Statistical Modeling for COVID-19 Related Depression: Prediction, Classification, and Intervention

Nosheen Faiz[1], Soofia Iftikhar[2], Beenish Khurshid[3], Saira Farman[3], Bushra Ismail[1]

[1]*Department of Statistics, Abdul Wali Khan University Mardan, Pakistan*
[2]*Department of Statistics, Shaheed Benazir Women University, Peshawar, Pakistan*
[3]*Department of Biochemistry, Abdul Wali Khan University Mardan, Pakistan*
Email: nosheenfaiz@awkum.edu.pk

Abstract: The globe has been in a chaos state since a corona-virus (SARSCoV2) first appeared in December 2019. It was helpful to utilize an isolation strategy with quarantine to slow down the spread of disease. As a result, individuals stayed indoors instead of engaging in their regular daily activities outside. The goal of this study is to examine the connections and potential mediatory pathways between mental health issues, how people perceive their illnesses, and disorders of anxiety and depression. This aim of this study is to use various machine learning approaches to predict, classify, and detect depression risk factors in two districs of Khyber Pukhtunkhwa (KPK), Pakistan. In this paper, machine learning methods i.e., Random Forest and LASSO have been used for feature selection. Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Least Absolute Shrinkage Selection Operation (LASSO), and Random Projection ensembles (RP) have been used to assess the performance of the LASSO and Random Forest by identifying important features. The results show that LASSO has performed better than the other methods. Additionally, the clustering technique is also utilized to detect different hot spots in the population by considering the data as an unsupervised issue.

Keywords: Machine learning, Covid-19, Feature selection, Depression, Risk factors.

1. Introduction

Severe Acute Respiratory Syndrome (SARS), also known as COVID-19 is a global epidemic that has raised major public health concerns. On December 31, 2019, instances of pneumonia of unidentified origin were stated in Wuhan, China. The outbreak was linked to seafood exposure in one of Wuhan's marketplaces, and was subsequently confirmed as a novel strain of Corona virus [1, 2]. Beginning as a local transmission from the Chinese city of Wuhan, COVID-19 has evolved to be one of the greatest disasters of the century. On February 11, 2020, the WHO formally recognized COVID-19 as a "pandemic" from its prior designation as a global health emergency (WHO, 2020). There were more than 3 million infections worldwide affecting more than 3 million individuals, and COVID-19 was blamed for more than 200,000 fatalities with a mortality rate of about 2%. The first imported case was discovered in Pakistan on February 26, 2020. Though the mortality rate in Pakistan is comparatively low, it is high in other nations such as Italy, Iran, and the United States [3, 4]. According to UNESCO, in April 8, 2020, instructional activities at educational establishments in 188 nations have been postponed globally [5]. With the growing understanding of the virus' unpredictable spread and the societal responses it may provoke, societies faced a great deal of uncertainty [6]. Due to the extraordinary COVID-19 pandemic's global reach and the widespread adoption of numerous severe infection control measures i.e., lock-down, and psychological discomfort. During the pandemic, there have been many mental health issues. In addition, fear of COVID-19 is one of the main causes of mental health issues during this time. More specifically, COVID-19 is a novel strain of infection, and to combat the disease's

effect, various stakeholders (including governments, healthcare professionals, policy makers, and scientists) need knowledge and data [7]. The COVID-19 epidemic, which affected 862 million children and adolescents, or roughly half of the world's student population, was estimated to have resulted in state school closures in 177 countries in March 18, 2020, according to the United Nations Educational, Scientific, and Cultural Organization. Before a week, this urgent situation has risen from twenty-nine states with nationwide school, college, or university closures. Less social contact, forced isolation, and quarantine all contributed to an increase in psychological anguish brought on by this pandemic. Recent research has identified a number of personality attributes as risk factors for mental health during the COVID-19 outbreak, including youth, female gender, blue-collar employment in "critical" jobs, low income, unemployment, having a history of mental illness, and physical inactivity [8]. Worldwide, mental illness is also surging at an epidemic rate, and COVID-19 anxiety made it particularly worse. Recent research has revealed that demographics, behavioral, and educational characteristics, such as gender, place of residence, marital status, socioeconomic level, loneliness, and future plans, are the key determinants of mental illness. Hence, anxiety is likely to spread among a large number of people due to the potentially fatal effects of COVID-19 and the fact that the several strategies used to reduce infection rates have had varying degrees of effectiveness [9, 10]. Given that it is doubtful that the COVID-19 infection and its intensity would be under control in the near future. It is crucial to gather scientific proof on COVID-19 phobia and its connections to elements affecting mental health [11]. It is challenging to distinguish between a serious major depressive episode and a temporary physiological reaction to the unexpected worldwide crisis since most surveys employ self-report scales that assess only likely occurrences of major depression. Nonetheless, it is still feasible that some susceptible people may have experienced depression during the epidemic and required clinical attention. It is generally established that being exposed to traumatic experiences or environmental and social stressors can have a variety of negative effects on one's mental health, including the start of depression or a worsening of an already existing depression [12]. COVID-19 was brought into Pakistan by foreign Pakistanis (students, zaireen, visitors, and pilgrims) traveling from other parts of the world. Pakistan has 7025 confirmed cases as a result. Because the main family structure of Pakistan, the world's fifth most populated nation, is a wider family system with several generations cohabiting, social isolation poses a risk to everyone in the family's mental health. Residents in communities rely on social support from their families and cultural activities to combat loneliness, negative moods, and psychological stress. During this pandemic outbreak, people may experience anxiety, rage, depressed symptoms, dread of dying, worry of contracting the disease themselves or passing it on to their family members, and other mental health issues [13–15]. Amidst the chaos caused by the emergence of SARSCoV2, isolation and quarantine measures significantly altered daily activities worldwide. This study explores the interconnections between mental health, illness perception, anxiety, and depression. Using machine learning techniques, specifically Random Forest and LASSO, depression risk factors were predicted and classified in Khyber Pukhtunkhwa (KPK), Pakistan. LASSO outperformed other methods, and clustering techniques identified distinct population hotspots.

2. Related study

In accordance with previous literature, other researchers have undertaken similar studies on a related subject, with stress, anxiety, and depression emerging as the most commonly observed mental health conditions among individuals in Pakistan. Numerous researchers are grappling with the myriad factors contributing to the increased prevalence of anxiety and depression among general practitioners and front line healthcare providers in Pakistan and other developing countries. Even the most advanced healthcare systems in the industrialized world have encountered shortcomings, suggesting that a potential factor could be the healthcare system's inability to effectively manage disease outbreaks. Pakistani doctors may worry more about getting sick and spreading infections to their family members as a result of inadequate infection control practices at their workplaces, and social isolation may make them more stressed out and raise their risk of developing psychological illness [16]. Focusing on the controlled actions such as improving hygiene, eating a healthy diet, exercising, sleeping, introspection, meditation, practicing minimalism, painting, writing, dancing, learning instruments, learning new languages, knitting, gardening, cooking, watching movies/serials, playing games, and journaling the personal observation and experiences can help reduce anxiety [16]. Although the WHO has classified the epidemic as a pandemic and Europe as the corona virus's center, Pakistan, a considerably less affected third-world

nation plagued by poverty, believed that corona virus was the least of their issues, possibly both by the government and the general public. The corona virus crisis in Pakistan could not be stopped by complete lock down and curfew because of the fear of retaliation from religious organizations and the absence of a comprehensive ban on prayer time in mosques [13]. Front line physicians and general practitioners in Pakistan and other developing countries may be seeing an upsurge in anxiety and sadness for a variety of reasons. Despite the fact that even the most cutting edge healthcare systems in the world have failed, one of the likely factors could be a belief that the healthcare system is unable to handle the epidemic. Due to poor infection control practices at work, Pakistani doctors may be more concerned about getting sick and infecting their 96 family members, while social isolation may increase stress and result in mental illness [17]. Psychological crisis intervention [18] is urgently required during the COVID-19 pandemic for affected, suspected, susceptible, and at-risk patients, caretakers, families, staff, and the general public in order to quickly prevent enormous risks from secondary mental health crisis. The goal of psychological crisis intervention is to manage the psycho social side effects and aftereffects of an infectious disease and lessen its psychological impact through prompt assessment, management, and prevention. Concerning and incites xenophobia among healthcare professionals and the common population. People's fear-related behaviors are heightened during outbreaks, especially in pandemic situations, and there is always an elevated risk of mental health issues [19]. Although self-isolation and quarantine have been advocated by experts, the necessary restrictions could have a negative impact on mental health in the short and possibly even the long term. Pakistan, a collectivist culture that places a high value on socialization, has been particularly vulnerable to self-isolation, social-distancing, and quarantine, and is reluctantly dealing with the uncertain and unpredictable emotional, psychological, psycho social, and social effects of this crisis. The psychological effects of quarantine include confusion, frustration, and post-traumatic stress disorder [20]. Psychiatric and psychological institutions offer counseling to quarantined patients, families, self-isolated individuals, health-care workers of medical and social-service personnel, and personnel in hospitals, laboratories, the field, and in quarantine. The actual application of the intervention is the main emphasis of psychological crisis intervention both during and after the outbreak phase [21, 22]. Additionally, patients cannot be transferred right away from the hospital to counseling or psychological intervention departments because the physical health department's assessment of the mental health states of those affected by the corona virus would still be confused even after the pandemic had ended. A fundamental principle in managing emotional distress and the public mental health emergencies brought on by the pandemics is to use professionally experienced and standard well-trained mental health practitioners, counseling psychologists, practicing psychotherapists, psychiatrists, and psychiatrist nurses who are familiar with the complicated case structures and work procedure [23]. This paper take in to account 2 district of KPK, Pakistan i.e., Mardan, Nowshehra dealing with mental illness created by pandemic i.e., COVID- 19. Mardan is one of the country's 19 high-burden Covid-19 districts. Covid-19 rate was found at 17% in Mardan and 7.3% in Nowshehra. From July 2021 to September 2021, patients in the district Mardan and Nowshehra who had verified clinical recovery and biological clearance after hospitalization for COVID-19 disease were included in this cross-sectional study.

Data Collection and Sample Size

This section provides a detailed explanation of research area, data collection, sample size, study design, and data analysis tools.

Participants and Data Collection

The cross-sectional study was conducted to predict mental illness in two districts i.e., Mardan and Nowshehra, two cities in Khyber Pukhtunkhwa, Pakistan from December 2021 to February 2022. Patients in the district Mardan and Noweshra who had verified clinical recovery and biological clearance after hospitalization for COVID-19 disease were 150 included in this cross-sectional study.

Data is acquired using a convenient sampling method [25]. A convenience sample is a non-probability sampling approach that chooses a sample from a group of people who are easy to contact or reach. A total of 1000 pre-designed questionnaires were distributed to people after informed consent who had previously been diagnosed with Covid-19. The consent of all participants involved in this study was taken in accordance with the Declaration of Helsinki, 1964. Approval for the study was obtained from

institutional Advanced Studies and Research Board of Abul Wali Khan University, Mardan, KP, Pakistan [26].

Sample Size

The overall number of participants in a research is referred to as the sample size [27], and in order to ensure that the sample as a whole accurately represents the entire population, this number is frequently divided into subgroups based on criteria such as, age, gender, and area. This can be presented as,
where,
n = sample size, Z = level of confidence for Z statistic, P = anticipated prevalence or percentage, and d = precision as a percentage.

The results of this investigation are presented with 95% confidence intervals (CI) using Z statistics (Z). The expected percentage (P) is the proportion of prevalence. It's important to remember that P has a scale from 0 to 1, and the sample size varies with P. The percentage of the KPK covid-19 affected population to the overall population of KPK, Pakistan is P in this case. KPK has a total population of 30,523371 with a covid-19 population of 166,000, resulting in a P value of 0.54. Consequently, the precision (d) value is set to 0.05. The required sample size is 398 according to the previously mentioned setup, although this paper is based on 622 participants. The data used in this paper is composed of 39 factors, the first 10 of which are related to the patient's medical history and the remaining 29 to depression. The degree of depression is computed as a response variable in this data by adding up the scores for each of the 29 questions by counting the number to the right of each one, which ranges from 0 to 2.

3. Methods

The methods used in analysis are describes as follows.

Least Absolute Shrinkage Selection Operation

Least Absolute Shrinkage Selection Operation (LASSO) [28] main operation is Regularization and model selection. The LASSO method restricts the entire sum of the model parameters to an absolute value that must be below a given threshold (higher bound). To do this, the technique shrinks (regularizes) regression coefficients that reduce some of these variables to zero. During the feature selection stage, variables that already have a non-zero coefficient are chosen to be a component of the model. Reducing and eliminating coefficients, reduces variance without significantly increasing bias, which is especially useful when there are a few occurrences and a large number of characteristics/features. Additionally, by eliminating redundant variables that are unrelated to the study variable, the LASSO reduces over-fitting and enhances model interpretability.

Random Forest

Random Forest (RF) [29] is an ensemble method used for classification and regression problems that uses the ideas of building numerous random trees, bootstrapping sample aggregation, voting systems, and other random variables in each decision split to improve its prediction capacity and efficacy [30]. It also backs up a strategy for determining the significance of variables. Random Forest outperforms to other sophisticated design learning approaches in terms of prediction accuracy. For implementing RF in this paper, R implementation given in package randomForest https://www.stat.berkeley.edu/~breiman/Usin HYPERLINK "https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf" HYPERLINK "https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf" HYPERLINK "https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf"g_random_forests_V3.1.pdf is used with the default values of its hyper parameters.

Support Vector Machines

Support Vector Machines (SVM) is an effective approach for binary classification, regression, and ranking [31]. SVM is built on a basic powerful principle of mathematics by selecting a hyperplane that, while allowing for a given amount of error, best fits the data points. It is a frequent machine learning method utilized by health care researchers in many categorization challenges due to its appealing characteristics as well as the handling of complicated nonlinear data points. To implement SVM, the R implementation given in package e1071 is utilized.

k-Nearest Neighbors

k-Nearest Neighbors (k-NN) [32, 33] is a supervised learning method that categorizes new data points into specified classes based on the characteristics of its neighbors. Observation characteristics are maintained for both the training and assessment datasets. For the k-NN algorithm, the k parameter, which specifies the number of neighbors, is selected. The R implementation provided in package caret is used with the default euclidean distance to implement k-NN.

Random Projection Ensembles

The Random Projection (RP) Ensemble [34] is a simple machine-learning method that minimizes data size without increasing model error, time, or size. The reduced-dimension random projections are used to project the data into a number of non-overlapping blocks, each of which is subsequently subjected to a base classifier. From each block, the best forecast in terms of error rate is chosen. To vote on a new observation, the preferred projections are merged into an ensemble.

Cluster Analysis

Cluster analysis [35–38] is a multivariate data mining technique that aims to classify items based on a set of user selected traits or attributes. It is applied in a variety of fields, including data compression, machine learning, pattern recognition, knowledge discovery, etc. High levels of both internal and external heterogeneity should be present in clusters.

Logistic Regression

It belongs to the class of generalised linear models known as the logit model [39–43] since it makes use of the logit link function. For binary classification techniques, it is utilized in a variety of areas, including biological, social science, and market research. This enables probabilistic interpretation; It is easy to modify and interpret the model with new data, in contrast to decision making. Its limitations include the fact that it is not linear and that interaction is challenging to define. Only discrete values may exist for the study variable.

4. Results And Discussion

The dataset has been partitioned into two segments: 70% of the data is utilized as training data, while the remaining 30% serves as testing data. A split-sample analysis consisting of 500 runs has been conducted for every combination of feature selection method and the specified classifiers, employing the 70% training and 30% testing partitions. A 1000 iterations process with 70% training and 30% testing is used to evaluate the model's efficacy. We have used various training set sizes to assess the method's performance under limited data conditions. The best model is determined based on the highest accuracy, sensitivity, specificity, and lowest Brier score. In the subsequent analysis stage, significant variables are selected using Random Forest (RF) and LASSO. The algorithm is executed 1000 times to compute average values for sensitivity, specificity, Brier score (BS), and error rate. These averages are used to assess and compare the algorithm's performance.

It is strongly advised to employ a model distinguished by its high accuracy, sensitivity, specificity, and a notably low Brier score when predicting depression in COVID-19 patients. The results of the preliminary stage are meticulously detailed in Table 1, shedding light on the efficacy of the LASSO model. Impressively, LASSO demonstrates better performance, boasting a remarkably low Brier score of 0.0033, coupled with impressive accuracy (0.9667), sensitivity (0.9618), and specificity (0.91963). To further visualize the robustness of these results, the outcomes from the extensive set of 1000 experimental runs have been vividly depicted in the accompanying box plots, as illustrated in Figure 1.

Table 1. Accuracy, Sensitivity, specificity and Brier score of the data using 70% Training part and 30% testing part.

Models	Metrics			
	Accuracy	Sensitivity	Specificity	BS
RF	0.958	0.924	0.979	0.058
k-NN	0.899	0.755	0.990	0.069
RP	0.937	0.958	0.921	0.013
SVM	0.959	0.966	0.956	0.464

LASSO	0.967	0.962	0.920	0.003

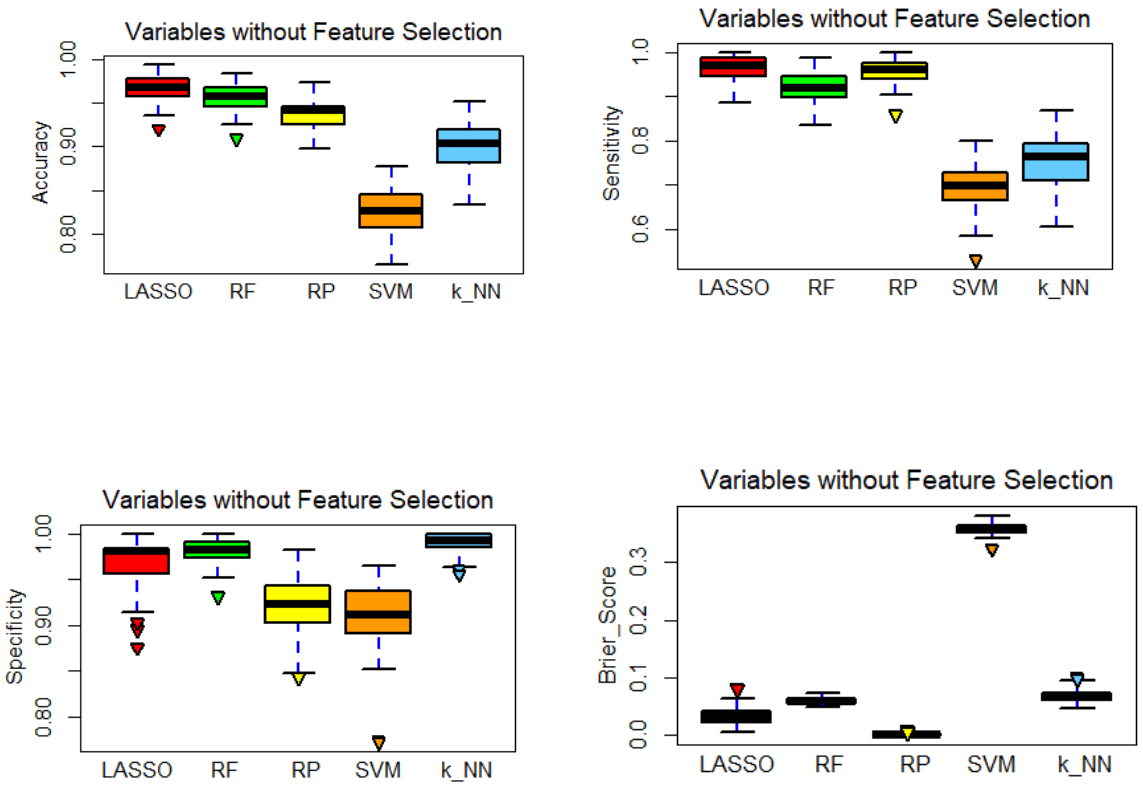

Figure 1: Accuracy, sensitivity, specificity and Brier score of LASSO, Randon Forest (RF), Random Projection Ensembles (RP), Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN) for 70% training and 30% testing size.

Feature Selection Using Random Forest

In this paper, the importance of features has been discussed through Random Forest and features with the highest degree of predictability are chosen. A declining Gini score represents the relative importance of a variable to the model. A breakdown of the mean decline in Gini impurity brought on by each variable is displayed in Table 2. According to the Gini index importance calculated using Random Forest, the variables are ordered in descending order of magnitude. As indicated in Table 3, these variables have been employed in subsets of the provided models, such as, Top 3, 6, 9, 12, and 15 selected features to investigate their effect on depression prediction in COVID-19 patients. The performance of each model is evaluated following a second run with selected feature subsets. It is clear that Random Forest outperforms other models in this scenario.

Table 2: Gini impurity corresponding to each feature.

No	Features	G.I	No	Features	G.I
1	Stick	11.99	21	Uncertainty paralyses	4.74
2	Uncomfortable	11.79	22	Visiting	4.71
3	Heart Races	10.96	23	Absent	4.35
4	Doubt	9.98	24	weight	3.89
5	Death Costs	9.38	25	Loss mind	3.88
6	Dangerous	9.29	26	Irritation	3.82
7	Distress	8.69	27	physiological	3.48
8	Afraid	8.13	28	Beloved suffers	3.14
9	Unpleasant death	8.05	29	Social media	2.57
10	Sleeping	8.02	30	Transplant	1.84
11	Hand clammy	7.55	31	Immune Compromised	1.83
12	Plan ahead	7.24	32	Smoker	1.46

13	Goals complete	6.98	33	Hypertension BP	1.41
14	Drill everything	6.97	34	Respiratory Gender	1.32
15	Weeping	6.82	35	Diabetic	1.29
16	Unknown fear	6.47	36	Asthma	1.29
17	Loss of life	6.36	37	Syndrome	0.99
18	Future	5.73	38	Cardiovascular diseases	0.96
19	Function	5.73	39	Age	0.74
20	Death	5.23			

Table 1: Top 3 to 15 selected features.

Top 3 Selected Features	Top 6Selelcted Features	Top 9 Selected Features	Top 12 Selected Features	Top 15 Selected Features
Stick	Stick	Stick	Stick	Stick
Uncomfortable	Uncomfortable	Uncomfortable	Uncomfortable	Uncomfortable
Heart Races	Heart Races	Heart Races	Heart Races	Heart Races
-	Doubt	Doubt	Doubt	Doubt
-	death Costs	death Costs	death Costs	death Costs
-	Dangerous	Dangerous	Dangerous	Dangerous
-	-	Distress	Distress	Distress
-	-	Afraid	Afraid	Afraid
-	-	Unpleasant death	Unpleasant death	Unpleasant death
-	-	-	Sleeping	Sleeping
-	-	-	Hand clammy	Hand clammy
-	-	-	Plan ahead	Plan ahead
-	-	-	-	Goals complete
-	-	-	-	Drill everything
-	-	-	-	Weeping

Table 4: Selection of important features by Logistic Regression.

Coefficients	Estimator	Std Error	z-value	Pr (>|z|)	Coefficients	Estimator	Std Error	z-value	Pr (>|z|)
(Intercept)	-1.30	1.39	-0.009	0.993	Weeping	3.58	1.22	0.003	0.998*
Gender	-5.31	1.12	0.000	1.000	Drill everything	1.71	1.20	0.000	1.000
Age	4.69	2.38	0.000	1.000	Uncertainty paralyses	2.72	1.29	0.002	0.998*
Immune compromised	-2.59	4.80	-0.001	1.000	Function	3.18	1.81	0.002	0.999
Smoker	3.51	8.88	0.000	1.000	Future	2.22	1.45	0.002	0.999
Hypertension BP	8.68	1.26	0.000	1.000	Doubt	1.99	1.96	0.001	0.999
Diabetic	1.51	1.51	0.001	0.999	Dangerous	1.62	1.06	0.002	0.999
Asthma	-1.04	1.84	-0.001	1.000	Plan ahead	2.90	1.73	0.002	0.999
Respiratory syndrome	1.09	1.47	0.000	1.000	Death	2.82	1.22	0.002	0.998*
Cardiovascular diseases	-4.33	2.80	0.000	1.000	Goals complete	2.56	1.55	0.002	0.999
Transplant	-6.56	5.09	-0.001	0.999	Stick	1.40	1.93	0.001	0.999
Afraid	2.45	1.52	0.002	0.999	Beloved suffers	3.62	1.35	0.003	0.998*
Uncomfortable	2.97	2.19	0.001	0.999	Unpleasant death	2.27	1.26	0.002	0.999
Hand clammy	2.16	1.02	0.002	0.998*	Absent	2.23	9.54	0.002	0.998*
Loss of life	3.09	1.36	0.002	0.998*	loss Mind	1.06	7.95	0.001	0.999
Social media	1.29	1.59	0.001	0.999	Death Costs	3.90	1.21	0.003	0.997*
Heart races	2.30	7.52	0.003	0.998*	Visiting	9.61	1.55	0.001	1.000
Distress	2.18	1.52	0.001	0.999	Weight	2.01	1.39	0.001	0.999
Sleeping	3.80	1.23	0.003	0.998*	Physiological	2.80	1.38	0.002	0.998*
Unknown fear	1.42	1.38	0.001	0.999	Irritation	2.74	1.07	0.003	0.998*

Some of the features that are death cost, physiological, absent, beloved suffers, death, Uncertainty paralyses, weeping, irritation, sleeping, heart races, loss of life and hands clammy are highly significant features in terms of their P-values. A total of 622 (243 under depress and 379 normal) participants are included in the study. From Table 4 death cost is 0.003 and the P-value is 0.997 which is highly significant and have strongly influence on depression. The Z-value for physiological is 0.002 and p-value

is 0.998 which is highly significant and have strongly influence on depression. Similarly, Z-value for absent, death, uncertainty, paralyses, loss of life and hand clammy is 0.002 and p-value is 0.998 which are highly significant and have strongly influence on depression. Also, the Z-value for beloved suffers, death, heart races, irritation and weeping is 0.003 and p-value is 0.998 which are highly significant and have a strong influence on depression. Table 4 contains information about significant predictors that are all very important and highly significant. The model's effectiveness based on the top 3 and top 6 features is shown in Table 5 and it shows that LASSO performs the best among all the models in terms of accuracy. Table 6 shows the output of top 9 and top 12 features selected by Random Forest indicating that LASSO is the best model. Similarly, Table 7 displays the outcomes obtained from the selection of the top 15 features utilizing the Random Forest. These results unequivocally underscore the superiority of the LASSO model in the given context, further affirming its status as the most effective model for this particular dataset. Upon careful examination of the tables, it becomes glaringly apparent that LASSO surpasses the performance of other methods, namely k-NN, SVM, Random Projection Ensembles, and Random Forest. The robustness and efficiency demonstrated by LASSO emphasize its superiority over the competing techniques, establishing it as the method of choice for this specific analysis.

Table 5: Models result on top 3 and 6 variables selected by Random Forest in terms of average Brier score, accuracy, specificity and sensitivity.

Models	Top 3 Selected Features				Top 6 Selected Features			
	Accuracy	Sensitivity	Specificity	BS	Accuracy	Sensitivity	Specificity	BS
RF	0.825	0.690	0.909	0.135	0.889	0.813	0.938	0.086
SVM	0.820	0.691	0.903	0.361	0.889	0.807	0.942	0.410
k-NN	0.820	0.637	0.937	0.132	0.868	0.771	0.926	0.099
RP	0.796	0.783	0.808	0.003	0.864	0.817	0.901	0.002
LASSO	0.826	0.680	0.922	0.174	0.890	0.833	0.927	0.110

Table 6. Models result on top 9 and 12 variables selected by Random Forest in terms of average Brier score, mean error rates, specificity and sensitivity.

Models	Top 9 Selected Features				Top 12 Selected Features			
	Accuracy	Sensitivity	Specificity	BS	Accuracy	Sensitivity	Specificity	BS
RF	0.899	0.847	0.933	0.075	0.920	0.862	0.957	0.067
SVM	0.903	0.859	0.931	0.429	0.920	0.869	0.946	0.438
k-NN	0.892	0.828	0.934	0.082	0.899	0.807	0.958	0.076
RP	0.887	0.870	0.900	0.002	0.896	0.900	0.894	0.002
LASSO	0.914	0.880	0.937	0.086	0.922	0.883	0.947	0.078

Table7: Top 15 features selected by the Random Forest in terms of average Brier score, accuracy, specificity and sensitivity.

Models	Top 15 Selected Features			
	Metrics			
	Accuracy	Sensitivity	Specificity	BS
RF	0.935	0.901	0.959	0.059
SVM	0.937	0.907	0.956	0.453
k-NN	0.898	0.783	0.970	0.073
RP	0.923	0.934	0.915	0.001
LASSO	0.938	0.908	0.958	0.062

Features Selection Using LASSO

In Table 8, the variables selected by LASSO with estimated coefficients are listed. LASSO has also been used as a feature selection method. The concept behind using LASSO regression to choose features is simple. We run a LASSO regression on a scaled version of our dataset, and we only consider features with coefficients greater than 0. Table 9 demonstrates that utilizing the chosen features, the LASSO approach beats Random Projection, SVM, k-NN, and Random Forest. The LASSO model is identified as the best choice, indicated by its minimal Brier score and maximal accuracy values.

Table 8: Features selection by LASSO.

No	Features		No	Features	
1	Unpleasant death	1.89	21	Sleeping	0.62
2	Future	1.64	22	Social media	0.61
3	Heart Races	1.49	23	Uncertainty paralyses	0.58
4	death	1.36	24	Hand clammy	0.55
5	Afraid	1.31	25	Loss of life	0.37
6	visiting	1.24	26	Distress	0.29
7	Plan Ahead	1.15	27	Loss mind	0.18
8	weight	1.15	28	Immune compromised	0.17
9	Stick	1.09	29	Transplant	0.10
10	Absent	1.07	30	Dangerous	0.01
11	Unknown fear	0.97	31	Gender	0.00
12	death Costs	0.92	32	Smoker	0.00
13	Physiological	0.91	33	Hypertension BP	0.00
14	Drill everything	0.78	34	Diabetic	0.00
15	Function	0.77	35	Cardiovascular Diseases	0.00
16	Irritation	0.75	36	Goals complete	0.00
17	Uncomfortable	0.73	37	Respiratory syndrome	-0.54
18	Weeping	0.71	38	Asthma	-0.68
19	Doubt	0.64	39	Age	-0.74
20	Beloved Suffers	0.63			

Table 9: Models performance in terms of average Brier score, accuracy, specificity and sensitivity after selecting features using LASSO.

Models	Accuracy	Sensitivity	Specificity	BS
RF	0.961	0.927	0.983	0.057
SVM	0.961	0.975	0.953	0.466
k-NN	0.903	0.761	0.993	0.067
RP	0.942	0.960	0.929	0.011
LASSO	0.971	0.966	0.973	0.003

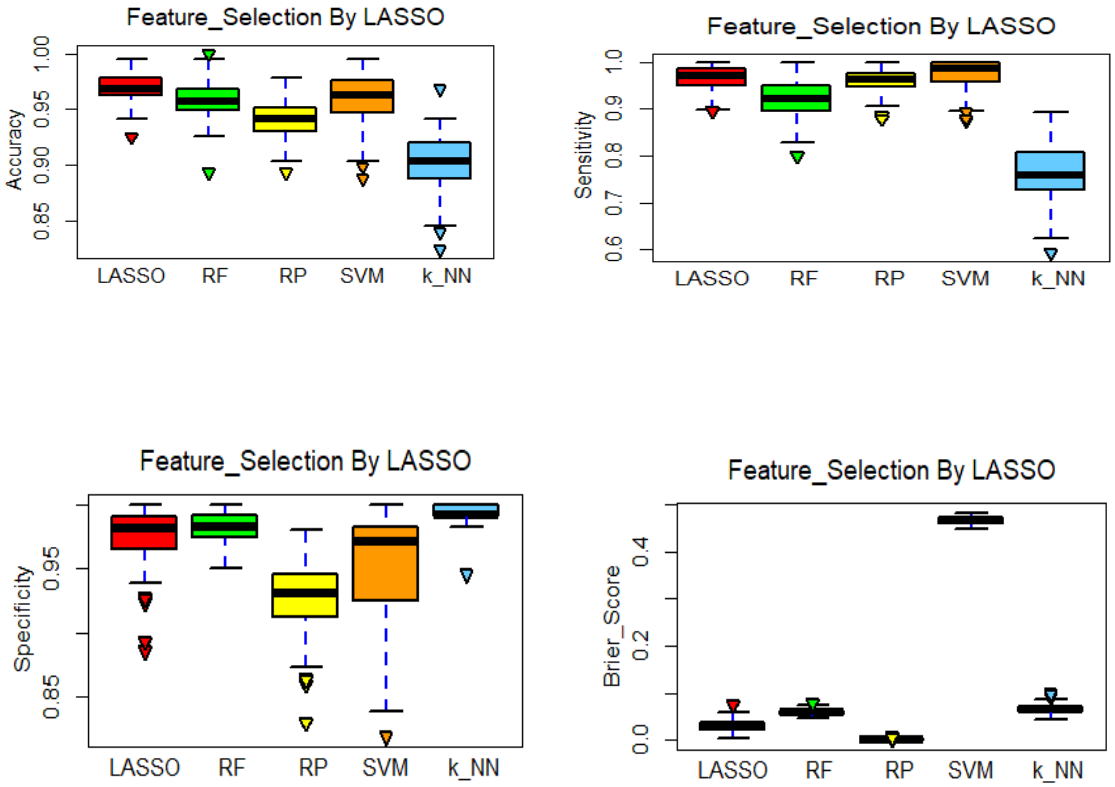

Figure 2: Accuracy, sensitivity, specificity and Brier score of LASSO, Randon Forest (RF), Random Projection Ensembles (RP), Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN) for 70% training and 30% testing size.

Boxplots have been generated to conduct a comprehensive analysis of various classifiers efficacy with feature selection and without feature selection. These boxplots illustrate accuracy metrics like sensitivity and specificity for 15 genes across k-NN, SVM, Random Forest, and Random Projection Ensembles classifiers, utilizing 70% of the data for training and 30% for testing. In statistical research, sensitivity boxplots, as depicted in Figures 1-2, are instrumental in identifying specific areas within a model or dataset that react sensitively to distinct inputs. They provide valuable insights into the model's responsiveness to varying input factors and offer a guide for refining the model or dataset to enhance performance. The boxplot's whiskers denote the minimum and maximum values within a specified range from the box, while the box itself delineates the 25th and 75th percentiles. The median is represented by a line inside the box. Similarly, the specificity boxplot (shown in Figures 1-2) illustrates how alterations in output values influence input variables, rather than how changes in inputs affect the model's output or dataset. For predictive purposes, a recommended model exhibits high accuracy, sensitivity, specificity, and a low error rate. Figure 1 depict the box-plots showing the LASSO performs better than any other technique used in the paper.

k-Means Clustering Analysis

This paper has considered binary classification based on the given depression levels. However, one could also access a more reasonable number of groups into which the respondents could be divided based on their depression levels. For this purpose, k-means clustering could be done by finding the optimal number of natural groups. The plot given below in the figure shows that the optimal number of groups is 6. As a result, compared to other k's, the between ss/total ss ratio for k = 6 tends to change more gradually. Figure 3 shows the sum of the squares on the y-axis, coupled with the number of clusters on the x-axis. A knee in the plot, which shows that adding another cluster does not significantly enhance the partition, is used to estimate the ideal number of clusters. This method strongly suggests 6 clusters. It may, therefore, be advisable to apply multi-classification with different levels of depression, such as normal

depressed, sliding depressed, moderately depressed, and severely depressed. By using the relationship between class labels, multi-classification has the potential to improve their performance.

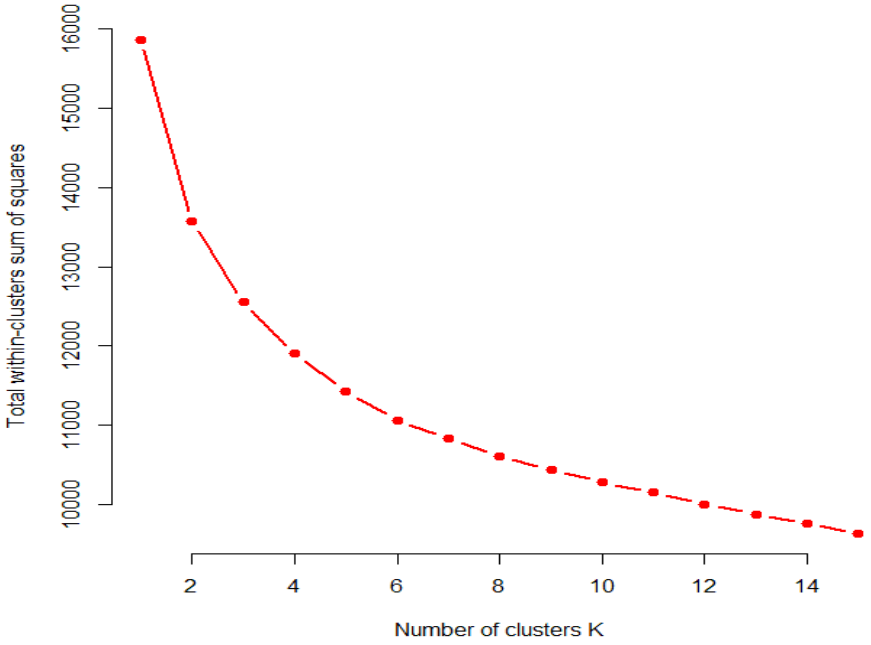

Figure 3: Elbow method plot for obtaining optimum number of clusters.

5. Conclusion

This paper investigated machine learning tools for prediction and classification of depression patient during fourth wave of COVID-19 in Pakistan. Pakistan experienced severe outbreaks, with KPK emerging as a hotspot for the pandemic. In the KPK region, two noticeable cities, Mardan and Nowshehra, have been chosen. This study covered patients who had clinical recovery and biological clearance after being hospitalised for COVID-19 illness. The present study focuses on a cross-sectional study conducted from July 2021 to September 2021. The goal of this study was to use methods based on machine learning to select, predict, and describe the effect of depression in covid-19 patients. A variety of depression prediction models have been used in the study, including k-NN, Random Forest, Random Projection Ensembles, Support Vector Machine (SVM), and LASSO. Random Forest and LASSO have been used to select the most important features. Features are ranked according to the estimations of the LASSO coefficients, LASSO outperforms all other approaches in terms of accuracy and Brier score. Furthermore, the majority of the significant variables selected by LASSO and logistic regression were the same. To establish the optimal number of natural groups for the unsupervised learning strategy, a k-mean clustering analysis was performed. The binary grouping based on the given depression levels was examined in the present study. As a result, when k = 6 is compared to different k's, the ratio moves slowly and remains constant.

This study could be beneficial for the government's health-care agencies to diagnose depression earlier and identify risk factors for depression. Using feature selection methods, the key factors associated with depression and anxiety are discovered. As a result, specialised interventions may be more helful. In the future, adding more informative variables to the approaches could allow for the inclusion of additional questions related to the medical history of depression. Furthermore, in the quest for a significantly improved model, alternative machine learning techniques like artificial neural networks could be explored.Employing multi-classification to categorize individuals based on varying degrees of depression has the potential to enhance overall performance. In this context, binary classification is utilized to assess different levels of depression.

References

1. Reynolds EH, Wilson JVK. Depression and anxiety in Babylon. Journal of the Royal Society of Medicine. 2013;106(12):478–481.
2. Khurshid B, Farman S, Parveen Z, Assad M, Shams U, Faiz N, et al. COVID-19 AND SARS-COV-2: WHAT WE KNOW SO FAR. Journal of Population Therapeutics and Clinical Pharmacology. 2023;30(17):1745–1760.
3. Organization WH, et al. COVID-19 and violence against women: what the health sector/system can do, 7 April 2020. World Health Organization; 2020. 364
4. Lee J. Mental health effects of school closures during COVID-19. The Lancet Child & Adolescent Health. 2020;4(6):421.
5. Alimoradi Z, Ohayon MM, Griffiths MD, Lin CY, Pakpour AH. Fear of COVID-19 and its association with mental health-related factors: systematic review and meta-analysis. BJPsych Open. 2022;8(2):e73.
6. Amon JJ. COVID-19 and detention: respecting human rights. Health and human rights. 2020;22(1):367.
7. Brink A. Depression and Loss: A Theme in Robert Burton's "Anatomy of Melancholy"(1621). The Canadian Journal of Psychiatry. 1979;24(8):767–772.
8. Tesaˇrov´a D. Aulus Cornelius Celsus and a regimen. Casopis Lekaru Ceskych.2018;157(5):263–267.
9. De Girolamo G, Ferrari C, Candini V, Buizza C, Calamandrei G, Caserotti M, et al. Psychological well-being during the COVID-19 pandemic in Italy assessed in a four-waves survey. Scientific Reports. 2022;12(1):17945.
10. Nkwayep CH, Bowong S, Tewa J, Kurths J. Short-term forecasts of the COVID-19 pandemic: a study case of Cameroon. Chaos, Solitons & Fractals. 2020;140:110106.
11. Pak A, Adegboye OA, Adekunle AI, Rahman KM, McBryde ES, Eisen DP. Economic consequences of the COVID-19 outbreak: the need for epidemic preparedness. Frontiers in public health. 2020;8:241.
12. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic. JAMA network open. 2020;3(9):e2019686–e2019686.
13. Shams S. Coronavirus: Is Pakistan taking COVID-19 too lightly. Bonn: DW Akademie. 2020;18.
14. Lanfredi M, Dagani J, Geviti A, Di Cosimo F, Bussolati M, Rillosi L, et al. Risk and protective factors associated with mental health status in an Italian sample of students during the fourth wave of COVID-19 pandemic. Child and Adolescent 392 Psychiatry and Mental Health. 2023;17(1):78.
15. Pavliuk O, Kolesnyk H. Machine-learning method for analyzing and predicting the number of hospitalizations of children during the fourth wave of the COVID-19 pandemic in the Lviv region. Journal of Reliable Intelligent Environments. 2023;9(1):17–26.
16. Rogers JP, Chesney E, Oliver D, Pollak TA, McGuire P, Fusar-Poli P, et al. Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic. The Lancet Psychiatry. 2020;7(7):611–627.
17. Senthilkumar D, Paulraj S. Prediction of low birth weight infants and its risk factors using data mining techniques. In: Proceedings of the 2015 international conference on industrial engineering and operations management; 2015. p.186–194.
18. Kang L, Li Y, Hu S, Chen M, Yang C, Yang BX, et al. The mental health of medical workers in Wuhan, China dealing with the 2019 novel coronavirus. The Lancet Psychiatry. 2020;7(3):e14.
19. Lai J, Ma S, Wang Y, Cai Z, Hu J, Wei N, et al. Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019. JAMA network open. 2020;3(3):e203976–e203976. 411
20. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. The lancet. 2020;395(10227):912–920.
21. Rana W, Mukhtar S, Mukhtar S. Mental health of medical workers in Pakistan during the pandemic COVID-19 outbreak. Asian journal of psychiatry. 2020;51:102080.
22. Banerjee D. The COVID-19 outbreak: Crucial role the psychiatrists can play. Asian journal of psychiatry. 2020;50:102014.
23. Bai Y, Lin CC, Lin CY, Chen JY, Chue CM, Chou P. Survey of stress reactions among health care workers involved with the SARS outbreak. Psychiatric services. 2004;55(9):1055–1057.
24. Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR)[Internet]. 2020;9(1):381–386.
25. Sedgwick P. Convenience sampling. Bmj. 2013;347.
26. Association WM, et al. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. Jama. 2013;310(20):2191–2194.
27. Israel GD. Determining sample size. 1992;.

28. Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. In: 2016 IEEE international conference on advances in computer applications (ICACA). IEEE; 2016. p. 18–20.
29. Breiman L. Random forests. Machine learning. 2001;45:5–32.
30. Khan Z, Gul N, Faiz N, Gul A, Adler W, Lausen B. Optimal trees selection for classification via out-of-bag assessment and sub-bagging. IEEE Access. 2021;9:28591–28607.
31. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16(10):906–914.
32. Silverman BW, Jones MC. E. fix and jl hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). International Statistical Review/Revue Internationale de Statistique. 1989; p. 233–238.
33. Cover T, Hart P. Nearest neighbor pattern classification. IEEE transactions on information theory. 1967;13(1):21–27.
34. Schclar A, Rokach L. Random projection ensemble classifiers. In: Enterprise Information Systems: 11th International Conference, ICEIS 2009, Milan, Italy, May 6-10, 2009. Proceedings 11. Springer; 2009. p. 309–316.
35. Edwards AW, Cavalli-Sforza LL. A method for cluster analysis. Biometrics. 1965; p. 362–375.
36. Frades I, Matthiesen R. Overview on techniques in cluster analysis. Bioinformatics methods in clinical research. 2010; p. 81–107.
37. Duran BS, Odell PL. Cluster analysis: a survey. vol. 100. Springer Science & Business Media; 2013.
38. Romesburg C. Cluster analysis for researchers. Lulu. com; 2004.
39. Wright RE. Logistic regression. 1995.
40. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395–2399.
41. Nick TG, Campbell KM. Logistic regression. Topics in biostatistics. 2007; p. 273–301.
42. Menard S. Applied logistic regression analysis. 106. Sage; 2002.
43. Field A. Logistic regression. Discovering statistics using SPSS. 2009;264:315.