC score. Random Forest and CatBoost performed well and had the same accuracy, albeit CatBoost had a higher recall, ROC-AUC score, and F1 score than Random Forest. But CatBoost had a lower precision value than Random Forest. But when compared to other methods, the Decision Tree provided slightly less performance. Therefore, it can be said that Decision Tree was less effective than other classification algorithms, and XGBoost was the best for our dataset.

3.3 Feature Importance Analysis

Feature importance has been evaluated for the Random Forest, Decision Tree, CatBoost, and XGBoost models. Table 7 displays a comprehensive breakdown of the importance assigned to each feature. Notably, there is a remarkable overlap in the top-ranking features across all models, with a predominant focus on the questions posed. This table directly compares the importance values assigned by each ML model to the identified features.

"Time of infection" is the top feature for all models. It can be seen that the top 5 features of Random Forest lie between 0.67 and 0.023, whereas, in the case of the

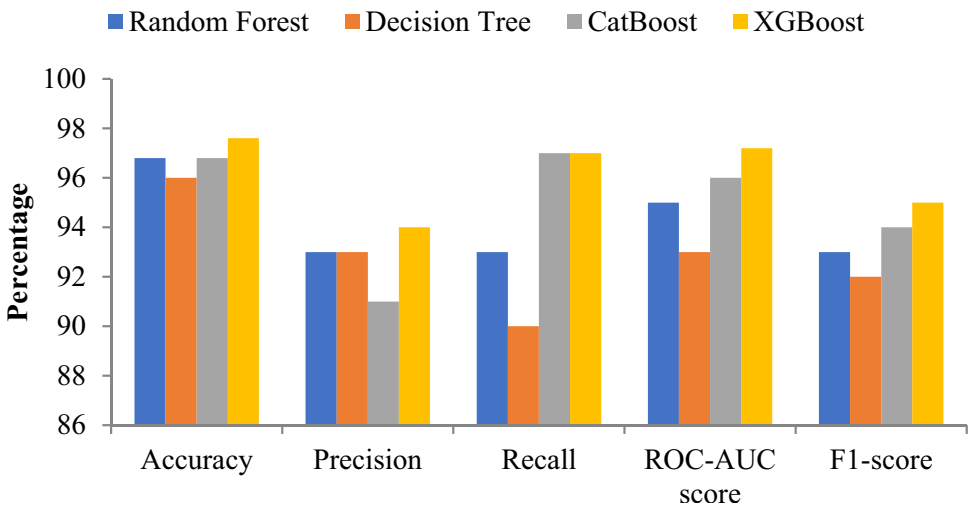

Fig. 3 Graphical representation of comparative analysis among all algorithms

Table 7 Feature importance scores obtained for the features of the COVID-19 infection

Features	Random forest	Decision tree	CatBoost	XGBoost
Time of infection	0.675	0.865	87.87	0.78
First and second dose gap	0.034	0.041	0.767	0.014
Vaccine name	0.023	0.003	1.162	0.018
Profession of respondents	0.025	0.007	0.767	0
Vaccine doses	0.017	0.009	0.698	0.008
Birth division	0.022	0	0.806	0.007
Vaccination division	0.020	0.005	1.286	0.015
Mask uses	0.023	0.015	1.875	0.05
Sanitizer uses	0.019	0.026	1.269	0.022
Side effect duration	0.031	0.018	0.112	0.03
Family	0.01	0.001	0.675	0.008
Area	0.015	0	0.332	0.005
Above 30 years	0.008	0.004	0.175	0.01
21–25 age group	0.010	0	0.008	0.599
26–30 age group	0.007	0	0	0.124
16–20 age group	0.007	0	0	0
Gender	0.018	0	0.00	0.078

Decision Tree, they lie between 0.87 and 0.015, and in the case of the XGBoost model, they lie between 0.78 and 0.02. Finally, in the case of the CatBoost model, they lie between 87.87 and 1.16. In each instance, the CatBoost model assigns more excellent importance scores to the variables than the other algorithms. Feature selection endeavors to mitigate overfitting, enhance model selection, and provide a deeper understanding of the underlying process. Accordingly, better-fitted models, in line with the objectives of feature selection, tend to exhibit higher feature importance scores compared to less well-fitted models. Table 7 shows that the CatBoost model outperforms the other models in terms of overall accuracy across all cases.

3.4 Discussion

The COVID-19 pandemic has affected the world unprecedentedly, with millions of people infected and hundreds of thousands of deaths reported. The pandemic has highlighted the importance of understanding the social and behavioral factors contributing to the disease's spread. Our findings suggest several social and behavioral factors are associated with COVID-19 infection. Increased age is a significant risk factor for severe COVID-19 infection [54]. The Chi-square test in our proposed study demonstrated a statistically significant relationship between age and COVID-19 infection in Bangladesh, as evidenced by a χ^2 value of 12.832 at a 5% significance level with $p < 0.05$.

A previous study identified that adults aged 20 to 49, with a particular focus on those between 35 and 49, played a pivotal role in the resurgence of COVID-19 epidemics in the United States, both before and after the reopening of schools in 2020. The risk of virus transmission was higher among those aged 20 to 49 [55]. Infection rates are high in working women between 60 and 69 years old. Women at ages 20–29 are less likely to have the infection. Women of working age in health and care occupations are also more likely to have increased infection rates [56]. The age group with the highest COVID-19 cases was adults aged 18 to 49. COVID-19 cases were less common in children aged 0–17 and adults aged 50 and older [57]. During the tumultuous pandemic, a meticulously orchestrated and extensive vaccination strategy was diligently implemented to mitigate the number of infected individuals. This stratagem aimed to achieve an optimal level of immunization, equivalent to approximately 47 vaccine doses administered per 100 inhabitants. However, as the relentless waves of the pandemic surged in May 2021, the optimal threshold of vaccines escalated to a staggering count of approximately 90 doses [58]. Notably, during this critical juncture, regions plagued by abysmal vaccination rates experienced a disproportionately sharp upsurge in the incidence of new COVID-19 cases, with rural areas particularly susceptible to the pernicious effects of the virus [59]. It has been observed that a distinct correlation exists between the prevalence of surgical morbidity and lower rates of COVID-19 immunization [60, 61]. Indeed, these vaccines have exhibited remarkable prowess in safeguarding against harmful consequences stemming from diverse SARS-CoV-2 variants. Our study observed that most respondents who had received two vaccine doses experienced a lower rate of COVID-19 infection. This finding suggests a significant relationship between the number of vaccine doses administered and the likelihood of virus infection ($\chi^2 = 11.845$, $p < 0.05$). A seminal study has postulated that the particular implementation of a vaccination strategy culminated in a remarkable decline in the number of COVID-19 patients spanning all age cohorts, accompanied by notable reductions in both hospitalization and mortality rates. Encouragingly, the hospitalization rate for individuals over 80 years experienced a staggering reduction of 80%, while those aged 50 to 70 exhibited a hospitalization rate on par with the pre-vaccination era [62]. The binding nature of vaccination in curtailing the proliferation of COVID-19 must be balanced, particularly among demographic segments with lower vaccination rates, such as young adults. In Israel, where comprehensive vaccination efforts have been implemented, the vaccine has demonstrated remarkable effectiveness in preventing SARS-CoV-2 infections, mitigating COVID-19 cases, averting hospitalizations, and averting deaths. Statistical analyses have revealed that the vaccine boasts an impressive efficacy rate of 92% in thwarting COVID-19 cases and an 87% efficacy rate in preventing COVID-19 hospitalizations [63].

In the United States, augmenting vaccination rates while concurrently adhering to non-pharmaceutical interventions, such as mask usage and physical distancing, can substantially diminish the incidence of COVID-19 cases, hospitalizations, and fatalities. Moreover, research has also underscored the potential role of elevated vaccination rates in improving the deleterious consequences precipitated by emerging COVID-19 variants [64]. Exploratory investigations have further shed light on the prevailing sentiments among the populace during the pandemic. Notably, over 80% of individuals harbored COVID-19-related concerns, with approximately 72% emphasizing the necessity of sanitizers as a preventive measure [65]. Our study identified statistically significant associations between COVID-19 infection and certain preventive measures. Specifically, using sanitizer was significantly associated with COVID-19 infection ($\chi^2 = 7.284$, $p < 0.05$), indicating its impact on disease transmission. Additionally, the type of masks worn, including surgical, cotton, or other varieties, exhibited a significant relationship with the likelihood of COVID-19 infection ($\chi^2 = 8.605$, $p < 0.05$). Consequently, emphasizing the utilization of sanitizers, mask-wearing, practicing safe distancing, and embracing vaccination become paramount in staving off the perils of the virus [66]. A repertoire of preventive measures, including a reasonable selection of surface cleaners, rigorous adherence to proper hand sanitization protocols, and consistent mask usage, represents tangible interventions that can effectively avert the onset of COVID-19 [67]. The majority of contagious diseases are transmitted by interaction, airborne transmission. Following the COVID-19 outbreak, the Taiwanese government began implementing various preventive measures such as social isolation and using masks and hand sanitizer. There is a good chance that this public reaction substantially affected the decrease [68]. The administration of the two-dose COVID-19 vaccine has demonstrated significant efficacy in reducing the incidence of infection, hospitalizations, and fatalities within the United States. Remarkably, the vaccine has contributed to a substantial decline in the infection rate, which now stands at a mere 4.6%. Moreover, the vaccine has proven instrumental in mitigating adverse outcomes, yielding a notable 63.5% reduction in non-intensive care unit (ICU) admissions, ICU hospitalizations, and deaths [69].

As the omicron variant made its initial impact in 2022, educational institutions that implemented stringent hygiene measures and comprehensive testing strategies accounted for approximately 20% of population-based infections. In contrast, this figure dwindled to a mere 2% during vacations and school closures. Notably, Germany witnessed a significant

proportion of both teachers and students, approximately one-third, falling prey to the omicron variant during its initial phase in 2022 [70]. During the first eight weeks of school reopening, a marginal upswing in infections was observed among both adults (2.2%) and children (4.5%) [71]. It is imperative to recognize that students, like anyone else, can contract COVID-19 if they encounter an infected individual. Close contact can occur within classrooms, hallways, or during extracurricular activities within a school setting. Consequently, it is paramount for students to adhere to recommended guidelines, including the consistent use of masks, practicing social distancing, and maintaining regular hand hygiene to curb the transmission of COVID-19.

Many traditional approaches identify COVID-19 cases, including the RT-PCR test and viral antigen detection [72]. However, many experimental techniques are expensive, time-consuming, and require special instructions [73]. It is only sometimes possible to gather samples from a small number of people in a short length of time. As a result, a large number of people need to be noticed. So, these techniques are not feasible for the people of developing countries. When people are infected, community transmission is one of the most critical topics to consider [74]. Furthermore, it is necessary to discover positive cases and detect negative cases more appropriately. It can spread around the neighborhood if some positive cases are mistaken for negatives. When many patients are admitted to the hospital, physicians and healthcare workers cannot take appropriate action, and policymakers may find it challenging to deal with the problem. If these patients are identified early on based on socio-demographic and behavioral factors, they can be isolated from the rest of the population and treated as soon as feasible, lowering the COVID-19 transmission rate. As a result, we needed a suitable model to distinguish COVID-19 positive and negative cases more efficiently. The proposed Machine Learning model can more reliably investigate COVID-19 infection cases.

This study highlights the importance of understanding social and behavioral factors affecting COVID-19 infection. Our findings suggest that promoting adherence to public health guidelines, addressing misinformation and stigma, and targeting interventions for vulnerable groups may effectively reduce the spread of the disease.

4 Conclusion

The COVID-19 pandemic has brought to light the intricate relationship between social and behavioral factors and the spread of infectious diseases. These factors have played a significant role in shaping the trajectory of the pandemic. Factors such as population density and lack of awareness have contributed to the rapid spread of COVID-19 worldwide. Additionally, reluctance to wear masks and the use of sanitizer has also fueled the spread of the virus. Vaccination is an essential tool to control the spread of COVID-19, and addressing social and behavioral factors is crucial to achieving high vaccination rates and maintaining the pandemic. This study uses a comprehensive approach to address the social and behavioral factors affecting COVID-19 infection. This research employed a diverse set of socio-demographic and behavioral factors, including area of residence, gender, age, profession, family size, birth division, vaccination division, vaccinated status, vaccine type, vaccine dosage, mask usage, mask type, sanitizer usage, side effects duration, coronavirus severity level, taste disturbance, fever, headache, fatigue, muscle pain, pain at the injection site, absence of symptoms, and time of infection, to comprehensively analyze the dynamics of COVID-19 infection. Pearson Chi-Square tests were rigorously conducted to unveil associations between the target variable and these diverse factors. In tandem, various ML classification techniques, namely Decision Tree, Random Forest, CatBoost, and XGBoost were harnessed to predict COVID-19 infections within the realm of public health, drawing insights from the intricate web of socio-demographic variables and vaccine-related information. Furthermore, Machine Learning-based feature importance analysis was instrumental in identifying the most significant factors contributing to COVID-19 infection risk. Below, the key findings derived from this comprehensive approach are summarized:

4.1 Statistical Analysis

The Pearson Chi-Square test revealed significant associations between various factors and COVID-19 infection. Among the many factors examined, three stood out with particularly noteworthy results. A χ^2 value of 8.835 ($p<0.05$) highlighted a significant relationship between the time gap between the first and second vaccine doses and the likelihood of COVID-19 infection. The use of sanitizer exhibited a statistically significant association ($\chi^2=7.284$, $p<0.05$) with COVID-19 infection. The type of masks worn, whether surgical, cotton, or other, also demonstrated a significant relationship ($\chi^2=8.605$, $p<0.05$) with the likelihood of COVID-19 infection.

4.2 Machine Learning-based COVID-19 Prediction

The XGBoost algorithm provides 97.6% accuracy, with 94% precision, 97% recall, 95% F1-score, and 97.2% ROC-AUC score, which is more promising than other algorithms and can benefit physicians in making final predictions about COVID-19 patients.

Table 8 Abbreviation table

AI	Artificial intelligence
ML	Machine learning
WHO	World health organization
TP	True positive
TN	True negative
FP	False positive
FN	False negative

4.3 Feature Importance Analysis

In feature importance analysis, "Time of infection" consistently ranked as the top predictor across all models. CatBoost exhibited the highest importance scores, outperforming other algorithms, ranging from 0.87 to 1.16. This underscores CatBoost's superior performance in feature selection and overall accuracy.

Our study underscores the intricate interplay between social and behavioral factors and the transmission of infectious diseases. It advances that controlling pandemics like COVID-19 necessitates a holistic understanding of human behavior, healthcare disparities, and public health interventions. By examining various socio-demographic and behavioral variables, we contribute to a deeper conceptual understanding of the factors influencing disease spread. The practical implications of our research are substantial. Policymakers and healthcare professionals can utilize our findings to formulate evidence-based strategies for pandemic control. For instance, our identification of the time gap between vaccine doses, sanitizer usage, and mask types as significant factors can guide the development of targeted interventions. Practical applications include optimizing vaccination schedules, promoting sanitizer use, and recommending specific mask types based on risk assessments. The value of our research lies in its potential to save lives and mitigate the societal impact of pandemics. By revealing significant predictors of COVID-19 infection, we provide actionable insights that can guide public health efforts. Our ML-based prediction model, especially the high-performing XGBoost algorithm, offers a valuable tool for healthcare professionals to identify high-risk individuals and allocate resources efficiently. This value extends to the improved well-being of individuals and communities. Our findings extend beyond specific geographical boundaries or barriers, holding broader international relevance. The methodologies we employed, including statistical analyses, ML, and feature importance techniques, can be applied in diverse settings to effectively study and combat infectious diseases. This enhances the generalizability of our study, making it a valuable resource for researchers and policymakers worldwide. Future work could expand the dataset to include a more extensive and diverse sample, allowing for a deeper understanding of the global implications of these socio-demographic and behavioral factors on infectious disease spread and facilitating the development of tailored public health strategies.

Abbreviations All abbreviations are listed in Table 8.

Acknowledgements All authors are thanked for their contributions to this research, and on behalf of all authors, we would like to thank Anupam Kumar Bairagi for supervising and significantly contributing to this study.

Author Contributions KMAU: conceptualization, investigation, supervision, writing—review, and editing. FSP: draft manuscript preparation, literature review management, writing—review, and editing. MT and SNJ: data collection, conceptualization, model training, analysis, and interpretation of results. MOF: study conception, design, supervision, and investigation of challenges. ABC, ABC, and ABC: study conception, design, and supervision. All authors reviewed the results and approved the final version of the paper.

Funding This research received no external funding.

Data Availability Data will be provided on request.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any authors.

Consent to Participate Not Applicable.

Consent for Publication Not Applicable.

References

1. Rajpal S, Agarwal M, Rajpal A, Lakhyani N, Saggar A, Kumar N. Cov-elm classifier: an extreme learning machine based identification of covid-19 using chest x-ray images. Intell Decis Technol. 2022;16(1):193–203.
2. World Health Organization. Situation by Region, Country, Territory & Area. WHO Coronavirus (COVID-19) Dashboard. 2021.
3. COVID—coronavirus statistics—Worldometer. https://www.worldometers.info/coronavirus/. Accessed 10 Apr, 2023.
4. Countries where Coronavirus has spread—Worldometer. https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/. Accessed 10 Apr, 2023.

5. Varela-Santos S, Melin P. A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. Inf Sci. 2021;545:403–14.
6. United States COVID—coronavirus statistics—worldometer. https://www.worldometers.info/coronavirus/country/us/. Accessed 10 Apr, 2023.
7. Which countries are impacted the most by the COVID-19 coronavirus? | World Economic Forum. https://www.weforum.org/agenda/2020/03/infographic-coronavirus/. Accessed 10 Apr, 2023.
8. COVID-19 vaccine doses administered by country worldwide 2023 | Statista. https://www.statista.com/statistics/1194934/number-of-covid-vaccine-doses-administered-by-county-worldwide/. Accessed 10 Apr, 2023.
9. Ochi S, So M, Hashimoto S, Denda K, Sekizawa Y. Behavioral factors associated with COVID-19 risk: a cross-sectional survey in Japan. Int J Environ Res Public Health. 2021;18(22):12184.
10. Hamer M, Kivimäki M, Gale CR, Batty GD. Lifestyle risk factors, inflammatory mechanisms, and COVID-19 hospitalization: a community-based cohort study of 387,109 adults in UK. Brain Behav Immun. 2020;87:184–7.
11. Ezati Rad R, Mohseni S, Kamalzadeh Takhti H, Hassani Azad M, Shahabi N, Aghamolaei T, Norozian F. Application of the protection motivation theory for predicting COVID-19 preventive behaviors in Hormozgan, Iran: a cross-sectional study. BMC Public Health. 2021;21(1):1–11.
12. Šuriņa S, Mārtinsone K, Upesleja G, Perepjolkina V. Factors associated with COVID-19 vaccination behaviour in Latvian population: cross-sectional study. Health Psychol Behav Med. 2022;10(1):514–36.
13. Romero Starke K, Petereit-Haack G, Schubert M, Kämpf D, Schliebner A, Hegewald J, Seidler A. The age-related risk of severe outcomes due to COVID-19 infection: a rapid review, meta-analysis, and meta-regression. Int J Environ Res Public Health. 2020;17(16):5974.
14. Janairo GIB, Yu DEC, Janairo JIB. A machine learning regression model for the screening and design of potential SARS-CoV-2 protease inhibitors. Netw Model Anal Health Inform Bioinform. 2021;10:1–8.
15. Chaudhuri AK, Sinha D, Banerjee DK, Das A. A novel enhanced decision tree model for detecting chronic kidney disease. Netw Model Anal Health Inform Bioinform. 2021;10:1–22.
16. Khan MAR, Afrin F, Prity FS, Ahammad I, Fatema S, Prosad R, Hasan MK, Uddin M. An effective approach for early liver disease prediction and sensitivity analysis. Iran J Comput Sci. 2023; 1–19.
17. Barragán-Montero A, Javaid U, Valdés G, Nguyen D, Desbordes P, Macq B, Willems S, Vandewinckele L, Holmström M, Löfman F, Michiels S. Artificial intelligence and machine learning for medical imaging: a technology review. Physica Med. 2021;83:242–56.
18. Kavas PÖ, Bozkurt MR, Kocayiğit İ, Bilgin C. Machine learning-based medical decision support system for diagnosing HFpEF and HFrEF using PPG. Biomed Signal Process Control. 2023;79:104164.
19. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201–8.
20. Rana M, Bhushan M. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools Appl. 2023;82(17):26731–69.
21. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, Marsh J, Devylder J, Walter M, Berrouiguet S, Lemey C. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res. 2021;23(5):e15708.
22. Khanbhai M, Anyadi P, Symons J, Flott K, Darzi A, Mayer E. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. BMJ Health Care Inform. 2021;28(1).
23. Manhas J, Gupta RK, Roy PP. A review on automated cancer detection in medical images using machine learning and deep learning based computational techniques: Challenges and opportunities. Arch Comput Methods Eng. 2021:1–41.
24. Allugunti VR. Breast cancer detection based on thermographic images using machine learning and deep learning algorithms. Int J Eng Comp Sci. 2022;4(1):49–56.
25. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.
26. Han Y, Huang J, Li R, Shao Q, Han D, Luo X, Qiu J. Impact analysis of environmental and social factors on early-stage COVID-19 transmission in China by machine learning. Environ Res. 2022;208: 112761.
27. Suthar AB, Wang J, Seffren V, Wiegand RE, Griffing S, Zell E. Public health impact of covid-19 vaccines in the US: observational study. BMJ. 2022;377.
28. Zhang Y, Khan MA, Zhu Z, Wang S. SNELM: SqueezeNet-guided ELM for COVID-19 recognition. Comput Syst Sci Eng. 2023;46(1):13.
29. Zhang Y, Satapathy SC, Zhu LY, Górriz JM, Wang S. A seven-layer convolutional neural network for chest CT-based COVID-19 diagnosis using stochastic pooling. IEEE Sens J. 2020;22(18):17573–82.
30. Wang SH, Fernandes SL, Zhu Z, Zhang YD. AVNC: attention-based VGG-style network for COVID-19 diagnosis by CBAM. IEEE Sens J. 2021;22(18):17431–8.
31. Zietz M, Zucker J, Tatonetti NP. Associations between blood type and COVID-19 infection, intubation, and death. Nat Commun. 2020;11(1):5761.
32. Viswanath K, Bekalu M, Dhawan D, Pinnamaneni R, Lang J, McLoud R. Individual and social determinants of COVID-19 vaccine uptake. BMC Public Health. 2021;21(1):818.
33. Saeed BQ, Al-Shahrabi R, Alhaj SS, Alkokhardi ZM, Adrees AO. Side effects and perceptions following Sinopharm COVID-19 vaccination. Int J Infect Dis. 2021;111:219–26.
34. Pradhan A, Prabhu S, Chadaga K, Sengupta S, Nath G. Supervised learning models for the preliminary detection of COVID-19 in patients using demographic and epidemiological parameters. Information. 2022;13(7):330.
35. de Souza WM, Buss LF, Candido DD, Carrera JP, Li S, Zarebski AE, Pereira RH, Prete CA Jr, de Souza-Santos AA, Parag KV, Belotti MC. Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. Nat Hum Behav. 2020;4(8):856–65.
36. Iwendi C, Huescas CG, Chakraborty C, Mohan S. COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients. J Exp Theor Artif Intell. 2022:1–21.
37. Prity FS, Nath N, Nath A, Uddin KA. Neural network-based strategies for automatically diagnosing of COVID-19 from X-ray images utilizing different feature extraction algorithms. Netw Model Anal Health Inform Bioinform. 2023;12(1):28.
38. Cafri G, Li L, Paxton EW, Fan J. Predicting risk for adverse health events using random forest. J Appl Stat. 2018;45(12):2279–94.
39. Kaur P, Kumar R, Kumar M. A healthcare monitoring system using random forest and internet of things (IoT). Multimedia Tools and Applications. 2019;78:19905–16.
40. Santos LI, Camargos MO, D'Angelo MF, Mendes JB, de Medeiros EE, Guimarães AL, Palhares RM. Decision tree and artificial immune systems for stroke prediction in imbalanced data. Expert Syst Appl. 2022;191: 116221.
41. Qiu X, Miao J, Lan Y, Sun W, Li G, Pan C, Wang Y, Zhao X, Zhu Z, Zhu S. Artificial neural network and decision tree

models of post-stroke depression at 3 months after stroke in patients with BMI≥ 24. J Psychosom Res. 2021;150: 110632.
42. Sahoo S, Subudhi A, Dash M, Sabut S. Automatic classification of cardiac arrhythmias based on hybrid features and decision tree algorithm. Int J Autom Comput. 2020;17(4):551–61.
43. Behadada O, Chikh MA. An interpretable classifier for detection of cardiac arrhythmias by using the fuzzy decision tree. Artif Intell Res. 2013;2(3):45–58.
44. Nasser FK, Behadili SF. Breast cancer detection using decision tree and k-nearest neighbour classifiers. Iraqi J Sci. 2022:4987–5003.
45. Chaudhuri AK, Sinha D, Banerjee DK, Das A. A novel enhanced decision tree model for detecting chronic kidney disease. Netw Model Anal Health Inform Bioinform. 2021;10:1–22.
46. Gupta H, Kumar P, Saurabh S, Mishra SK, Appasani B, Pati A, Ravariu C, Srinivasulu A. Category boosting machine learning algorithm for breast cancer prediction. REVUE ROUMAINE DES SCIENCES TECHNIQUES—SÉRIE ÉLECTROTECHNIQUE ET ÉNERGÉTIQUE. 2021;66(3):201–6.
47. Almars AM, Alwateer M, Qaraad M, Amjad S, Fathi H, Kelany AK, Hussein NK, Elhosseini M. Brain cancer prediction based on novel interpretable ensemble gene selection algorithm and classifier. Diagnostics. 2021;11(10):1936.
48. Sudharson D, Retheneka SO, Yogeswari L, Logita SJ, Shankari S, Surraya SN. Enhancing the Efficiency of Lung Disease Prediction using CatBoost and Expectation Maximization Algorithms. In: 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2022 Sep 21; p. 57–61.
49. Kabiraj S, Raihan M, Alvi N, Afrin M, Akter L, Sohagi SA, Podder E. Breast cancer risk prediction using XGBoost and random forest algorithm. In: 2020 11th international conference on computing, communication and networking technologies (ICCCNT). IEEE 2020 Jul 1; pp. 1–4.
50. Sinha NK, Khulal M, Gurung M, Lal A. Developing a web based system for breast cancer prediction using xgboost classifier. Int J Eng Res Technol (IJERT). 2020;9(6):852–6.
51. Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, Liu X, Yang Y, Xu S. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. Thoracic cancer. 2020;11(1):95–102.
52. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Trans Biomed Eng. 2020;68(1):148–60.
53. Binson VA, Subramoniam M, Sunny Y, Mathew L. Prediction of pulmonary diseases with electronic nose using SVM and XGBoost. IEEE Sens J. 2021;21(18):20886–95.
54. Monod M, Blenkinsop A, Xi X, Hebert D, Bershan S, Tietze S, Baguelin M, Bradley VC, Chen Y, Coupland H, Filippi S. Age groups that sustain resurging COVID-19 epidemics in the United States. Science. 2021;371(6536):eabe8372.
55. Sobotka T, Brzozowska Z, Muttarak R, Zeman K, Di Lego V. Age, gender and COVID-19 infections. MedRxiv. 2020:2020-05.
56. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med. 2020;26(8):1205–11.
57. Coccia M. Optimal levels of vaccination to reduce COVID-19 infected individuals and deaths: a global analysis. Environ Res. 2022;204: 112314.
58. Cuadros DF, Miller FD, Awad S, Coule P, MacKinnon NJ. Analysis of vaccination rates and new COVID-19 infections by US county, July–August 2021. JAMA Netw Open. 2022;5(2):e2147915.
59. Prasad NK, Lake R, Englum BR, Turner DJ, Siddiqui T, Mayorga-Carlin M, Sorkin JD, Lal BK. COVID-19 vaccination associated with reduced postoperative SARS-CoV-2 infection and morbidity. Ann Surg. 2022;275(1):31.
60. Fiolet T, Kherabi Y, MacDonald CJ, Ghosn J, Peiffer-Smadja N. Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review. Clin Microbiol Infect. 2022;28(2):202–21.
61. Roghani A. The influence of COVID-19 vaccination on daily cases, hospitalization, and death rate in Tennessee, United States: case study. JMIRx med. 2021;2(3): e29324.
62. Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, Brooks N, Smaja M, Mircus G, Pan K, Southern J. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. The Lancet. 2021;397(10287):1819–29.
63. Borchering RK, Viboud C, Howerton E, Smith CP, Truelove S, Runge MC, Reich NG, Contamin L, Levander J, Salerno J, Van Panhuis W. Modeling of future COVID-19 cases, hospitalizations, and deaths, by vaccination rates and nonpharmaceutical intervention scenarios—United States, April–September 2021. Morb Mortal Wkly Rep. 2021;70(19):719.
64. Roy D, Tripathy S, Kar SK, Sharma N, Verma SK, Kaushal V. Study of knowledge, attitude, anxiety & perceived mental healthcare need in Indian population during COVID-19 pandemic. Asian J Psychiatr. 2020;51: 102083.
65. Kunduru KR, Kutner N, Nassar-Marjiya E, Shaheen-Mualim M, Rizik L, Farah S. Disinfectants role in the prevention of spreading the COVID-19 and other infectious diseases: the need for functional polymers! Polym Adv Technol. 2022;33(11):3853–61.
66. Pradhan D, Biswasroy P, Naik PK, Ghosh G, Rath G. A review of current interventions for COVID-19 prevention. Arch Med Res. 2020;51(5):363–74.
67. Galvin CJ, Li YC, Malwade S, Syed-Abdul S. COVID-19 preventive measures showing an unintended decline in infectious diseases in Taiwan. Int J Infect Dis. 2020;98:18–20.
68. Moghadas SM, Vilches TN, Zhang K, Wells CR, Shoukat A, Singer BH, Meyers LA, Neuzil KM, Langley JM, Fitzpatrick MC, Galvani AP. The impact of vaccination on coronavirus disease 2019 (COVID-19) outbreaks in the United States. Clin Infect Dis. 2021;73(12):2257–64.
69. Heinsohn T, Lange B, Vanella P, Rodiah I, Glöckner S, Joachim A, Becker D, Brändle T, Dhein S, Ehehalt S, Fries M. Infection and transmission risks of COVID-19 in schools and their contribution to population infections in Germany: a retrospective observational study using nationwide and regional health and education agency notification data. PLoS Med. 2022;19(12): e1003913.
70. Yuan P, Aruffo E, Gatov E, Tan Y, Li Q, Ogden N, Collier S, Nasri B, Moyles I, Zhu H. School and community reopening during the COVID-19 pandemic: a mathematical modelling study. R Soc Open Sci. 2022;9(2): 211883.
71. Ahammed K, Satu MS, Abedin MZ, Rahaman MA, Islam SM. Early detection of coronavirus cases using chest X-ray images employing machine learning and deep learning approaches. MedRxiv. 2020;10(2020.06):07-20124594.
72. Dhar BC. Diagnostic assay and technology advancement for detecting SARS-CoV-2 infections causing the COVID-19 pandemic. Anal Bioanal Chem. 2022;414(9):2903–34.
73. Nicastri E, D'Abramo A, Faggioni G, De Santis R, Mariano A, Lepore L, Molinari F, Petralito G, Fillo S, Munzi D, Corpolongo A. Coronavirus disease (COVID-19) in a paucisymptomatic

patient: epidemiological and clinical challenge in settings with limited community transmission, Italy, February 2020. Eurosurveillance. 2020;25(11):2000230.
74. Byambasuren O, Cardona M, Bell K, Clark J, McLaws ML, Glasziou P. Estimating the extent of asymptomatic COVID-19 and its potential for community transmission: systematic review and meta-analysis. Off J Assoc Med Microbiol Infect Dis Canada. 2020;5(4):223–34.