JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MENTAL HEALTH PREDICTION USING MACHINE LEARNING

MS. DHARSHANASHRI G,
Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

MS. SANTHIYA A K
Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

MS. HEMAVATHI R
Assistant Professor
Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

Abstract – *In today's healthcare landscape, mental health issues have become increasingly prominent, presenting a pressing concern for society. One significant contributing factor to this issue is the widespread lack of awareness among people from all walks of life. Our goal with this initiative is to foster greater awareness among individuals about various mental health conditions, including depression, anxiety, PTSD, and insomnia, through the application of compassionate machine learning techniques. To achieve this, we have gathered data from individuals of diverse backgrounds, encompassing varying ages, professions, genders, and lifestyles, using a survey form. These questions are thoughtfully designed to mirror those often used by psychologists to delve deeply into their patients' experiences. We envision that the implementation of such a system could serve as a beacon of hope, offering timely diagnosis and support to individuals, thus potentially mitigating the onset of a widespread "mental health epidemic.".*

Key Words: **MENTAL HEALTH PREDICTION, MACHINE LEARNING ALGORITHMS, DEPRESSION, ANXIETY, PTSD, INSOMNIA.**

Throughout history, mental health challenges have been present, dating back to as early as the 5th century BC. However, in the modern era, these issues have become more widespread. In India alone, government statistics suggest that a staggering 130 million people may be grappling with various forms of mental illness. This significant number can be attributed to a combination of factors, including a fractured healthcare system and inadequate government support for mental health initiatives. Unfortunately, discussing mental health remains taboo in Indian society, leading to only a small percentage of individuals - approximately 8 to 10 percent - receiving the support and treatment they need. As a result, many individuals silently battle conditions such as depression, PTSD, anxiety, insomnia, and bipolar disorder, contributing to alarmingly high suicide rates. Medical professionals estimate that a substantial portion of those seeking medical assistance - around 35 percent - may be experiencing these mental health challenges. Additionally, the lack of affordability further exacerbates the situation, hindering access to essential support and care for those in need.

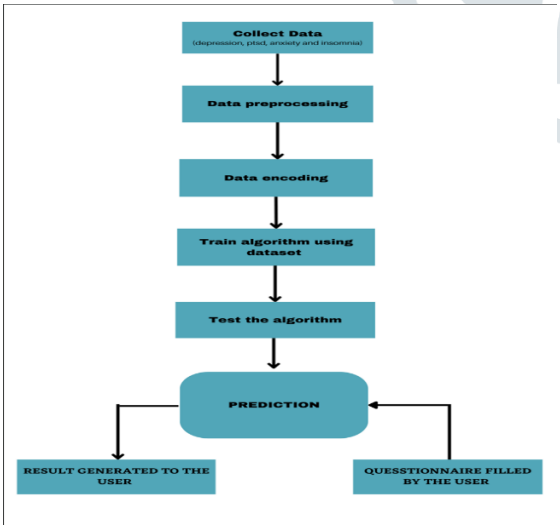
A significant portion of India's population resides below the poverty line, lacking access to basic necessities such as shelter, food, clean water, and healthcare. Consequently, receiving adequate treatment for mental illness remains a distant aspiration for many. Even among the more affluent top 10 percent of the population, accessing treatment is often hindered by the high costs involved.
According to data from the World Health Organization, India has a mere 0.75 psychologists and psychiatrists per 100,000 people, a stark contrast to countries like

Argentina, which boasts 106 psychologists per 100,000 people. To address the looming threat of a mental health epidemic, it is imperative for the government to implement robust healthcare policies and allocate sufficient funding towards mental health There has been many studies and researches where people have been predicting mental health problems like depression and anxiety using the algorithms of machine learning, like decision tree, support vector machine, random forest and convolution neural network for the collection and classification of data from blog posts. For converting text into meaningful vectors like Bag-of-words, topic modeling etc. these techniques are used. In some cases, python programming has also been used for modeling experiments, with the best result among all the classifiers [2] being generated by CNN with the accuracy of 78 percent. In one study 470 se amen were questioned about their occupation, socio-economic background and health condition along sixteen other parameters like age, weight, family earning, marital status, etc. Different machine learning algorithms like logistic regression, naïve bayes, random forest, Catboost and SVM were applied for classification [7]. On getting the result Catboost showed the highest accuracy and precision of 82.6 percent and 84.1 percent respectively. Sau et al. (2017) manually collected data from the Medical College and Hospital of Kolkata, West Bengal on 630 elderly individuals, 520 of whom were in special care. After applying different classification methods Bayesian Network, logistic, multiple layer perceptron, Naïve Bayes, random forest, random tree, J48, sequential random optimization, random sub-space and K star they observed that random forest produced the best accuracy rate of 91% and 89% among the two data sets of 110 and 520 people, respectively. For feature selection and classification, WEKA tool was used in [1]. Change in heart rate, change in blood pressure and acoustics of speech [8],[3] are some of the symptoms of depression and weak emotional state. Diagnosis of Ptsd through speech has been done in recent times. A typical speech-based PTSD diagnostic system consists of three components including data acquisition, feature extraction and classification [6]. In the data acquiring stage a patient is asked questions and the speech dialogue of that patient is recorded. The feature extraction component then processes the speech data and extracts features for the classification component to predict whether or not the subject being interviewed has any level of PTSD. Though other modalities such as EEG, fMRI and MRI were also studied for PTSD diagnosis [5], [4], the data collection process for these modalities is expensive and cannot meet the growing need. Speech is non-invasive and the interview can be conducted remotely via telephone or recording media so that privacy of the patient is strictly protected, making the speech-based method an ideal diagnostic tool for diagnosis and treatment monitoring. In January 2019 research was published about insomnia being predicted through ML algorithms where fourteen parameters were considered. Multiple classification algorithms were applied like DT, random forest, etc. among all the models SVM came out to have the best accuracy of 91.634 percent and the f measure score was 92.13. They have further applied to a dataset of 100 patients where the SVM comes with a good accuracy of 92%. They have declared mobility problems, vision problems as primary factors [9].

initiatives. To diagnose a patient's problem, the doctor may ask the patient to fill out a questionnaire.

The objective of this research paper is to help people understand about their problems and give doctors an overview into their patient's psyche. All of this could only be possible when we use models with the most accuracy.

DESIGN



The system goes through multiple stages before the final value could be predicted accurately. These stages are data collection, data preprocessing, data encoding, training and testing of the algorithm. Once the desired accuracy is obtained, we can integrate the system with an application for real world use.

MACHINE LEARNING ALGORITHMS

To ensure the best possible working of machine learning algorithms it needs to work with some key parameters. Each and every task requires a different model based on the type of data and work is being dealt with. Hence, it is crucial to adjust the model's parameters to increase its utility and accuracy. In our work we have tried to ensure to tune all the models with adequate parameter values and plump for the foremost value for our models

Once the right parameters are selected, we move towards applying machine learning algorithms on our collected dataset of depression, anxiety, Ptsd, insomnia. The collected dataset is usually split into two subsets namely training and testing. It is done to avoid overfitting. In an ideal situation the training and testing dataset is split in the ratio of 80:20 i.e., 80 percent of it goes for training the model and the rest 20 percent is used to test the accuracy of the model. Through research we have selected the following machine learning algorithms to find the best possible algorithm that could give us the most accuracy.

A) Random forest (RF): It is an algorithm that comes under supervised form of learning. The working principle is to create multiple decision trees and all of them are combined to get precise predictions. Hence, it is considered a popular machine learning algorithm.

B) Decision tree (DT): A decision tree comes under supervised learning algorithms where data is continuously split according to the parameter. The tree consists of two

things i.e., decision nodes and leaves. Decision node is the stage where data is split and all the choices made are the leaves.

C) Logistic regression (LR): Is also a part of supervised learning algorithms group used for solving the classification problem. Logistic regression model works with binary variables like 0 and 1, yes and no, etc. It uses sigmoid function or logistic function which is a complex cost function.

D) Support vector machine (SVM): is a prominent algorithm used for both regression and classification. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

E) K-nearest neighbor (KNN): Also known as a lazy or non-parametric algorithm. The algorithm is actually based on feature similarity. The prediction is done according to the calculation of the nearest data points. As it stores all of the training data, it can be computationally expensive when working on a large dataset.

F) Naive bayes (NB): It is a classifier which is based upon conditional probability models. These classifiers are a set of classification algorithms that are based on Bayes Theorem. It's a group of algorithms where a common principle is shared between them. In our study, we have applied Gaussian Naïve Bayes.

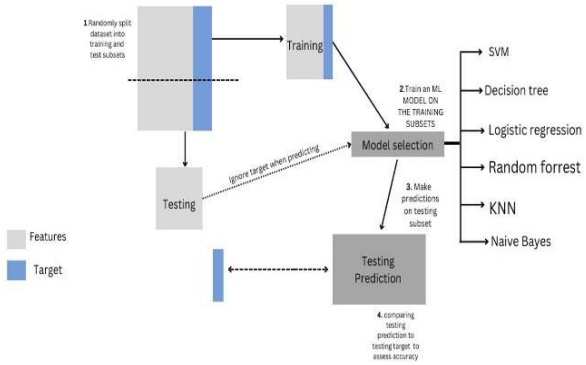
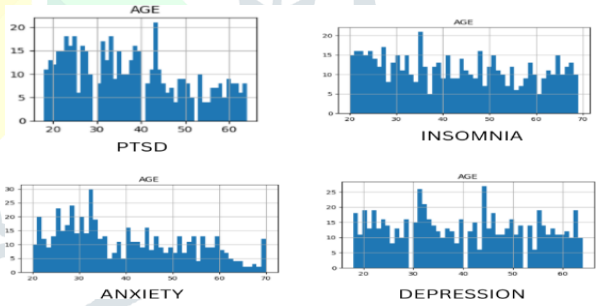


Fig -4.1: Methodological Framework

1. IMPLEMENTATION

The initial step is data collection. We have tried to collect data from different places. There was no standard dataset available which could match our requirements. Hence, we had to collect all the data ourselves. We made a survey form for each disease and distributed, both online and offline for people to fill it. The nature of our questions was objective and situational. We also included people who are currently suffering from some kind of mental illness and are seeing doctors for it and taking some kind of medications. Once the data collection is done, the user's response is converted using numeric values of 0 to 3, and in some cases 0 to 4. Once we had enough data collected, it was moved to preprocessing and is split into two subsets i.e., training and test data set.



It is important to fill out the missing values in the dataset or modify it to increase the quality of the dataset. Once the preprocessing of data is completed, it then moved to feature extraction thenceforth prediction of mental illness.

Fig -5.1: Dataset Overview

WORKFLOW OF THE SYSTEM

In order to put our work in real world use we have deployed our work on web applications. In our application users can take a test for whichever disease out of the four they want, based on the inputs received, our model predicts the severity of the problem they are facing.

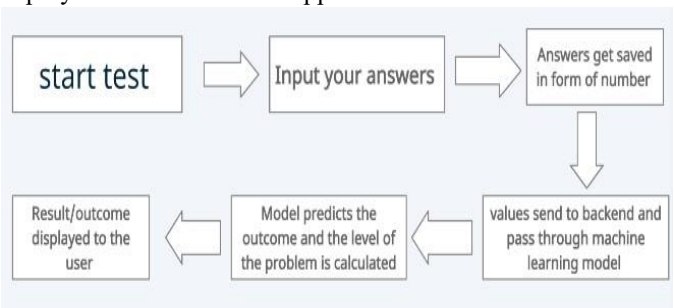


FIG 6- WORKFLOW

RESULTS

In order to achieve high accuracy with the model the data needs to be properly cleaned and preprocessed until it is well fitted. To do this we used python libraries like NumPy, pandas and matplotlib. In order to get the best result for our work we had to pass each of our datasets through multiple ML algorithms like logistic regression, SVM, random forest, k-neighbors etc. Example: - for anxiety, we ran the above-mentioned algorithms and achieved accuracy of 97.27%, 94%, 81%, 80% etc. respectively. Same was the case for the other three diseases which had different levels of accuracy. For our system we chose the algorithm which gave us the true and highest accuracy. We also tried to finetune the hyperparameter to check if the accuracy could be increased more.

CONCLUSION AND FUTURESCOPE

We believe we were able to achieve a good accuracy for each of the four diseases. furthermore, in future we can add more

disease and combine multiple method along with questionnaire to make this process more robust and stronger.

REFERENCES:

[1] Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology. “Healthcare Technology Letters 4 (6): 238-43.

[2] Tyshchenko, Y. (2018)"Depression and anxiety detection from blog posts data."Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia.

[3] R.A.Calvo and S. D’Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Trans. Affective. Comput., 1(1):18-37, 2010.

[4]Q. Zhang, Q. Wu, H. Zu, L. He, H. Huang, J. Zhang and W. Zhang. Multimodal MRI-Based Classification of Trauma Survivors with and without Post-Traumatic Stress Disorder. Frontiers in Neuroscience, 2016.

[5] X. Zhuang, V. Rozgic, M. Crystal and B. P. Marx. Improving Speech Based PTSD Detection via Multi- View Learning. IEEE Spoken Language Technology Workshop. 260-265, 2014.

[6] B. Knoth, D. Vergyri, E. Shriberg, V. Mitra, M. Mclaren, A. Kathol, C. Richey and M. Graciarena. Systems for speech-baseD assessment of a patient’s state-of-mind. WO2016028495 A1. 2015.

[7] Sau, A., Bhakta, I. (2018) "Screening of anxiety and depression among the seafarers using machine learning technology."Informatics in Medicine Unlocked :100149.

[8] S. R. Krothapalli and S. G. Koolagudi. Characterization and recognition of emotions from speech using excitation source information. Int. J. Speech Technol., 16(2):181-201, 2012.

[9] R. Ahuja, V. Vivek, M. Chandna, S. Virmani and A. Banga, "Comparative Study of Various Machine Learning Algorithms for Prediction of Insomnia", 2019.

[10] Y. Kaneita et al., "Insomnia Among Japanese Adolescents: A Nationwide Representative Survey", Sleep, vol. 29, no. 12, pp. 1543-1550, 2006.

[11] P. Singh, "Insomnia: A sleep disorder: Its causes, symptoms and treatments

SVC()

	precision	recall	f1-score	support
0	0.90	0.94	0.92	4252
1	0.94	0.89	0.91	4142
accuracy			0.92	8394
macro avg	0.92	0.92	0.92	8394
weighted avg	0.92	0.92	0.92	8394

RandomForestClassifier()

	precision	recall	f1-score	support
0	0.89	0.90	0.90	4252
1	0.90	0.88	0.89	4142
accuracy			0.89	8394
macro avg	0.89	0.89	0.89	8394
weighted avg	0.89	0.89	0.89	8394

AdaBoostClassifier()

	precision	recall	f1-score	support
0	0.84	0.92	0.88	4252
1	0.91	0.83	0.86	4142
accuracy			0.87	8394
macro avg	0.88	0.87	0.87	8394
weighted avg	0.88	0.87	0.87	8394