Evaluation of Mental Stress using Predictive Analysis

Prof. Yogesh Pingle
Information Technology Department
Vidyavardhini's College of Enginnering and Technology
University of Mumbai

***Abstract*—Mental health issues like depression, anxiety sleep disorders can start at the young age due to the mental stress that the students go through in their day-to-day lives. In India, more than 5 Cr individuals experience the ill effects of mental stress. According to the survey done in the University of Melbourne, on an average every four children out of five are suffering from the mental disorders. Psychological health plays a very important role in molding the development characteristics and academic performance of the child. To solve this problem and develop a cost-effective solution considering the expensive fees for counselling that cannot be affordable by everyone, we built a robust model using machine learning algorithms to determine the stress level of the students. To incorporate this, we have collected the real time survey data of students belonging to different streams from various colleges based on questionnaire. We have used the ensemble learning method to build the model that will predict the stress level based on most votes.**

Keywords— Spatial data, Psychological factors, Ensemble Learning, Machine Learning, Mental Stress, Stacking, CNN-Adaboost

I. INTRODUCTION

Stress is dined as seen trouble brought about by a communication between an individual and their condition. While individuals can deal with pressure better or more relies upon their general adapting aptitudes, encountering pressure also often is known to influence prosperity and physical and emotional well-being adversely. Stress is considered as a important concern, as stress related chronic diseases has a negative impact on the mental as well as physical health of the people.

To determine and manage the mental stress among the individuals, the traditional approach is to concern the psychiatrists. Individuals can go to psychiatrists who will help them to identify the causes of stress and manage the stress levels accordingly. But due to the expensive fees that cannot be affordable by everyone, we proposed the technical approach for determining the mental stress by using the machine learning model approach. This technical approach of determining stress eliminates the need of individuals to visit the psychiatrists physically.

Stress could be due to various problems in an individual's life. It could be due to the factors like work pressure, parental or society pressure, attempting to achieve unrealistic goals, school activities pressure or psychological trauma. We focus on three psychological concerns - Depression Disorder, Anxiety Disorder & Sleep Disorder. Depression disorders can be described as different mood swings like anger or sadness which may cause stress. Anxiety disorders like feeling of nervousness or worry or a concern of something uncertain can happen may lead to stress of an individual. Sleep disorders like insomnia also causes stress for an individual.

Machine learning includes the investigation of calculations that can separate data naturally. It utilizes information mining procedures and another learning calculation to manufacture models of what is going on behind collected data with the goal that it can foresee the future results. The key idea behind this is to collect and manage huge amounts of real time data and then use it to predict the stress level among the students. To make the collected data consistent, the process of data cleaning is done.

The following sections are divided as follows. Literature review has been discussed in Section 2; the proposed system is defined in Section 3. Algorithms used for building the system has been discussed in Section 4. Results from the system have been obtained in Section 5. Finally, In Section 6, the paper is concluded.

II. LITERATURE REVIEW

Paper [1] uses CNN-UDRP algorithm which is Convolution Neural Network based on Unimodal disease risk prediction (CNN-UDRP) which is predicting the risk of heart diseases. The accuracy of CNN-UDRP for heart disease is approximately 60 - 65%. Hence the performance is better for risk prediction of heart diseases. The result of the system which has been shown are in the form of low, medium as well as high risk.

Paper [2] focuses on the linguistic characteristics of Facebook and Twitter data. It evaluates how models trained on Facebook data perform at predicting stress on the twitter language in a cross-domain setting. Hence the need for transfer learning is determined to apply the stress model on Twitter language which is trained on Facebook language.

In paper [3], the proposed system is divided in three modules - Sensor (hardware circuit) Device connected to microcontroller, Mobile App for showing the results & Cloud Storage where data is being stored. Electroencephalography

(EEG) brain waves are considered as an important parameter for detecting the levels of stress. It identifies stress and alert is provided to the user on a mobile application which is used to help the user for alerting the stress level anytime. The system depends machine learning algorithms which are used as well as on the datasets which are used.

Occupational Stress is an important health-related concern. Paper [4] focuses on predicting the occupational stress. The data is collected by conducting a survey among the people working at different sectors of occupation. The survey focuses on three factors, i.e., psychosocial, environmental, and physical factors. In this paper, the analysis is done by using metric (Support Vector Machine and neural Network) and non-metric (Decision tree and Random Forest) approaches. It concludes that the metric approaches provide good accuracy.

Paper [5] predicts the comorbid risk of the patients by analyzing the longitudinal EHR data from the other patients by using a novel deep learning framework. It performs the prediction of risk using the data of patients with single chronic disease of interest. With the help of these contributions, it determines the heterogeneous characteristics within the EHR data of the patients. Finally, the proposed model is validated quantitatively and qualitatively on the EHR data of the patients.

In paper [6], Chen proposed CNN-based (CNN-MDRP) which is also called as Multi-modal Disease Risk Prediction algorithm which uses SQL as well as no SQL data (in JSON format) from hospitals in China. Chen concluded that the proposed system for disease risk prediction provides an accuracy of about 94.80% which is faster as compared to CNN-UDRP algorithm [1].

Paper [7] focuses on predicting the next level of stress which is based on the current stress level, road or environment conditions and the driving actions of the drivers. The results describe that every user handles the stress is different for other users. It is stated that it is easy to provide the level of stress based on the current behavior of that user and the conditions for a single-user, but these results cannot be generalized in case for the cross-user scenario.

Paper [8] conducts a study of stress, mobile usage as well as data which is captured from sensors. This study is based on different observations. It makes the use of an Android application The Stress Collector (TSC) to collect the stress data. After installing the application on an Android mobile using (.apk) file, it continuously runs in the background to collect the mobile usage and sensor data periodically. This paper describes significant correlations among stress and smartphone data, and it performs better than the back reported levels. It helps to encourage for further investigations on stress prediction using smartphones.

Paper [9] considers the interrelated features and the possible effects of the future events by predicting future stress level of the teen age students from the usage of micro-blog. The effectiveness of the methodology is verified by the experiments. The paper also helps to encourage the work towards teenager's future stress level trends.

Paper [10] presents a system for predicting various stress levels from (ECG - electrocardiogram) data. In this, the Galvanic Skin Response (GSR) signals are being recorded at a 31 Hz rate and ECG signals were continuously collected at a rate of 496 Hz. The paper states that using a blend of Heart Rate Variability and electrocardiogram features, prediction is difficult of GSR if it is in the highest percentile or lowest 20 percentile with >97% accuracy. It suggests that by using consumer ECG devices, we can predict whether the user is stressed or not without the requirement of GSR estimations.

III. PROPOSED SYSTEM

The main thought of the paper is to foresee the stress level among the students. This stress could be due to various factors like depression, anxiety, or sleeplessness. To determine the stress of individuals at these factors mentioned above, a questionnaire focusing on these factors is prepared and circulated to generate survey data. The data is collected by conducting the survey and the dataset is generated. While conducting the survey, students were asked to fill their personal details like Name, College Name, Stream, Age. But some of the introverts might leave these fields empty and hence data cleaning is done to remove the missing information from the dataset.

The Proposed System Architecture is as shown in figure 1. After Data Collection via Questionnaire, the dataset is uploaded to perform data cleaning. Then the algorithms like KNN, Random Forest & CNN-Adaboost are applied to the refined dataset which will generate an ensemble learning model for future stress prediction. Finally, the built model will be used for predicting the final stress level as Low, Medium, or High.

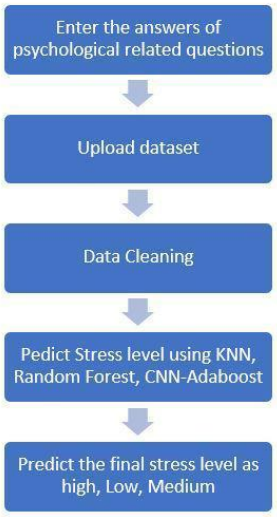

Fig. 1. Proposed System Architecture

A. Dataset

The data is collected by conducting survey among the students studying in different streams like Engineering, Medical, Law, etc. belonging to different colleges in Mumbai. The dataset collected is of the students be-longing to age group 15-30 years. For conducting the survey, a questionnaire consisting of 15 questions was circulated that focused on three categories (Depression, Anxiety, Sleep) of disorders for stress prediction. Each of questions has four options of answers. The options are given as -" Not at all";" Several Days";" More than half the days";" Nearly Every Day". Along with this, the other attributes are Student Name, College Name, Stream of Study and Age.

The dataset consists of the questions as an attribute and the class label that is the outcome of the stress prediction. Each tuple in the dataset consists of integer values for the attributes. The class label values are provided by a psychiatrist depending upon the input values provided by the students as a part of survey data.

B. Data Cleaning

While conducting survey, the attributes Student Name, College Name, Stream of Study and Age are not mandatory. Hence these fields consist of missing values. Data Cleaning is done to remove the unwanted data from the dataset.

C. Prediction of Stress level

The main purpose of this system is to predict the stress level of students. The target attribute class=f0,1,2g predicts the stress level in which 0 defines low stress, 1 defines medium stress and 2 defines high stress.

IV. ALGORITHMS

The combination of machine learning as well as deep learning algorithms is used to find the stress level which is used further for predicting diseases. The (CNN) Convolutional Neural Network which is deep learning algorithm is used first to cast vote for the level of stress. Then various machine learning algorithms like K- Nearest Neighbours (KNN), Random Forest and CNN-ADABOOST then predicts the class labels. The final class label is decided based on majority of number of votes.

A. CNN

CNN also known as Convolutional Neural Network which is a deep learning neural networks is also sometimes called as space invariant or shift invariant neural networks. It is composed of convolutional layer which is the important building block of CNN, pooling layer and a number of fully connected layers. Each time bias and weights are updated to predict the class labels through fixed number of epochs. Although the interpretation of the knowledge learned by the Convolutional neural networks is difficult for the human user, they perform well in prediction of the class labels.

B. KNN

KNN is the K- Nearest neighbour algorithm which is lazy learning algorithm that classifies the given unknown instance of the query into the correct target class by performing some mathematical calculations based on the query fired and the tuples in the dataset. KNN is useful because it involves simple calculations to predict the class labels mainly Euclidean distance.

We are going to use in our proposed methodology. It is a non-parametric technique of classification basically divided into structured or structureless Neural Network.

C. RANDOM FOREST

Random Forest, an efficient ensemble learning method comprises of a series of decision tree classifiers casting vote to predict the class label. Each decision tree classifier takes the random sample data with or without replacement and predicts the class label. The final class label is decided based on the majority voting. This algorithm is more accurate in predicting the class labels, more error prune than normal decision trees and less likely to overfit the training model.

D. CNN-ADABOOST

Adaboost is the short form of adaptive boosting. It is mainly referred as one of the main techniques to increase the efficiency of the algorithm and can be implemented in combination with any of the machine learning or deep learning algorithms. CNN- adaboost updates the weights iteratively by focusing on the data training tuples which are misclassified during the previous iteration and as-signs them higher weights. Due to this nature, CNN-adaboost shows more accuracy than that of the normal CNN algorithm.

V. IMPLEMENTATION / RESULTS

A. Implementation Setup

For the purpose of implementing the proposed system, we divided the dataset into two parts – 70% of data in training dataset which contains nearly 1600 tuples and 30% of data in testing/validation dataset which comprises of nearly 600 tuples. The model is first developed and trained using CNN algorithm and then tested using the same. For the same dataset, we developed stacking model and tested it. The accuracy of both the models is measured and its comparison is described in Figure 2.

B. Parameters of Comparison

The following bar chart in figure 2 clearly shows that the stacking outperforms over the CNN.

Further, to compare the two models we need to understand the concept of confusion matrix. The following table 1 represents the standard format of confusion matrix.

The terms are explained as follows:

TruePositives (TP): The classifier is being correctly classified class tuples as positive.
TrueNegatives (TN): The classifier is being correctly classified class tuples as positive.

FalsePositives (FP): The classifier is mistakenly being classified class tuples as negative.

FalseNegatives (FN): The classifier is mistakenly being classified class tuples as positive.

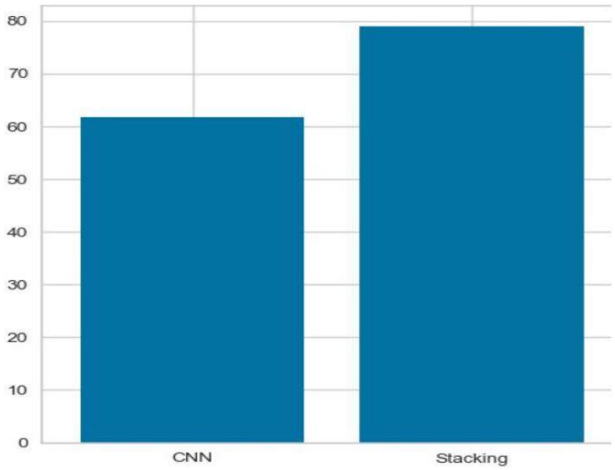

Fig. 2. Accuracy Comparison of CNN and Stacking algorithms

TABLE 1. THE STANDARD CONFUSION MATRIX

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

As the important factor of interest is "High Stress" that is our model should be able to predict the stress level as high more accurately whenever the stress level is high. we must incorporate the different types of measuring the efficiency for comparison of both the models. We measure the efficiency of both the algorithms using the following parameters.

Accuracy: Accuracy defines the overall correctness of the model that is built. It is important factor to compare the efficiency of various machine learning models or algorithms.

Calculation of Accuracy is as follows:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalNumberofTuples}$$

Precision: Precision defines how often the model determines the Positive examples (the main class of interest) correctly. It can be measured as the total true positives divided by the total (False as well as True) positive examples classified.

Calculation of Precision is as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Recall: Recall is the measure of completeness in the classification report. It defines the total positive tuples predicted so far by the model developed.

Recall is being calculated as follows:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

F1-Score: One way of integrating Precision as well as Recall unit is F1 Score. It is represented by following formula:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Support: Support is used to measure of actual occurrences of the main class of interest in the dataset. Support determines the total percentage of the dataset instances which follows the predictions of the classifier model which we have built. Higher the value of support, better the classifier model we built.

TABLE 2. COMPARISON OF VARIOUS EFFICIENCY PARAMETERS

Parameters	Classes	CNN	Stacking
PRECISION	Low	0.32	0.76
	Medium	0.61	0.69
	High	0.9	0.92
RECALL	Low	0.8	0.84
	Medium	0.66	0.86
	High	0.56	0.88
F1-SCORE	Low	0.79	0.8
	Medium	0.64	0.9
	High	0.69	0.92
SUPPORT	Low	61%	68%
	Medium	62%	77%
	High	76%	86%

We can see in the above table 2 that Stacking performs better than that of the Convolutional neural network. This is mainly because of the reason that rather than taking into

consideration a single classifier, we consider various classifiers to predict the class labels and Ada-boost further more focuses on the data tuples which are classified wrongly during previous iterations in the next iterations. This nature of Ada-boost increases the overall efficiency of the stacking classifier build.

VI. CONCLUSIONS

In this paper, we compared the model which was trained using the CNN algorithm and stacking algorithms mainly – KNN, Random Forest and CNN-Adaboost. We compared the accuracy of both the models and found that the accuracy of the Stacking model was much better in the prediction of the stress level amongst the students. We got highest accuracy nearly about 88.86% using stacking algorithms and got the correct risk level prediction of all the classes most of the times. We feed the student data to the model using the questionnaire-based approach and got stress predictions into three major classes namely High, Low & Medium. The model gives us low cost and low time complexity and is affordable to all people with varying financial backgrounds. In future we may add more detailed description and try to figure out the root causes of stress.

REFERENCES

[1] Sayali Ambekar, Rashmi Phalnikar," Disease Risk Pre-diction by Using Convolutional Neural Network", IEEE 2018.

[2] Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, Lyle H. Ungar," Understanding and Measuring Psychological Stress Using Social Media", ICWSM 2019.

[3] Chrandrasekar Vuppalapati, Mohamad S Khan, Nisha Raghu, Priyanka Veleru, Suma Khursheed," A System To Detect Mental Stress Using Machine Learning And Mobile Development", 2019

[4] S.K. Yadav, Arshad Hashmi," An Investigation of Occu-pational stress Classification by using Machine Learning Techniques", Vol.-6, Issue-6, 2018.

[5] Jinghe Zhang, Jiaqi Gong, Laura Barnes, "HCNN: Hetero-geneous Convolution Neural Networks for Comorbid Risk Prediction with Electronic Health Records", IEEE 2017.

[6] Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", IEEE 2017.

[7] Mario Munoz˜-Organero, Victor Corcoba-Magana, ˜" Pre-dicting Upcoming Values of Stress While Driving", IEEE 2016.

[8] Thomas Stutz1, Thomas Kowar, Michael Kager, Martin Tiefengrabner, Markus Stuppner, Jens Blechert, Frank H. Wilhelm, Simon Ginzinger," Smartphone Based Stress Prediction", UMAP 2015, LNCS 9146, pp. 240–251, 2015.

[9] Yiping Li1, Jing Huang, Hao Wang, Ling Feng," Pre-dicting Teenager's Future Stress Level from Micro-blog", IEEE 2015.

[10] David Liu, Mark Ulric," Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors", Stanford CS 229: Machine Learning, Autumn 2013-2014.

[11] Prof. Dhomse Kanchan, B Mr. Mahale Kishor," MStudy of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis" IEEE 2017.

[12] Lu Zhang, Xiaopeng Fan, hengzhong Xu, "A Fusion Financial Prediction Strategy Based on RNN and Rep-resentative pattern Discovery" IEEE 2017.

[13] Lee Ker Xin, Nur'Aini Abdul Rashid,"Predicting Gen-eralized Anxiety Disorder Among Women Using Random Forest Approach" IEEE 2016.

[14] Prof. Dhomse Kanchan, Mr. Mahale Kishor," Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", IEEE 2017.

[15] S. Dilli Arasu, Dr. R. Thirumalaiselvi, "A novel impu-tation method for effective Prediction of coronary kidney disease", IEEE 2017.

[16] Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avi-dan" Predicting Hospital Readmission via Cost sensitive Deep Learning", IEEE 2017.