An Efficient Early Diagnosis and Healthcare Monitoring System for Mental Disorder using Machine learning Approach

C Kishor Kumar Reddy
Stanley College of Engineering & Technology for Women
Pullisani Satvika (pullisanisatvika@gmail.com)
Stanley College of Engineering & Technology for Women

Research Article

Keywords: Mental Health, Accuracy, Precision, Machine Learning, Deep Learning, Random forest, Mental illness Detection

Posted Date: May 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2912222/v1

Additional Declarations: No competing interests reported.

An Efficient Early Diagnosis and Healthcare Monitoring System for Mental Disorder using Machine learning Approach

Kishor Kumar Reddy C[1], Pullisani Satvika[2*], Marlia M. Hanafiah[3] and Srinath Doss[4]

[1,2]Department of Computer Science & Engineering, Stanley College of Engineering & Technology for Women, India

[3]Department of Earth Science and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia

[3]Centre for Tropical Climate Change System, Institute of Climate Change, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Selangor, Malaysia

[4]Dean, Faculty of Engineering & Technology, Botho University, Botswana

Abstract

Recently, the increase in internal health problems in society has led to an increase in research on the development of mechanistic capacity models to detect or predict internal mental health. The effective use of internal health assessments or discovery models allows internal health interpreters to redefine internal suffering more objectively than ever before, and in the early stages when interventions may be more effective.In this work, we aim to apply a bias mitigation system based on multitasking literacy to perform a fairness analysis and to fear the predicted model using the Reddit dataset. This article employs an efficient technique for machine learning random forests. The proposed model was evaluated against various performance metrics and the model showed 91.00% accuracy. This is an advantage compared to existing approaches.

Keywords:

Mental Health, Accuracy, Precision, Machine Learning, Deep Learning, Random forest, Mental illness Detection

1. Introduction

Mental illness, also known as internal diseases, is a outspread term used to describe a comprehensive of conditions that affect a person's mood, thinking, and geste. These conditions can vary from mild to severe and can be as a consequence of variety of factors, including inheritable predilection, environmental factors, and life gests. Some common types of internal ails include depression, anxiety diseases, bipolar complaint, schizophrenia, and eating diseases. Mental illness can be treated with a variety of curatives, including drug, psychoanalysis, and life changes. [1]

Diagnosing internal illness generally involves a comprehensive evaluation by a good internal health professional. These techniques can be used to analyze large datasets of various forms of data, such as electronic health records, medical images, and social media posts .The process may include the Clinical Interview, Physical Examination, Cerebral Assessments, Individual Criteria and Cooperative Decision.

In Clinical Interview, the internal medic will stance a thorough interview with the solitary to gather information about their traits, medical history, and family history. This may involve asking questions about the existent's studies, feelings, actions, and overall functioning.

In Physical Examination, occasionally , a physical examination may be accomplished to rule out any beginning medical conditions that could be causing or contributing to the internal health symptoms.

In Cerebral Assessments, the Formalized cerebral assessments, similar as questionnaires or tests, may be administered to assess the existent's cognitive, emotional and behavioral functioning. These assessments can help give fresh information to prop in the opinion. In Individual Criteria, the internal health professional will compare the existent's symptoms and history in opposition to the criteria outlined in the Diagnostic and Statistical Manual of Mental diseases

(DSM- 5), which is a extensively used reference companion for diagnosing internal ails.

In Cooperative Decision- Making, the internal health professional and the individual work together to reach a opinion grounded on the information gathered. This may involve agitating different implicit judgments, considering the inflexibility and duration of symptoms, and exploring treatment options.

The success rate of diagnosing internal illness can fluctuate depending on several factors, including the type and inflexibility of the internal illness, the accuracy of the information

handed by the existent, the experience and moxie of the internal health professional, and the tools and styles used for assessment.[11] Diagnosing internal illness can be complex and grueling as internal health conditions frequently present with lapping symptoms and can have different situations of inflexibility. Also, individualities may have different ways of expressing their symptoms, and there may be private rudiments involved in the individual process. [2]Still, with a thorough evaluation by a good internal health professional using established individual criteria and assessment tools, the delicacy of diagnosing internal illness can be fairly high.

The use of standardized and validated assessment measures, along with a comprehensive clinical interview and careful consideration of the existent's history, can ameliorate the delicacy of the opinion. It's important to note that internal health judgments aren't stationary and may evolve over time as new information becomes available or as symptoms change. Mental health professionals may also use a trial- and- error approach to determine the most accurate opinion and knitter treatment consequently.[3]The success rate in diagnosing mental illness is a complex issue that depends on multiple factors, including the diagnostic criteria used, the clinician's experience and also the techniques they use.

There are several ways used for the vaticination of internal illness, including

1. Behavioral and Symptom Monitoring: This involves monitoring changes in geste and symptoms over time to descry patterns or changes that may indicate the onset of a internal illness.

2. Inheritable Testing: This involves assaying a person's DNA for inheritable labels associated with specific internal ails.

3. Brain Imaging: This involves using neuroimaging ways, similar as MRI or CT reviews, to examine the structure and function of the brain and identify abnormalities that may be associated with internal illness.

4. Cerebral Assessment: This involves using standardized tests and interviews to assess a person's internal health and identify any symptoms or actions that may be reflective of a internal illness.

5. Machine Learning Algorithms :This involves using artificial intelligence and machine literacy ways to dissect large quantities of data and identify patterns and connections that may be prophetic of internal illness. Advancements in machine learning (ML) techniques, specifically

deep learning for qualitative data and natural language processing for quantitative data, have revolutionized the diagnosis and prediction of mental illness.[4]

The emerging field of computational psychiatry is at the forefront of using ML techniques to gain better insights into cognitive and neural processes, as well as developing predictive models for mental illness using complex data. In recent studies, the usage of machine learning techniques like K nearest neighbour (KNN), Support Vector Machine (SVM), and Random Forest have demonstrated an accuracy rate of up to 90% in predicting mental health conditions. It's important to note that while these ways may be useful in prognosticating internal illness, they aren't reliable and should always be used in confluence with clinical judgment and professional evaluation.[5]

Model evaluation for mental illness vaticination generally involves measuring the performance of the model using colorful criteria similar as accuracy, perfection, recall, F1 score, and area under the receiver operating characteristic wind(AUC- ROC).To understand the statistics of the internal health conditions, the WMH Survey Initiative of the World Health Organization(WHO) designed a check to help countries all over the world carry out and dissect epidemiological checks of the frequency and supplements of internal diseases. A crucial end of the WMH checks is to aid countries that would not else have the moxie or structure to apply finest community epidemiological checks by furnishing centralized instrument development, training, and data analysis which is given in the Table 1 and Table 2 below.

Table 1: Any Mental Illness in Past Year Among People progressed 18 or Aged; by Gender and Detailed Age order, figures in Thousands, 2021

Age	Total	Male	Female
18-25	11,265	4,283	6,983
26-34	13,275	5,557	7,719
35-44	11,234	4,350	6,884
45-54	8,530	3,338	5,142
55-59	3,531	1,322	2,208
60 or older	3,469	1,238	2,231

Table 2: Any Mental Illness in Past Year Among People progressed 18 or Aged; by Gender and Detailed Age order, figures in Thousands, 2022

Age	**Males (%)**	**Females (%)**	**Persons (%)**
16–24	31.2	46.6	39.6
25–34	21.9	32.1	27.1
35–44	18.5	20.7	19.7
45–54	16.5	24.3	21
55–64	14.6	18.9	17.1
65–74	9.5	13.1	11.4
75–85	2.8	4.9	3.7

The paper has several contributors who have worked on different aspects of the design. Then is a brief description of each donation's part in the design:

1. Dataset: The donation responsible for the dataset would have gathered and prepared the data that was used to train and test the machine literacy prototype. This would include relating applicable sources of data, drawing and pre processing the data, and conceivably performing point engineering to prize useful information from the data.

2. Analysis: The donation responsible for analysis would have conducted a detailed analysis of the paper, comparing and differing different machine literacy models to determine which one is more effective. This analysis would have included assessing the accuracy, precision, recall, and F1 score of the models, as well as assessing their computational effectiveness.

3. Model structure: This donation responsible for model structure would have developed the machine literacy models that were estimated in the paper. This would include opting an applicable algorithm or armature, choosing hyper parameters, and training the model on the set dataset.

4. Model Evaluation: This donation responsible for model evaluation would have assessed the performance of the trained models using colourful criteria , similar as accuracy, precision, recall, F1 score, and area under the ROC curve. They would also have conducted trials to test the robustness of the models, similar as by introducing noise or varying the dataset size. Overall,

each donation plays an major part in the blossoming and evaluation of machine literacy models, and their benefactions are necessary for producing a high- quality exploration paper.[6]
The result of the paper is recognized as follows: Chapter 2 which discusses Relevant Works, Chapter 3 which discusses Proposed Methodologies, Chapter 4 which discusses Results and Discussion , Chapter 5 that is Conclusion which is followed by References.

2. Related work

To make better guesses about mental health, scientists have been using machines to learn since the 1980s.Sijia Zhou et al. [1] proposed New and advanced technologies such as machines that can learn, analyzing large amounts of data, and artificial intelligence are being used to create personalized medicine and improve mental health treatments. This is causing big changes in the way people think about healthcare. This review of literature is about how artificial intelligence is used in brain operations and people's thoughts on it. This review looks at how new technology like deep literacy and AI is changing the way we assess and treat mental health issues. To do this, we searched for articles on PUBMED and Web of Science using words like "mental health treatments," "opinions on mental health," "artificial intelligence," and "deep literacy." The information we gathered for this study is only from a few people, and we don't have enough information about how certain symptoms might affect people in the long run. The accuracy percentage 87% is obtained for the CEASE Method.

T. E. Ramya et al. [6] proposed a new system was suggested where comments from Twitter are collected and analyzed using a special type of computer program called a convolution neural network. This can figure out what kind of people have certain features and is quick at figuring out how a person is feeling or acting.They retrieved information by using a tool called Twitter API. The Kaggle dataset needs to be cleaned up using methods like tokenization and removing unnecessary words. Then, different models like Support Vector Machine and decision trees can be used to classify the markers into different categories. The connection between scores and feelings has been figured out accurately and saved for further study of emotions. The accuracy level is 83.76%

U. Sairam et al. [7] are interested in online communities for data. They've crawled data from 247 online communities of 80,000 users. Also they've uprooted the user- verbal posts grounded on motifs, which was used as an input in a computer program. They used machine learning

techniques to create a standard system for analyzing our health's internal features.Ten different classifiers were employed and estimated to elect the optimal bone with the named set of features. These are Bayesian Network(BN), logistic, multiple subcaste perceptron(MLP), Naïve Bayes(NB), arbitrary timber(RF), arbitrary tree(RT), J48,successional minimum optimisation(SMO), arbitrarysub-space(RS), and K Star(KS). MLP Has reported the stylish performance among other algorithms with an accuracy of 97.8%.

Iqra Ameer et al. [8] proposed a model that analyzes unshaped stoner data on the Reddit platform and classifies five common internal ails depression, anxiety, bipolar complaint, ADHD and PTSD. They trained machines to identify illnesses inside people by using three types of literacy models: traditional, deep, and transfer. The RoBERTa model, which already learned a lot, was better than other computer programs at reading and getting things right 80% of the time i.e,the accuracy.

Mohsin kamal et al. [9] proposed a new way of using machine learning that states to identify a person's illness based on their social media posts and comments on Reddit. They planned to use a method called "XGBoost" to sort data into four different groups: Schizophrenia, Autism, Obsessive-Compulsive Disorder, and Post-Traumatic Stress Disorder. This method is very commonly used and accurate. They compared their new way of doing things to other methods that are already well-known in the same field. These methods are called Naïve Bayes and Support Vector Machine. The exploration community has already tested these methods and reported their results. The test shows that their new way of sorting information works better than other ways. They got 68% accuracy of the classifications , which means our method is good.

Pravin Karmore et al. [11] proposed a model where they made use of recordings from the University of Stanford. The recordings are conversations between people and computers. There are more than 300 recordings that last from 56 seconds to 16 minutes. They used two models ResNet- 34 and ResNet- 50 among which ResNet- 50 had an loftiest accuracy of 82.4%.

Faruk Hossen et al. [12] ,In the AiPsych system, they used two sets of information: one about people's complaints and another about suggestions for getting better. Both sets of information were obtained from Kaggle. They created a mobile app called AiPsych. It uses computer programs to predict how students are feeling inside during the COVID-19 pandemic. Their test showed that being monitored while learning to read and write is better than advanced learning, with a 97% accuracy rate for predicting internal illnesses using a Support Vector Machine

(SVM). Random Forest algorithm has shown the most accurate results of 91% for predicting recovery suggestions.

Konda Vaishnavi et al. [13] proposed a model where they collected data from social media information to predict people's health. They gathered information from 55 different studies. This research looked at five ways machines can read information and checked if they were accurate in identifying health problems inside the body by using different measures of how accurate they were. There are five different ways machines can learn: Logistic Regression, K-NN Classifier, Decision Tree Classifier, Random Forest, and Stacking. We looked at different methods, tried them out, and found the most accurate way to shape something. It works well and we can predict its success 81.75% of the time.

Xiaofeng Wang et al. [14] stated that knowing about a person's internal health can help public health workers make plans to help them with their mental health. This document predicts the physical wellbeing of healthcare employees based on their ability to use machines and 32 different factors. They asked 5,108 Chinese medical workers questions and found 32 things that affect their health. They used the answers to see how healthy they are on the inside. In this research, they suggest a new way to predict the health of medical workers using a special computer program that selects and prioritizes the most important factors. To predict the health of medical workers, they use different methods like a logistic regression model or computer algorithms like the double club or mongrel bettered dragonfly. They also have a new model that they use to make predictions called the vaticination model. The test showed that the new model is very accurate (92.55%). It works better than other methods that are currently used.

Salih Tutun et al. [15] proposed a model that uses smart technology and computer programs to create a helpful tool that can spot and identify different illnesses inside our bodies. To create a tool for evaluating something, we use the NEPAR algorithm to figure out what questions need to be answered. Colourful machines are taught to predict what's wrong with a person by using information from their answers and other data. Actors help with this training. The study found that a computer program called DSS can diagnose internal diseases with 89% accuracy by asking 28 questions without the need for human input.

Min Ju Kang et al. [16] proposed a model that collected information for SNSB from a research study done in South Korea on memory problems at three different hospitals. Data was collected from 14,926 people who completed a test called the Seoul Neuropsychological

Screening Battery. The results were put into three categories based on memory and thinking ability: normal, a mild problem called mild cognitive impairment, and a more serious problem called Alzheimer's disease. They trained a machine how to read and understand using a computer program called TensorFlow. Before using it on SNSB, different models such as CI, MCI or ADD, logistic regression, and neural network algorithms with multilayer structures were compared. The logistic regression method was used to classify a balanced dataset into two groups (NC vs CI), and it was 85.9% accurate. However, when three groups (NC vs MCI vs ADD) were classified, the accuracy dropped to 79.0%.

Chen Dongrui et al. [17] proposed a model where they used information from health and personality tests given to students in the 2022 class at a university for their study. These tests were given online with the help of the university's counseling center, and they gathered 2,780 usable responses. The system takes information about a person's health from internal tests. After making sure the information is correct, it uses a decision-making tool to analyze the data. The Adaboost method is used to improve a decision-making tool called the decision tree. This helps it work better when there are many duplications of data. The improved tool is used to make predictions about someone's health. Out of all the ways to make decisions, Decision Tree is the most accurate with a value of 84.8%.

Mostafa Rezapou et al. [18] tproposedd a model where they used data from a research group and different mathematical models to study how COVID-19 has affected the mental health of workers in the United States. Out of all the ways to do something, using Random Forest is the most accurate at 92.0%.

Harleen Kaur et al. [19]proposed a model whose dataset is gathered from a place called the IEEE data harbor, where people can easily find information from Twitter about COVID-19. This information lists the ID numbers and how people feel about the COVID-19 outbreak on Twitter. They looked at some information using computers and special programs to see if it was positive or negative. We used different ways of teaching the computer, like Support Vector Machine and intermittent Neural Network, with a tool called R. There are lots of ways to teach computers to do this, like SVM, RNN, CNN, Random Forest, Naïve Bayes, and LSTM. Out of these, SVM was the most accurate, with 93% success.

Syed Nasrullah et al. [20] proposed a model Syed Nasrullah and his colleagues. A new way of collecting information from Reddit was suggested. It uses a special computer system called a

"deep literacy model" made up of two parts: a convolutional neural network and an intermittent neural network. In this study, they predicted different types of internal illness, such as anxiety. They used a method called multiclass bracket to do this. Reduce anxiety, comparing bipolar disorder and anxiety. Different types of mental health conditions can be classified into categories such as dementia, non-dementia, and psychotic. They talk about how different computer models are tested to see how well they work. The testing looks at different things like how accurate and complete the models are. One of the models was better than the others at predicting different categories, but it wasn't always right.

.

3. Proposed Model For Mental Illness Prediction

Random Forest is a frequently used computer program for solving problems in prediction and regression. It's a system that uses a lot of decision trees to guess things more accurately. In a Random Forest model, many trees are made using different parts of the training information. Every tree in the group is taught on a different part of the practice information, and a random part of features is utilized to solve each difficulty in the tree.

When making a vaticination, each tree in the timber votes on the prognosticated class, and the most popular class is chosen as the final vaticination. Random Forest has a wide range of operations in different disciplines, including finance, healthcare, marketing, and more. It can be used for tasks similar as prognosticating client churn, credit threat analysis, medical opinion, and image bracket. One of the main advantages of Random Forest is that it's lower prone to overfitting compared to other machine literacy models. Each tree uses a different set of information and only some of the information, which helps to avoid problems and errors. Random Forest can manage many input variables and can provide significance estimates for point selection, which is helpful. In simple words: Random Forest is a useful and adaptable tool for many different types of tasks in machine learning. Data scientists and machine learning interpreters like to use it because it can prevent overfitting and handle many input variables.

Key Characteristics of Random Forest

1. Diversity: When making a tree, some attributes are left out and every tree is unique..

2. Immune to the curse of dimensionality: Since each tree doesn't contemplate all the features, the point space is narrowed.

3. Parallelization: Every tree is made differently with unique information and characteristics. This means we can use the computer's processor to create any arbitary value.

4. Train- Test split: We don't need to separate the data into training and testing for a certain value, because 30% of the data is always hidden from the decision tree.

5. Stability: Stability happens when the outcome is based on mature voting or averaging.

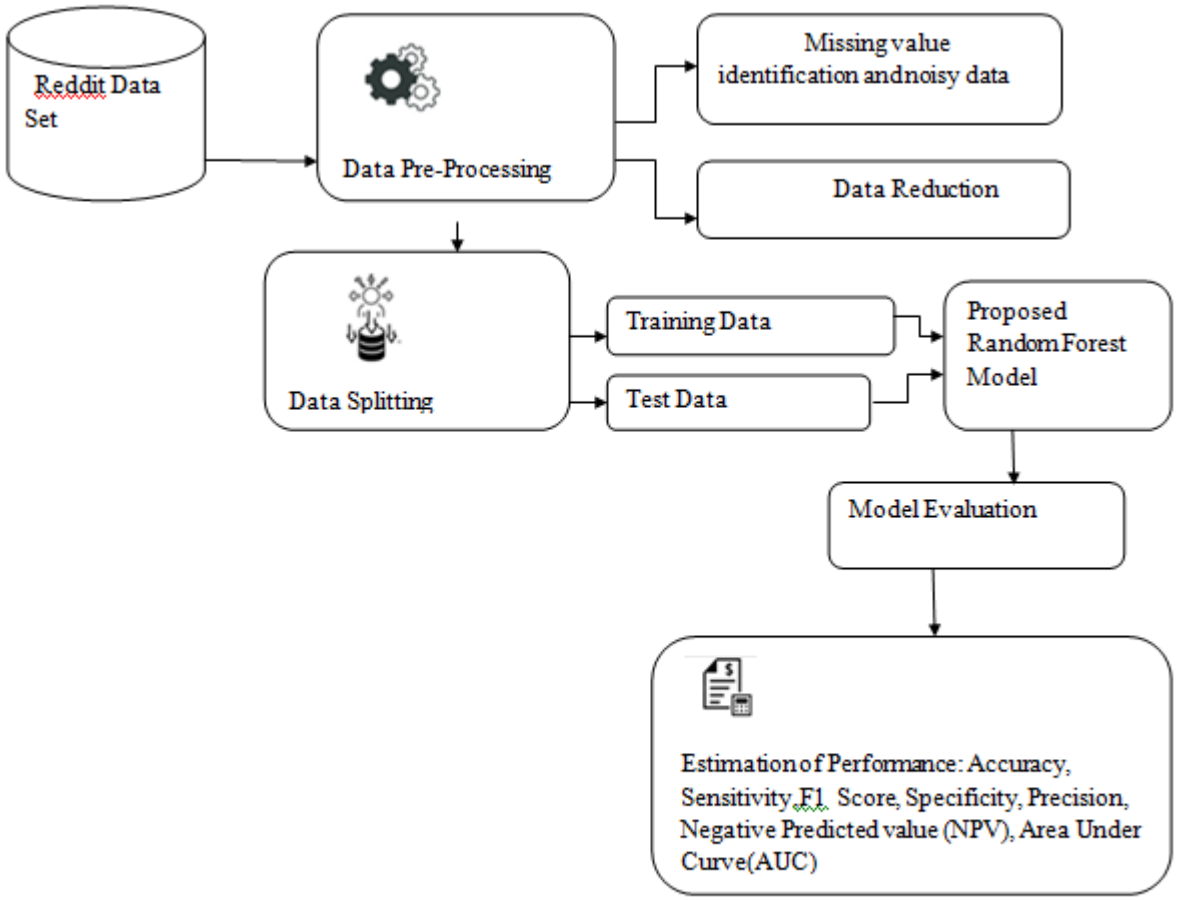


Fig 1: Proposed Random Forest model

Steps Involved in the proposed model

3.1Data collection

Data collection is when you gather information from different places and put it together to make sense of it all. This process involves relating and penetrating the different data sources, mapping the data to a common format, and coordinating any inconsistencies or disagreement between the sources. The thing of data integration is to make it easier to pierce and dissect data that's spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data. Data integration can be challenging due to the variety of data formats, structures, and semantics used by different data sources. Different data sources may use different data types, naming conventions, and schemas, making it delicate to combine the data into a single view. Data integration is the process of bringing together different information and using it for various activities similar as business intelligence, data warehousing, master data operation, and analytics. The datset that is used in the random forest algorithm is the Reddit Dataset. Reddit is a social media platform that allows users to post, bandy and bounce on content in colourful communities called subreddits. Reddit's data set includes a vast quantum of stoner- generated content, similar as posts, commentary, and stoner information. This data set has come precious resource for experimenters and data scientists studying colourful fields, including social media analysis, natural language processing, and machine literacy.

The Reddit data set is intimately available through the Reddit API and colourful third- party websites that give access to the data in a structured format. Experimenters and data scientists can use this data to conduct colourful analyses, similar as sentiment analysis of stoner commentary, content modelling of subreddit content, and stoner geste analysis.

Table 3: Dataset Description

S.No	**Title Of The Paper**	**Dataset Used**	**No Of Instances**	**No Of Attributes**
1	[1.]	-70 University students - Layered, various leveled show for interpreting raw sensor information into markers of	17592	30

		behaviors and states related to mental wellbeing - 449 Subjects -366 Patients - 44 participants - 1,600 community-dwelling adults -15 clinically anxious youths matured 12 to 18 - 17,572 patients (90,934 treatment session transcripts) - Studies that included randomized controlled trials (RCTs) analyzing the affect of telephone-administered psychotherapy on depressive side effects.		
2	[2 .]	-RSDD -UCLA student tweets -Private dataset collected from Twitter -Private dataset collected from Twitter -eRisk 2018 task -eRisk 2019 task -eRisk 2017 task -eRisk 2020 task -eRisk 2021 depression task	-(9,210 depression, 107,274 control) -(8,068 rest deprived tweets, 10,326 control) -(3,984 self-destructive tweets, 30,376 control) -(2,880 positive, 1,883 unbiased, 2,605 negative)	8

		- Depression dataset curated - Reddit data -UMD dataset -private SWMH dataset -DepSign shared - Dreaddit -Goemotion - CEASE -TTDD dataset	-(134 anorexia, 1,153 control) -(135 depression, 752 control) -(145 self-harm, 618 control) -(negligible 20, mellow 40, direct 49, serious 61) -(1,402 depressed, 5,160 control) -(1,293 depression posts, 5,49 control) -(363 low, 503 serious, 866 control) -(distinctive mental disorders 54,412) - task (not 3,081, direct 8,325,severe 1261) -#self-destructive dataset(3,984suicidal tweets, 30,322 control) -(3,553) -(58,009 emotion) -(4,932) -(1,402 depression, 1,402 control)	

3	[3.]	The Dataset consists of Twitter information of individuals whose mean age for practical members was 30.3, a long time	For depression:147, For PTSD :174	30
4	[4.]	The Dataset is taken from Reddit API which was utilized to crawl 13 subreddits and 17159 posts content and title content to get information for this inquire about work. Out of the 13 subreddits, 5 can be specifically related with a mental ailment.	3000 posts beneath the subreddits managing with mental ailment and almost 300 posts for each of the common point subreddits.	13
5	[5.]	There are two sets of information from 2011 to 2019: 1) A dataset of clients with one of the three enthusiastic disarranges of intrigued (BD, MDD, Advertisement); and 2) A dataset of clients without known mental disarranges serving as the control gather	100	5
6	[6.]	The data set consists of a portion of	200	9

		the TILES dataset, physiological information		
7	[7.]	Twitter API	1500	7
8	[8.]	247 online communities of 80,000 users.	80,000	9
9	[9.]	Reddit's post dataset The dataset comprises a add up to of 16,930 posts. The posts were assist separated into prepare, dev, and test bunches with 13,726 posts within the training set, 1,716 posts within the dev set, 1,488 posts within the test set. multi-class dataset from Reddit social media platform	16,930 posts	5
10	[10.]	Reddit dataset The dataset has created almost 32330 comments on 4000 posts on clinical subreddit.	32330 comments on 4000 posts	3
11	[11.]	SinaWeibo information from college understudies to identify depression among college understudies (IRIS datasets)	16 solid subjects and 7 psychiatric patients with depression or somatoform disorder	7
12	[12.]	The dataset prepared by the College of Stanford which contains more than	300	3

		three hundred sound clips with each sound clip containing a recorded intelligently session by patient-computer of term extending from least fifty six seconds to greatest sixteen minutes. The dataset contains sound fragments of patients extending from age of sixteen sixty four with cruel age calculated to be thirty three point five years, with a deviation of fifteen point three years		
13	[13.]	There are two datasets within the AiPsych framework (APS), one for the conclusion of mental illness and the other for the forecast of recuperation suggestions. Both datasets were gotten from kaggle.com.	DMIS dataset:1182 (6 features-categorical, 10-numerical) With 61 positive instances and 1021 negative instances RSS dataset:335(290- positive (treatment) instances, 45 are negative (counselling instances)	19
14	[14.]	The dataset is collected from the "Mental Wellbeing Status of Restorative Laborers Amid COVID-19" overview conducted in Changchun, Jilin Territory, China from June 1, 2020 to June 7, 2020, this paper predicts the mental wellbeing of Chinese therapeutic specialists amid COVID-19.	220	32

		(survey last test measure is 5,108)		
15	[15.]	Symptom Checklist 28-Artificial Intelligence (SCL-28-AI)	-	28
16	[16.]	Clinic-based dataset	17,139	46
17	[17.]	The information utilized for the test in this paper were collected from the mental wellbeing test and identity test information of understudies within the class of 2022 at a college. Symptom Self-Rating Scale SC.	2,780 valid questionnaire	90
18	[18]	The information is obtained from the College of Michigan's Inter-university Consortium for Political and Social Investigate (ICPSR).	518	29
19	[19.]	The data set consists of Twitter information almost COVID-19 contain IDs and opinion scores of the tweets related to the COVID-19 widespread.	600	-
20	[20.]	The dataset has been collected from the reddit utilizing NLTK python library.	3,01,746	4

3.2 Data processing

Data processing means taking information that is not useful yet and changing it into information that can be used. Raw information is gathered and organized, then analyzed and saved in a way that makes it easy to read. Data mining is usually done by a group of smart people who work with data. They do it in steps, and the next step is to make sure the data is clean. This is

important because if we use dirty data in mining, it can make the results wrong by confusing the procedures. This step removes data that is not useful because it is either too noisy or has problems. Some ways of tidying up data can do it alone, but they might not be very strong.

Data drawing does its work in this way.

(i) To replace the missing information, you can use different methods like filling them with bright colours similar to doing it by hand, using statistical measures, ignoring the incomplete data or filling in with the most likely value.

(ii) Get rid of the noisy data which is the same as random error. The system of binning can help take away this loud sound. Binning styles mean organizing things by putting them into lockers or pails based on their value. Making something smooth requires looking at nearby values.

Data Reduction:

Data Reduction means reducing the amount of data that is being processed or analyzed by using different techniques to make it easier to understand and work with.

The representation is smaller in size but still accurate. Data is made smaller in a way similar to how Naive Bayes, Decision Trees, Neural Networks, and so on work. Data reduction means making the data smaller. One way to do this is by reducing the number of things we look at in the data, such as the number of attributes. This is called "Dimensionality Reduction". Numerosity reduction means using simpler ways of showing data instead of the original large amount of data.

3.3Data Splitting

When you begin creating a model, one of the first things to decide is how to use the information that is available. A common thing in data science is to split the data into two groups called the training and testing sets. The training set is like a starting point to make models and points that help estimate parameters and compare different models. It's important to do this to create a final model. The test set is only used at the end of the training to give an honest evaluation of how well the model works. It's important that you don't use the test before this point. If you look at the test results, it might mess up things because the testing data is already a part of making the model. There are different ways to split the data into sets for practice and testing.

The usual way is to use an understanding of a random piece. A fully arbitrary slice is an easy method to use and helps protect the process from being affected by any particular piece of data. This method can cause problems when the outcome is not spread out evenly. A safer way to split would be to use a random sample based on the growth. For models with brackets, you can choose samples randomly from each group. This method makes sure that the number of results are about the same in both the practice and testing groups. If the growth is a number, we can make artificial layers based on four equal parts of the information we have.

3.4 Proposed Random Forest Algorithm

Algorithm : Pseudocode for the Random Forest algorithm,

Precondition: A training set S := (x1, y1), . . . ,(xn, yn), features F, and number of trees in forest B.

1 function RandomForest(S , F)

2 H ← ∅

3 for i ∈ 1, . . . , B do

4 S (i) ← A bootstrap sample from S

5 hi ← RandomizedTreeLearn(S (i) , F)

6 H ← H ∪ {hi}

7 end for

8 return H

9 end function

10 function RandomizedTreeLearn(S , F)

11 At each node:

12 f ← very small subset of F

13 Split on best feature in f

14 return The learned tree

15 end function

3.5 Model Evaluation

To evaluate a Random Forest model, common metrics include accuracy, precision, recall, F1 score, confusion matrix, ROC curve, and AUC. Choosing the appropriate metric(s) depends on the specific task and goals.

4. Results and Discussion

Having good mental health is really important for being healthy overall, but it's hard for doctors to know the right way to help people with mental health problems.However, recent advancements in machine literacy and deep literacy ways have shown promising results for the opinion of internal health diseases. These ways use large datasets to discern patterns and identify crucial features that are reflective of specific disorders. This approach allows for a more objective and data- driven opinion, reducing the eventuality for bias or private interpretation. Machine literacy and deep literacy ways also have the eventuality to ameliorate the effectiveness of internal health opinion by automating certain aspects of webbing and assessment. Likewise, these ways can help to identify individualities at threat of developing internal health diseases before symptoms indeed arise, allowing for earlier intervention and treatment. Still, it's important to note that these technologies aren't a relief for mortal moxie and should be used in confluence with clinical assessment.

The accuracy of a model is the number of correct predictions compared to the total number of predictions. The accuracy score measures how well a model works. It ranges from 0 to 1, with 1 being the best possible score.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad [1]$$

Table 4 shows proposed Random Forest (RF) has achieved the highest accuracy at 91.00% and HLBDA-BPNN-14 has achieved the second highest at 90.98%. The worst accuracy has been performed by Naïve Bayes at 43.52. The average accuracy of other models are obtained as 72.73% . The results are graphically depicted in figure 2

Table 4: Comparison of Accuracy measures

Methods Used	**Accuracy(%)**
NB [18]	43.52

GRU [8]	62.00
bi-GRU [8]	63.00
CNN [8]	64.00
MLR [18]	64.71
SVM [18]	70.00
CEASE [2]	74.40
LSTM [8]	76.00
BERT [8]	78.00
bi-LSTM [8]	78.00
DT [17]	78.00
linear SVM [8]	79.00
LR [8]	79.00
XLNet [8]	79.00
ResNet-34 model [11]	80.00
NN [18]	80.59
Stacking [13]	81.75
XGBoost [9]	82.00
KNN [17]	82.30
GTB [18]	82.35
ResNet-50 model [11]	82.40
RoBERT-a [8]	83.00
tf-idf technique [10]	83.10
ANN [17]	83.60
Boosting [17]	84.30
GBDT [17]	84.80
ensemble method [1]	87.00
stacked light GBM[18]	87.10
Bayestheorem [10]	88.50
BBA-BPNN-8[14]	89.02
IG-bBOA-BPNN-4[14]	90.00

MLP [7]	90.05
GCBA-BPNN4[14]	90.20
HLBDA-BPNN-4[14]	90.39
IG-bBOA-BPNN-10[14]	90.39
HIDA-BPNN-4[14]	90.78
HIDA-BPNN-4GCBA-BPNN-9 [14]	90.98
HLBDA-BPNN-14 [14]	90.98
PROPOSED MODEL RF	91.00

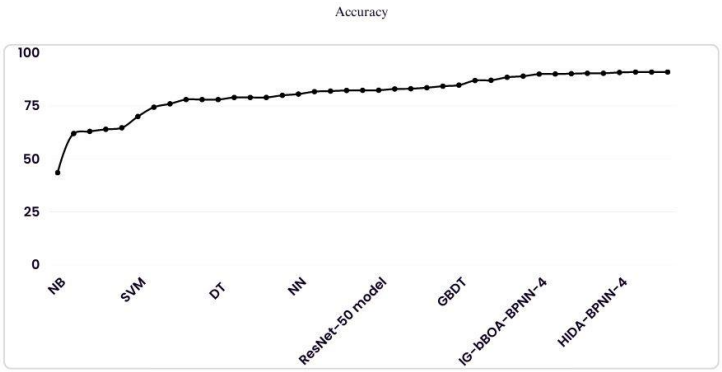


Figure 2: Comparison of Accuracy with various methods.

Sensitivity means the chance of getting a positive test result if someone really has the condition being tested for.

Specificity =TN/(TN+FP) [2]

Table 5 shows proposed Random Forest (RF) has achieved the highest sensitivity at 90.00% and RoBERTa has achieved the second highest at 83.00%.The worst sensitivity has been performed by Neural Network(NN) at 37.00%. The Average Sensitivity of other models are obtained as 62.35%.The results are graphically depicted in figure 3.

Table 5: Comparison of Sensitivity measures.

Methods Used	**Sensitivity**
NN [19]	37.00
CNN [20]	41.44
SVM [19]	49.00
ensemble DL model [20]	50.14
DT [12]	71.00
LR [12]	73.00
KNN [12]	76.00
NB [12]	76.00
ANN [12]	79.00
DL [11]	80.00
Stacked Light GBM [18]	81.00
RoBERTa [8]	83.00
PROPOSED MODEL RF	90.00

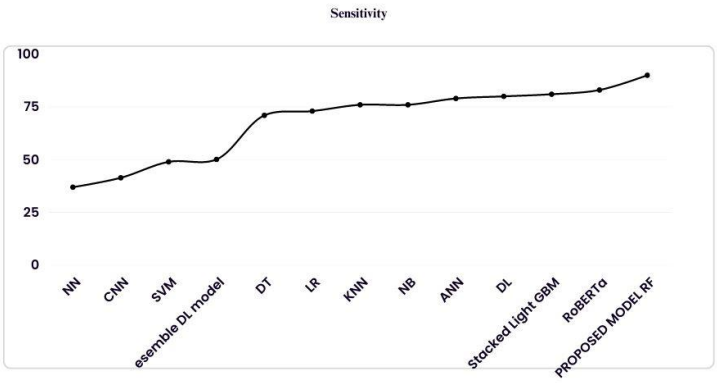


Figure 3: Comparison of Sensitivity with various methods.

The F1 score is a way of measuring how accurate and complete something is by combining two other measures called precision and recall. The F1 score is a number from 0 to 1. An F1 score of 1.0 means that the precision and recall are perfect. If the F1 score is 0, it means that either the precision or the recall is 0. The F1 score tells us the average of precision and recall. The F1 score is a number that we calculate by taking the average of two other numbers called precision and recall. The number F1 can be between 0 and 1. 1 is the best and 0 is the worst value.

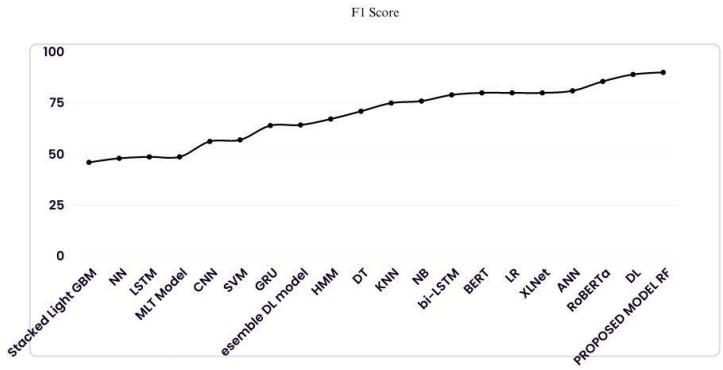
F1= 2xprecisionxrecall/precision+recall [3]

Table 6 shows proposed Random Forest (RF) has achieved the highest F1 Score at 90.00% and Deep Learning (DL) has achieved the second highest at 89.00%.The worst F1 Score has been performed by Stacked Light Gradient Tree(Stacked Light GBM) at 46.00%. The Average F1 Score of other models are obtained as 69.33% . The results are graphically depicted in figure 4.

Table 6: Comparison of F1 Score measures.

Methods Used	**F1 Score**
Stacked Light GBM	46.00

[18]	
NN [19]	48.00
LSTM [5]	48.70
MLT Model [5]	48.70
CNN [20]	56.25
SVM [19]	57.00
GRU [8]	64.00
ensemble DL model [20]	64.30
HMM [3]	67.20
DT [12]	71.00
KNN [12]	75.00
NB [12]	76.00
bi-LSTM [8]	79.00
BERT [8]	80.00
LR [8]	80.00
XLNet [8]	80.00
ANN [12]	81.00
RoBERTa [4]	85.60
DL [11]	89.00
PROPOSED MODEL RF	90.00



The specificity of a test tells us how many people who don't have the disease will test negative. This formula explains how specific something is.

Specificity=TN/(TN+FP) [4]

Table 7 shows proposed Random Forest(RF) has achieved the highest Specificity at 90.00% and Hidden Markov Model (HMM) has achieved the second highest at 81.30%.The worst Specificity has been performed by Logistic Regression(LR)at 27.40%. The Specificity of other models are obtained as 51.78%.The results are graphically depicted in figure 5.

Table 7: Comparison of Specificity measures.

Methods Used		**Specificity**
LR	[17]	27.40
KNN	[17]	35.20
ANN	[17]	38.00
Boosting	[17]	46.30
DT	[17]	47.20
GBDT	[17]	48.00
HMM	[3]	81.30

PROPOSED MODEL RF	90.90

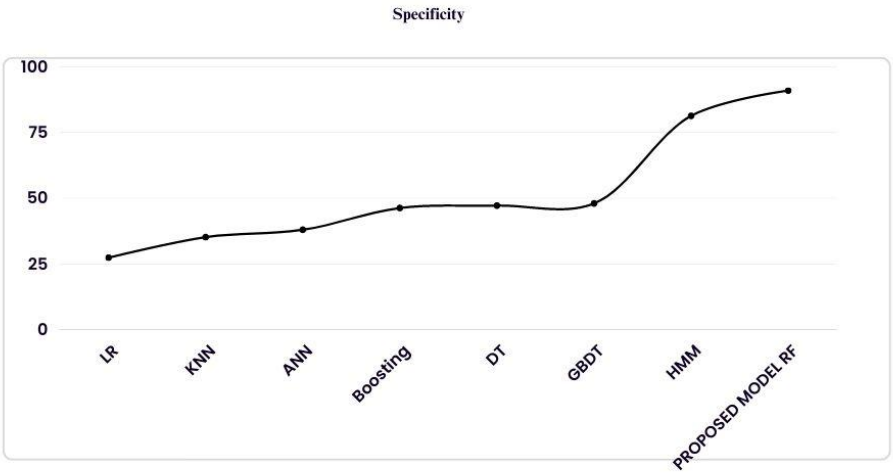


Figure 5: Comparison of Specificity with various methods.

Precision is a way to tell how well a model can correctly find positive results. You can find precision with a specific math formula:

Precision= TP/(TP+FP) [5]

Table 8 shows proposed Random Forest (RF) has achieved the highest Precision at 91.00% and RoBERTa has achieved the second highest at 85.00%.The worst Precision has been performed by Hidden Markov Model (HMM) at 42.00%. The Precision of other models are obtained as 70.90% .The results are graphically depicted in figure 6.

Table 8: Comparison of Precision measures.

Methods Used	**Precision**

HMM [3]	42.00
Stacked light GBM [18]	46.00
SVM [19]	69.00
DT [12]	71.00
NN [19]	72.00
LR [12]	73.00
ANN [12]	75.00
NB [12]	76.00
KNN [12]	80.00
RoBERTa [8]	85.00
PROPOSED MODEL RF	91.00

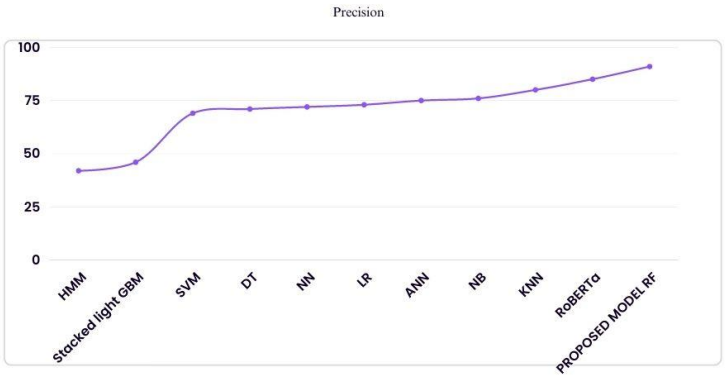


Figure 6: Comparison of Precision with various methods.

Negative Predicted Value (NPV) could be a measurable degree that makes a difference to assess the execution of a specific test. NPV is the suggestion of true negative comes about among all the people who test negative amid the method.

Negative predicted value=TN/(TN+FN) [6]

Table 9 shows proposed Random Forest (RF) has achieved the highest NPV value at 89.04% and (MCI) has achieved the second highest at 85.80%.The worst NPV value has been performed by Hidden Markov Model (HMM) at 60.20%.The results are graphically depicted in figure 7.

Table 9: Comparison of Negative predicted value (NPV) measures.

Methods Used	**Negative Predicted Value**
HMM [3]	60.20
MCI [16]	85.80
PROPOSED MODEL RF	89.04

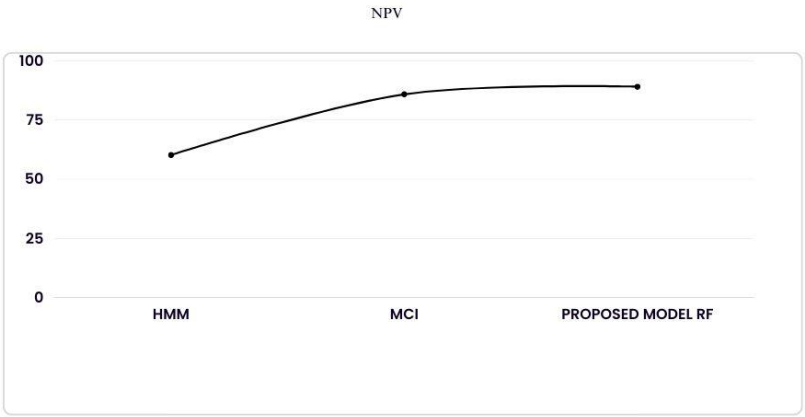


Figure 7: Comparison of Negative predicted value with various methods.

The area under a curve is a principal concept in calculus that alludes to the degree of the locale encased by the curve and the x-axis. It speaks to the overall esteem of a work over a given

interim and can be approximated utilizing numerical strategies such as Riemann wholes or Simpson's run the show. The region beneath a curve has numerous applications in science, material science, building, and other fields, such as finding the full separate traveled by a moving protest or the entire sum of work done by a constrain.

AUC= (Recall+Sensitivity)/2 [7]

Table 10 shows proposed Random Fores (RF) has achieved the highest AUC value at 88.00% and Gradient-boosted decision trees (GBDT) has achieved the second highest at 87.70%.The worst AUC value has been performed by Decision Tree(DT) at 66.70% . %. The AUC value of other models are obtained as 83.24%.The results are graphically depicted in figure 8.

Table 10: Comparison of Area Under Curve (AUC) measures.

Methods Used	**Area Under Curve**
DT [17]	66.70
LR [17]	83.50
KNN [17]	84.00
ANN [17]	85.20
Boosting [17]	87.60
GBDT [17]	87.70
PROPOSED MODEL RF	88.00

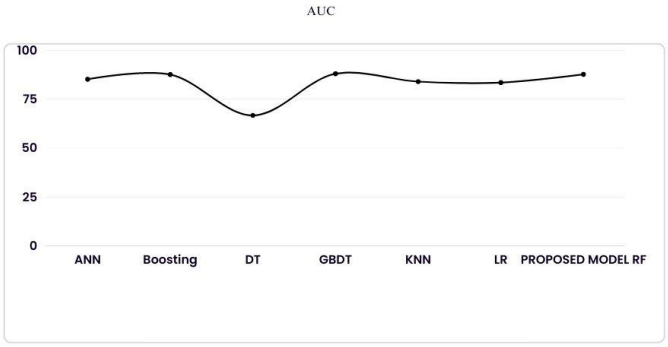


Figure 8: Comparison of AUC Value with various methods.

5. Conclusions

Healthcare checking and anticipation fabrics play an critical part in sparing multitudinous lives, especially when cases are ever set up.In this work, an Random Forest(RF) – grounded healthcare checking frame for internal illness discovery. The proposed show for highlight choice fulfilled the most elevated heartiness utility for all cycles. The proposed RF strategy gives advanced utilities for perceptivity, fineness, review, F1- score, accuracy, True Positive rate(TPV), AUC(Area Under Curve), and Negative Predicted Value(NPV) when compared with the being models. [8] From these comes approximately, we are going state that the proposed RF works well in accurately feting the inner wellbeing condition of cases. Grounded on the linked internal health condition, the specialist can incontinently deliver encourage treatment in the event that essential.

In development, a offer assistance consider will be carried out with other point- determination and optimization methods to move forward the practicality of visionary classifiers for the conclusion of inside wellbeing vaticination.[12]Either, the examine about proposed will be realized working out the wearable progresses and available particulars inside the appear. It's important to note that the use of Random Forest or any other machine learning algorithm in internal illness vaticination should be done with caution, considering ethical considerations

similar as data sequestration, bias, and interpretability, and should always be used in confluence with clinical judgment and moxie.[18]

References

1. Zhou S, Zhao J, Zhang L, Application of Artificial Intelligence on Psychological Interventions and Diagnosis: An Overview,Front Psychiatry, 2022 Mar 17

2. Tianlin Zhang, Kailai Yang, Shaoxiong Ji, Sophia Ananiadou, Emotion fusion for mental illness detection from social media: A survey, Information Fusion, Volume 92,2023

3. Andrew G. Reecea* , Andrew J. Reaganb,Katharina L.M. Lixd , Peter Sheridan Doddsb,c , Christopher M. Danforthb,c* , Ellen J. Langer, Forecasting the onset and course of mental illness with Twitter data,Sci Rep 7, 13006 (2017)

4. Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. (2020), Detection and Classification of mental illnesses on social media using RoBERTa, arXiv Prepr. arXiv2011.11226,23 Nov,2020

5. Khadija Zanna, Kusha Sridhar, Han Yu, Akane Sano, Bias Reducing Multitask Learning on Mental Health Prediction,10TH International Conference On Affective Computing & Intelligent Intteraction (ACII 2022),77 Aug,2022

6. T. E. Ramya ,S.Sindhupriya, An Effective Approach for Mental Health Prediction Using Machine Learning algorithm, International Journal of Engineering Research & Technology (IJERT) CCICS –2022 Conference Proceedings (Volume 10 – Issue 13), 3 Oct,2022

7 .U. Sairam1 , Santhosh Voruganti, Mental Health Prediction Using Deep Learning, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 10, Issue II, Feb 2022

8. Ameer, Iqra & Arif, Muhammad & Sidorov, Grigori & Gomez Adorno, Helena & Gelbukh, Alexander. (2022).,Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning. 10.48550/arXiv.2207.01012.,July,2022

9. Kamal, Mohsin & Khan, Saif Ur Rehman & Hussain, Shahid & Nasir, Anam & Aslam, Khurram & Tariq, Subhan & Ullah, Mian. (2020).,Predicting Mental Illness using Social Media Posts and Comments, International Journal of Advanced Computer Science and Applications. 11. 607-613. 10.14569/IJACSA.2020.0111271.,2020

10. Prof. P.M. Kamde1 , Shantanu Ramteke2 , Ruturaj Kadam3 , Gunj Hundiwala4 , Suryansh5,SVM Classification Technique to Analyze Mental Health and Stress Levels,International Journal of Advanced Research in Science, Communication and Technology (IJARSCT),7, May 2022

11. Karmore, Pravin. (2020). Mental Illness Detection Through Audio Signal Processing,Bioscience Biotechnology Research Communications. 13. 305-309. 10.21786/bbrc/13.14/71,2020

12.Faruk Hossen1 , Sajedul Talukder2 and Refatul Fahad3, AIPSYCH: A Mobile Application-Based Artificial Psychiatrist For Predicting Mental Illness And Recovery Suggestions Among Students, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.13, No.1, March 2022

13. Vaishnavi, Konda & Kamath, U & Rao B, Ashwath & Reddy, N V Subba. (2022). Predicting Mental Health Illness using Machine Learning Algorithms,Journal of Physics: Conference Series. 2161. 012021. 10.1088/1742-6596/2161/1/012021. 2022

14.Xiaofeng Wang1 , Hu Li 1 , Chuanyong Sun1,2 *, Xiumin Zhang3 , Tan Wang1 , Chenyu Dong1 and Dongyang Guo1,Prediction of Mental Health in Medical Workers During COVID-19 Based on Machine Learning ,Frontier doi: 10.3389/fpubh.2021.697850Sep,2021

15. Tutun, Salih & Johnson, Marina & Ahmed, Abdulaziz & Albizri, Abdullah & Irgil, Sedat & Yeşilkaya, Ilker & Ucar, Esma & Sengun, Tanalp & Harfouche, Antoine. (2022). ,An AI-based Decision Support System for Predicting Mental Health Disorders, Information Systems Frontiers. 10.1007/s10796-022-10282-5., 21 April 2022

16. Kang, Min & Kim, Sang & Na, Duk & Kim, Byeong & Yang, Dong & Kim, Eun-Joo & Na, Hae & Han, Hyun & Lee, Jae-Hong & Kim, Jong Hun & Park, Kee & Park, Kyung & Han, Seol-Heui & Kim, Seong & Yoon, Soo & Yoon, Bora & Seo, Sang & Moon, So & Yang, Youngsoon & Youn, Young Chul. (2019). ,Prediction of cognitive impairment via deep learning trained with multi-center neuropsychological test data, BMC Medical Informatics and Decision Making. 19. 10.1186/s12911-019-0974-x.

17. Dongrui, Chen & Shengjie, Wang & Kate, Wen. (2022), Integrated Learning-Based Algorithm for Predicting Graduates' Employment Mental Health, Mathematical Problems in Engineering. 2022. 1-9. 10.1155/2022/5761815.

18. Rezapour, Mostafa & Hansen, Lucas. (2022),A machine learning analysis of COVID-19 mental health data,Scientific Reports. 12. 10.1038/s41598-022-19314-1.
19. Kaur, Harleen & Ahsaan, Shafqat & Alankar, Bhavya & Chang, Victor. (2021),A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets,Information Systems Frontiers. 23. 10.1007/s10796-021-10135-7.
20. Syed, Nasrullah & Jalali, Asadullah. (2022),Detection of Types of Mental Illness through the Social Network Using Ensembled Deep Learning Model, Computational Intelligence and Neuroscience. 2022. 6. 10.1155/2022/9404242.