EmoSupport: A Comparative Study for the Analysis of Mental Health of Undergraduate Students

1st Manisha Aeri
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
maeri@gehu.ac.in

2nd Anushka Naithani
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
anushka18jan2002@gmail.com

3rd Pratham Kashyap
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
prathamkashyap1980@gmail.com

4th Anmol Jaiswal
Computer Science and Engineering
Graphic Era Hill University
Dehradun, India
anmoljaiswal012001@gmail.com

***Abstract*—Mental health disorders, including anxiety, depression, and stress, profoundly impact individuals' well being and necessitate effective early detection for timely intervention. This research investigates the predictive capabilities of machine learning algorithms in assessing anxiety, depression, and stress levels based on questionnaire derived scores. Utilizing a dataset comprising self reported scores obtained through a tailored questionnaire designed for mental health assessment, we delve into the application of Decision Trees, Naive Bayes, Support Vector Machines (SVM), and Random Forests for prediction. Data pre processing involved comprehensive cleaning, encoding categorical variables, and careful feature selection, ensuring the relevance of features in the predictive models. Each algorithm underwent individual implementation, wherein we scrutinized their performances in predicting mental health conditions. Evaluation metrics such as accuracy, precision, and recall were employed to assess the models' proficiency in predicting anxiety, depression, and stress levels. The findings underscore the potential of machine learning in accurately predicting mental health conditions based on questionnaire responses, offering insights into personalized interventions and early detection systems. This study contributes to advancing the understanding of machine learning applications in mental health assessment, highlighting avenues for impactful interventions in mental health care.**

***Index Terms*—Mental Health Assessment, Decision Trees, Naive Bayes, Support Vector Machines (SVM), Random Forests, KNN, Logistic Regression.**

I. Introduction

Mental health disorders, including anxiety, depression, and stress related conditions, transcend geographical and cultural boundaries, impacting individuals of diverse demographics globally. The prevalence of these disorders highlights the critical need for proactive measures, emphasizing early detection and interventions crucial in preventing exacerbation, fostering improved mental well being, and curbing the societal burden associated with untreated mental health challenges [9]. These conditions not only affect individuals but also extend their ramifications to societal structures, emphasizing the necessity for collaborative efforts across sectors to establish robust support systems, promote awareness, and dismantle stigmas surrounding mental health issues, emphasizing the urgency to address these challenges comprehensively. The profound and far reaching impact of untreated mental health conditions not only impede individual lives but also reverberate through educational, workforce, healthcare, and social realms, emphasizing the collective responsibility to foster resilience and create healthier societies through holistic mental health initiatives [21] [23].

A person's overall wellness greatly depends on their mental health. However, pinpointing those who need medical assistance for mental health issues can be challenging, resulting in delayed or inadequate treatment [16].The conventional approaches to mental health assessments traditionally rely on subjective self reports and clinical evaluations, providing foundational insights into individuals' mental well being. However, while invaluable, these methodologies encounter limitations in their scalability and expediency. They often require extensive time, resources, and professional expertise, hindering their widespread application and timely interventions. Moreover, these methods might not capture the comprehensive spectrum of mental health conditions, as they often rely on observable symptoms and self reported experiences, potentially overlooking subtle nuances and underlying complexities [19]. Yet, the landscape of mental health assessment is undergoing a transformative shift propelled by technological advancements, particularly in the realm of machine learning and predictive analytics. These advancements represent an unprecedented opportunity to overhaul traditional assessment paradigms [7]. By leveraging the power of these innovations, mental health assessments can harness data driven insights culled from a myriad of sources, encompassing not only questionnaire based scores but also behavioral patterns and demographic information. [22] The integration of machine learning algorithms into these assessments holds the promise of revolutionizing the field. These algorithms are adept at discerning intricate patterns within vast datasets, potentially augmenting existing assessment methodologies. This integration could usher in a new era of precision, efficiency, and scalability in evaluating mental health conditions, enabling more nuanced and timely interventions tailored to individuals' unique needs [10]. The utilization of machine learning algorithms marks a promising step towards a future where mental health assessments are not only more precise but also more accessible, facilitating early interventions and support mechanisms for individuals experiencing mental health challenges. This research initiative embodies a dedicated exploration into the domain of machine learning driven mental health assessments, strategically focusing on the precise prediction of anxiety, depression, and stress levels rooted in comprehensive questionnaire scores [25]. The study utilizes a substantial dataset comprising behavioral and demographic information from both autistic and non autistic individuals to train and assess the performance of machine learning algorithms [15].The primary objective of this study is to delve deeply into a diverse spectrum of machine learning algorithms, encompassing Decision Trees, Naive Bayes, Support Vector Machines (SVM), and Random Forests, among others. The aim is to comprehensively scrutinize their individual capabilities in accurately predicting nuanced mental health conditions, embracing the complexity inherent in these disorders. By examining the predictive potential of a varied array of machine learning algorithms, this study seeks not only to evaluate their performance but also to unravel the intricate relationships between diverse datasets and mental health outcomes [21]. The overarching goal extends beyond mere prediction; it endeavors to contribute significantly to the ongoing discourse within mental health research. Through the elucidation of machine learning driven predictive models' potential [26], this study aspires to provide novel insights and methodological advancements that could revolutionize mental health assessment methodologies.

The envisioned outcomes of this research carry the promise of substantial implications for mental health care practices. The potential development of predictive models could pave the way for personalized interventions tailored to individual needs, offering more targeted and effective treatments [28]. Moreover, the envisaged outcomes also hold the potential to foster the creation of accessible tools that can be seamlessly integrated into the arsenal of mental health professionals [18]. Such tools, harnessing the power of machine learning, have the capacity to transform mental health care delivery by facilitating early detection, informed decision making, and responsive interventions, ultimately enhancing the quality of care and support for individuals navigating mental health challenges.

II. Literature Survey

Machine Learning Approach to Prediction and Assessment of Depression and Anxiety: [1] A Literature Review: This paper reviews studies conducted between 2010 and 2022 on the use of Machine Learning (ML) techniques for screening and monitoring anxiety and depression, prevalent mental health conditions. The research emphasizes the efficacy of ML algorithms, including Convolutional Neural Network (CNN) and

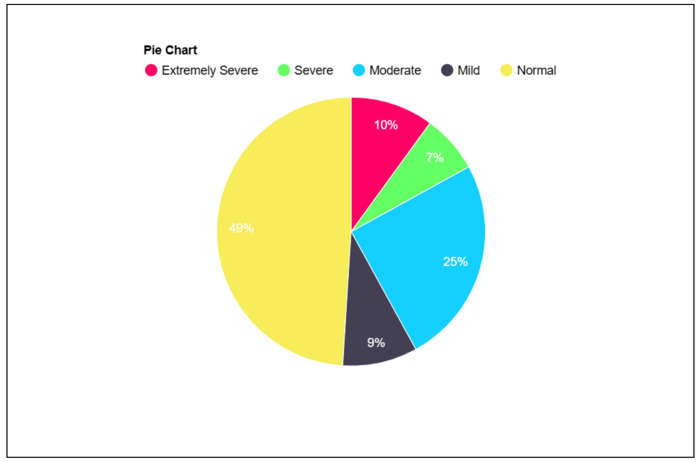

Fig. 1. Anxiety level distribution among Indian citizens

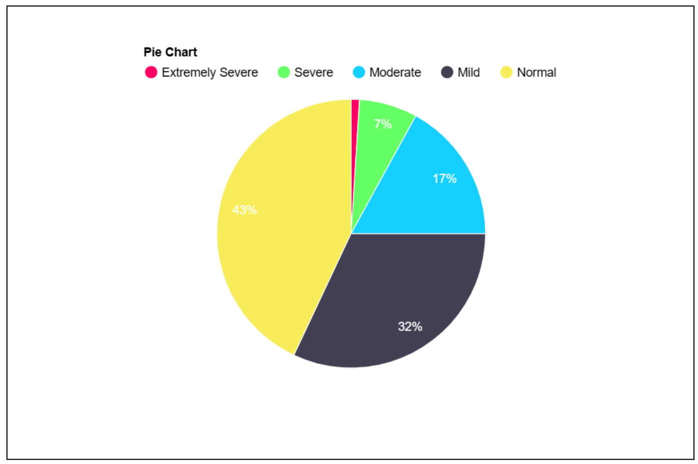

Fig. 2. Stress level distribution among Indian citizens

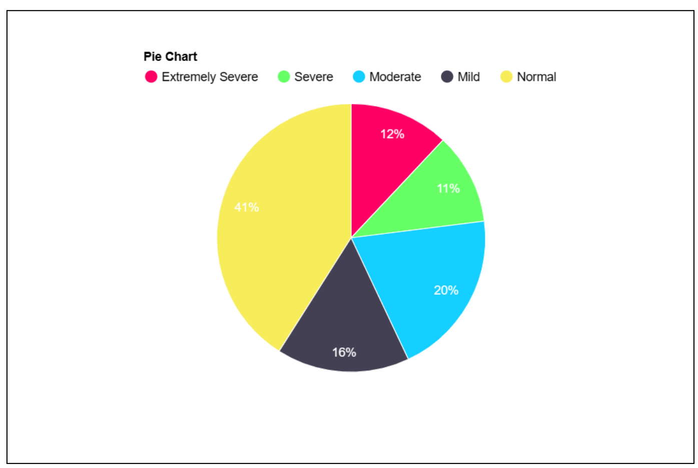

Fig. 3. Depression level distribution among Indian citizens

Support Vector Machine (SVM), in accurately anticipating the severity of anxiety and depression based on patient behaviors. The identified keywords, such as Machine Learning, anxiety, and depression, were utilized for data collection, while analysis employed various tools like K Nearest Neighbor, Random Forest, Decision Tree, and Naive Bayes. Notably, SVM and CNN algorithms demonstrated superior accuracy compared to other methods. The findings underscore the significant progress in leveraging ML for mental health assessments and highlight the potential for enhancing psychologists' ability to tailor effective treatment strategies based on individual disease levels.

A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis: [2] This comprehensive review delves into the application of machine learning (ML) and deep learning (DL) techniques for diagnosing seven prevalent mental health conditions. Spanning the years 2010 to 2022 and analyzing 33 articles using the PRISMA review methodology, the study covers methodologies such as supervised learning, ensemble learning, and transfer learning. It scrutinizes various ML models like SVM, decision trees, random forest, and XGBoost, alongside DL architectures including CNN, DBN, AE, and RNN. The paper underscores challenges like data limitations, language constraints, and ethical concerns in data anonymization. While recognizing the efficacy of ML and DL in diagnosing mental health disorders, the review advocates for further exploration of diverse data modalities. It acknowledges the potential of DL methods to predict multiple disorders simultaneously and concludes by urging researchers to address challenges for improved model performance, emphasizing the importance of high quality data and explainable DL models in enhancing diagnostic accuracy for real world application in mental health care.

Applications of artificial intelligence machine learning for detection of stress: [3] This review highlights the crucial role of artificial intelligence (AI) and machine learning (ML) in addressing biomedical challenges related to psychological stress. It emphasizes the impact of stress on human health and its association with conditions such as autoimmune diseases, metabolic syndrome, sleep disorders, and suicidal thoughts. The paper stresses the need for early detection and management of chronic stress for disease prevention. Drawing from previous studies, the review underscores the success of AI and ML in predicting stress levels, particularly in post traumatic stress disorder, achieving accuracy rates of approximately 90 percent. The authors advocate for a shift towards detecting prolonged distress using AI and ML driven technology, introducing Swarm Intelligence (SI) as a promising AI subcategory for stress detection in clinical settings while preserving privacy. The review concludes optimistically, suggesting the potential benefits of integrating AI and ML into standard clinical practices for stress diagnostics, calling for further research in this direction.

Depression Detection From Social Networks Data Based on Machine Learning and Deep Learning Techniques: [4] Published in the IEEE Transactions on Computational Social Systems, this study addresses the global health concern of underdiagnosed and undertreated depression. Utilizing user generated data from social networks, the paper explores the application of machine learning (ML) and deep learning (DL) techniques to enhance automated depression detection. Employing a systematic literature review (SLR) methodology, the study evaluates state of the art methods, considering diverse data forms such as images, text, and videos. Recognizing challenges associated with incomplete and inaccurate data, the review emphasizes the effectiveness of ML and DL in providing valuable insights for depression diagnosis. By highlighting critical challenges and presenting cutting edge techniques, the survey aims to assist readers and researchers in the ML and DL fields, contributing to a broader understanding of mental health issues and showcasing the potential of advanced technologies in improving diagnostic capabilities.

On The Robustness of Machine Learning Models for Stress and Anxiety Recognition From Heart Activity Signals: [5] This article delves into the generalizability of machine learning models in detecting negative affective states, particularly stress and anxiety, using physiological signals from body worn sensors. Focused on cardiac signals, specifically heart rate variability features from blood volume pulse (BVP) and electrocardiogram (ECG) data, the study employs open datasets from two large experimental studies. Emphasizing the importance of high quality training data and addressing artifacts, the research highlights the potential impact of training models on noisy proxies within lower quality data. The article underscores the necessity of incorporating a diverse range of emotional states in training data to enhance accuracy and reduce erroneous classifications. Evaluating the generalizability of BVP and ECG sourced data across different laboratory conditions, the findings indicate that signal quality significantly influences generalizability, with ECG data outperforming BVP data. The study suggests prioritizing data quality before capture to maximize model.

Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms: [6] The survey investigates anxiety, depression, and stress prevalence, using ma chine learning algorithms to predict them. Employing the DASS 21 questionnaire across diverse backgrounds, five algorithms—Decision Tree, Random Forest, Na¨ıve Bayes, SVM, K Nearest Neighbor—predicted severity. Challenges include class imbalances, using F1 scores for evaluation. Naıve Bayes shows highest accuracy; Random Forest excels in F1 score. Variable analysis highlights disorder contributors, emphasizing addressing imbalances and using apt metrics for mental health prediction.

Assessment of Anxiety, Depression and Stress using Ma chine Learning Models: [7] The article by Prince Kumara, Shruti Garga, and Ashwani Garg evaluates anxiety, depression, and stress using machine learning. Eight algorithms across different groups are used, including a hybrid model, on DASS42 data from online questionnaires (2017 2019). The hybrid algorithm excels, particularly the radial basis function network in accuracy. The study underscores the importance of using machine learning for assessing mental health issues, especially considering people's hesitancy to openly share feelings. It offers insights into multiclass classification for anxiety, depression, and stress severity. Machine learning models to detect anxiety and depression through social media: [8] A scoping review: This scoping review explores machine learning's use in detecting anxiety and depression from social media data. It analyzes 54 articles (2013 2021) focusing on ML models, data sources, and performance metrics. Most studies target depression on platforms like Twitter, Facebook, and others, predominantly in English but also in languages like Chinese and Bangla. Models include AdaBoost, CNN, GRU, KNN, LR, LSTM, MLP, NB, Random Forest, DT, SVM, and XGBoost. Performance metrics commonly involve F1 score, accuracy, and precision. The review emphasizes ML's potential to complement traditional mental health screening, stressing continuous analysis of social media data. Ethical considerations, reproducibility, and the gap between research and clinical care are highlighted for impactful mental health diagnostics. A survey of machine learning techniques in physiology based mental stress detection systems: [9]This paper conducts an extensive survey on automated/semi automated medical diagnosis systems, specifically focusing on detecting mental stress. Highlighting the global prevalence of stress, the study emphasizes the importance of early detection and management for individuals' well being. It explores physiological features known for their reliability in stress detection systems and covers aspects such as data collection, machine learning's role in emotion and stress detection, evaluation measures, challenges, and applications. Visual representations and dedi cated sections to emotions, physiology, and machine learning algorithms organize the research. Stress and emotions, sharing a physiological basis, are central to the study due to their transient nature. The paper identifies research gaps, offering insights into the relationships between physiological features, emotions, and stress, aiding in the development of effective stress detection systems. Computer assisted identification of stress, anxiety, depression (SAD) in students: [10] This paper delves into stress, depression, and anxiety (SAD) as physiological states ex pressed through speech, body language, and facial expressions. It focuses on these conditions within student life, highlighting the importance of early detection for overall well being. The study systematically reviews computerized techniques, especially machine learning algorithms, for identifying SAD using datasets including questionnaires, audio, and video in puts. It emphasizes AI and machine learning's effectiveness in detecting SAD parameters through various models and feature extraction methods. The interconnected nature of these psychological states is explored, emphasizing computer vision techniques like facial expressions for accurate recognition. The paper addresses challenges such as dataset availability and reviews existing models, offering insights into the potential of machine learning for detecting psychological disorders and suggesting future directions for research.

III. Proposed Methodology

The proposed methodology encompasses a meticulous approach aimed at comprehensive data acquisition, involving the creation of a detailed questionnaire comprising 30 questions exploring stress, anxiety, and depression facets across diverse demographics. Targeting students and individuals experiencing mental health challenges, both online platforms and in person interviews are utilized for data collection, ensuring a broad and diverse sample. Preprocessing the gathered data involves rigorous handling of missing values, outlier treatment, and standardization for suitability in machine learning algorithms. Subsequent feature engineering endeavors to extract pertinent features and transform qualitative responses into structured data. The study employs Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest algorithms for predictive modeling, training them iteratively on the prepared dataset to optimize performance. Rigorous evaluation using various metrics facilitates algorithm comparison, aiding in identifying the most suitable model. Interpretation of outcomes yields insights into significant predictors of mental health states, offering implications for intervention strategies.

A. *Dataset Description*

The dataset utilized in this study encompasses responses garnered from a focused group of 1200 undergraduate students, drawn from various colleges and universities, specifically targeting individuals within the age bracket of 18 to 24 years. These participants represent a spectrum of diverse backgrounds, reflecting the multifaceted nature of stress, anxiety, and depression prevalent among individuals at the cusp of adulthood and educational pursuits. The meticulously curated questionnaire, comprising 30 in depth inquiries, traverses the complexities of mental health, exploring personal experiences, coping mechanisms, lifestyle routines, and societal influences pertinent to this demographic. Leveraging a blend of online platforms and in person interactions, this dataset aims for a holistic representation, amalgamating qualitative nuances into structured, analyzable data. Rigorous preprocessing techniques have been applied, ensuring data quality by addressing missing values, handling outliers, and standardizing responses. This rich and comprehensive dataset serves as a robust foundation, poised to drive nuanced analyses and predictive modeling to unravel the intricacies of mental health challenges faced by this cohort.

The following describes the key points for dataset description:

- Data Collection: The initial phase involves comprehensive data acquisition employing a meticulously designed questionnaire comprising 30 questions exploring various dimensions of stress, anxiety, and depression. Participant recruitment strategies aim to target diverse demographics, potentially focusing on students or individuals facing mental health challenges. Utilizing both online platforms and in person interviews facilitates a broad reach and diverse sample representation.

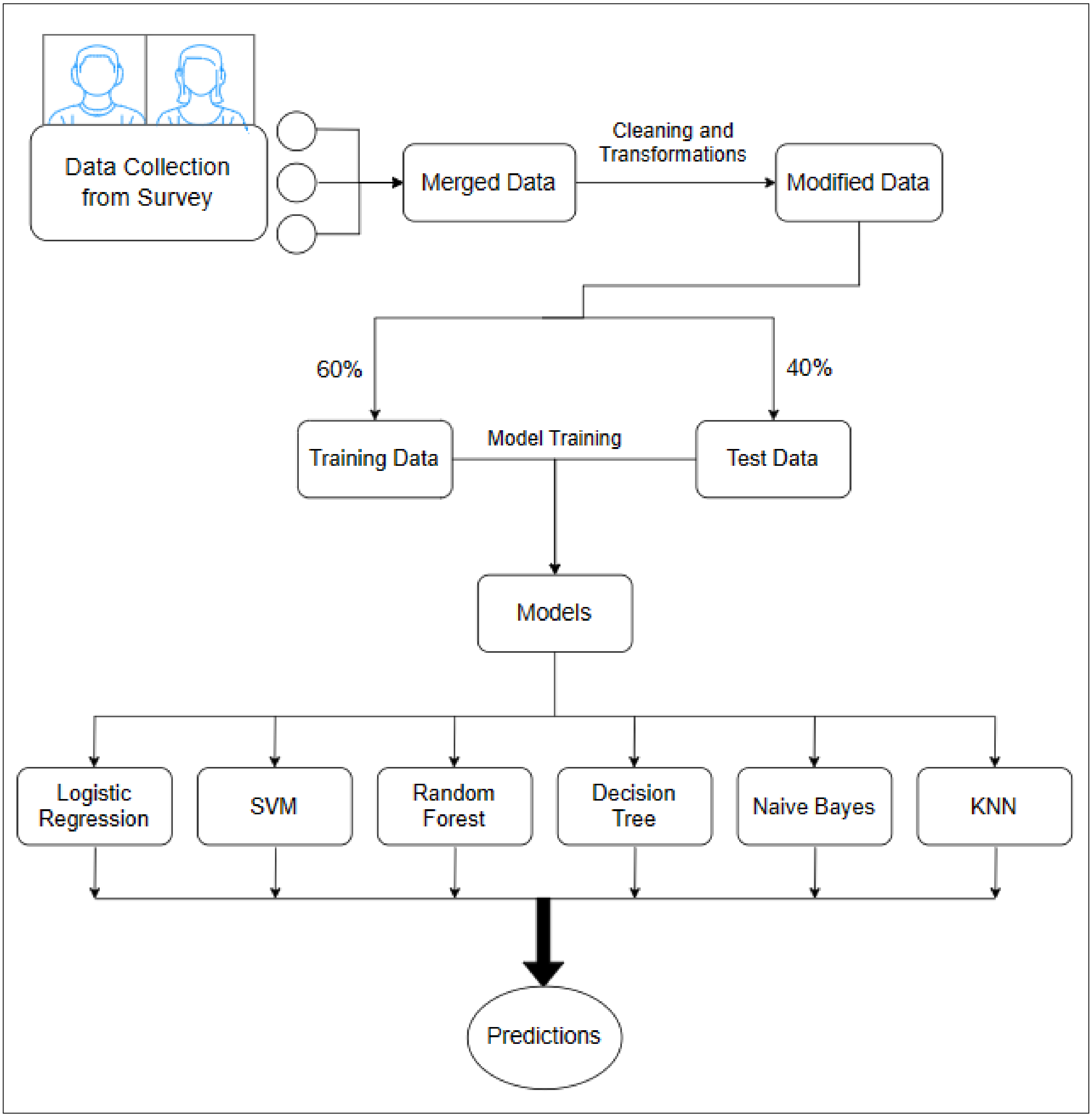

Fig. 4. Architecture of the Project

- Dataset Preparation: The gathered data undergoes meticulous preprocessing, encompassing tasks such as handling missing values, outlier treatment, and standardization. Categorization and encoding techniques are employed to translate qualitative responses into a structured format suitable for machine learning algorithms.
- Demographic and Psychographic Attributes: The dataset encapsulates a broad spectrum of demographic information, including age, gender, occupation, educational background, and geographical location. Additionally, psychographic elements such as lifestyle choices, coping mechanisms, and social support systems were included to enrich the dataset's context.
- Data Features: The dataset comprises multifaceted attributes reflecting mental health states, stress triggers, coping strategies, emotional well being, daily stressors, environmental influences, and behavioral patterns. Each question from the questionnaire represents a specific feature or attribute within the dataset.
- Structure and Format: Organized in a tabular format, the dataset consists of rows representing individual respondents and columns representing distinct features or questionnaire questions. The data types include categorical (e.g., gender, occupation), numerical (Likert scale ratings), and potentially textual (open ended responses) data.
- Dataset Size and Distribution: The dataset encompasses [Specify the total number of respondents or instances], providing a substantial sample for analysis. The distribution illustrates the prevalence of stress levels or mental health conditions, delineating the percentages or counts of individuals categorized across various stress severity levels.
- Data Quality and Preprocessing: Efforts were undertaken to manage missing values and ensure data integrity. Imputation techniques were applied where necessary, and preprocessing steps included the removal of duplicates, standardization of formats, and handling outliers or inconsistencies in responses.
- Ethics and Privacy Measures: Stringent measures were implemented to uphold respondent anonymity and confidentiality. Sensitive information was handled with utmost care and stored securely to maintain ethical standards.

B. Architecture of the project

For building a model of Mental Health based on the dataset collected from 1200 students, we have followed these steps: 1. Data Collection and Preprocessing Layer:

- Survey Design: Developing a comprehensive questionnaire is pivotal to capture a holistic view of stress related factors. Considering the 30 provided questions, the questionnaire should encompass diverse dimensions of mental health. This includes exploring emotional states, triggers, coping mechanisms, lifestyle influences, demographic specifics, and potentially relevant behavioral patterns.
  - Questionnaire Structure: Structure the questionnaire methodically, ensuring a blend of open ended and close ended questions. Open ended questions allow for detailed qualitative insights, while closed ended questions provide quantifiable data for analysis.
  - Covering Varied Aspects: The questionnaire should delve into various aspects such as emotional responses, environmental influences, coping strategies, lifestyle choices, and social support systems. It could cover topics like daily stressors, triggers, impact on daily life, mental health history, and available support mechanisms.
  - Demographic Considerations: Ensure the questionnaire incorporates demographic details like age, gender, occupation, educational background, geographical location, and any specific factors relevant to stress or mental health disparities.
  - Validation and Pilot Testing: Before the formal data collection, validate the questionnaire by pilot testing it with a small group. This helps refine questions, ensuring clarity and relevance to diverse individuals.
- • Participant Recruitment: Recruiting participants is crucial to obtaining a diverse and representative sample. The targeted sample group could include students, working professionals, or individuals from varying demographics experiencing stress or mental health challenges.
  - Targeted Outreach: Utilize multiple channels for recruitment, including university campuses, workplaces, mental health support groups, online forums, and social media platforms. These avenues offer access to a wide spectrum of potential participants.
  - Informed Consent: Prioritize informed consent, ensuring participants are aware of the study's purpose, confidentiality measures, and their rights as respondents. Clearly outline how their data will be used, stored, and anonymized.
  - Diversity and Inclusion: Aim for diversity in the sample, encompassing individuals from different age groups, socioeconomic backgrounds, cultural identities, and geographical locations. This diversity enriches the dataset, offering a comprehensive understanding of stress across varied demographics.
  - Data Collection Medium: Employ a mix of data collection methods, such as online surveys,face to face interviews, or phone interviews, to accommodate participants' preferences and accessibility.

2. Model Development:

- Algorithm Selection Rationale: We meticulously choose algorithms aligning with our project's objectives and dataset attributes. For instance, Logistic Regression is suitable for binary stress classification, capturing presence or absence of stress based on various factors. SVMs cater to intricate non linear stress patterns, exploring relationships that aren't linearly correlated. Decision Trees or Random Forests, adept at handling complex feature interactions, help unveil intricate dependencies among multiple variables contributing to stress.

- Structured Dataset Utilization: The selected algorithms are employed to train models using our meticulously structured dataset. This dataset incorporates responses from the comprehensive survey encompassing diverse facets of stress, anxiety, and depression.
- Train Test Split Technique: To ensure model generalization and mitigate overfitting, we utilize the train test split methodology. This technique partitions the dataset into two subsets: a larger training set and a smaller testing set. The model learns from the training data and then validates its performance on the unseen testing data. This approach helps assess how well the model generalizes to new, unseen data by evaluating its predictive accuracy on the testing set.
- Overfitting Mitigation: Both techniques, train test splits and k fold cross validation, play a crucial role in preventing overfitting. They enable us to assess the model's ability to generalize to new data while minimizing the risk of learning from noise or idiosyncrasies present in the training dataset.
- Decision Tree Based Algorithms: Decision trees, such as Random Forest or Gradient Boosting, evaluate feature importance by assessing their impact on reducing impurity within decision nodes. For instance, 'Social Support' or 'Workload' might emerge as critical features that effectively differentiate stress levels. The depth and splits in these trees demonstrate the hierarchy of feature importance, elucidating the pivotal role of specific factors in predicting stress. Example: In a Random Forest model, 'Workload' might be the first split, indicating that high workload directly correlates with increased stress. Subsequent splits further delineate how other factors interact, revealing their relative importance in predicting stress outcomes.
- Robustness and Validation: By employing these techniques, we aim to ensure the models are robust, reliable, and capable of making accurate predictions regarding stress levels. This validation process enhances the trustworthiness of our models' performance evaluations and their applicability to real world scenarios.

C. Algorithm

1. Logistic Regeression: Logistic Regression is a statistical method primarily used for binary classification. It's suitable for predicting categorical outcomes based on features. In the context of your project, it's applied to forecast stress, anxiety, and depression levels based on survey responses, offering the probability of an individual experiencing these conditions.

- Functionality:
  - Stress, Anxiety, and Depression Prediction: Logistic Regression is well suited for binary classification

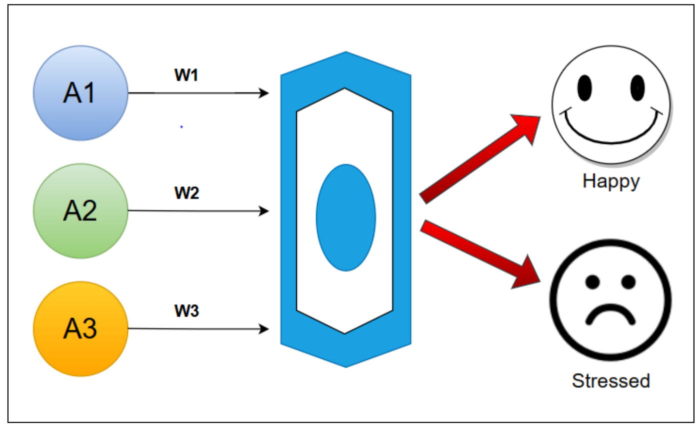

Fig. 5. Logistic Regression Representation

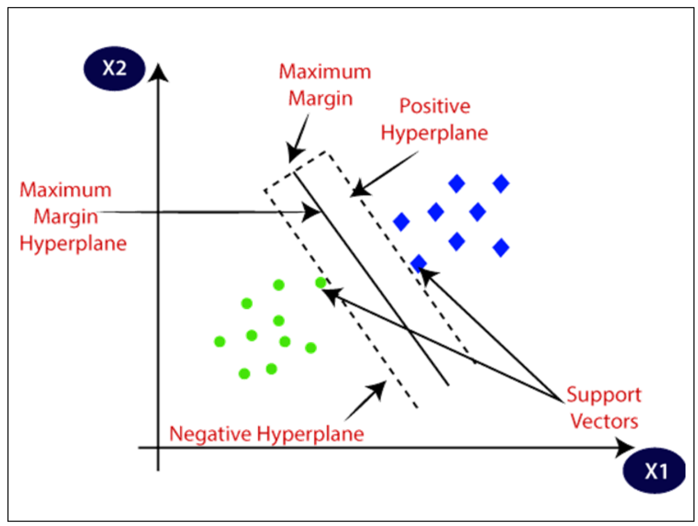

Fig. 6. SVM Representation

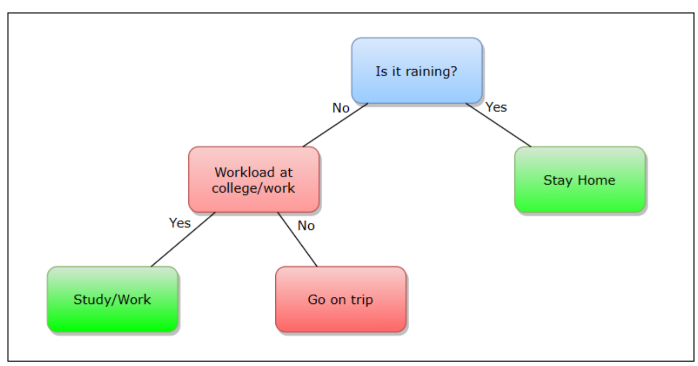

Fig. 7. Decision Tree Representation

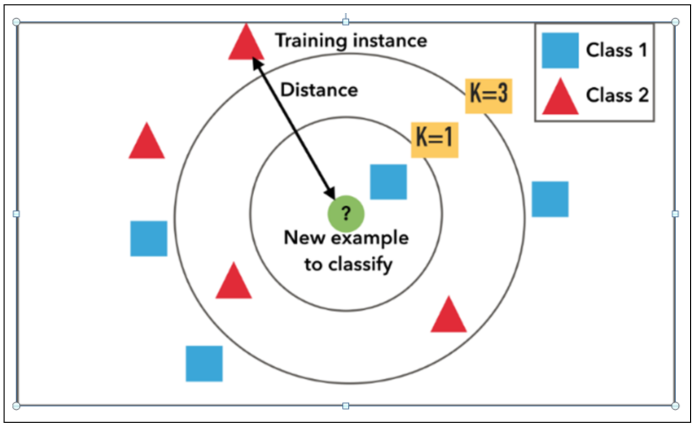

Fig. 8. KNN Representation

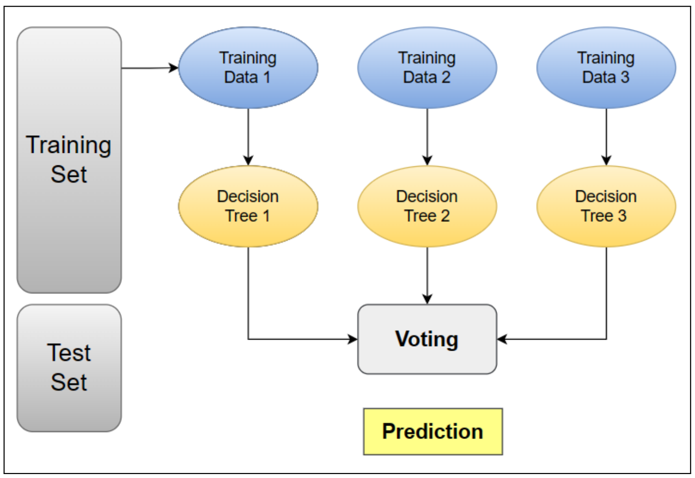

Fig. 9. Random Forest Representation

tasks like predicting stress levels. By analyzing survey responses related to stressors, it determines the probability of an individual being stressed or not stressed, anxious or not anxious, depressed or not depressed.
 – Interpretability: It provides insights into how each survey question or feature influences the likelihood of stress, anxiety, or depression. For instance, it might reveal that higher scores on questions related to social isolation correspond to increased odds of experiencing anxiety.

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

P(H|E) = This is posterior.
P(H) = This is the prior i.e. what you believed before you saw the evidence.
P(E|H) = this the likelihood of seeing that evidence if your hypothesis is correct.
P(E) = this is the normalizing of that evidence under any circumstances.

Fig. 10. Naive Bayes Representation

- Usage Example: Analyzing survey factors like workload, social support, and lifestyle habits, Logistic Regression would estimate the likelihood of an individual being stressed. For instance, higher scores in questions about heavy workload and lack of support might correlate with a higher probability of being stressed.
- Evaluation: Assessing model performance involves metrics like accuracy, precision, recall, and F1 score for each class (stressed or not stressed, anxious or not anxious, depressed or not depressed). This evaluates the model's effectiveness in correctly identifying stress, anxiety, and depression cases.

2. Support Vector Machines (SVM): SVM is a supervised learning algorithm suitable for both classification and regression tasks. It's adept at handling non linear relationships and is applied in your project to delineate boundaries between stress, anxiety, and depression levels based on survey features, ensuring a clear demarcation between different psychological states.

- Functionality:
  - Boundary Optimization: For stress, anxiety, and depression prediction, SVM aims to find an optimal boundary between classes based on survey responses. It is adept at capturing nonlinear relationships between features and can identify distinct patterns related to stress, anxiety, or depression.
  - Multiclass Classification: SVM can be extended to handle multiclass classification tasks, making it suitable for identifying varying levels of stress, anxiety, and depression.
- Usage Example: Utilizing survey questions about worry patterns, physical symptoms, and behavioral changes, SVM would delineate a decision boundary to classify individuals as anxious or not based on these factors.
- Evaluation: Similar to Logistic Regression, assessing SVM involves standard classification metrics to gauge its performance in correctly classifying stress, anxiety, and depression cases.

3. Decision Tree: Decision Tree construct tree like structures to derive rules for classification tasks. They're employed in your project to elucidate relationships between survey questions and stress related outcomes, creating hierarchical decision rules for predicting stress, anxiety, or depression based on survey responses.

- Functionality:
  - Hierarchical Decision Rules: Decision Trees utilize survey responses to construct a tree like structure where each node represents a feature, and branches emanating from nodes indicate feature values. These trees create decision rules to predict stress, anxiety, or depression based on the hierarchy of responses to survey questions.

- Information Gain: They split the data based on features to maximize the information gain at each level, identifying the most relevant questions for predicting psychological states.

- Usage Example: Stress Identification: Decision Trees might discern that individuals with a high workload, coupled with insufficient support, are more likely to experience stress. It creates a rule based structure to predict stress levels based on these factors.
- Evaluation: Assessing Decision Trees involves metrics such as accuracy, Gini impurity, or information gain. These metrics measure the tree's effectiveness in correctly predicting stress, anxiety, or depression based on survey responses.

4. Naive Bayes: Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes independence between features, hence "naive," which simplifies calculations. In your project, Naive Bayes predicts stress, anxiety, and depression levels by computing the probability of an individual experiencing these conditions based on survey responses. It's efficient, particularly with smaller datasets, and works well when features are conditionally independent.

- Functionality: Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are independent, hence the "naive" designation. It calculates the probability of a particular event based on prior knowledge of conditions related to that event.
- Feature Independence: Despite the assumption of feature independence, Naive Bayes can effectively handle complex relationships in the data. It is particularly suitable for text classification tasks, making it useful for analyzing open ended responses in the survey related to stress, anxiety, and depression.
- Usage Example: In the context of stress prediction, Naive Bayes could assess the likelihood of stress based on the co occurrence of certain keywords or phrases in the survey responses. For instance, frequent mentions of terms like "overwhelmed" or "pressure" might contribute to a higher probability of stress.
- Evaluation: Naive Bayes is evaluated using metrics such as accuracy, precision, recall, and F1 score. Its effectiveness lies in its simplicity, making it particularly valuable when the assumption of feature independence aligns with the characteristics of the data.

5. K Nearest Neighbors (KNN): K Nearest Neighbors is a simple yet effective algorithm for both classification and regression tasks. It works by identifying the 'k' nearest data points to a new instance and classifying it based on the majority vote or averaging the 'k' neighbors' values. In your project, KNN assesses stress, anxiety, and depression levels by finding similar patterns or responses within the survey dataset to predict an individual's mental health condition based on similarities with other respondents.

- Functionality: KNN is a non parametric and instance-based learning algorithm used for classification and regression tasks. It classifies a data point based on the majority class of its k nearest neighbors in the feature space. The choice of 'k' determines the number of neighbors considered.
- Feature Proximity: KNN operates on the assumption that similar data points share similar characteristics. It's effective in capturing local patterns and can adapt well to various types of features.
- Usage Example: In the survey context, KNN might predict stress levels by examining the responses of individuals with similar profiles in terms of demographics, lifestyle, or responses to specific survey questions. It considers the proximity of a person's characteristics to those of its k nearest neighbors.
- Evaluation: KNN's performance is typically assessed using metrics like accuracy and confusion matrix. The choice of 'k' is crucial, as too few neighbors might lead to noise sensitivity, while too many may oversimplify the model. Cross validation techniques help optimize 'k' for robust predictions.

6. Random Forests:

- Functionality:
  - Ensemble of Decision Trees: Random Forests aggregate multiple Decision Trees to improve predictive accuracy. Each tree is trained on a random subset of the dataset and makes individual predictions, and the final output is determined by aggregating these predictions.
  - Feature Importance: Random Forests evaluate the importance of each survey question or feature across multiple trees, providing insights into the most influential factors related to stress, anxiety, or depression.
- Usage Example: Depression Prediction: Random Forests analyze survey questions related to mood swings, sleep patterns, and social interactions across multiple decision trees. By combining these trees' outputs, it provides a more accurate prediction of depression based on these factors.
- Evaluation: Similar to Decision Trees, Random Forests are assessed using accuracy, Gini impurity, or information gain metrics, evaluating their collective performance in predicting stress, anxiety, and depression across multiple trees.

D. *Implementation Steps*

Here are the steps we have used for implementing machine learning:

- Data Collection: Gather survey responses from participants, focusing on stress, anxiety, and depression related queries.
- Data Cleaning: Preprocess the collected data by handling missing values, outliers, and standardizing formats.

- Exploratory Data Analysis (EDA): Analyze data distributions, correlations, and trends to understand the relationships between survey questions and stress indicators.
- Feature Selection: Use statistical measures and feature importance techniques to identify influential variables related to stress, anxiety, and depression.
- Model Selection: Choose appropriate algorithms (Logistic Regression, SVM, Decision Trees, Random Forests, Naive Bayes, and KNN) based on the project's objectives and data characteristics.
- Model Development: Train the selected models on the dataset using techniques like train test splits or crossvalidation to optimize their performance.
- Model Evaluation: Assess the models' performance using metrics like accuracy, precision, recall, and F1 score to determine their efficacy in stress prediction.
- Hyperparameter Tuning: Fine tune model parameters to enhance predictive accuracy and prevent overfitting.
- Ensemble Methods: If applicable, consider ensemble techniques to combine models for improved predictions.
- Deployment: Implement the best performing model into a production environment for stress prediction based on new data.
- Monitoring and Iteration: Continuously monitor the model's performance and iterate as needed to maintain accuracy and relevance.

E. Evaluation

1. Performance parameters: We also used performance indicators including accuracy, recall, precision, and F1 Score to determine how well detection of stress, anxiety and depression is working.

- Accuracy: Measures the overall correctness of predictions, the ratio of correctly predicted instances to the total instances.
- Precision: Indicates the accuracy of positive predictions, the ratio of correctly predicted positive observations to the total predicted positive observations.
- Recall (Sensitivity): Measures the ratio of correctly predicted positive observations to all actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

2) Confusion Matrix: It is a table used to understand the effectiveness of the classification and detection model. In regarding to botnet detection, a confusion matrix can be used to assess the accuracy of a botnet detection system by comparing real network traffic classifications.

The confusion matrix is often a two by two table that summarises the system's true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The rows of the matrix indicate the data's actual classification, while the columns represent the data's expected classification.

Various performance characteristics such as accuracy, recall, precision, and F1 score can be calculated by analysing the confusion matrix. For example, accuracy can be measured as (TN+TP)/(TN+FP+FN+TP), precision as TP/(TP+FP), and recall as TP/(TP+FN). These performance parameters can provide insights into the detection system's effectiveness and efficiency, as well as help lead system improvements.

Fig. 11. Logistic Regression Learning Curve

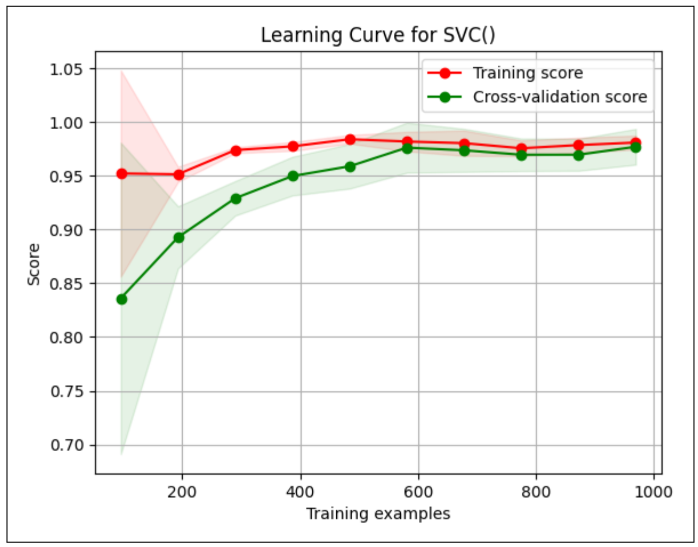

Fig. 12. SVM Learning Curve

IV. Result and Analysis

- Stress Detection
  - – Logistic Regression Metrics: Logistic Regression yielded a high accuracy of 98.14% in stress detection, showcasing robust performance. It demonstrated a precision of 97.98
  - SVM (Support Vector Machine) Metrics: SVM achieved an accuracy of 95.26% in stress detection, demonstrating strong performance. It showcased a precision of 95.83% and recall of 93.67%, indicating reliable identification of stress cases.

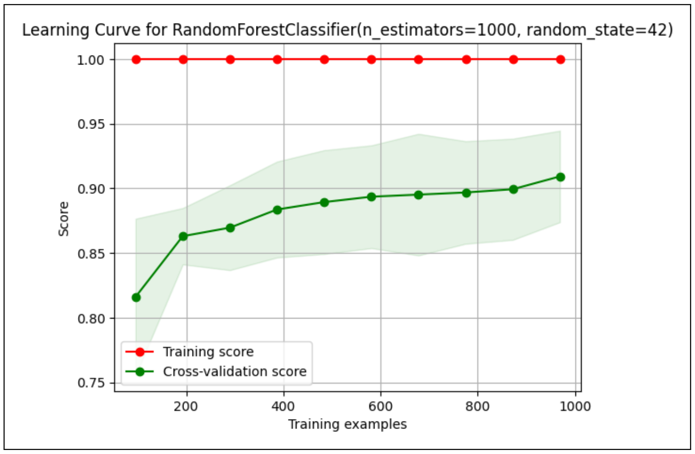

Fig. 13. Random Forest Learning Curve

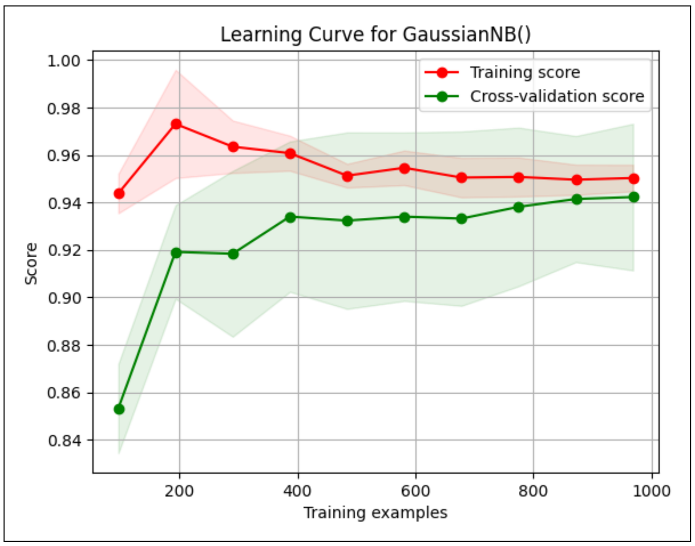

Fig. 14. Naive Bayes Learning Curve

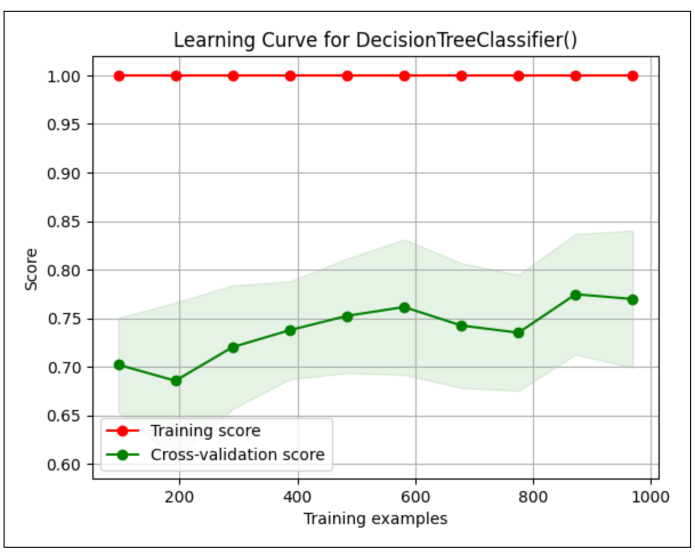

Fig. 15. Decision Tree Learning Curve

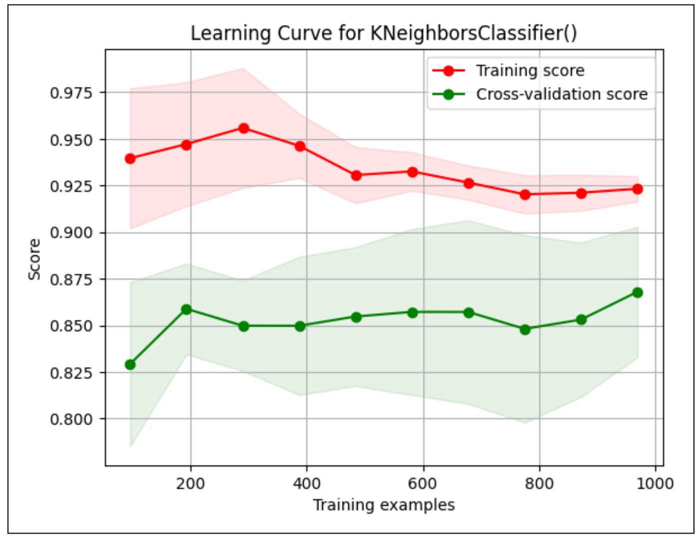

Fig. 16. KNN Learning Curve

  – Forest Metrics: The Random Forest model achieved an accuracy of 91.75% in stress detection. With a precision rate of 94.81% and recall rate of 88.66%, Random Forest demonstrated robust predictive capabilities for stress detection.
  – Decision Tree Metrics: The Decision Tree algorithm achieved an accuracy of 75.88% in stress detection, with a precision of 77.08% and recall of 74.90%. The confusion matrix indicated some misclassifications.
  – Naıve Bayes Metrics: Na¨ıve Bayes showcased a high accuracy of 93.20% in stress detection. It displayed exceptional precision (99.54%) but a slightly lower recall (87.04%) compared to other models.
  – KNN Metrics: The K Nearest Neighbors algorithm achieved an accuracy of 88.66% in stress detection. With balanced precision (89.67%) and recall (87.85%), KNN showcased a reliable performance in identifying stress cases.

- Anxiety Detection
  – Metrics: Logistic Regression achieved an accuracy of 97.94% in anxiety detection. It displayed high precision (99.55%) and recall (96.09%), showcasing strong predictive capabilities for identifying anxiety cases.
  – SVM (Support Vector Machine) Metrics: SVM achieved an accuracy of 95.05% in anxiety detection, demonstrating reliable performance. It showcased a precision of 100% and recall of 89.57% in identifying anxiety cases.
  – Random Forest Metrics: Random Forest demonstrated an accuracy of 88.45% in anxiety detection. With precision and recall rates of 94.85% and 80%, respectively, it exhibited robust predictive capabilities.
  – Decision Tree Metrics: The Decision Tree for anxiety

Models	Mental State	Accuracy	Precision	Recall	F1 Score
Logistic Regression	Stress	0.9814432989690721	0.9798387096774194	0.9838056680161943	0.9818181818181817
	Anxiety	0.979381443298969	0.9954954954954955	0.9608695652173913	0.9778761061946903
	Depression	0.9670103092783505	0.9596412556053812	0.9683257918552036	0.9639639639639639
Support Vector Machine	Stress	0.9525773195876288	0.9583333333333334	0.9366515837104072	0.947368421052631 6
	Anxiety	0.9505154639175257	1.0	0.8956521739130435	0.944954128440367
	Depression	0.9525773195876288	0.9583333333333334	0.9366515837104072	0.947368421052631 6
Random Forest	Stress	0.9175257731958762	0.948051948051948	0.8866396761133604	0.9163179916317992
	Anxiety	0.8845360824742268	0.9484536082474226	0.8	0.8679245283018867
	Depression	0.8969072164948454	0.8940092165898618	0.8778280542986425	0.8858447488584476
Decision Tree	Stress	0.7587628865979381	0.7708333333333334	0.7489878542510121	0.7597535934291582
	Anxiety	0.7546391752577319	0.7488789237668162	0.7260869565217392	0.7373068432671082
	Depression	0.7711340206185567	0.7370689655172413	0.7737556561085973	0.7549668874172185
Naïve Bayes	Stress	0.931958762886598	0.9953703703703703	0.8704453441295547	0.9287257019438445
	Anxiety	0.9381443298969072	0.9901960784313726	0.8782608695652174	0.9308755760368664
	Depression	0.9360824742268041	0.9702970297029703	0.8868778280542986	0.9267139479905437
K-Nearest Neighbors	Stress	0.8865979381443299	0.8966942148760331	0.8785425101214575	0.887525562372188
	Anxiety	0.8721649484536083	0.9158415841584159	0.8043478260869565	0.8564814814814816
	Depression	0.8577319587628865	0.8486238532110092	0.8371040723981901	0.8428246013667426

Fig. 17. Comparison of Machine learning Models

detection achieved an accuracy of 75.46%. It displayed a precision of 74.89% and recall of 72.61%, indicating some misclassifications.
  - Naıve Bayes Metrics: Na¨ıve Bayes achieved an accuracy of 93.81% in anxiety detection. It showcased high precision (99.02%) and good recall (87.83%), making it effective in identifying cases.
  - KNN Metrics: K Nearest Neighbors achieved an accuracy of 87.22% in anxiety detection. With precision and recall rates of 91.58% and 80.43% respectively, KNN showcased a reliable performance.
- Depression Detection
  - Logistic Regression Metrics: Logistic Regression achieved an accuracy of 96.70% in depression detection. With a precision of 95.96% and recall of 96.83%, it demonstrated robust performance in identifying depression cases.
  - SVM (Support Vector Machine) Metrics: SVM achieved an accuracy of 95.26% in depression detection, showcasing strong performance. It demonstrated a precision of 95.83% and recall of 93.67%, indicating reliable identification of depression cases.
  - Random Forest Metrics: Random Forest achieved an accuracy of 89.69% in depression detection. With a precision of 89.40% and recall of 87.78%, it exhibited robust predictive capabilities in identifying depression cases.
  - Decision Tree Metrics: The Decision Tree algorithm achieved an accuracy of 77.11% in depression detection. It displayed a precision of 73.71% and recall of 77.38%.
  - Na¨ıve Bayes Metrics: Na¨ıve Bayes achieved an accuracy of 93.61% in depression detection. With precision and recall rates of 97.03% and 88.69% respectively, Na¨ıve Bayes showed a robust performance.
  - KNN Metrics: K Nearest Neighbors achieved an accuracy of 85.77% in depression detection. With precision and recall rates of 84.86% and 83.71% respectively, KNN showcased a reliable performance

V. Conclusion

In conclusion, this research embarked on a journey to unravel the intricate web of stress, anxiety, and depression, employing a robust framework of machine learning algorithms. The primary goal was to develop predictive models capable of discerning and forecasting these psychological states based on multifaceted survey responses. By integrating sophisticated

	Stress	Anxiety	Depression
Logistic Regression	$\begin{bmatrix}233 & 5\\ 4 & 243\end{bmatrix}$	$\begin{bmatrix}254 & 1\\ 9 & 221\end{bmatrix}$	$\begin{bmatrix}255 & 9\\ 7 & 214\end{bmatrix}$
Support Vector Machine	$\begin{bmatrix}255 & 9\\ 14 & 207\end{bmatrix}$	$\begin{bmatrix}255 & 0\\ 24 & 206\end{bmatrix}$	$\begin{bmatrix}255 & 9\\ 14 & 207\end{bmatrix}$
Random Forest	$\begin{bmatrix}226 & 12\\ 28 & 219\end{bmatrix}$	$\begin{bmatrix}254 & 10\\ 46 & 184\end{bmatrix}$	$\begin{bmatrix}241 & 23\\ 27 & 194\end{bmatrix}$
Decision Tree	$\begin{bmatrix}183 & 55\\ 62 & 185\end{bmatrix}$	$\begin{bmatrix}199 & 56\\ 63 & 167\end{bmatrix}$	$\begin{bmatrix}203 & 61\\ 50 & 171\end{bmatrix}$
Naïve Bayes	$\begin{bmatrix}237 & 1\\ 32 & 215\end{bmatrix}$	$\begin{bmatrix}253 & 2\\ 28 & 202\end{bmatrix}$	$\begin{bmatrix}258 & 6\\ 25 & 196\end{bmatrix}$
K-Nearest Neighbors	$\begin{bmatrix}213 & 25\\ 30 & 217\end{bmatrix}$	$\begin{bmatrix}238 & 17\\ 45 & 185\end{bmatrix}$	$\begin{bmatrix}231 & 33\\ 36 & 185\end{bmatrix}$

Fig. 18. Confusion Matrix of Machine learning Models

algorithms such as Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Naive Bayes, and K Nearest Neighbors, this study aimed to harness the predictive potential of these methodologies within the realm of mental health diagnostics.

Through meticulous data collection, extensive exploratory data analysis, and feature engineering, this research delved deep into understanding the relationship between survey questions and the manifestation of stress related conditions. Despite the commendable success in developing predictive models, this study acknowledges certain limitations. The dataset, while this might benefit from additional dimensions and diverse demographic representations to enhance model generalization. Ethical onsiderations surrounding data privacy and biases inherent in self reported surveys underscore the need for caution in interpreting and deploying these models in realworld settings. However, the predictive power demonstrated by these algorithms signals a promising avenue for augmenting mental health diagnostics, provided ethical and procedural concerns are diligently addressed.

In essence, this research strides toward bridging the gap between traditional mental health assessments and contemporary machine learning techniques. The amalgamation of comprehensive survey data and advanced algorithms opens doors to a more nuanced understanding of psychological wellbeing, laying the foundation for future research and applications aimed at early detection and personalized interventions for stress, anxiety, and depression. As we navigate this inter section of mental health and technology, ethical frameworks and continual refinement of methodologies will be pivotal in harnessing the full potential of machine learning for the betterment of mental health diagnostics and interventions.

Examining the results reveals inherent limitations in the study. The dataset's predominant composition of responses from college students may introduce sample bias, potentially limiting the generalizability of findings to a broader demographic spectrum. Moreover, the reliance on self-reported data carries the risk of social desirability bias and inaccuracies, given variations in individuals' willingness to disclose accurate information about their mental health conditions. Additionally, the survey's primary focus on anxiety, depression, and stress might inadvertently sideline other mental health disorders, thus restricting the model's applicability to a comprehensive spectrum. The one-time nature of the survey raises concerns about its ability to capture the dynamic nature of mental health conditions over time. Finally, the absence of clinical validation and reliance on questionnaire-derived scores could impact the model's precision and diminish its clinical relevance.

These collective limitations emphasize the need for cautious interpretation and suggest potential avenues for future research and refinement.

REFERENCES

[1] Machine Learning Approach to Prediction and Assessment of Depression and Anxiety: A Literature Review. author=Vaanathy, Sabanayagam and Charles, Joseph and Lekamge, Sugeeswari, journal=IUP Journal of Computer Sciences(2023)

[2] A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis author=Iyortsuun, Ngumimi Karen and Kim, Soo Hyung and Jhon, Min and Yang, Hyung Jeong and Pant, Sudarshan (2023)

[3] Applications of artificial intelligence machine learning for detection of stress: a critical overview, author=Mentis, Alexios Fotios A and Lee, Donghoon and Roussos, Panos (2023)

[4] Depression Detection From Social Networks Data Based on Machine Learning and Deep Learning Techniques: An Interrogative Survey, author=Hasib, Khan Md and Islam, Md Rafiqul and Sakib, Shadman and Akbar, Md Ali and Razzak, Imran and Alam, Mohammad Shafiul (2023)

[5] On the robustness of machine learning models for stress and anxiety recognition from heart activity signals, author=Henry, John and Lloyd, Huw and Turner, Martin and Kendrick, Connah (2023)

[6] Priya, Anu, Shruti Garg, and Neha Prerna Tigga. ”Predicting anxiety, depression and stress in modern life using machine learning algorithms.” Procedia Computer Science 167 (2020)

[7] umar, Prince, Shruti Garg, and Ashwani Garg. ”Assessment of anxiety, depression and stress using machine learning models.” Procedia Computer Science 171 (2020)

[8] Ahmed, Arfan, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd Alrazaq, and Mowafa Househ. ”Machine learning models to detect anxiety and depression through social media: A scoping review.” Computer Methods and Programs in Biomedicine Update (2022)

[9] Panicker, Suja Sreeith, and Prakasam Gayathri. ”A survey of machine learning techniques in physiology based mental stress detection systems.” Biocybernetics and Biomedical Engineering 39.2 (2019)

[10] Singh, Astha, and Divya Kumar. ”Computer Assisted identification of Stress, Anxiety, Depression (SAD) in Students: A State of the art review.” Medical Engineering & Physics (2022)

[11] Aleem, Shumaila, Noor ul Huda, Rashid Amin, Samina Khalid, Sultan S Alshamrani, and Abdullah Alshehri. ”Machine learning algorithms for depression: diagnosis, insights, and research directions.” Electronics 11.7 (2022)

[12] Subhani, Ahmad Rauf, Wajid Mumtaz, Mohamed Naufal Bin Mohamed Saad, Nidal Kamel, and Aamir Saeed Malik. ”Machine learning framework for the detection of mental stress at multiple levels.” IEEE Access (2017)

[13] Jain, Swati, Suraj Prakash Narayan, Rupesh Kumar Dewang, Utkarsh Bhartiya, Nalini Meena, and Varun Kumar. ”A machine learning based depression analysis and suicidal ideation detection system using questionnaires and twitter.” (2019)

[14] arg, Prerna, Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. ”Stress detection by machine learning and wearable sensors.” 26th International Conference on Intelligent User Interfaces Companion. (2021)

[15] M. Bajaj, P. Rawat, A. Bhatt, V. Sharma, A. Jain and N. Kumar, ”Classification And Prediction of Brain Tumors and its Types using Deep Learning,” (2023)

[16] Teelhawod, Bibi Nushrina, Faijan Akhtar, Md Belal Bin Heyat, Pragati Tripathi, Rajat Mehrotra, Ashamo Betelihem Asfaw, Omar Al Shorman, and Mahmoud Masadeh. ”Machine learning in E health: a comprehensive survey of anxiety.” 2021 International Conference on Data Analytics for Business and Industry (ICDABI). IEEE, (2021)

[17] Bobade, Pramod, and Vani, M. ”Stress detection with machine learning and deep learning using multimodal physiological data.” 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, (2020)

[18] McGinnis, Ryan S, Ellen W McGinnis, Jessica Hruschak, Nestor L Lopez Duran, Kate Fitzgerald, Katherine L Rosenblum, and Maria Muzik. ”Rapid anxiety and depression diagnosis in young children enabled by wearable sensors and machine learning.” 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, (2018)

[19] Mohan, Yogeswaran, Sia Seng Chee, Donica Kan Pei Xin, and Lee Poh Foong. ”Artificial neural network for classification of depressive and normal in EEG.” 2016 IEEE EMBS conference on biomedical engineering and sciences (IECBES), IEEE, (2016)

[20] P. Rawat, M. Bajaj, S. Vats and V. Sharma, ”ASD Diagnosis in Children, Adults, and Adolescents using Various Machine Learning Techniques,” 2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT), Dehradun, India, (2023)

[21] M. Bajaj, P. Rawat, Diksha, S. Vats, V. Sharma and L. Gopal, ”Prediction of Mental Health Treatment Adherence using Machine Learning Algorithms,” 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, (2023)

[22] Masood, Khalid, and Mohammed A Alghamdi. ”Modeling mental stress using a deep learning framework.” IEEE access 7 (2019)

[23] Srinath, KS, K Kiran, S Pranavi, M Amrutha, P Deepa Shenoy, and KR Venugopal. ”Prediction of Depression, Anxiety and Stress Levels Using Dass 42.” 2022 IEEE 7th International conference for Convergence in Technology (I2CT), IEEE, (2022)

[24] Zhang, Shen, Mei Tu, Yueru Yan, Yimeng Zhuang, Likun Ge, and Gaoxia Wei. ”Detecting Depression, Anxiety and Mental Stress in One Sequential Model with Multi task Learning.” International Conference on Human Computer Interaction, Springer, (2022)

[25] Raval, Dhatri, and others. ”Stress detection using convolutional neural network and internet of things.” Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12.12 (2021)

[26] Ghosh, Shreya, and Tarique Anwar. ”Depression intensity estimation via social media: a deep learning approach.” IEEE Transactions on Computational Social Systems 8.6 (2021)

[27] Budiyanto, Setiyo, Harry Candra Sihombing, Fajar Rahayu IM. ”Depression and anxiety detection through the Closed Loop method using DASS 21.” TELKOMNIKA (Telecommunication Computing Electronics and Control)(2019)

[28] Qasrawi, Radwan, Stephanny Paola Vicuna Polo, Diala Abu Al Halawa, Sameh Hallaq, Ziad Abdeen, and others. ”Assessment and prediction of depression and anxiety risk factors in schoolchildren: machine learning techniques performance analysis.” JMIR Formative Research (2022)

[29] Bhatnagar, Shaurya, Jyoti Agarwal, and Ojasvi Rajeev Sharma. ”Detection and classification of anxiety in university students through the application of machine learning.” Procedia Computer Science 218 (2023)