National College of Ireland

Unleashing the power of Data: Personalized Anxiety Interventions

MSc Research Project
Data Analytics

Manoj Kumar Periyasamy
Student ID: x22153209

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Manoj Kumar Periyasamy
Student ID:	x22153209
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Arjun Chikkankod
Submission Due Date:	14/12/2023
Project Title:	Unleashing the power of Data: Personalized Anxiety Interventions
Word Count:	6583
Page Count:	17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Manoj Kumar Periyasamy
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Unleashing the power of Data: Personalized Anxiety Interventions

Manoj Kumar Periyasamy
x22153209

Abstract

This research aims to examine how real-time data analytics may lead to personalized anxiety treatments. We use machine learning models combined with fitness tracker data so as to discover such patterns and potential predictors of anxiety disorders. In this study, data is used at the minute's level of output of physical activities, heart rate recording and sleep monitoring. To ensure more credible and relevant insights, data pre-processing and feature selection techniques are utilized. Predictive machine learning models such as Linear Regression, Gradient Boosting, Decision Tree, and Random Forest are employed for the prediction of anxiety.The findings indicate that Random Forest outperforms other machine learning models, demonstrating the lowest MSE values across most features compared to other machine learning models. Although Decision Tree was competitive, it exhibited slightly higher MSE values. A comprehensive analysis involving Mean Squared Error (MSE) and R-squared further solidified Random Forest's superiority, showcasing the lowest MSE and an impressive R-squared value. These results suggest that Random Forest excels in capturing intricate relationships within the data, making it particularly useful for precise anxiety level predictions. Finally, different deep learning models are also implemented,likely Feed-forward Neural Network and an Long Short Term Memory(LSTM). While both demonstrated comparable results results with R square values and MSE values. This shows that Neural Network also excels in capturing intricate relationships within the data which is useful for Anxiety prediction. The insights derived from this analysis not only contribute to refining predictive models but also advance the understanding of anxiety dynamics, catering to the broader goal of improving intervention strategies

Keywords : Machine Learning,Deep Learning, Decision Tree, Linear Regression,Gradient Boost,Random Forest,Long Short Term Memory,Feed-forward Neural Network ,Fitness Tracking, Data Preprocessing, Health Monitoring.

1 Introduction

In an era characterized by the unprecedented ubiquity of technology and the proliferation of real-time data, harnessing this wealth of information holds transformative potential, particularly in the domain of mental health. Anxiety, a pervasive and often debilitating condition, stands as a significant target for innovative interventions. This study embarks on a journey to unleash the power of real-time data for the development of personalized anxiety interventions, recognizing the dynamic interplay between data science and mental

well-being. The contemporary landscape witnesses an increasing reliance on wearables and digital platforms that effortlessly capture a myriad of physiological and behavioural metrics. From metrics related to physical activity, such as MET, heart rate, calories, and steps, to temporal patterns represented by time range, the depth and breadth of real-time data offer a unique lens into the intricate nuances of an individual's daily life. Against this backdrop, machine learning models, including Decision Tree, Random Forest, Linear Regression, and Gradient Boosting, emerge as powerful tools capable of distilling meaningful insights from this wealth of data.

Research Question: *"How can data analytics techniques applied to real-time monitoring data from wearable devices and mobile applications be leveraged to identify patterns, predictive factors, and individualized indicators associated with anxiety disorders, and how can this knowledge be used to develop personalized intervention?"*

This research is driven by a dual purpose: first, to conduct a comparative analysis of the predictive capabilities of these machine learning models in estimating anxiety levels; and second, to pave the way for the development of personalized interventions tailored to an individual's unique behavioural and physiological profile. Anxiety, a complex and multifaceted psychological phenomenon, necessitates a nuanced understanding that extends beyond conventional diagnostic approaches. Real-time data, coupled with advanced machine learning algorithms, holds the promise of unravelling the intricate tapestry of anxiety dynamics. As we stand at the intersection of mental health and technological innovation, the potential to revolutionize anxiety management beckons. By scrutinizing the performance of different machine learning models and identifying the most effective predictors through a comprehensive analysis of Mean Squared Error and R-squared values, this study seeks to contribute valuable insights to the burgeoning field of personalized mental health interventions. The subsequent sections will delve into the methodologies employed, the results obtained, and the implications for the future development of ethically sound tools to alleviate anxiety and improve overall well-being.

In the contemporary era of digital health and wearable technology, the utilization of fitness trackers, exemplified by devices like Fitbit, has become ubiquitous in monitoring and optimizing physical well-being. These devices excel at capturing an extensive array of metrics, ranging from steps taken and heart rate to caloric expenditure and exercise intensity. However, amidst the wealth of physiological data they provide, the challenge arises when attempting to draw direct correlations to mental health indicators, such as anxiety or depression. Fitness trackers, designed primarily for monitoring physical activity, present a nuanced landscape when it comes to assessing mental health. The data they yield offers valuable insights into an individual's daily life, yet it inherently lacks the depth required for a comprehensive understanding of mental well-being. This distinction is crucial, as mental health encompasses intricate facets beyond the scope of physical activity metrics.

This study delves into the complexities of identifying anxiety or depression based solely on data from fitness trackers, acknowledging the inherent challenges posed by the indirect nature of these indicators. While there may exist subtle patterns and indirect indicators associated with mental health within the fitness tracker data, it is paramount to recognize that these devices are not intended for the diagnosis of mental health conditions. The limitations inherent in relying solely on fitness tracker data for mental health assessments underscore the need for caution and a broader evaluation strategy. Moreover,

this exploration emphasizes the importance of contextualizing fitness tracker data within a more comprehensive assessment framework, one that incorporates clinical evaluations and self-reported mental health measures. The integration of fitness tracker data into a larger, more holistic evaluation is essential for a nuanced understanding of an individual's overall well-being. However, this integration also prompts a consideration of the potential ethical implications associated with handling sensitive health-related data. As we navigate the intersection of technology and mental health, it becomes imperative to tread carefully, leveraging fitness tracker data as a valuable piece of a larger puzzle in mental health assessment. This introduction sets the stage for an exploration into the nuanced dynamics of using fitness tracker data in mental health assessments, acknowledging both its potential and the essential need for a judicious and comprehensive approach to understanding mental well-being.

2 Related Work

TusharP and colleagues Thakre et al. (2022) The study aims to utilize deep learning techniques to analyse polysomnographic data for the detection of anxiety and depression, which are common psychiatric conditions with a bidirectional relationship with sleep. The study suggests that machine learning techniques, specifically deep learning, have the potential to accurately detect the presence of anxiety and depression through the analysis of sleep study data. In the independent test set, the model achieved a high accuracy of 96.88% This paper highlights the promising application of deep learning in mental health assessment through the analysis of sleep data.
NicholasC and colleaguesJacobson et al. (2021) The study aims to assess the capacity of passive smartphone and wearable sensor data, collected throughout daily life, to predict long-term prognosis in generalized anxiety disorder and panic disorder symptoms. Out-of-sample cross-validated results indicated that wearable movement data could significantly predict which individuals would experience symptom deterioration, with an AUC (Area Under the Curve) of 0.696, a balanced accuracy of 68.7%, 84.6% sensitivity, and 52.7% specificity. This research paper highlights the potential of using deep learning models and wearable sensor data for predicting the long-term prognosis of anxiety disorder symptoms.
Malaika Pandit, and colleagues Pandit et al. (2023) the research aims to understand the factors contributing to anxiety and depression and implement predictive modelling to enhance the accuracy and efficiency of early mental health diagnoses. research on using data science and machine learning techniques for early detection and predictive modelling of anxiety and depression The study found that a Tabular Deep Neural Network (DNN) outperformed other machine learning classifiers, including Artificial Neural Networks (ANN), by approximately 30%.
Sahaja Dixit and colleagues Dixit et al. (2022) this research underscores the potential of data science and deep learning techniques in improving the early diagnosis of mental health conditions, ultimately facilitating better access to support for individuals in need. Early Detection and Classification of Alzheimer's Disease (AD), Parkinson's Detection Generalized Anxiety Disorder Stress Prediction for Patients. highlights the growing importance of deep learning and machine learning in the early detection and diagnosis of various neurocircuitry disorders and diseases, which can significantly impact the well-being of individuals. article reviews relevant literature on neurological diseases, emphas-

izing the role of deep learning and machine learning in diagnosing these diseases at their primary stages.

Goswami, S and colleagues Goswami et al. (2019) the paper describes a framework that employs machine learning techniques to partially automate the process of conducting systematic literature reviews by extracting specific data elements related to anxiety outcome measures from publications. This research contributes to the development of automated tools for conducting systematic literature reviews, making the process more efficient and accurate. The aim of this work is to develop an effective framework using machine learning to streamline the process of systematic literature reviews, specifically focusing on collecting data related to anxiety outcome measures data elements are extracted from relevant publications retrieved from databases such as Medline, EMBASE, CINAHL, AHMED, and Pyscinfo.

Arfan Ahmed and colleagues Ahmed et al. (2022) describes the research explores the use of machine learning models to detect anxiety and depression through analysis of social media data, particularly on platforms like Facebook and Twitter, aims to leverage machine learning models for detecting anxiety and depression in individuals through their social media activity. This approach becomes particularly relevant during the COVID-19 pandemic, where traditional screening methods may face limitations. he studies involved searching six bibliographic databases, following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) protocol. A total of 2,219 studies were retrieved, with 54 studies included in the review. The majority of the studies (70%, 38 out of 54) were conducted at the peak of the COVID-19 pandemic (2019-2020). underscores the importance of leveraging machine learning and social media data for detecting mental health conditions and offers a valuable approach during periods of public health crises.

N. Saddaf Khan and colleagues Khan et al. (2021) from the study that focuses on using electroencephalography (EEG) signals and machine learning techniques for stress assessment and anxiety mitigation in individuals with autism spectrum disorder (ASD) and neurotypical individuals. investigate the feasibility of utilizing EEG signals for stress assessment and identifying stress states in adolescents, including those with ASD, and to develop a closed-loop adaptation of respiration entrainment for anxiety mitigation. The study involved both neurotypical (n=5) and ASD (n=8) participants. Several machine learning classifiers were compared, including support vector machine (SVM) and deep learning methods. This research showcases the potential of utilizing EEG-based machine learning to improve the assessment and mitigation of mental stress, particularly in vulnerable populations like adolescents with ASD.

P. Bobade and colleagues Bobade and Vani (2020) describing the research on Human Activity Recognition (HAR) for detecting anxiety-related behaviours using smartphone motion sensors and Inertial Measurement Unit (IMU) data. develop a HAR-based method to recognize behaviours associated with anxiety, which is the most common form of mental disorder, affecting a significant portion of the adult population. A novel dataset of anxious behaviours is created, utilizing motion sensors from smartphones and IMU data. Unique features are extracted to represent these behaviours. The study highlights the performance of a deep learning model comprising Convolution Neural Network (CNN) and Long-Short Term Memory (LSTM). This model is shown to outperform other algorithms, achieving over 92% accuracy in recognizing anxiety-related behaviours. This research underscores the potential of HAR and deep learning for identifying anxiety-related behaviours, which can have significant implications for the field of psychiatry and

the early detection of anxiety disorders.
Astha Singhe and colleagues Singh and Kumar (2021) for the three-class classification problem (amusement vs. baseline vs. stress), machine learning techniques achieved accuracies of up to 81.65%. For the binary classification problem (stress vs. non-stress), machine learning techniques achieved accuracies of up to 93.20%. Deep learning techniques outperformed machine learning techniques, achieving accuracies of up to 84.32% for the three-class classification problem and 95.21% for the binary classification problem. potential of leveraging wearable sensor data and advanced machine learning and deep learning techniques for stress detection, which can be a significant contribution to stress management and mental health.
Wanderley Espinola et.al Espinola et al. (2022)Random forests with 300 trees achieved the best classification performance, with an accuracy of 75.2%, a kappa score of 69.08%, a sensitivity of 75.30%, and a specificity of 93.80%. These results were obtained for the simultaneous detection of major depressive disorder, schizophrenia, bipolar disorder, and generalized anxiety disorder. This study highlights the potential of vocal acoustic analysis and machine learning as a promising tool for supporting the diagnosis of mental disorders, offering a more objective and accessible approach in the field of psychiatry. The research aims to develop a methodology to assist in the diagnosis of major depressive disorder, bipolar disorder, schizophrenia, and generalized anxiety disorder using vocal acoustic features and machine learning. This approach provides an objective and non-invasive method for psychiatric diagnosis.
G.G.Rajput, et.al.Rajput and Alashetty (2022) describes address the issue of early readmissions of diabetic patients, which can lead to increased healthcare costs and reduced treatment quality. The study employs machine learning techniques to analyse a dataset of 100,000 medical records from 70,000 diabetic patients. A unique preprocessing approach is demonstrated to reduce the 55 characteristics of the dataset to 21, making it more manageable for machine learning analysis. Gradient Boosting outperformed other machine learning methods, achieving an accuracy of 77.4% in predicting early readmission of diabetic patients. contributes to the use of machine learning for addressing challenges in healthcare, particularly related to early readmission of diabetic patients.
Faisal Mashel Albagmi a, et.al. Albagmi et al. (2022)from the study that focuses on using machine learning to classify anxiety problems during the Covid-19 pandemic in Saudi Arabia. classify two-class and three-class anxiety problems early using a dataset collected during the Covid-19 pandemic in Saudi Arabia. The study focuses on identifying factors influencing anxiety levels and screening for Generalized Anxiety Disorders (GAD-7). Experimental results showed promising outcomes for the early classification of both two-class and three-class anxiety problems. The Support Vector Machine classifier achieved a classification accuracy of 100%, precision of 1.0, recall of 1.0, and f-measure of 1.0 using 10 features. The comparison between the Support Vector Machine and J48 Decision Tree indicated that the Support Vector Machine outperformed the Decision Tree in this context. identifying anxiety problems during public health crises and showcases the effectiveness of the Support Vector Machine classifier.

3 Methodology

The research methodology consists of distinct stages, as illustrated in the accompanying figure (Refer Figure 1): data collection, data preprocessing, feature selection, data

modelling, and evaluation metrics. This study is guided by the CRISP methodology. The dataset was generated through responses obtained from a distributed survey conducted via Amazon Mechanical Turk between December 3, 2016, and December 5, 2016. Thirty eligible Fitbit users provided consent for the submission of personal tracker data, encompassing minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed using either the export session ID (column A) or timestamp (column B). The variability in output reflects the use of various Fitbit tracker types and individual tracking behaviours/preferences.

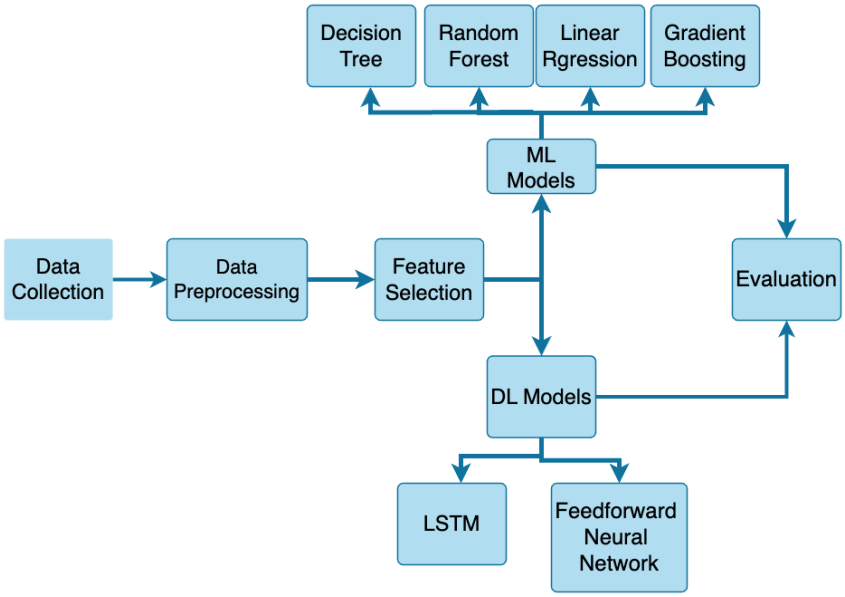

Figure 1: Research Methodology Flow Chart for Prediction of Anxiety

The foundation of our investigation pertaining to the forecasting of physical activity intensity is the systematic gathering and organization of data. The 'ActivityMetrics' dataset is employed in order to gain a more comprehensive understanding of the wide range of physical activities. A multitude of factors, including heart rate, calories expended, intensity, number of steps, and time range, are incorporated for every category of activity represented in this dataset. This dataset documents the variety and unpredictability of numerous physical activities, which is quite valuable

Data collection commences with a statistically valid sample that encompasses a broad spectrum of physical activities, ranging from sedentary pursuits such as yoga and meandering to more physically demanding activities like weightlifting and running. It is essential that the prediction models be diverse in order to account for the vast array of physical activity patterns observed in the real world.

Once the 'ActivityMetrics' dataset has been obtained, the subsequent phase is data preparation, which is an essential step in preparing the unprocessed data for analysis. Before commencing any analysis, it is critical to ascertain that the dataset is devoid of errors, anomalies, and discrepancies. Data cleansing encompasses the resolution of inconsistencies in measurements, the maintenance of data consistency, and the improvement of the information's overall dependability. Following normalization, the pipeline for preprocessing proceeds. Normalization is an essential process when dealing with metrics that exhibit variation in scale, including heart rate, steps, and metabolic equivalents (MET). Its purpose is to mitigate the impact of magnitude-based fluctuations on the model's

training process. By doing so, one can establish a sense of assurance regarding the consistency of the dataset, thereby enhancing the dependability of comparisons and averting the exaggeration of particular measurements on account of their magnitude. Inclusion of absent values is an additional component of data preparation. When fragmentary data cannot be avoided in practical situations, a systematic approach to completing missing data points is required. By replacing missing values with imputation methods such as median or mean substitution, one can guarantee that the predictive models are trained on the complete dataset.

Following the completion of feature selection and model training, the 'ActivityMetrics' dataset has been cleansed, normalized, and missing value removal has been performed. Numerous machine learning models have endeavoured to forecast metrics that reflect the intensity of physical activity; by meticulously collecting and preparing data, we can scrutinize their endeavours. Gradient Boosting, Decision Tree, Random Forest, and Linear Regression are some of these models. By incorporating an exhaustive dataset and meticulous preprocessing techniques, this methodology aims to improve prediction models pertaining to the intensity of physical activity. The objective is to enhance and broaden the applicability of the models.

The stage of feature selection and engineering guides our analytical exploration of the intricate terrain pertaining to the evaluation of physical exercise intensity. In order to ensure that our variables accurately reflect the complexities of physical activity patterns, we employ a meticulous selection process for attributes from the 'ActivityMetrics' dataset. Feature selection entails a thorough examination and assessment of the significance and contribution of every statistic within the dataset in order to make

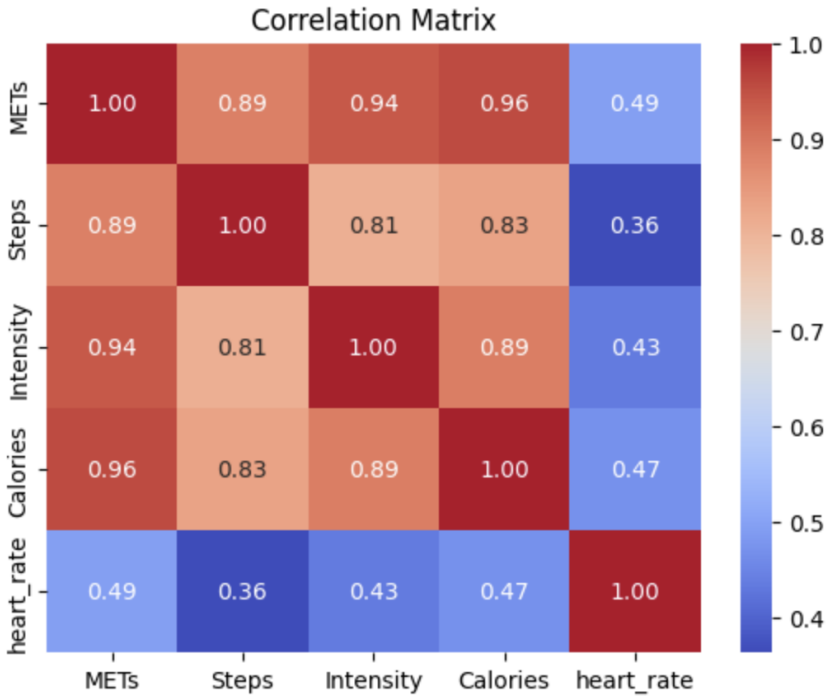

Figure 2: Correlation Matrix

precise predictions regarding the intensity of physical activity.In figure 2, a correlation matrix has been plotted for a a Quick and complete knowledge of variable connections.The

matrix shows relationships between various indicators, with values ranging from -1 to 1. A positive correlation suggests a positive linear relationship, whereas a negative correlation shows a negative linear relationship. A number of 1 or -1 indicates complete correlation, whereas 0 indicates no linear correlation.

The 'ActivityMetrics' dataset has been enhanced through the implementation of engineering principles and rigorous feature selection to incorporate variables that represent various forms of physical activity. Once these characteristics have been meticulously selected, machine learning algorithms such as Linear Regression, Gradient Boosting, Decision Tree, and Random Forest are prepared to comprehend the complexities of physical activity intensity prediction. Through the astute optimization of characteristics and the seamless integration of the dataset's abundance, our objective is to develop a predictive model that comprehensively and precisely encapsulates the fundamental nature of anxiety.Figure 1 depicts the system architecture.

4 Design Specification

This specific design is the combination of machine learning and deep learning frameworks architecture. This system design consists of deep learning, and in machine learning regression part is discussed. For predicting the anxiety from the people, we have used machine learning model, by providing a corresponding result like MSE and Accuracy values. To ensure the future Anxiety population prediction is accurate, 90%. deep learning, and traditional machine learning methods i.e., decision tree, random forest, linear regression and gradient boosting algorithms have been used to predict the heart rate, intensity, calories as shown in Figure 1. The components of machine learning and deep learning forecasting model included. Well-Architected Machine Learning LifeCycle for Prediction of Anxiety(Refer Figure 3) and further analysis of Anxiety is mentioned in below subsections. This life cycle explains the CRISP Methodology structure.

4.1 Machine Learning Life Cycle prediction of Anxiety

The Well-Architected Machine Learning (ML) Lifecycle for the Prediction of Anxiety is designed to ensure a robust and effective framework for developing and deploying predictive models in the context of anxiety detection. The lifecycle encompasses key stages, starting with comprehensive data collection that includes diverse sources of relevant information related to anxiety. Data preprocessing techniques are applied to ensure data quality, handle missing values, and standardize formats. Feature selection is conducted to identify the most relevant variables for prediction. The data modeling phase involves the implementation of machine learning algorithms tailored to anxiety prediction, with a focus on model interpretability and performance. Rigorous evaluation metrics are employed to assess the model's MSE and R-squared (R2). Ethical considerations, including privacy and bias mitigation, are integrated into the design to ensure responsible deployment. Continuous monitoring and adaptation are emphasized, fostering an agile approach that accommodates changes in the data distribution and user needs over time. This well-architected ML lifecycle aims to contribute to the development of reliable and ethical anxiety prediction models.

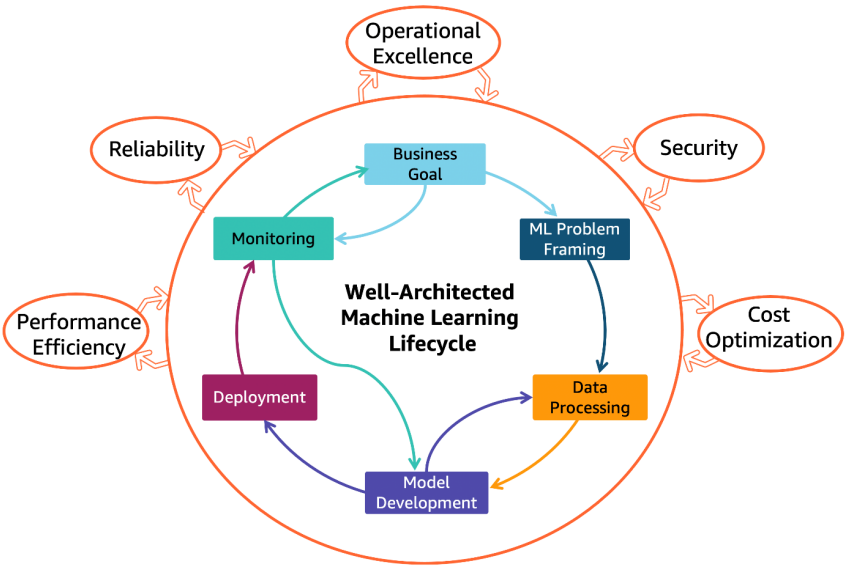

Figure 3: Well-Architected Machine Learning LifeCycle for Prediction of Anxiety

4.2 Neural Network for Regression

In the implemented feedforward neural network using TensorFlow and Keras, a three-layer architecture was employed with 64 and 32 units in the first and second hidden layers, respectively. The model utilized the ReLU activation function and linear activation for regression output. The Adam optimizer and Mean Squared Error (MSE) loss function were applied during compilation. Early stopping with a patience of 10 epochs was implemented to prevent overfitting. Training for 100 epochs resulted in an impressive performance. Refer Figure 4.

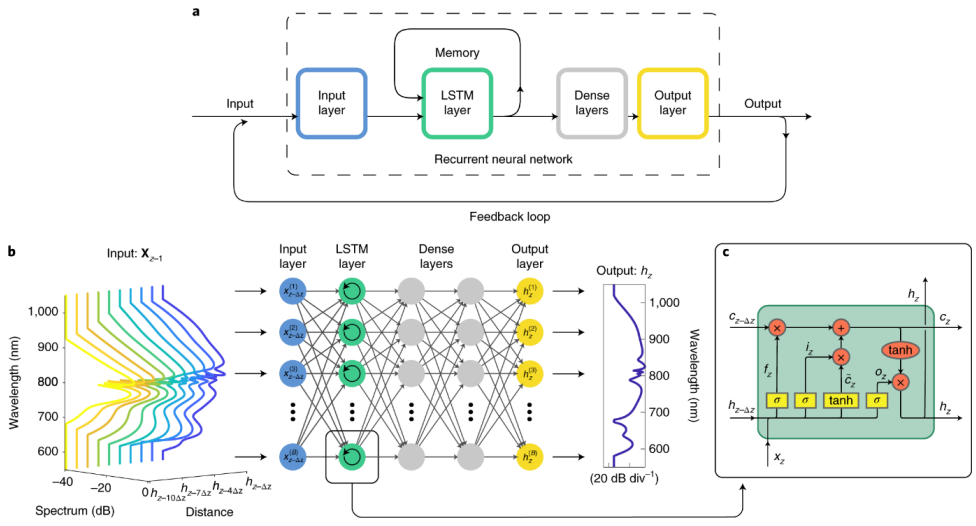

Figure 4: Neural Network Structure

4.3 LSTM Model for Regression

The Long Short-Term Memory (LSTM) model was designed with a single LSTM layer of 64 units using the ReLU activation function, followed by a dense layer with linear

activation for regression. The input shape was specified as (1, features), indicating the model's expectation of sequences with a length of 1. It is noteworthy that the code contains a duplicated section for building the LSTM model. This model is ready for compilation, training, and evaluation for time-series or sequential data, building on the strengths of LSTM architecture.

5 Implementation

In the Machine Learning Life Cycle for Anxiety prediction , various regression models were employed, including Decision Tree, Random Forest, Linear Regression, and Gradient Boost. Additionally, a Deep Learning approach was implemented, featuring a basic Neural Network alongside an LSTM forecasting model.

This comprehensive approach demonstrates the versatility of regression techniques in addressing anxiety prediction within the machine learning framework. Implemented for forecasting anxiety, the Decision Tree algorithm navigates the 'ActivityMetrics' dataset, revealing determinants impacting predictions. Its interpretability allows understanding the reasoning behind predictions. Employing iterative feature evaluation, the tree discerns critical parameters distinguishing activity levels. Pruned to prevent overfitting, it ensures robust and dependable predictions. Expanding into Gradient Boosting, Random Forest, and Linear Regression, this methodology explores diverse prediction perspectives. Random Forest, a robust predictive methodology for physical activity intensity, employs ensemble learning to enhance accuracy. In contrast to Gradient Boosting, it constructs decision trees independently, capturing diverse dataset characteristics. Each tree contributes unique insights, and their integration forms a comprehensive prediction model. While Gradient Boosting excels in capturing intricate correlations, Random Forest excels in versatility and simplicity, demonstrating its effectiveness in modeling anxiety intensity. Linear Regression, a fundamental technique for physical activity intensity prediction, explores linear correlations between input features and metrics. Using the 'ActivityMetrics' dataset, it establishes proportionality, weighting characteristics to predict outcomes. Coefficients quantify the influence of each feature on activity level, adjusted during training. Positive and negative coefficients reveal relationships, aiding nuanced comprehension. While proficient in characterizing linear relationships, Linear Regression has limitations in complex scenarios, addressed by regularization techniques like Ridge or Lasso regression.

Gradient Boosting, a powerful methodology, integrates Decision Trees to forecast physical activity intensity. Utilizing a hierarchical structure, Decision Trees iteratively render decisions based on attributes in the 'ActivityMetrics' dataset. The approach, leveraging interpretability, reveals determinants impacting predictions, offering a greater understanding of the reasoning mechanism. Decision Trees, while remarkable, may experience overfitting, addressed by hyperparameter pruning. The model predicts anxiety levels logically and dependably.

6 Evaluation and Results

This study conducted multiple experiments to achieve an accurate regression analysis for forecasting anxiety rates, incorporating both deep learning and supervised machine learning models. The experiments involved a combination of traditional supervised machine

learning models, such as Decision Tree, Random Forest, Linear Regression, and Gradient Boost, alongside deep learning models. The deep learning models comprised a basic Neural Network and an LSTM forecasting model. Each experiment aimed to optimize model accuracy and effectiveness in predicting anxiety rates, offering a comprehensive exploration of both traditional and advanced techniques in regression analysis.

6.1 Experiment 1: Exploratory Data Analysis

The objective of this experiment is to analyse patterns in anxiety data, including heart rate, calories expended, intensity levels, steps taken. The primary focus is to understand the impact of anxiety on calorie expenditure. After conducting an analysis, it was observed that the total daily calories burned in the month of May 2016 ranged from a highest of 25,000 Kcal to a lowest of 5,000 Kcal. This study provides valuable insights for decision-making. (Refer to Figure 5)

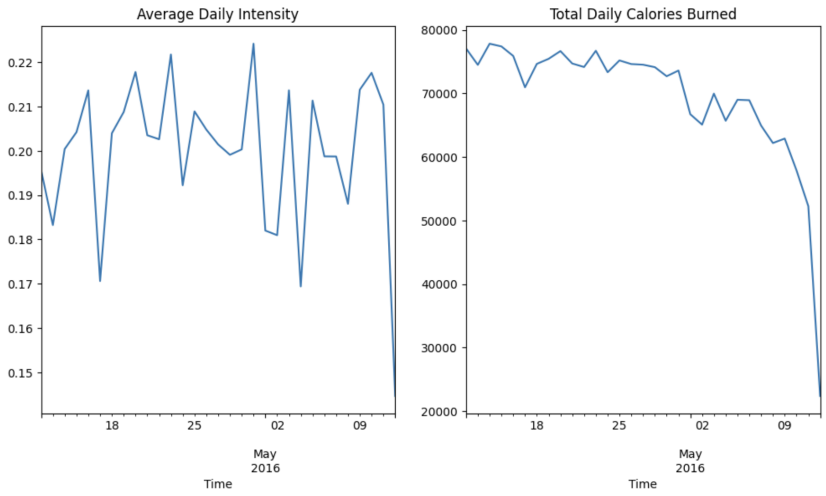

Figure 5: Analysis of Average Daily Intensity and Calories Burned

Similarly, from Figure 5 it is the graph illustrates the average daily intensity over time in May 2016. On the x-axis, the dates are represented as "18," "25," "02," and "09," likely indicating specific days of the month. The y-axis spans from 0.275 to 0.425, depicting the range of average daily intensity values. The upward or downward trends in the graph reveal how the average daily intensity levels fluctuate across the specified dates. It provides a visual representation of the variations in intensity throughout May 2016, allowing for insights into potential patterns or changes in anxiety levels during that period. The specific values on the y-axis indicate the average intensity level at different points in time, providing a quantitative measure of activity intensity.
visualize the daily trends of METs, steps, intensity, calories, and heart rate in a clear and organized manner, making it easier to identify patterns or fluctuations in these variables over the specified time period. specific colours to each variable. Now, METs will be represented in blue, steps in orange, intensity in green, calories in red, and heart rate in purple in Figure 6. The resulting Figure 7 The x-axis represents the actual calorie values (y_test), and the y-axis represents the predicted calorie values (y_pred). Each point on the plot corresponds to a data point from the test set. The transparency of points is set

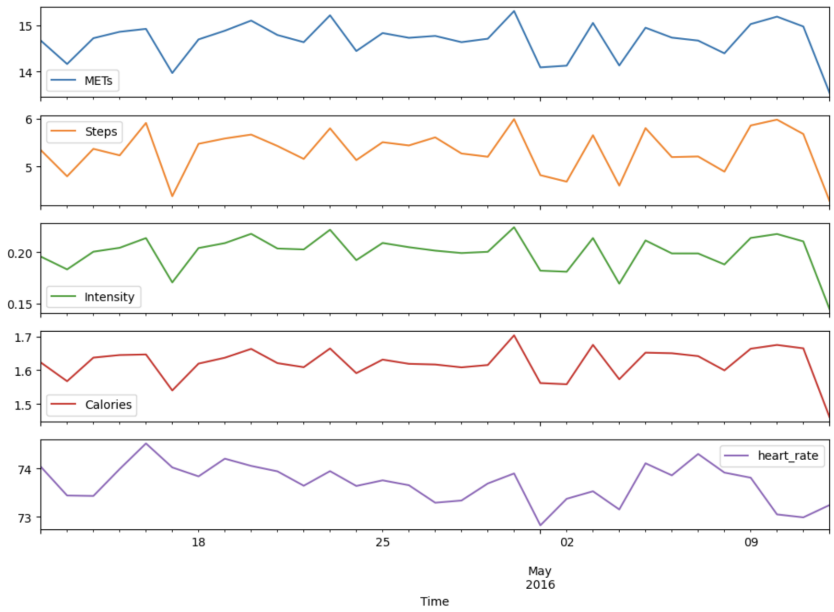

Figure 6: Time complexity with respect to the features

to 0.5 (alpha=0.5) for better visibility. Similar to the first subplot, the second one is for the Decision Tree Regressor. It compares actual versus predicted calories using a scatter plot Random Forest Regressorin the third subplot, showing how well it predicts calorie values compared to the actual values. the fourth subplot is for the Gradient Boosting Regressor, illustrating the relationship between actual and predicted calorie values. each subplot represents a different regression model, and the scatter plots within each subplot visually compare the actual and predicted calorie values. The degree of scatter and the pattern of points provide insights into how well each model performs in predicting calorie values based on the given features.

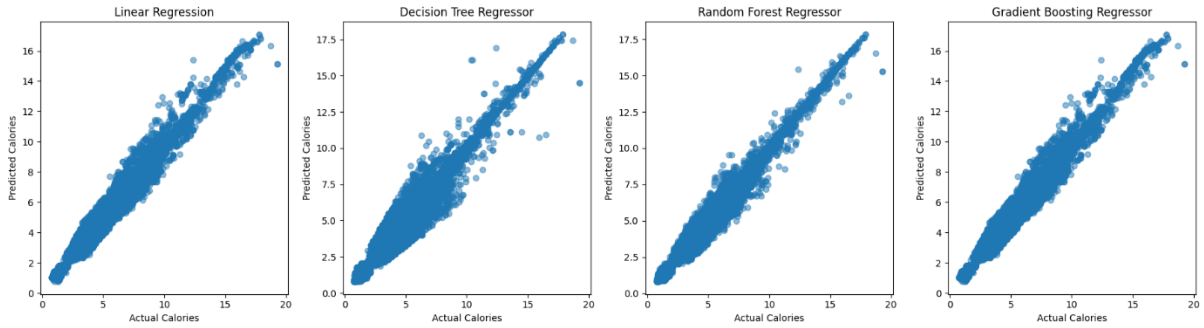

Figure 7: Actual Calories vs Predicted Calories

6.2 Evaluation and comparison of ML and DL

outlines the process of predicting anxiety levels in a given dataset using machine learning models. Two different models were employed: a Long Short-Term Memory (LSTM) network and a Feedforward Neural Network. The dataset contains features such as METs, steps, intensity, calories, and heart rate. The dataset comprises information related to physical activity, including METs, steps, intensity, calories, and heart rate. The primary

objective is to predict anxiety levels based on these features. The results of the training processes for both models are visualized above. The training and validation loss plots provide insights into the convergence and performance of each model. Further analysis and evaluation metrics such as Mean Squared Error (MSE) and R-squared can be derived from the predictions to assess the models' accuracy in predicting anxiety levels. Training

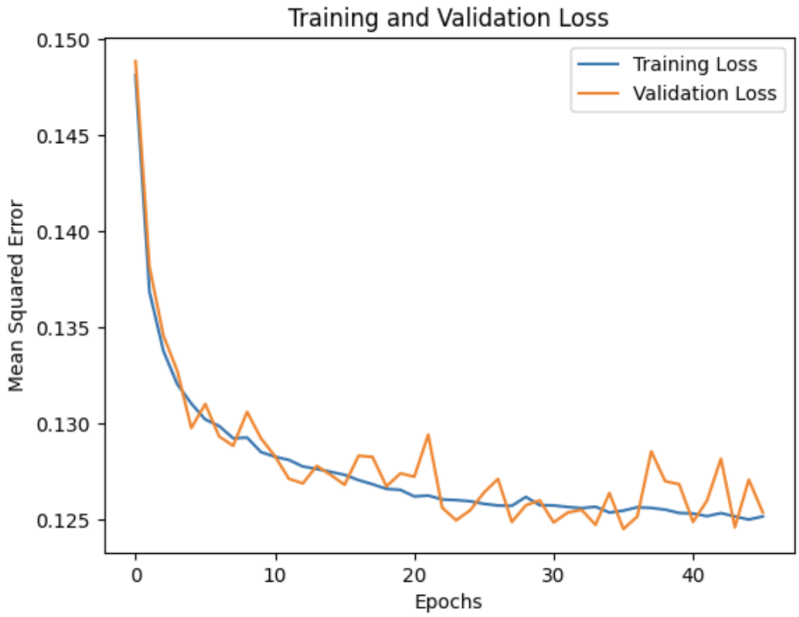

Figure 8: Training and Validation Graph for Feed Forward Network

loss in neural networks, whether in conventional or specialized designs such as Long Short-Term Memory (LSTM) networks, Feed Forward Network(FFN), is an important measure that reflects the model's capacity to learn and adapt during the training phase.. It is quantified by a loss function that calculates the disparity between the predicted outputs and the actual values in the training dataset. The optimization process aims to minimize this training loss by iteratively adjusting the model's internal parameters. Monitoring the training loss over epochs serves as a diagnostic tool, offering insights into the model's convergence. A decreasing training loss suggests that the model is effectively capturing the underlying patterns in the data. However, care must be taken to avoid overfitting, where the model becomes excessively specialized to the training data and struggles to generalize to new, unseen data. Conversely, a persistently high training loss may indicate underfitting, signalling that the model is not adequately learning the complexities of the data.In Figure 8, Over the training epochs, the MSE of the FFN model decreased, showing that the model continuously improved its performance. The training loss was closely followed by the validation loss, indicating that the model did not overfit the training data. The convergence of the training and validation loss curves indicates that the model is able to generalize effectively to previously unknown data, which is essential for making accurate predictions.

In the realm of LSTM networks Figure 9, which excel in handling sequential data and capturing long-term dependencies, the training loss extends its significance to the temporal dynamics of the input sequences. Backpropagation Through Time (BPTT) is a key component, allowing LSTMs to learn from past and future time steps. These networks are particularly useful in tasks where understanding the order of the data is crucial, such as

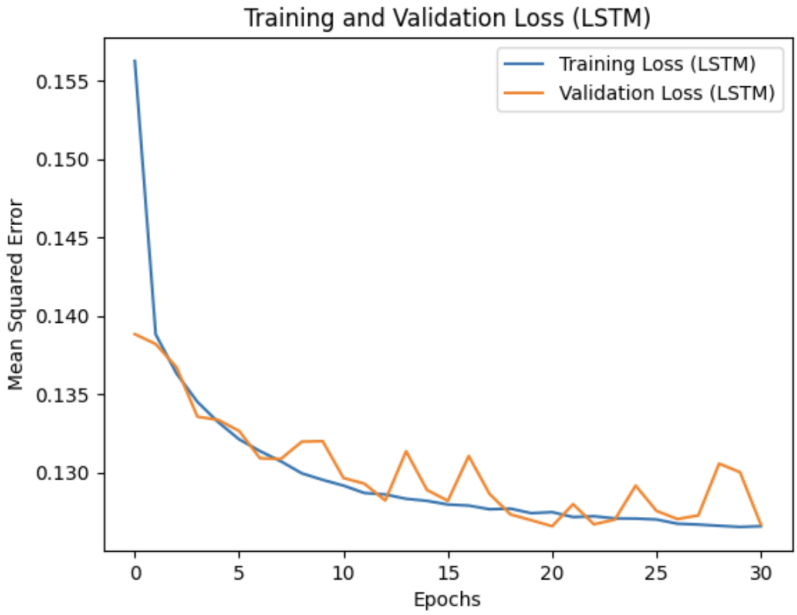

Figure 9: Training and Validation Graph for LSTM

time series prediction or natural language processing. The training loss in LSTMs reflects their ability to remember and utilize information across extended sequences, addressing challenges like the vanishing gradient problem associated with traditional recurrent neural networks. Monitoring training loss in LSTMs is crucial for assessing their proficiency in handling temporal dependencies and ensuring effective learning throughout sequential data.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

$$R^2 = 1 - \frac{RSS}{TSS} \tag{2}$$

From equation (1) MSE (Mean Squared Error): Calculates the average of squared differences between predicted and actual values. From equation (2) R-squared (Coefficient of Determination): Represents the proportion of predictable variance in the dependent variable. To determine which model has provided the best results, we can analyze the Mean Squared Error (MSE) and R-squared values from both tables. In general, lower MSE and higher R-squared values indicate better performance in terms of prediction accuracy and goodness of fit.

Table 1: Comparative Analysis of Mean Squared Error and R-squared

	Linear Regression	Decision Tree	Random Forest	Gradient Boosting
Mean-squared Error	0.152	0.127	0.121	0.134
R-squared	0.916	0.929	0.933	0.925

Table 1, compares various models, including Gradient Boosting, Random Forest, and Decision Tree, regarding their Mean Squared Error and R-squared values. Random Forest exhibits superior prediction accuracy and model fit compared to alternative methods, as

evidenced by its minimal mean-squared error of 0.121 and maximum R-squared value of 0.933.

Figure 10, clearly depicts the effect of numerous characteristics on the Random Forest Regressor's predictions. Notably, METs appear as the major element, indicating its significant effect on the model's prediction accuracy.

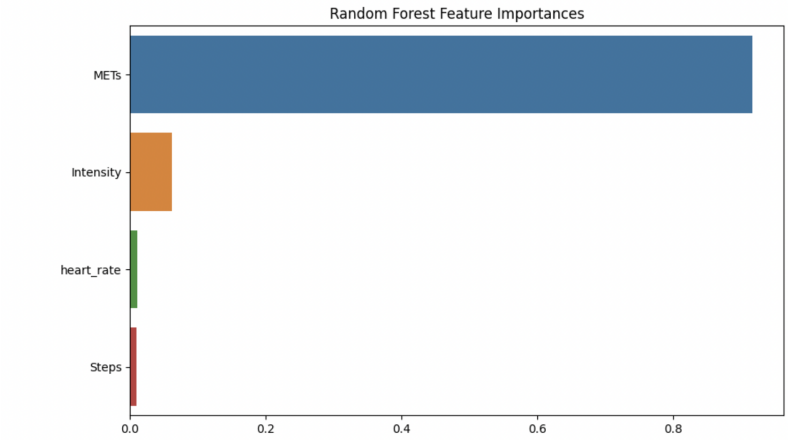

Figure 10: Random Forest Feature Importance

Therefore, based on the consistently lower MSE values across all features, Ramdom Forest can be considered the best-fit model among the machine learning models compared in the table. The results suggest that Ramdom Forest is particularly effective in capturing the complex relationships within the data and providing more accurate predictions of anxiety levels based on the given features. It showcases superior performance compared to Decision Tree, Gradian boosting, and Linear Regression in the context of this study.

Table 2: Comparative Analysis of Mean Squared Error and R-squared for Deep Learning models

	Feedforward Neural Network	LSTM
Mean-squared Error	0.124	0.126
R-squared	0.931	0.930

Based on the results on Table 2, both the deep learning models indicating similar results in terms of MSE and R squared values, demonstrates their equivalent efficiency in capturing and predicting the variance within the dataset.

6.3 Machine Learning vs. Deep Learning

After evaluating the models, it appears that the Random Forest technique is the efficient among all machine learning models. It has the lowest Mean Squared Error and highest R squared value indicating predictive accuracy. The Decision Tree model also performs well although it has higher error metrics and lower R squared value. It can be a choice if interpretability and simplicity are considered. Despite error metrics the Gradient Boosting model still shows a decent level of predictive effectiveness.
Similarly for deep learning, both the models performed efficiently with almost similar R

square value and Mean Square Error. In comparing Machine Learning to Deep Learning, both exhibit high accuracy. When it comes to interpretability, machine learning models generally offer more transparency compared to deep learning models like neural networks. In terms of complexity, Deep Learning models, such as Feedforward Neural Network and Long Short-Term Memory (LSTM), excel at capturing intricate patterns and relationships within the data, making them well-suited for complex tasks. Considerations for model selection include task complexity, where deep learning models may have an advantage in capturing complex patterns and relationships. If interpretability is crucial, machine learning models might be preferred. Additionally, the size of the dataset is a crucial factor; deep learning models often require large amounts of data to perform optimally, while certain machine learning models can be effective even with smaller datasets. The observed consistency in neural network accuracy across numerous iterations shows that the neural network model is solid and stable under varying conditions. Because neural networks are very adaptable and capable of learning complicated patterns, they may display consistent accuracy when given with varying datasets or variances in input characteristics.

The decline in accuracy found in machine learning models on numerous iterations on the other hands, might be due to their inherent limitations in managing complicated connections and capturing non-linear patterns in data.

7 Conclusion and Future Work

In the prediction of anxiety levels using various machine learning models and deep learning modles, deep learning models emerged as the most promising performer, exhibiting almost similar mean-squared error and R-squared value compared to the traditional supervised machine learning models. Other models such as Decision Tree, Random Forest, Linear Regression, and Gradient Boosting, demonstrated varying degrees of success in predicting different features related to anxiety, such as MET, heart rate, calories, intensity, steps, and time range. The results suggest that the deep learning models with its consistency and ability to capture intricate patterns and relationships within the data, is well-suited for modelling the complexities associated with anxiety prediction. It showcases a robust performance across multiple metrics, indicating its potential for accurate estimations of anxiety levels based on the provided features.

To enhance the predictive capabilities of anxiety models, future research can explore several avenues. Feature engineering should be further investigated to identify and incorporate additional relevant features, fostering a deeper understanding of the nuanced relationships between input features and anxiety levels for improved model performance. Ensemble approaches, involving the combination of predictions from multiple models, present an opportunity to leverage the strengths of diverse algorithms, enhancing the overall robustness and accuracy of predictions by mitigating individual model biases. Continued research and refinement of predictive models, coupled with a comprehensive understanding of contextual factors influencing anxiety, will contribute to the development of robust and ethically sound tools for anxiety prediction and management.

References

Ahmed, A., Aziz, S., Toro, C. T., Alzubaidi, M., Irshaidat, S., Serhan, H. A., Abd-alrazaq, A. A. and Househ, M. (2022). Machine learning models to detect anxiety and depression through social media: A scoping review, *Computer Methods and Programs in Biomedicine Update* **2**: 100066.

Albagmi, F. M., Alansari, A., Al Shawan, D. S., AlNujaidi, H. Y. and Olatunji, S. O. (2022). Prediction of generalized anxiety levels during the covid-19 pandemic: A machine learning-based modeling approach, *Informatics in Medicine Unlocked* **28**: 100854.

Bobade, P. and Vani, M. (2020). Stress detection with machine learning and deep learning using multimodal physiological data, *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*.

Dixit, S., Gaikwad, A., Vyas, V., Shindikar, M. and Kamble, K. (2022). United neurological study of disorders: Alzheimer's disease, parkinson's disease detection, anxiety detection, and stress detection using various machine learning algorithms, *2022 International Conference on Signal and Information Processing (IConSIP)*.

Espinola, W. C., Gomes, J. and Silva Pereira, M. J. (2022). Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study, *Research in Biomedical Engineering* **38**: 813–829.

Goswami, S., Pal, S., Goldsworthy, S. and Basu, T. (2019). An effective machine learning framework for data elements extraction from the literature of anxiety outcome measures to build systematic review, *in* W. Abramowicz and R. Corchuelo (eds), *Business Information Systems*, Vol. 353 of *Lecture Notes in Business Information Processing*, Springer, Cham.

Jacobson, N. C., Lekkas, D., Huang, R. and Thomas, N. (2021). Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years, *Journal of Affective Disorders* **282**: 104–111.

Khan, N. S., Ghani, M. S. and Anjum, G. (2021). Adam-sense: Anxiety-displaying activities recognition by motion sensors, *Pervasive and Mobile Computing* **78**: 101485.

Pandit, M., Azwaan, M., Wani, S., Rawad, A. A. I., Abdulghafor, A. A. and Gulzar, Y. (2023). Examining factors for anxiety and depression prediction, *International Journal on Perceptive and Cognitive Computing* **9**(1): ARTICLE PAGES. Articles.

Rajput, G. G. and Alashetty, A. (2022). A machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique, *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*.

Singh, A. and Kumar, D. (2021). Identification of anxiety and depression using dass-21 questionnaire and machine learning, *First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT)*.

Thakre, T. P., Kulkarni, H., Adams, K. S., Mischel, R., Hayes, R. and Pandurangi, A. (2022). Polysomnographic identification of anxiety and depression using deep learning, *Journal of Psychiatric Research* **150**: 54–63.