INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

* **Corresponding author**.

Salma.jce@gmail.com

Competing Interests: None

Decoding Minds: Estimation of Stress Level in Students using Machine Learning

Salma S Shahapur[1]*, **Praveen Chitti**[1], **Shahak Patil**[1],
Chinmay Abhay Nerurkar[2], **Vijay Shivaram Shivannagol**[1],
Vinayak C Rayanaikar[1], **Vishwajit Sawant**[1], **Vadiraj Betageri**[1]

1 Department of Electronics and Communication, Jain College of Engineering, Karnataka, India
2 Principle Software Engineer at Microsoft, New York, United States of America

Abstract

Objectives: Develop a predictive model to categorize student's stress levels and support early interventions based on self-reported data, academic performance, and study load. This will help to receive early diagnosis and treatment. **Methods**: In this work the data set used was downloaded from a website called KAGGLE. The dataset has more than 6000 samples, the parameters considered in this dataset are Anxiety level, self-esteem, mental_health_history, depression, headache, blood pressure, sleep_quality, breathing_problem, noise_level, living conditions, Safety, basic needs, academic performance, study_load, teacher_student_relationship, future_career_concerns, social support, peer_pressure, extracurricular_activities and bullying which directly or indirectly has an effect on the mental health of the students, so basically here 20 different types of factors are taken into consideration. This specific Research Work employs Machine Learning (ML) approaches to analyze stress levels in students from stress-level text data. Logistic Regression (LR) with 89.46%, KNeighbors with 92.8%, Decision Tree with 94.5%, Random Forest with 95%, and Gradient Boosting with 90.15%, algorithms are used to determine stress levels. **Findings:** Several significant findings have emerged in this research on predicting mental stress levels in students using machine learning. Studies on feature importance emphasize the importance of sleep quality, depression, mental_health_history, academic performance, and participation in extracurricular activities and several other parameters as critical criteria for accurate prediction. Multimodal techniques that integrate data from mental health history, family history, and academic records provide a more complete picture of a student's life. Temporal dynamics are important, as stress levels fluctuate throughout time as a result of academic and personal events. Some research goes beyond prediction, investigating intervention options based on tailored stress management suggestions. **Novelty:** In order to anticipate student's mental stress, this study presents a novel machine-learning architec-

ture. This methodology attempts to give early identification of students' mental health at risk by leveraging diverse data sources and using different machine learning algorithms with a very high accuracy level.

Keywords: Stress Level; Students; Machine Learning; Decision Tree; Physio Bank

1 Introduction

Stress among students has increased in the academic setting, resulting in a range of psychological and physical symptoms. Unmanaged stress has negative effects that go beyond the classroom and affect people's general well-being and long-term mental health. Conventional techniques for measuring stress frequently depend on self-reporting, which is biased and subjective. In addition, prompt intervention is hampered by the absence of real-time monitoring tools. The limitations of current student stress measurement techniques include their reliance on self-reporting, infrequent assessment frequency, and inability to offer real-time insights. These drawbacks highlight the necessity of a more continuous, objective, and non-intrusive method for tracking and estimating stress levels. In order to develop such a solution, machine learning presents a promising path because it can analyze various data sources and identify patterns that indicate stress.

- **Current Stress Assessment Challenges**

Prejudice and Subjectivity: Self-reported stress evaluations are prone to subjectivity and individual interpretation, which may cause stress levels to be overestimated or underestimated. This subjectivity makes it extremely difficult to assess students' stress levels with any degree of accuracy.

Time-Limited Resolution: Conventional methods of assessment frequently take place at regular intervals and only give a glimpse of a student's stress level. This temporal restriction makes it more difficult to identify extended periods of elevated stress or abrupt spikes, which hinders prompt intervention.

Privacy Issues: There are ethical and privacy issues with many of the stress assessment techniques currently in use since they require invasive procedures or the gathering of sensitive data. One of the biggest challenges in this field is striking a balance between the requirement for accurate assessments and the defense of students' privacy.

- **Research's Significance**

This study is important for a number of reasons:

Improved Prompt Intervention: When stress levels rise, educators and mental health professionals can act quickly to prevent the worsening of mental health problems by using machine learning algorithms for real-time stress estimation.

Constant and Objective Surveillance: Relying less on subjective self-reporting and providing a more thorough understanding of a student's stress patterns, machine learning offers an objective and continuous method of stress assessment.

A Method for Preserving Privacy: The ethical issues with conventional stress assessment techniques are resolved by creating a machine learning model that respects privacy and produces accurate stress estimates.

Mental health refers to a medical state that impacts an individual's thoughts, emotions, and social interactions. These issues have shown that mental illness requires innovative protective and therapeutic measures and has significant societal ramifications. A critical first step in putting such strategies into practice is early mental health detection. Machine learning uses advanced statistical and probabilistic techniques to create adaptive systems that improve with time. It enables

multiple researchers to extract valuable insights from the data, design personalized experiences, and develop solutions for artificial intelligence systems. Working with unsupervised data is the main objective of unsupervised learning.

Presenting a systematic research study on machine learning techniques for predicting and identifying mental health problems in students is the main ambition of this work. In order to predict mental health, machine learning models are made to review a variety of data sources, such as study load, sleep_ quality, physiological signals, and behavioural patterns. These models forecast an individual's mental health state based on past data.

The World Health Organisation (WHO) defines mental disorders or mental health issues as an aggregate of atypical conduct, emotions, thoughts, and interpersonal relationships. Intellectual health troubles can affect people's everyday sports and interpersonal relationships. The current state of machine learning could aid in knowledge extraction and enhance the standard of medical procedures. Everybody occasionally feels anxious, depressed, or worried. But the variety of those who are afflicted by mental illness is much less in numbers. Many mental illnesses have been identified and classified. These encircle a wide range of conditions, such as generalized anxiety disorder, anxiety, mood, attention, and behaviour disorders. The illness itself can influence the signs and symptoms of mental illness.

- **Typical symptoms include**

1. Protracted feeling of sorrow.
2. Overindulging in emotional fluctuations.
3. Abandoning social interactions, family, or activities.
4. Depleted energy or encountering troubles in regularly dozing.
5. Encountering animosity, fury, or aggression.
6. Encountering hallucinations, auditory hallucinations, or feeling suspicious.
7. Mental illness often elicits contemplations of suicide or demise; its impacts are all-encompassing.

It affects people of all ages, both males and females of all economic and educational levels are also affected. Thankfully, there is usually a treatment. That's why, we try to make it stronger by diverting attention or using drugs, alcohol, or suicidal behavior as a means of self-medication when we ignore emotional mistakes. In an effort to hide our issues from others, we bottle up our issues. Hopefully, it will be things will get better in time. Or we give up and convince ourselves this sounds like "how we are." You can develop some habits that will lift and strengthen your spirit. Make your patience and life more interesting but mental health is like physical fitness. It takes work to create and preserve it. These days, because of all the ways life can be negative it affects our ability to regulate our emotions and requires more work to maintain good mental.

Mental health is generally self-reported by the patient and is difficult to use psychological questionnaires for specific emotional and social processes. Many people with mental health or emotional problems can recover with the right help and supervision of care. Machine learning is an approach to sophisticated statistics and probability Methods of building systems that can be learned from experience. It is considered a very helpful tool to predict mental health. This allows many researchers to perform passive construction of Intelligent systems, providing a personalized experience, and extracting valuable information issues.

Future events have been predicted and classified using popular machine learning algorithms such as vector support machines, random forests, and artificial neural networks. The most famous one in the process of research, learning, and examinations, learning is automatically supervised Teaching, especially in medicine, is about disease prognosis. All the data Examples should reflect the terminology, attributes, and values of supervised learning. More specifically, Training data sets are used in supervised learning, which is a classification technique. The next one is in other words; unsupervised learning makes unsupervised predictions. Without using data, the primary purpose of unsupervised learning is supervision.

An estimated 450 million human beings suffer from mental fitness problems along with despair, schizophrenia, interest deficit hyperactivity sickness (ADHD), autism spectrum disease (ASD), and other conditions. Children and teenagers Eighteen people also experienced mental health disorders in addition to adults. One plus Among the most important and widespread public health problems are mental health disorders. For example, depression increases the risk of suicidal thoughts and attempts and is one of the main thoughts that cause disability. Today scientific and medical advances have led to tremendous growth Effective medical treatment and technology have made it possible to diagnose the disease itself in the Early stages of development.

- **Purpose of the Research**

Creating a Model for Machine Learning: The main goal is to create and hone a machine learning model that can reliably determine stress levels from a variety of data sources.

Monitoring in Real Time: Allow for the real-time tracking of stress levels in order to deliver prompt feedback and assistance.

Maintaining Privacy: Reduce the amount of sensitive data that is collected and used in the developed model to ensure privacy is respected.

In [1] researchers have used machine learning algorithms to predict the severity of depression, stress, and anxiety in college students. Data from four hundred students were gathered using DASS21. This standard questionnaire is used to gauge the typical symptoms of depression, stress, and anxiety. Mild, normal, moderate, severe, and extremely severe were the different severity levels. The Support Vector Machine, KNN, logistic regression, decision tree, and naive Bayes classification algorithms were used.

In [2] using machine learning algorithms, the author has presented a study on earlier research on stress detection. Provided a framework for classifying stress levels and analyzing them using the PhysioBank dataset. The suggested gradient boost algorithm has been successfully used to implement stress level classification, according to the statistical analysis that was performed for feature selection and extraction. The assessed outcomes demonstrated that the suggested model obtained 83.33% accuracy, 75% specificity, 75% sensitivity, 90% positive predictive value, 90% negative predictive value, 16.66% error rate, 83.33% F1_Score, and 75% recall.

In [3] The results show that all three algorithms accurately predicted and classified stress levels. Out of the three algorithms evaluated, the Random Forest algorithm demonstrated the best performance in terms of accuracy, with the Support Vector Machine and Artificial Neural Network algorithms following closely behind. The analysis results indicate that EDA signals were the most informative features for predicting stress levels. Comparing the result from previous studies suggests that the approach employed here outperforms existing methods. In conclusion, the study demonstrates the feasibility and effectiveness of using machine learning techniques to predict and classify mental stress. The proposed system has potential applications to improve stress management and promote mental health in various fields, including healthcare, sports, and workplace environments.

In [4] the study's author discussed earlier findings on machine learning-based stress detection research studies. Our system for identifying user stress is predicated on the random forest (RF) algorithm and support vector machine (SVM). In this study, we examined the ways in which different factors impact users' stress levels. The dataset, which includes 270 participants overall, was gathered from surveys given to individuals of various ages. Based on the structured questionnaire is the data set. We employed RF and SVM classifiers for the classification. Parameters for accuracy, precision, and recall were used to gauge performance. It was found that SVM outperforms RF in accuracy (80.2%).

In [5] the information came from the replies of 3561 (62.58%) of the 5690 undergraduate students at University A, a national university in Japan, who finished the health survey in 2020 and 2021. They conducted two analyses: the first predicted a mental health issue in 2020 based on demographics, health survey responses, and response time in the same year; the second predicted a mental health issue in 2021 based on the same input variables as the first analysis. They contrasted the outcomes of several machine learning models, including XGBoost, LightGBM, random forest, elastic net, and logistic regression. Using the selected model, the outcomes with and without answering time conditions were compared.

In [6] to predict stress using data from Fitbit smart watches, researchers have employed random forests, decision trees, support vector machines, naive bayes, logistic regression, and k-nearest neighbor models with k-fold cross-validation and voting ensemble learning. The highest binary class prediction of stress accuracy, 94°, was achieved by the support vector machine model. According to our research, the combination of voting ensemble and support vector machine may be a more useful tool for detecting stress in smart watch users. Their work closes a number of methodological gaps in earlier research on smart watch-based automatic stress prediction. It has also been observed that using ensemble learning techniques improves the model's performance.

In [7] it develops a novel hybrid convolutional neural network (CNN) and long short-term memory (LSTM) model to enable early warning of common mental health risks like depression in teenagers. The model is trained on a large clinical dataset with over 50,000 adolescents encompassing electronic health records and neuroimaging data. This allows discovering subtle predictive patterns within clinical encounters and brain structure or function at the population level. The methodology centers on a tailored CNN and LSTM architecture to comprehensively model the multivariate clinical time series data, combining CNN spatial feature learning with LSTM temporal sequence modeling. The hybrid approach leverages complementary techniques to analyze the neuroimaging and electronic health record data. The model achieves an accuracy of 95%, AUC of 97%, precision of 94%, recall of 91%, and F1 score of 92% on held-out test data, significantly outperforming prior state-of-the-art models which had lower accuracy. Detailed predictive feature analysis and clinical validation confirm the model's utility for early mental health risk screening and targeted intervention in teenagers. Overall, the study demonstrates an effective approach to integrating big

data, deep learning, and rigorous evaluation for developing more accurate, explainable and useful models to support adolescent mental healthcare through data-driven early warning systems.

In [8] it was analyzed, a weekly association between estimated sleep measures and perceived stress for participants (N = 525). Through mixed-effects regression models, we identified consistent associations between perceived stress scores and average nightly total sleep time (TST), resting heart rate (RHR), heart rate variability (HRV), and respiratory rate (ARR). These effects persisted after controlling for gender and week of the semester. Specifically, for every additional hour of TST, the odds of experiencing moderate-to-high stress decreased by 0.617 or by 38.3% (p<0.01). For each 1 beat per minute increase in RHR, the odds of experiencing moderate-to-high stress increased by 1.036 or by 3.6% (p<0.01). For each 1 millisecond increase in HRV, the odds of experiencing moderate-to-high stress decreased by 0.988 or by 1.2% (p<0.05). For each additional breath per minute increase in ARR, the odds of experiencing moderate-to-high stress increased by 1.230 or by 23.0% (p<0.01). Consistent with previous research, participants who did not identify as male (i.e., female, nonbinary, and transgender participants) had significantly higher self-reported stress throughout the study. The week of the semester was also a significant predictor of stress. Sleep data from wearable devices may help us understand and to better predict stress, a strong signal of the ongoing mental health epidemic among college students.

In [9] the goal of this research is to better understand how to create predictive algorithm models for client Dialectical Behavior Therapy (DBT) skill selection in Digital Counselling Environments (DCEs) by utilizing neurocognitive data obtained through functional near-infrared spectroscopy (fNIRS). This study specifically aims to investigate the usefulness of combining machine learning algorithms with neurocognitive data to predict the success of client selection strategies in a virtual environment. The study recruited fifty participants from rural schools in the United States. Three conditions for developing DBT skills were used by student participants: face-to-face, virtual reality-based, and time-delay-control. In each condition, neurocognitive responses were gathered during the learning phase and utilized to forecast the likelihood that a client would successfully choose DBT techniques when given the chance during the assessment phase. The resulting algorithm's average predictive accuracy was approximately 83%. The findings also demonstrated the potential to record cognitive alterations in almost real-time (~300 ms) during DBT skill development. The results of this study show that neurocognitive data in DCEs can be used to better predict client outcomes, improve the caliber and dependability of AI counselors, and enhance counselors' use of client-based analytics in both in-person and online counseling settings.

In [10] the real-time dataset that is being used is based on surveys and provides information about the daily routines and circumstances of different people. A variety of questions are included in the survey questionnaire to gauge respondents' psychological states and levels of stress. The models are trained on this dataset in order to ascertain the frequency of any psychological instability. A comparison is made between different bipolar classification techniques and their accuracy of performance using the real-time dataset. One important factor in lowering the likelihood of severity is the detection of psychological instability.

So, as of now from the various previous research work done published it can be clearly visualized that in some cases there is a high accuracy but the parameters involved are very less in number and whereas in some cases the parameters are in higher number, but the accuracy is not reaching the expected value. Also, there is a large gap for involving technological advancement and there useful applications in the medical sector so by using such modern machine learning technology we can easily predict the outcome in a very early stage.

2 Methodology

An online survey was used to collect an open data set. Many elements are involved. The survey included respondent's attitudes about mental health. The purpose of this research is to find out the respondent's perception of mental health at work. The data set used for this project has many missing, inconsistent, and redundant values generally Therefore, data cleansing is necessary to prepare the data for the machine learning processing model. It is observed that the data in the columns has unusual and exaggerated values. The process or steps followed in this research work is shown in terms of flow chart in Figure 1. This work is superior to others because it is using more than 15 different parameters which have a major impact on the student mental health mainly the stress level. Here the unique predictive system is used which takes the multiple inputs and display the stress level of a particular student with a very high accuracy level.

2.1 Different Stages Used in This Work

In this research work a range of machine learning techniques, including Logistic Regression, Random Forest, Decision Tree, Kneighbors Classifiers, Gradient Boosting, and deep learning-based classifiers.
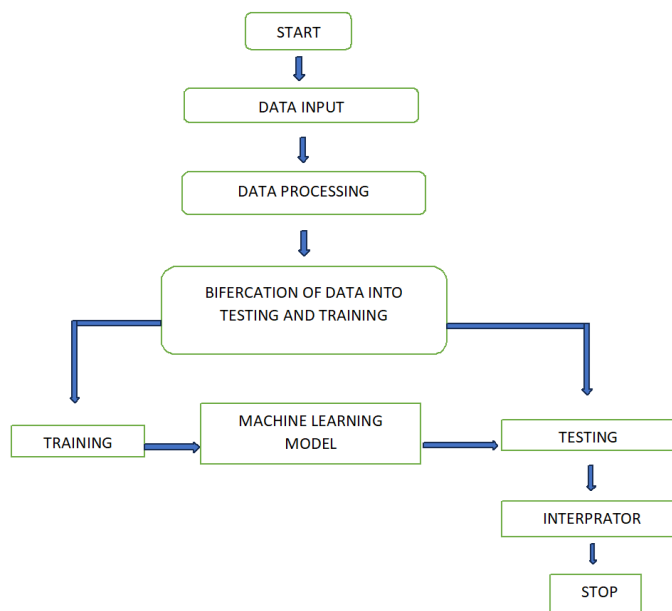
Fig 1. Flow chart of proposed work

2.1.1 Data Collection
In this stage, various websites are surveyed to collect the required data set for the research work. The data set obtained has a total of 20 different parameters with one TARGET column.

2.1.2 Preprocessing
The collected data undergoes several preprocessing steps to improve the accuracy of the models. These steps include feature selection, data cleaning, data balancing, removing NaN values, etc.

2.1.3 Model Selection
In this research work five distinct algorithms are used namely:

- **Logistic Regression:** A Method for Discrete Probability Modelling events with multiple input variables. One of the maximum famous binary outcomes. A logistic regression model is either real or binary false, yes or no, etc. Popular models for binary problems logistic regression classification. It helps to predict the probability control will fall into one of two groups. Real conversion characteristic returns with a probability between 0 and 1 is a central component regression. The S-curve-like shape of the function makes it ideal for modelling probability. The hypothesis function in this model determines the probability that an event example will be categorized into positive categories using a weighted combination of input Features. Applications for logistic regression can be found in marketing, fraud detection, emotional analysis, and diagnosis.
- **Random Forest:** Leo Breiman and Adele Cutler are the trademark holders of this widely used machine learning algorithm, which aggregates the output of several decision trees to produce a single result. Because it can handle both regression and classification problems, its flexibility and ease of use have increased its adoption. Ensemble learning approaches that are robust and flexible are machine learning techniques like Random Forest. It functions by constructing a collection of decision trees, each of which is constructed using a different subset of the training data and features. This randomization helps to reduce overfitting, which is a significant issue with individual decision trees. Bagging is the process of continuously sampling and replacing subsets of data in order to produce a diverse collection of trees. Moreover, Random Forest selects the features of each tree at random, reducing the likelihood that any one feature will influence the decision-making process. In order to generate predictions, Random Forest combines the output from these individual trees using majority voting for classification tasks and averaging for regression tasks.

- **Decision Tree:** In a decision tree, every internal node represents a "test" on an attribute (such as whether a coin flips heads or tails), every branch represents the test's result, and every leaf node represents a class label (the choice made after calculating all attributes). The form resembles a flowchart. Class guidelines are represented by the paths from the root to the leaf. Decision trees represent a class of machine learning techniques that are frequently applied to both regression and classification applications. It is a tree-like graphic representation of a decision-making process, where each internal node represents a choice or test on a specific feature, each branch represents the test's outcome, and every leaf node represents the very last selection or prediction. Decision trees are useful for many tasks, such as diagnosing medical conditions and forecasting customer attrition. Furthermore, they are easily interpreted.
- **Kneighbors Classifiers**: Simple but effective supervised machine learning K-neighbors (K-NN) algorithm for regression and classification problems. K-NN is a non-parametric data classification method depending on the number of adjacent neighbors in the feature space. In K-NN, "K" indicates the number of data points to consider; so far hyperparameters are to be modified depending on the specific problem at hand. Algorithms often use interval measures for Euclidean intervals to Determine the proximity of the data point. The nearest neighbor algorithm sometimes called KNN or k-NN, non-parametric control Individual data points were grouped and classified based on distance for classification or predictive information. Although it can be applied to classification problems or regression problems it is often used as a classification scheme based on the concept of similar nodes, they should be spaced close together.
- **Gradient Boosting:** Gradient boosting is a powerful machine-learning technique for applications involving both regression and classification. It works by gradually combining weak predictive models, typically decision trees, to create a strong ensemble model. The basic idea behind gradient boosting is to minimize the errors of the previous model in the ensemble by fitting each new model to the residual errors of the combined ensemble. To do this, the weights of each weak model are changed, and the best fit is then found through gradient descent optimization. Popular gradient boosting implementations include the algorithms AdaBoost, XGBoost, LightGBM, and CatBoost; each has unique benefits and variations.

2.1.4 Model Training and Evaluation:
The data is split into a training set and a testing set. The training set is used to train the models, and the testing set is used for their performance.

2.1.5 Model Comparison:
Comparison between the results of the different algorithms used in the testing and training process based on both accuracy & time performance.

2.2 Algorithm

A rough format of the model code used in the research work.
    (i) Firstly, import the required header files which are included in the code
      Import numpy as num
      Import pandas as pd
      From sklearn.model_selection import train_test_split
      From sklearn.linear_model import LogisticRegression
    From sklearn.metrics import accuracy_score
    (ii) Read the input dataset required in CSV format using the read function.
    (iii) Get the data info using the info function and label the X and Y parameters respectively.
    (iv) Split the dataset into the train and test as per the requirement (i.e., 80:20 or 70:30).
    (v) Give the test and train data as the input to different algorithms as mentioned earlier (Logistic Regression, Random Forest, Decision Tree, Kneighbors Classifier, Gradient Boosting).
    (vi) Obtain the result i.e., the accuracy level of the above-mentioned algorithm.
    (vii) Now by giving the input parameters to the model "Predictive System" the stress level of the student is determined.

2.3 Experimental Setup

The experiment is conducted in the Windows 64-bit operating system. Additionally, anaconda with Python 3.8.8 has been used. Also, to implement deep learning algorithms TensorFlow and Keras frameworks have been installed. The rest environmental information is shown in Table 1.

Table 1. The environment of tests

No	Used resource	Resource information
1	Operating System	Windows 11, 64-bit
2	Computer CPU	11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz
3	Type of hard disk drive	SSD
4	Tensor Flow Version:	2.14.0
5	Keras Version:	2.11.0
6	Pandas:	2.1.2
7	Python Version	3.12.0

2.4 Dataset

The dataset in the experiments is based on the different social media platforms. Many texts are not appropriate for the experiments, such as being either too long, having complex emotional tendencies, or having too many special characters. Hence, the original textual data are pre-processed. The dataset was taken from various social media sites KAGGLE which presents the parameters related to mental health. It consists of two attributes the first one is the parameters and the second one is the target. Lastly, there are about 6000 records that meet the criteria of the algorithms which consist of positive, negative, and neutral data. The output of the system is of three types:

1. If the output is 0, it indicates that the mental health of students is good.
2. If the output is 1, it indicates that the students are not in a good mental state and have stress of level 1.
3. If the output is 2, it indicates that the students are not in a good mental state and have a stress of level 2.

However, the selected 70% of the dataset is the training set, and 30% is the test set.

In mental health conditions, there are basically two stress levels:

1. In stress level one the mental condition of the student can be treated by some changes in their daily activities or maintaining a healthy balanced diet, reducing the screening time and similarly some basic steps or precautions can be taken to treat the mental health.
2. In stress level two the mental health of the student will be greatly affected and can have to undergo some counselling or therapies prescribed by well-qualified mental health doctors. Also, their respective parents have to take care that the student is not facing any kind of stress from the outside world.

Figure 2 shows a comparison between the actual value and the predicted value, for which the data is taken from the 'cm' variable which is mentioned in the Random Forest algorithm.

3 Results and Discussion

The research conducted on " Decoding Minds: Estimation of Stress Level in Students using Machine Learning" has yielded promising and transformative results. By harnessing the power of machine learning techniques, this study offers a novel approach to address the increasingly critical issue of mental health among students. The utilization of diverse data sources, including Anxiety level, self-esteem, mental_health_history, depression, headache, blood pressure, sleep quality, breathing_problem, noise_level, living conditions, Safety, basic needs, academic performance, study_load, teacher_student_relationship, future_career_concerns , social support, peer_pressure, extracurricular_activities and bullying which directly or indirectly has an effect on the mental health of the students, so basically here 19 different types of factors are taken into consideration which may be considered descent approach where artificial intelligence along with machine learning are together combine to get a highly accurate value, the maximum and minimum values of the respective parameters used is shown in Table 2.

Recent research on predicting student mental health conditions using machine learning algorithms has yielded a medium level of accuracy, and they lack in using more number of parameters or features, and also they have used very less number of samples. In the research work of [4] only 270 participants or samples were used, whereas compared to this research paper there are more than 6000 samples. Some other paper shows the usage of only one or two algorithms. In [2] the research work shows the usage of only one algorithm that is Gradient Boosting algorithm, whereas in this research paper five different algorithms are used
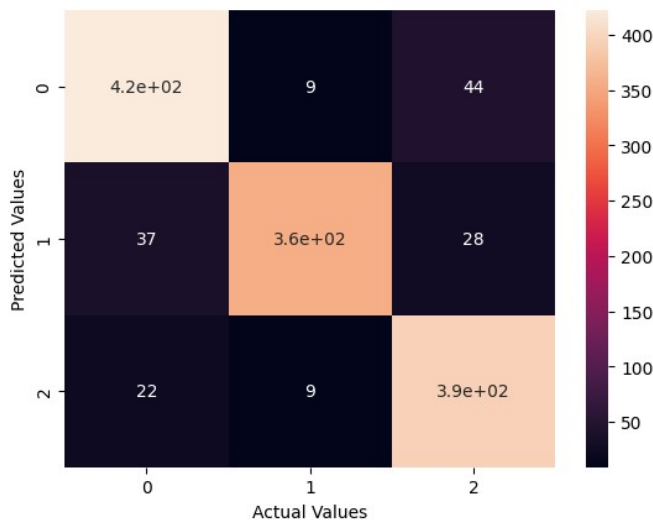
Fig 2. Correlation graph between the actual value and the predicated values

Table 2. Parameters used with their corresponding maximum and minimum values

Features	Max and Min Values
anxiety_level	Min:0 and Max:21
self_esteem	Min:0 and Max:30
mental_health_history	Min:0 and Max:1
depression	Min:0 and Max:27
headache	Min:0 and Max:5
blood_pressure	Min:1 and Max:3
sleep_quality	Min:0 and Max:5
breathing_problem	Min:0 and Max:5
noise_level	Min:0 and Max:5
living_conditions	Min:0 and Max:5
Safety	Min:0 and Max:5
basic_needs	Min:0 and Max:5
academic_performance	Min:0 and Max:5
study_load	Min:0 and Max:5
teacher_student_relationship	Min:0 and Max:5
future_career_concerns	Min:0 and Max:5
social_support	Min:0 and Max:3
peer_pressure	Min:0 and Max:5
extracurricular_activities	Min:0 and Max:5
bullying	Min:0 and Max:5

which are showing a very high accuracy level. Through the utilization of advanced models and neural networks, scientists have examined a variety of datasets that include elements like social interactions, behavioural patterns, and academic performance. These algorithms provide more ease in understanding the dynamics of mental health among students by spotting minute patterns and correlations that may elude traditional methods. Deep learning models' predictive powers allow for early support and intervention, which may enhance general well-being and academic achievement. However, it's important to recognize the privacy issues and ethical issues that come with handling sensitive data.

Figure 3 shows a heatmap which is a graphical data representation in Python that uses colours to represent the values in a matrix. When analysing the size of relationships or patterns in a two-dimensional dataset, this is especially helpful. The seaborn library is a popular Python heatmap tool, built on top of Matplotlib. The data matrix is shown as a grid, where the colour intensity of each cell represents its numerical value. Cooler shades like blue indicate lower values, while warmer shades like red or yellow frequently represent higher values. Heatmaps are adaptable and have uses in many domains, including biology, finance, and machine learning in addition to data analysis. In this particular case, the heat map shows the relation between the parameter and its effect on the mental health of the student using some numerical value along with different shades as mentioned in Figure 3. The result obtained in this research paper shows the high accuracy value of the various algorithms used and this is clearly depicted in Figure 4, which shows the comparison between the accuracy level in terms of percentage of different algorithms used in the paper.
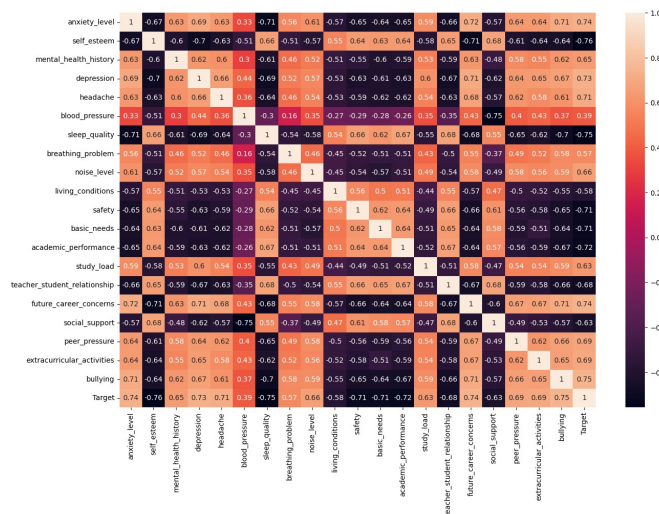

Fig 3. Correlation between mental health and parameters likely to have effect on mental health
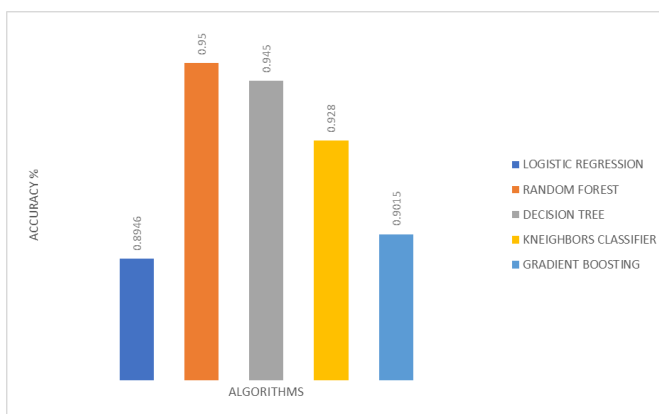

Fig 4. Graphical representation of accuracy of algorithms

4 Conclusion

This study has successfully developed a model to predict students' stress level that shows promising accuracy based on multiple input features which are 19 in number by employing sophisticated algorithms namely Logistic Regression, Random Forest, Decision Tree, Kneighbors Classifiers, Gradient Boosting, and thorough data analysis. The results highlight the potential of machine learning as a useful instrument for recognizing and resolving stress-related issues in learning environments. Accurately estimating stress levels makes early intervention strategies possible, enabling teachers and support personnel to actively help students manage their well-being. Moreover, the knowledge acquired from this study can be used to build customized interventions that are suited to the requirements of specific students. Educational institutions can put targeted strategies into place to create a more supportive and conducive learning environment by identifying patterns and trends in stress-related data. It is imperative to recognize the limitations of our research, including the necessity for additional validation across a range of educational contexts and populations. Subsequent investigations ought to concentrate on enhancing the model, integrating real-time data, and investigating supplementary elements that contribute to students' stress. To sum up, using machine learning to estimate stress levels is a promising way to encourage mental health awareness and intervention in educational settings. In addition to adding to the expanding corpus of research on the subject of technology and mental health, this study opens the door for real-world application in educational settings where students' welfare is given top priority. Researchers, teachers, and tech developers will need to work together more closely as we advance to improve and apply these creative fixes, which will ultimately help to create a safer and more encouraging learning environment for students.

References

1) Malik SS, Khan A. Anxiety, Depression and Stress prediction among College Students using Machine Learning Algorithms. In: 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE. 2023. Available from: https://doi.org/10.1109/ICEEICT56924.2023.10157693.
2) Kene A, Thakare S. Mental Stress Level Prediction and Classification based on Machine Learning. In: 2021 Smart Technologies, Communication and Robotics (STCR). IEEE. 2021. Available from: https://doi.org/10.1109/STCR51658.2021.9588803.
3) Verma H, Kumar N, Sharma YK, and PV. StressDetect: ML for Mental Stress Prediction. In: Optimized Predictive Models in Healthcare Using Machine Learning. Wiley. 2024. Available from: http://dx.doi.org/10.1002/9781394175376.ch20.
4) Kene A, Thakare S. Prediction of Mental Stress Level Based on Machine Learning. In: Machine Intelligence and Smart Systems. Algorithms for Intelligent Systems;Singapore. Springer. 2022;p. 525–536. Available from: https://doi.org/10.1007/978-981-16-9650-3_41.
5) Baba A, Bunji K. Prediction of Mental Health Problem Using Annual Student Health Survey: Machine Learning Approach. *JMIR Mental Health*. 2023;10. Available from: https://dx.doi.org/10.2196/42420.
6) Verma P, Singh R. Mental Stress Prediction Using Wrist Wearable Through Machine Learning Approaches. In: 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET). IEEE. 2023. Available from: http://dx.doi.org/10.1109/icseiet58677.2023.10303348.
7) Zhang Z. Early warning model of adolescent mental health based on big data and machine learning. *Soft Computing*. 2024;28(1):811–828. Available from: https://doi.org/10.1007/s00500-023-09422-z.
8) Bloomfield LSP, Fudolig MI, Kim J, Llorin J, Lovato JL, McGinnis EW, et al. Predicting stress in first-year college students using sleep data from wearable devices. *PLOS Digital Health*. 2024;3(4):1–16. Available from: https://dx.doi.org/10.1371/journal.pdig.0000473.
9) Lamb R, Almusharraf N, Choi I, Firestone J, Kavner A, Owens T, et al. Machine learning prediction of mental health strategy selection in school aged children using neurocognitive data. *Computers in Human Behavior*. 2024;156. Available from: https://doi.org/10.1016/j.chb.2024.108197.
10) Radhika C, Shraddha N, Vaishnavi P, Shirisha K. Prediction of Mental Health Instability using Machine Learning and Deep Learning Algorithms. *Journal of Computer Science and Applications*. 2023;15(1):47–58. Available from: https://dx.doi.org/10.37624/jcsa/15.1.2023.47-58.