The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): 655-682, 2026 www.thebioscan.com

Identification of Health Crisis During COVID-19 Using Machine Learning Techniques

Ria Pyne[1], Piyali De[2], Avijit Kumar Chaudhuri[3], Payel Sengupta[4], Debmalya Mukherjee[5], RANJAN BANERJEE[6], AMARTYA GHOSH[7], PRANAB GHARAI[8], Sudip Pandit[9]

[1,2,4,5,6,7]Assistant Professor, [3]Professor & HOD, [8]Assistant Professor/ Research Scholar
[1,2,3,4,5,6,7,8,9]Computer science and engineering, Brainware University
pyneria.gmail.com, me.piyalide@gmail.com, c.avijit@gmail.com, payel9433@gmail.com, dbml.mukherjee@gmail.com
rbkpccst@gmail.com, com.amartya@gmail.com, pranab.g10@gmail.com

0009-0007-8323-8354, 0009-0007-2631-615X, 0000-0002-5310-3180, 0000-0003-3981-5971, 0009-0006-9946-0964
0009-0003-1950-7530, 0009-0002-2504-3325, 0009-0004-3602-5223
[9]Student, CSE 4th year, Techno Engineering College, Banipur
panditsudip969@gmail.com

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.pp655-682

KEYWORDS

Logistic regression, Mental Health, Prediction of Depression, Classification, Dataset, Machine Learning

Received on:
10-03-2026

Accepted on:
08-04-2026

Published on:
03-05-2026

Abstract

Psychiatry began to develop empirical approaches only in the past 30 years to conceptualize, assess, and document positive mental health. In society, we accept mental disorders as normal and try to hide them. In this article, the authors came across the fact of how severe Mental Illness affects our lives during COVID-19 and the measures that can be taken to detect the same. Authors have developed methods to predict an individual's depressive factors. Primary data have been collected from people aged 18 to 59. Neurotic suffering cannot be fully understood without understanding real health and a sound approach towards life so, the authors have taken the approach of all the factors including depression, anxiety, stress, eating disorder, short temper, hallucination[1], loneliness, suicidal attempt, constant guilt feeling, fearful all these and more, realistic disturbing state and trait that results in various mental diseases have been considered and analysed. The authors used Gain Ratio with Random Forest, Naïve Bayes, Decision Tree, and SVM, and Information Gain with Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, and SVM to predict an individual's depression.

2. INTRODUCTION

In our country, most people suffered from depression during COVID-19. A depressed person can be a family person or an employee, and the severe effects can be seen in every aspect of life due to depression. Authors have taken 90-100 attributes on which it has been observed that one goes into depression, including one or multiple attributes. Authors have also observed that many of the depressed people went to psychiatrists and have healed themselves completely, but at the same time, the new generation took the help of our traditional techniques like Yoga and meditation, and have gained freedom from

depression. However, some people face depression throughout their lives [2], and no healing approach can rectify the same. So, it is obvious that depression is becoming a very challenging mental health issue, and the authors have found all these attributes, including childhood abuse, heartbreak, divorce, past trauma, and many more realistic challenges, lie behind depression. Nowadays, machine learning techniques (MLT) are widely used across various fields, including [3,5] science, medicine, and the social sciences. MLTs help to determine a mathematical relation between the independent and dependent variables. We aim to develop a prediction of the factors of depression that would predict the probability of a person getting depressed based on their mental state and the response that they generate as a result of those attributes, like emotional overeating, sleeplessness [4], and many more. In this study, the prediction of a health crisis is made using 5 state-of-the-art classification methods.

3. Relevant Literature:

Year	Method	classification accuracy(%)	Sensitivity/Specificity	Precision	ROC/AUC Area	Kappa Statistic	Features Selected
[6]	Bagging Algorithm, J48DT	Bagging(81.41%)	Bagging(74.93/86.64)	X	X	X	
		J48 DT(78.90%)	J48 DT (72.01/84.48)				
[7]	One Dependency Augmented	ODANB(80.46%)	X	x	X	x	

	Naïve Bayes Classifier (ODANB), NB	Naive Bayes(84.14%)					
[8]	Binary Particle Swarm Optimization (BPSO)and Genetic Algorithm (GA)	81.46	x	x	x	x	
[9]	Cost-Sensitive Case-based Reasoning (CSCBR)	X	0.90/0.87	x	x	X	
[10]	Artificial Neural Networks	70%	x	x	x	x	
[11]	Weighted Fuzzy Rules	62.35%	0.766/0.766	x	x	x	
[12]	Rotation Forest, Levenberg-Marquardt	91.20%	✓	X	✓	x	
[13]	Without Voting K-NearestNeighbor	97.10%	0.935/0.987	x	x	X	13
[14]	DT	83.90%	81.6/83	X	x	x	
[15]	CART	83.49%	✓	✓		✓	
[16]	NB ,DT, J48	NB(83.40%)	✓	x	x	x	

		DT(76.20%) J48(77.50%)					
[17]	Majority (Vote Based)	81.82%	0.7368/0.9286	x	X	X	
	Ensemble Technique						
[18]	Feature selection based on the Least Squares Twin Support Vector Machine (LSTSVM)	85.59%	0.8571/0.9091	X	X	x	11
[19]	J48	56.76%	x	x	X	x	
[20]		91.10%	1/0.84	X	✓	x	
[21]	Dynamic time warping (DTW)	x	95.9/94	x	X	X	
[22]	Minimum Redundancy Maximum Relevance Feature Selection (MRMR/FS)	84.85%	x	x	X	x	
	LR&SVM						
Our Study		**100**	**0.80/1**	**x**	**X**	**0.8**	**6**

4. Dataset Description:

No.	**Attributes**	**Description**

1	Your Age	Age of the person
2	Your Gender	Gender of the person
3	Teacher	Teacher of the person
4	Home Maker	Is person homemaker or not
5	College Student	is the person a college student or not
6	School Student	is the person School student or not
7	Business Person	Does the person do any business
8	Unemployed or job seeker	is the person is unemployed
9	Health Service Provider	Is the person's health service provider
10	Service Provider	Is the person service provider
11	Other Occupations	is the person do any other occupations
12	Single	is the person is single
13	Married	is the person is married
14	Divorced	is the person is divorced
15	Separated	is the person is separated
16	Widowed	is the person is widowed
17	In a Relationship	is the person is in any relationship
18	Anxiety	is the person have any anxiety
19	Depression	is the person is depressed
20	Post-Traumatic Stress	faced by the person
21	Eating Disorders	is the person have eating disorders
22	Neurodevelopmental Disorder	faced by the person
23	Others Problems	faced by the person
24	Faced this situation In your past	faced by the person
25	Faced this situation currently	faced by the person
26	Happened only once	faced by the person
27	Happens frequently	faced by the person

28	Facing these problems for less than 1 month	faced by the person
29	Facing these problems for more than 1 month and less than 6 months	faced by the person
30	Facing these problems for more than 6 months and less than 1 year	faced by the person
31	Facing these problems for more than a year	faced by the person
32	Insecurity	faced by the person
33	Irritability	faced by the person
34	Loneliness	faced by the person
35	Convolution	faced by the person
36	Hallucination	faced by the person
37	Thought of self-harm or suicide	faced by the person
38	Aggression	faced by the person
39	Nervousness or excessive sweating	faced by the person
40	Lack of concentration or confidence	faced other problems by the person
41	Anything doing more than its needed	faced other problems by the person
42	Difficulty in berating or shortness of breath	faced other problems by the person
43	Fatigue	faced other problems by the person
44	Insomnia	faced other problems by the person
45	Restlessness	faced other problems by the person
46	Loss of appetite	faced other problems by the person
47	Intrusive thoughts	faced other problems by the person
48	Guilt without any reasons	faced other problems by the person
49	Sudden weight gain or loss	faced other problems by the person
50	Excessive crying or eating or sleeping	faced other problems by the person

51	Hopelessness or loss of interest in activities	faced other problems by the person
52	Other symptoms	faced other problems by the person
53	Drug and alcohol misuse	faced other problems by the person
54	Excessive family pressure	faced other problems by the person
55	Severe or long-term stress	faced other problems by the person
56	Social isolation or loneliness	faced other problems by the person
57	Homelessness or poor housing	faced other problems by the person
58	Unemployment or losing your job	faced other problems by the person
59	Being the victim of a violent crime	faced other problems by the person
60	Excessive mental stress for studies	faced other problems by the person
61	Childhood abuse trauma or neglect	faced other problems by the person
62	Social disadvantage poverty or debt	faced other problems by the person
63	Being a long-term career for someone	faced other problems by the person
64	Insecurity for excessive self-obsession	faced other problems by the person
65	Bereavement	faced other problems by the person
66	Having a long-term physical health condition	faced other problems by the person
67	Domestic violence bullying or other abuse as an adult	faced other problems by the person
68	Experiencing discrimination and stigma including racism	faced other problems by the person
69	Being involved in a serious incident in which you feared for your life	faced other problems by the person
70	Physical causes – for example a head injury or a neurological condition	faced other problems by the person
71	Other causes	faced other problems by the person

72	Have you ever visited a professional regarding your mental health?	faced other problems by the person
73	Is your problem diagnosed?	Reasons of the problem
74	Are your problem resolved?	Reasons of the problem
75	Are you still in treatment?	Reasons of the problem
76	Less than 1 month in treatment	Reasons of the problem
77	More than 1 month and less than 6 months in treatment	Reasons of the problem
78	More than 6 months and less than a year in treatment	Reasons of the problem
79	More than a year in treatment	Reasons of the problem
80	Is your problem fixed by consulting with a doctor or psychiatrist or therapist	Reasons of the problem
81	Is your problem fixed with helped by friends or relatives	Reasons of the problem
82	Is your problem fixed by self-realization or self-motivation	Reasons of the problem
83	Is there a history of mental disorder in your family?	Reasons of the problem
84	Do your grandfather had any kind of mental illness?	Reasons of the problem
85	Do your grandmother had any kind of mental illness?	Reasons of the problem
86	Do your father had any kind of mental illness?	Reasons of the problem
87	Do your mother had any kind of mental illness?	Reasons of the problem
88	Do your brother had any kind of mental illness?	Reasons of the problem

89	Do your sister had any kind of mental illness?	Reasons of the problem
90	Other family members who had mental illness	Reasons of the problem
91	Health Crisis	faced by the person

5. Research Methodology:

I. Feature selection methods

In machine learning, several methods are used to choose features. These feature selection methods remove outdated or redundant features from the given dataset. There are two types of feature selection methods: wrapper and filter methods.

The wrapper checks and selects attributes based on the target learning algorithm's accuracy estimates. The wrapper checks the function space by omitting certain characteristics and evaluating the effect of element omission on predictive metrics.

On the other hand, the filter method used the general functions and features of data. The filter explicitly uses the number correlation between a set of characteristics and the target function. The similarity between the characteristic and the target variable is determined by the target variable value. Filter-based solutions are quicker and more flexible than wrapper-based methods. We also achieve low computational complexity using wrapper-based methods.

This paper uses two significant filter approaches-

1)Information Gain and 2) Gain Ratio

The following section describes the characteristics of these approaches.

Information Gain:

Information gain is used in a variety of applications to evaluate features. This method controls all functions according to user-defined target values. Information gain (relative entropy) is a function that depicts the difference between two probability distributions in probability theory and information theory [23]. This approach assesses a feature M, then by measuring the amount of information obtained from the factor N of a class or group defined as follows-

algorithms and 3 Feature Selection (FS) methods.

$$I(M) = H(P(N)) - H\left(P\left(\frac{N}{M}\right)\right)$$

In particular, it tests the difference between the marginal distribution of measurable N on the assumption that it is independent of **H (P (N))** and

the conditional distribution of N on the assumption that it is dependent on $H\left(P\left(\frac{N}{M}\right)\right)$.

If M is not expressed differently, N will be independent of M, which means that M will have limited benefit value for data, and vice versa.

Gain ratio:

To reduce bias in information gain, we use the gain ratio method, which accounts for the number and size of divisions. To correct information gain, the gain ratio recognizes the inherent knowledge of a split[23]. The attribute value decreases as the information intrinsically increases.

The gain ratio is calculated using the following equation.

$$GR(attribute) = \frac{Gain(attribute)}{Intrinsic - Info(attribute)}$$

II. Classification Algorithms

Different classification algorithms have different strengths and weaknesses. Here we use 5 classification algorithms described below-

Naïve Bayes:

This classifier relies on Bayes' theorem and primary probabilistic classification, making the strong assumption that all variables are independent. That is, the presence or absence of a particular variable is assumed to be unrelated to the presence or absence of other variables in the data set. A Naive Bayes classifier deals only with conditional probabilities. Bayes' theorem calculates the probability that an event will occur, given the likelihood that another event has already occurred [23].

In Naïve Bayes, if X is an event dependent on event Y, Bayes' theorem is used to determine the conditional probability of X given Y.

The algorithm determines the probability by counting the cases in which X and Y occur together and dividing by the cases in which either X or Y occurs. Naïve Bayes efficiently estimates parameters with limited training data.

Decision Tree(DT)

It is a more popular and widely used data mining algorithm derived from the ID3 algorithm. These algorithms make the set of rules comprehensible, and such is preferred in many applications.

Training data is a compilation of Sd = sd1, sd2, sd3, sd4 … … .. of already classified samples. Each sample sdi consists of an m-dimensional vector x1,i , x2,i , x3,i , … … … … xm,i , where the xj represents attribute values or features of the sample, as well as the class in which Sdi falls. J48 selects the data attribute at each tree node that most effectively splits the sample set into subsets enriched in either class. The splitting criterion for the sub-tree is the average information gain (difference in entropy) [23]. The attribute having the highest uniform value of knowledge is selected when making the decision.

Then the J48 algorithm recurs on partitioned sub-lists.

Support Vector Machine (SVM):

An SVM is a machine learning method that analyses data and learns patterns for classification and regression analysis. Compared with other classifiers, SVM has shown better performance across a range of applications. In an infinite-dimensional space, SVM creates a set of hyperplanes, or a single hyperplane, used to classify, regress, or perform other functions. It is very useful for classifying data. SVMs are also known as linear classifiers divides the data into classes by maximizing the interclass distance using an ideal hyperplane [23].

Classifying data from different sets is the most common use of machine learning techniques. In SVM, a data point is set as an x-dimensional vector, and it differentiates such points with an (x-1) dimensional hyperplane.

During training, SVM learns to map vectors to their classes in high-dimensional space. The main aim of SVM is to reduce error rates through its minimization-based approach.

4) Logistic Regression(LR):

The LR classifier uses linear regression. It sets the logistic model parameters. It connects the values of explanatory variables with the probability of a case. LR primarily aims to provide a simple way to classify things based on their probability [23].

The main drawback of LR is that it cannot handle variables with nonlinear or interactive effects. LR estimates a result that can only be one of two values. It uses a method called maximum likelihood ratio to find which variables are important.

LR s useful when we can predict the presence or absence of a dependent feature based on the values of a set of independent features.

Random Forest (RF):

RF is a machine learning method that builds several decision trees using random sampling. It uses voting to make the final prediction. This group vote is less risky and less affected by unusual data than a single decision tree, and it reduces uncertainty when information is limited or unreliable. RF can handle many input values without dropping any important data.

Rf uses permutation test ideas to assess the importance of each factor by comparing prediction errors. Each tree gets a score, and the group's overall score is averaged. The result is then divided by the standard deviation to create useful differences in decision trees. This method also involves picking features at random. The forest's growth depends on the trees themselves [23].

III. Clasificassion

A. Sensitivity and Specificity Analysis:

All health care, diagnoses, and procedures are imprecise and subject to error. The best method for

evaluating this medical error is the sensitivity and specificity test. A medical examination and diagnosis should be differentiated. A test is one of many terms used to describe the medical condition; a diagnosis is a mixture of observations and numerous tests that demonstrate the patient's pathophysiology. Sensitivity/Specificity assessment is necessary for testing and diagnosis, as shown in equations 3 and 4. There are test outcomes and an objective indicator of truth or theory in the medical dataset, and an analysis of sensitivity and specificity. The consistency of test checks performed by if-then rules governs the similarity between a new test result and the expected value. The best-fit rule specifies the class composition of the study when defining a test example. True Positive (TP) specifies the number of correct positive predictions (classifications); True Negative (TN) indicates the number of correct negative predictions; False Positive (FP) represents the number of incorrect positive predictions; False Negative (FN) identifies the number of incorrect negative predictions.

The two measures are:

$$\text{Sensitivity} = \frac{True\ Positive\ (T_P)}{True\ Positives(T_P)+False\ Negatives(F_N)} \quad (3)$$

$$\text{Specificity} = \frac{True\ Negatives(T_N)}{False\ Positives(F_P)+True\ Negatives(T_N)} \quad (4)$$

Sensitivity measures a test's tendency to be positive when the condition is present, or how many positive examples of tests are remembered [66]. Sensitivity measures, in other words, test how often people find what they are looking for. It falls within several near-synonyms: false-negative rate, recall, type I error, type II error, omission error, or alternative hypothesis. Specificity tests a test's tendency to be negative if the condition does not exist, or how many negative test instances are omitted.

In other words, specificity measures how often they search for what anyone can find. It falls within various near-synonyms: false-positive rate, accuracy, type I error, type II error, commission error, or null hypothesis. Predictive precision provides an overall assessment. Only findings with high values across all three measures can be assigned a high confidence level.

B. Kappa Statistics:

Kappa error, or the interpretation of Cohen's Kappa Statistics, is used to determine the classifier's effectiveness. Output comparisons in classification algorithms can yield incorrect results when the percentage of misses is used as the sole predictor of precision [68]. Also, the focus must be given to the possibility of mistakes when making these decisions. In this sense, the Kappa statistic is a fair predictor of classification performance, which may be due to chance. The Kappa statistic usually ranges from −1 to +1. If the classifiers measured the Kappa value

reaches '1', it is presumed the classifier's output is more realistic than by chance. Hence, the Kappa statistic is a recommended performance metric for classifier analysis. Thus, the kappa statistic is a statistical measure of two data sets 'inter-rate' conformity. The significant difference between 0 and 1 indicates better interrater agreement. For a chance of agreement, Kappa represents the convention's usual meaning. The agreement calculation is more rigorous than a simple percentage. Equation 5 describes the Kappa statistic.

$$KAPPA(K) = \frac{Percentage\_of\_agreement(P_A) - Chance\_of\_agreement(C_A)}{1 - Chance\_of\_agreement(C_A)} \quad (5)$$

Where K=1 implies the full inter-rate agreement between the classifier and ground truth, and K=0 is a certain probability of consensus. Comparing classification algorithm success rates can yield misleading results when the percentage of misses is used as the sole metric. Even when making such calculations, the cost of error needs to be considered.

IV. Result and Discussions:

The authors select two feature methodologies: Information Gain and Gain Ratio. They use several state-of-the-art classifiers, including 1) Random Forest, 2) Naïve Bayes, 3) Logistic Regression, 4) Decision Tree, and 5) SVM (Support Vector Machine).

The authors aim to achieve maximum classification accuracy (100%) using selected feature subsets and classifiers, while reducing the number of attributes from 90. They employed various train-test splits (50-50, 66-34, 80-20) and 10-fold cross-validation.

The author first selects the feature method Information gain with Random Forest, using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 75.7282%, Sensitivity of 0.75, Specificity of 0, and a Kappa value of 0. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 78.5714%, Sensitivity of 0.78, Specificity of 0, and a kappa value of 0. Then the author selects an 80-20 training-testing split on 90 attributes, yielding a total accuracy of 88.5714%, a sensitivity of 0.8, a specificity of 0, and a kappa value of 0. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 77.2947, Sensitivity of 0.77, Specificity of 0, and a kappa value of 0. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 7 attributes, yielding a total accuracy of 92.233%, Sensitivity of 0.89, Specificity of 1, and a Kappa value of 0.8167. Then the author selects a 66-34 training-test split across 7 attributes, yielding a total accuracy of 88.5714%, Sensitivity of 0.84, Specificity of 1, and a kappa value of 0.75. Then the author selects an 80-20 training-test split across 7 attributes, resulting in a total accuracy of 100%,

Sensitivity of 1, Specificity of 1, and a kappa value of 1. Then the author selects a 10-fold cross-validation split on 7 attributes, yielding a total accuracy of 100%, a sensitivity of 1, a specificity of 1, and a kappa value of 1.

Then the author selects the Information gain feature with Naïve Bayes, using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 76.699%, Sensitivity of 0.787, Specificity of 0.555, and a Kappa value of 0.3681. Then the author selects a 66-34 training-test split on 90 attributes, so the total accuracy is 74.2857%, Sensitivity is 0.813, Specificity is 0.363, and the kappa value is 0.1544. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 68.2927%, Sensitivity of 0.8125, Specificity of 0.222, and a kappa value of 0.0362. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 71.9807%, Sensitivity of 0.786, Specificity of 0.31, and a kappa value of 0.0769. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 6 attributes, yielding a total accuracy of 86.4087%, Sensitivity of 1, Specificity of 0.857, and a Kappa value of 0.3681. Then the author selects a 66-34 training-test split on 6 attributes, yielding a total accuracy of 90%, Sensitivity of 1, Specificity of 0.888, and a kappa value of 0.6154. Then the author selects an 80-20 training-test split across 6 attributes, yielding a total accuracy of 97.561%, Sensitivity of 1, Specificity of 0.972, and a kappa value of 0.8951. Then the author selects a 10-fold cross-validation training-testing split on 6 attributes, yielding a total accuracy of 93.2367%, Sensitivity of 1, Specificity of 0.925, and a kappa value of 0.6849.

Then the author selects the Information gain feature with Logistic Regression using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 63%, Sensitivity of 0.822, Specificity of 0.3841, and a Kappa value of 0.1757. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 62%, Sensitivity of 0.795, Specificity of 0.238, and a kappa value of 0.037. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 65%, Sensitivity of 0.851, Specificity of 0.285, and a kappa value of 0.1534. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 68%, Sensitivity of 0.836, Specificity of 0.363, and a kappa value of 0.2171. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 13 attributes, yielding a total accuracy of 91%, Sensitivity of 0.8, Specificity of 0.958, and a Kappa value of 0.7819. Then the author selects a 66-34 training-test split across 13 attributes, yielding a total accuracy of 90%, Sensitivity of 0.9, Specificity of 0.9, and a kappa value of 0.7657. Then the author selects an 80-20 training-test split across 13 attributes, yielding a total accuracy of 92%,

Sensitivity of 0.846, Specificity of 0.964, and a kappa value of 0.8275. Then the author selects a 10-fold cross-validation training-testing split on 13 attributes, yielding a total accuracy of 90%, Sensitivity of 0.818, Specificity of 0.934, and a kappa value of 0.7524.

Then the author selects the Information gain feature using a Decision Tree with a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 75%, Sensitivity of 0.795, Specificity of 0.4, and a Kappa value of 0.1783. Then the author selects a 66-34 training-testing split on 90 attributes, yielding a total accuracy of 71%, Sensitivity of 0.851, Specificity of 0.437, and a kappa value of 0.2959. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 78%, Sensitivity of 0.852, Specificity of 0.428, and a kappa value of 0.2664. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 71%, Sensitivity of 0.784, Specificity of 0.29, and a kappa value of 0.0614. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 13 attributes, yielding a total accuracy of 88%, Sensitivity of 0.758, Specificity of 0.932, and a Kappa value of 0.7059. Then the author selects a 66-34 training-test split across 13 attributes, yielding a total accuracy of 87%, Sensitivity of 0.818, Specificity of 0.895, and a kappa value of 0.7053. Then the author selects an 80-20 training-test split across 13 attributes, yielding a total accuracy of 90%, Sensitivity of 0.75, Specificity of 1, and a kappa value of 0.7853. Then the author selects a 10-fold cross-validation training-testing split on 13 attributes, yielding a total accuracy of 91%, Sensitivity of 0.849, Specificity of 0.935, and a kappa value of 0.7745.

Then the author selects the Information gain feature with SVM (Support Vector Machine) using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 74%, Sensitivity of 0.787, Specificity of 0.347, and a Kappa value of 0.1313. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 68%, Sensitivity of 0.862, Specificity of 0.421, and a kappa value of 0.3039. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 68%, Sensitivity of 0.833, Specificity of 0.272, and a kappa value of 0.1161. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 72%, Sensitivity of 0.828, Specificity of 0.4, and a kappa value of 0.2328. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 13 attributes, yielding a total accuracy of 93%, Sensitivity of 0.884, Specificity of 0.948, and a Kappa value of 0.8222. Then the author selects a 66-34 training-test split on 13 attributes, yielding a total accuracy of 91%, Sensitivity of 0.904, Specificity of 0.918, and a kappa value of 0.8013. Then the author

selects an 80-20 training-test split across 13 attributes, yielding a total accuracy of 92%, Sensitivity of 0.846, Specificity of 0.964, and a kappa value of 0.8275. Then the author selects a 10-fold cross-validation training-testing split on 13 attributes, yielding a total accuracy of 91%, Sensitivity of 0.849, Specificity of 0.935, and a kappa value of 0.7745.

Then the author selects the Gain Ratio feature with a Random Forest using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 75.7782%, Sensitivity of 0.757, Specificity of 0, and a Kappa value of 0. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 78.5714%, Sensitivity of 0.785, Specificity of 0, and a kappa value of 0. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 80.4878%, Sensitivity of 0.804, Specificity of 0, and a kappa value of 0. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 77.2947%, Sensitivity of 0.772, Specificity of 0, and a kappa value of 0. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 16 attributes, yielding a total accuracy of 90%, Sensitivity of 0.971, Specificity of 0.97, and a Kappa value of 0.9355. Then the author selects a 66-34 training-test split on 16 attributes, yielding a total accuracy of 95.7143%, Sensitivity of 0.956, Specificity of 0.957, and a kappa value of 0.9039. Then the author selects an 80-20 training-test split across 16 attributes, yielding a total accuracy of 90.2439%, Sensitivity of 0.846, Specificity of 0.928, and a kappa value of 0.7747. Then the author selects a 10-fold cross-validation training-testing split on 16 attributes, yielding a total accuracy of 93.7198%, Sensitivity of 0.9125, Specificity of 0.952, and a kappa value of 0.8673.

Then the author selects the Gain Ratio feature with Naïve Bayes using a 50-50 training-test split on 90 attributes, yielding a total accuracy of 76.699%, Sensitivity of 0.787, Specificity of 0.555, and a Kappa value of 0.19. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 74.2857%, Sensitivity of 0.813, Specificity of 0.363, and a kappa value of 0.1544. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 68.2927%, Sensitivity of 0.8125, Specificity of 0.222, and a kappa value of 0.0362. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 71.9807%, Sensitivity of 0.786, Specificity of 0.31, and a kappa value of 0.0769. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-testing split across 15 attributes, yielding a total accuracy of 92.233%, Sensitivity of 0.913, Specificity of 0.925, and a Kappa value of 0.7892. Then the author selects a 66-34 training-test split on 15 attributes, yielding a total accuracy of 85.7143%, Sensitivity of 0.882,

Specificity of 0.849, and a kappa value of 0.6531. Then the author selects an 80-20 training-test split across 15 attributes, yielding a total accuracy of 87.8049%, Sensitivity of 0.818, Specificity of 0.9, and a kappa value of 0.6981. Then the author selects a 10-fold cross-validation training-testing split on 15 attributes, yielding a total accuracy of 87.9227%, Sensitivity of 0.84, Specificity of 0.889, and a kappa value of 0.6694.

Then the author selects the Gain Ratio feature with Logistic Regression using a 50-50 training-test split on 90 attributes, yielding a total accuracy of 63.1068%, Sensitivity of 0.822, Specificity of 0.341, and a Kappa value of 0.1757. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 62.8571%, Sensitivity of 0.795, Specificity of 0.238, and a kappa value of 0.037. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 65.8537%, Sensitivity of 0.851, Specificity of 0.285, and a kappa value of 0.1534. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 68.599%, Sensitivity of 0.836, Specificity of 0.363, and a kappa value of 0.2171. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 15 attributes, yielding a total accuracy of 91.2621%, Sensitivity of 0.8, Specificity of 0.958, and a Kappa value of 0.7819. Then the author selects a 66-34 training-test split on 15 attributes, yielding a total accuracy of 90%, Sensitivity of 0.9, Specificity of 0.9, and a kappa value of 0.7656. Then the author selects an 80-20 training-test split across 15 attributes, yielding a total accuracy of 92.6829%, Sensitivity of 0.846, Specificity of 0.964, and a kappa value of 0.8275. Then the author selects a 10-fold cross-validation training-testing split on 15 attributes, yielding a total accuracy of 90.3382%, Sensitivity of 0.818, Specificity of 0.934, and a kappa value of 0.7524.

Then the author selects the Gain Ratio feature for a Decision Tree using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 71.8447%, Sensitivity of 0.795, Specificity of 0.4, and a Kappa value of 0.1783. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 75.7143%, Sensitivity of 0.851, Specificity of 0.437, and a kappa value of 0.2959. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 78.0488%, Sensitivity of 0.852, Specificity of 0.428, and a kappa value of 0.2644. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 71.0145%, Sensitivity of 0.784, Specificity of 0.29, and a kappa value of 0.0614. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 14 attributes, yielding a total accuracy of 86.4078%, Sensitivity of 0.921, Specificity of 0.804, and a Kappa value of 0.7227. Then the author selects

a 66-34 training-test split on 14 attributes, yielding a total accuracy of 91.4286%, Sensitivity of 0.909, Specificity of 0.923, and a kappa value of 0.8193. Then the author selects an 80-20 training-test split across 14 attributes, yielding a total accuracy of 90.2439%, Sensitivity of 0.956, Specificity of 0.8333, and a kappa value of 0.7995. Then the author selects a 10-fold cross-validation training-testing split on 14 attributes, yielding a total accuracy of 91.3043%, Sensitivity of 0.918, Specificity of 0.905, and a kappa value of 0.821.

Then the author selects the Gain Ratio feature with SVM (Support Vector Machine) using a 50-50 training-test split on 90 attributes, resulting in a total accuracy of 68.932%, Sensitivity of 0.787, Specificity of 0.34, and a Kappa value of 0.1313. Then the author selects a 66-34 training-test split on 90 attributes, yielding a total accuracy of 74.2857%, Sensitivity of 0.862, Specificity of 0.421, and a kappa value of 0.3039. Then the author selects an 80-20 training-test split on 90 attributes, yielding a total accuracy of 68.2927%, Sensitivity of 0.833, Specificity of 0.272, and a kappa value of 0.1161. Then the author selects a 10-fold cross-validation training-testing split on 90 attributes, yielding a total accuracy of 72.4638%, Sensitivity of 0.828, Specificity of 0.4, and a kappa value of 0.2328. Then the authors reduced the number of attributes from 90 to achieve the highest accuracy. The author selects a 50-50 training-test split across 15 attributes, yielding a total accuracy of 93.2039%, Sensitivity of 0.884, Specificity of 0.948, and a Kappa value of 0.822. Then the author selects a 66-34 training-test split across 15 attributes, yielding a total accuracy of 91.4286%, Sensitivity of 0.904, Specificity of 0.918, and a kappa value of 0.8013. Then the author selects an 80-20 training-test split across 15 attributes, yielding a total accuracy of 92.6829%, Sensitivity of 0.846, Specificity of 0.964, and a kappa value of 0.8275. Then the author selects a 10-fold cross-validation training-testing split on 15 attributes, yielding a total accuracy of 91.3043%, Sensitivity of 0.849, Specificity of 0.935, and a kappa value of 0.7745.

Assessment of FS Methods for the Dataset:

Classifiers	**FS Method**	**Testing Training partition**	**Number of Features**	**Accuracy**	**Kappa Statistic**
Random Forest	Information Gain	50-50	90	75.7282	0
			7	92.233	0.8167

		66-34	90	78.5714	0
			7	88.5714	0.75
		80-20	90	80.4878	0
			7	100	1
		10-cross validation	90	77.2947	0
			7	100	1
	Gain Ratio	50-50	90	75.7282	0
			16	90	0.9355
		66-34	90	78.5714	0
			16	95.7143	0.9039
		80-20	90	80.4878	0
			16	90.2439	0.7747
		10-cross validation	90	77.2947	0
			16	93.7198	0.8673
Naïve Bayes	Information Gain	50-50	90	76.699	0.3681
			6	86.4087	0.3681
		66-34	90	74.2857	0.1544

			6	90	0.6154
		80-20	90	68.2927	0.0362
			6	97.561	0.8951
		10-cross validation	90	71.9807	0.0769
			6	93.2367	0.6849
	Gain Ratio	50-50	90	76.699	0.19
			15	92.233	0.7892
		66-34	90	74.2857	0.1544
			15	85.7143	0.6531
		80-20	90	68.2927	0.0362
			15	87.8049	0.6981
		10-cross validation	90	71.9807	0.0769
			15	87.9227	0.6694
Logistic Regression	Information Gain	50-50	90	63	0.1757
			13	91	0.7819
		66-34	90	62	0.037
			13	90	0.7657
		80-20	90	65	0.1534
			13	92	0.8275
		10-cross validation	90	68	0.2171
			13	90	0.7524

	Gain Ratio	50-50	90	63.1068	0.1757
			15	91.2621	0.7819
		66-34	90	62.8571	0.037
			15	90	0.7656
		80-20	90	65.8537	0.1534
			15	92.6829	0.8275
		10-cross validation	90	68.599	0.2171
			15	90.3382	0.7524
Decision Tree	Information Gain	50-50	90	75	0.1783
			13	88	0.7059
		66-34	90	71	0.2959
			13	87	0.7053
		80-20	90	78	0.2664
			13	90	0.7853
		10-cross validation	90	71	0.0614
			13	91	0.7745
	Gain Ratio	50-50	90	71.8447	0.1783
			14	86.4078	0.7227
		66-34	90	75.7143	0.2959
			14	91.4286	0.8193

		80-20	90	78.0488	0.2644
			14	90.2439	0.7995
		10-cross validation	90	71.0145	0.0614
			14	91.3043	0.821
SVM (Support Vector Machine)	Information Gain	50-50	90	74	0.1313
			13	93	0.8222
		66-34	90	68	0.3039
			13	91	0.8013
		80-20	90	68	0.1161
			13	92	0.8275
		10-cross validation	90	72	0.2328
			13	91	0.7745
	Gain Ratio	50-50	90	68.932	0.1313
			15	93.2039	0.822
		66-34	90	74.2857	0.3039
			15	91.4286	0.8013
		80-20	90	68.2927	0.1161
			15	92.6829	0.8275
		10-cross validation	90	72.4638	0.2328

			15	91.3043	0.7745

Comparison of sensitivity and specificity:

Algorithm	Feature Selection Method	Training Testing Partition	No. of features	Sensitivity	Specificity
Random Forest	Information Gain	50-50	90	0.75	0
			7	0.89	1
		66-34	90	0.78	0
			7	0.84	1
		80-20	90	0.8	0
			7	1	1
		10-cross validation	90	0.77	0
			7	1	1
	Gain Ratio	50-50	90	0.757	0
			16	0.971	0.97
		66-34	90	0.785	0
			16	0.956	0.957
		80-20	90	0.804	0
			16	0.846	0.928
		10-cross validation	90	0.772	0
			16	0.9125	0.952
Naïve Bayes	Information Gain	50-50	90	0.787	0.555
			6	1	0.857
		66-34	90	0.813	0.363
			6	1	0.888
		80-20	90	0.8125	0.222
			6	1	0.972

		10-cross validation	90	0.786	0.31
			6	1	0.925
	Gain Ratio	50-50	90	0.787	0.555
			15	0.913	0.925
		66-34	90	0.813	0.363
			15	0.882	0.849
		80-20	90	0.8125	0.222
			15	0.818	0.9
		10-cross validation	90	0.786	0.31
			15	0.84	0.889
Logistic Regression	Information Gain	50-50	90	0.822	0.341
			13	0.8	0.958
		66-34	90	0.795	0.238
			13	0.9	0.9
		80-20	90	0.851	0.285
			13	0.846	0.964
		10-cross validation	90	0.836	0.363
			13	0.818	0.934
	Gain Ratio	50-50	90	0.822	0.341
			15	0.8	0.958
		66-34	90	0.795	0.238
			15	0.9	0.9
		80-20	90	0.851	0.285
			15	0.846	0.964
		10-cross validation	90	0.836	0.363
			15	0.818	0.934
Decision Tree	Information Gain	50-50	90	0.795	0.4
			13	0.758	0.932
		66-34	90	0.851	0.437
			13	0.818	0.895

		80-20	90	0.852	0.428
			13	0.75	1
		10-cross validation	90	0.784	0.29
			13	0.849	0.935
	Gain Ratio	50-50	90	0.795	0.4
			14	0.912	0.804
		66-34	90	0.851	0.437
			14	0.909	0.923
		80-20	90	0.852	0.428
			14	0.956	0.8333
		10-cross validation	90	0.784	0.29
			14	0.918	0.905
SVM (Support Vector Machine)	Information Gain	50-50	90	0.787	0.347
			13	0.884	0.948
		66-34	90	0.862	0.421
			13	0.904	0.918
		80-20	90	0.833	0.272
			13	0.846	0.964
		10-cross validation	90	0.828	0.4
			13	0.849	0.935
	Gain Ratio	50-50	90	0.787	0.34
			15	0.884	0.948
		66-34	90	0.862	0.421
			15	0.904	0.918
		80-20	90	0.833	0.272
			15	0.846	0.964
		10-cross validation	90	0.828	0.4
			15	0.849	0.935

V. Conclusion:

This paper uses Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, and SVM (Support Vector Machine) to predict the health crisis. It has proved to be a powerful statistical tool for measuring mental health conditions. By using these methods, researchers can identify risk factors and predictors of various mental health conditions, thereby developing more effective prevention and treatment strategies. We have collected the data by circulating Google among our relatives, colleagues, teachers, and friends. Based on that, we established a relationship between the dependent and independent variables. After that, we constructed a confusion matrix comparing the actual target values with those predicted by the machine learning model. After checking the Confusion Matrix, we move to cross-validation, where we evaluate the accuracy of 10 sub-lists and obtain the Confusion Matrix for each sub-list. We predict accuracy, sensitivity, specificity, and Kappa Value for user-choice test data and the 10 sub-lists. However, we predict the highest accuracy, which is 100%, using the Information Gain feature with the Random Forest Classifier. The results of this study suggest that Random Forest can be a reliable tool for assessing mental health conditions. However, it is important to note that the accuracy of the results depends heavily on the quality of the collected data, the sample size, and the choice of variables. Thus, further research is needed to enhance the predictive power of the Random Forest model for mental health conditions.

VI. References:

[1] Das, S., Chaudhuri, A. K., & Ghosh, P. (2026). Medical Diagnosis Through Improved Feature Selection and Advanced Ensemble Techniques: A Study on Breast Cancer and Chronic Kidney Disease. *SN Computer Science*, *7*(3), 222.

[2] Das, S., Chaudhuri, A. K., Paul, D., & Ghosh, P. (2026). Sequential Attribute Designator (SAD): A Novel Feature-Selection Framework for Pulmonary Disease Research. In *Next-Generation Bioinformatics for Pulmonary Disease Research* (pp. 413-436). IGI Global Scientific Publishing.

[3] https://www.geeksforgeeks.org/supervised-machine-learning/

[4] https://www.yalemedicine.org/conditions/depression

[5] Das, S., Chaudhuri, A. K., Das, S., & Ghosh, P. (2025). Multistage feature selection and stacked generalization model for cancer detection. *Scientific Reports*, *15*(1), 38124.

[6Tu, M. C., Shin, D., & Shin, D. (2009, October). Effective diagnosis of heart disease through bagging approach. In *2009 2nd international conference on biomedical engineering and informatics* (pp. 1-4). IEEE.

[7] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal*

on Computer Science and Engineering (IJCSE), *2*(02), 250-255.

[8] Babaoglu, İ., Findik, O., & Ülker, E. (2010). A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. *Expert Systems with Applications*, *37*(4), 3177-3183.

[9] Park, Y. J., Chun, S. H., & Kim, B. C. (2011). Cost-sensitive case-based reasoning using a genetic algorithm: Application to medical diagnosis. *Artificial intelligence in medicine*, *51*(2), 133-145.

[10] Rajeswari, K., Vaithiyanathan, V., & Neelakantan, T. R. (2012). Feature selection in ischemic heart disease identification using feed forward neural networks. *Procedia Engineering*, *41*, 1818-1823.

[11] Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University-Computer and Information Sciences*, *24*(1), 27-40.

[12] Karabulut, E. M., & İbrikçi, T. (2012). Effective diagnosis of coronary artery disease using the rotation forest ensemble method. *Journal of medical systems*, *36*(5), 3011-3018.

[13] Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology*, *2*(3), 220-223.

[14] Shouman, M., Turner, T., & Stocker, R. (2012). Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. In *Proceedings of the international conference on data science (ICDATA)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[15] Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. *Caribbean Journal of Sciences and Technology*, *1*(1), 208-217.

[16] Hari Ganesh, S., & Gajenthiran, M. (2014). Comparative study of data mining approaches for prediction heart diseases. *IOSR Journal of Engineering*, *4*(7), 36-9.

[17] Bashir, S., Qamar, U., & Javed, M. Y. (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In *International conference on information society (i-Society 2014)* (pp. 259-264). IEEE.

[18] Tomar, D., & Agarwal, S. (2014). Feature selection based least square twin support vector machine for diagnosis of heart disease. *International Journal of Bio-Science and Bio-Technology*, *6*(2), 69-82.

[19] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.

[20] Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P., & Li, G. (2017). An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert systems with applications*, *68*, 163-172.

[21] Varatharajan, R., Manogaran, G., & Priyan, M. K. (2018). Retracted article: A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. *Multimedia Tools and Applications*, *77*(8), 10195-10215.

[22] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th international bhurban conference on applied sciences and technology (IBCAST)* (pp. 619-623). IEEE.

[23] Ray, A., Chaudhuri, A. K., & Das, S. (2023). A novel improved logistic regression model for diagnosing heart disease. In Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem (pp. 215-240). CRC Press.