International Journal on Robotics, Automation and Sciences

Mental Health Problems Prediction Using Machine Learning Techniques

Jia-Pao Cheng, Su-Cheng Haw*

***Abstract* - Mental health problems encompass a range of conditions that can impact an individual's emotions and behaviors. The conventional methods of mental illness prediction often suffer from the issue of either over-detection or under-detection and the time-consuming manual review process of patients' data during screening sessions. Therefore, this research aims to utilize machine learning techniques to predict mental health problems, complementing the traditional clinical screening and diagnosis process. The proposed models in this project: Logistic Regression, K-Nearest Neighbors, and Random Forest leverage relevant factors from the dataset concerning mental health survey published by Open Source Mental Disorders in 2014 to predict mental health problems. Feature selection and hyperparameter fine-tuning are employed to identify the factors contributing to mental health problems and enhance the performance of the models. The evaluation of these models is measured using accuracy, recall, precision, F1 score, and AUROC. Experimental evaluation results indicated that the Random Forest model utilizing hyperparameters derived from the RandomizedSearchCV method outperforms during model selection using cross-validation. When predicting test set data, it exhibits a good generalization with an accuracy of 83.23%, recall of 89.87%, precision of 78.02%, F1 score of 83.53%, and AUROC of 83.57%.**

Keywords*—*Mental Health Problems, Logistic Regression, K-Nearest Neighbors, Random Forest

I. INTRODUCTION

Mental health problems are conditions that may affect a person's feelings and behavior, resulting in psychosocial impairments or loss of ability which require treatment [1]. The World Health Organization reported that in 2019, among the population of 970 million people, there was 1 in every 8 people suffered from a mental disorder [2]. One in three Malaysians, according to the National Health and Morbidity Survey conducted by the Ministry of Health [3], struggles with mental health issues, with the highest prevalence in teenagers aged 16 to 19 and from low-income families. In truth, there are numerous mental illnesses for which the majority of individuals do not have access to the necessary care, including anxiety, depression, post-traumatic stress disorder, and bipolar disorder due to reasons such as lack of knowledge, ignorance in treatment access, and more [4]. A person's social interactions and daily activities might be disrupted by mental health issues, which can also lead to poor work performance. In the worst-case scenario, self-harm and suicide may also emerge [5].

To emphasize, mental health diagnosis is a complex process that is not straightforward. Traditional methods of predicting mental health entail a series of clinical screenings that include a face-to-face interview between a patient and a human doctor, completing

*Corresponding author, Email: sucheng@mmu.edu.my

Su-Cheng Haw is a professor under Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia 63100, Cyberjaya, Malaysia

Jia-Pao Cheng is a student under Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia 63100, Cyberjaya, Malaysia

(email: sucheng@mmu.edu.my, 1191101533@student.mmu.edu.my)

International Journal on Robotics, Automation and Sciences (2023) 5,2:59-72
https://doi.org/10.33093/ijoras.2023.5.2.7
Manuscript received: 13 July 2023 | Accepted: 2 August 2023 | Published: 30 September 2023

Published by MMU PRESS. URL: http://journals.mmupress.com/ijoras

questionnaires, and taking psychological tests. The process is prone to misdiagnosis, especially with a higher number of false positive cases [6]. Moreover, manually reviewing patient data for decision-making is ineffective and time-consuming, leading to a delay in early diagnosis in which early intervention can bring positive impacts on mental health [7]. Costly consultation and lengthy screening tests can also burden patients with packed schedules or limited financial resources, leading to the neglect of mental illness treatment. Therefore, this research strives to develop Machine Learning models using Logistic Regression, K-Nearest Neighbors, and Random Forest to predict whether an individual has mental health problems.

II. Background and Related Work

A. Overview of Machine Learning

Making good use of data in the healthcare sector contributes to pattern detection of an illness, timely diagnosis, real-time patient monitoring, and more treatment enhancement [8]. Greater contributions and higher work performance are to be brought through the implementation of Machine Learning. To understand what is Machine Learning, one must first understand that Machine Learning is a branch of Artificial Intelligence that emphasizes data and algorithms to simulate a human's learning process and gradually improve accuracy. Machine Learning is divided into four categories: supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning. Each of these categories is widely used in specific aspects.

Supervised learning is an approach where a dataset is trained with labels to solve classification problems or prediction of future events. When a new dataset without any labels is fed to the machine, it classifies the new data or makes predictions based on what it has learnt during the training process. This approach is classified further into classification and regression. Classification is to categorize new data into specific classes, which a class is also known as a label or a target. The output of a classification is a category (e.g., red or blue, spoon or fork, cat or dog) rather than a value which is the output of a regression problem. It is useful in detecting email spam, recognizing speech, identifying cancer cells, and so on.

Unsupervised learning is the opposite of supervised learning where zero human-intervene is involved. The machine is trained with unlabeled data to discover underlying patterns without human supervision. It is categorized into clustering, association, and dimensionality reduction. Clustering aims to group data points based on similarities, and it is normally used to identify products that should be placed closely for better organization. On the other hand, association is using rules to calculate the dependency of one data on another, and the association rule is widely used to find a group of products that customers buy together. Meanwhile, dimensionality reduction is to shrink high dimensions datasets to a lesser dimension while maintaining data integrity.

Reinforcement learning is the learning of a machine from trial and error through the exploration in an environment. Desired behavior is rewarded with positive value while negative action is punished by assigning negative value as feedback until reaching the maximum reward. This method is widely used in robotics and gaming.

Semi-supervised learning is the combination of supervised and unsupervised learning that is introduced to solve the downside of supervised and unsupervised learning where it is time costly to hand-labeled data; while unsupervised learning is limited in terms of application spectrum. In this method, a limited amount of labeled data is used to train the machine with a vast amount of unlabeled data. At first, similar data is clustered with unsupervised learning, then existing labeled data is used to label the remaining unlabeled data. It is widely used in speech analysis and web content classification. Table 1 summarizes the four categories of Machine Learning algorithms.

TABLE I. Summary of Machine Learning Categories

	Characteristic	***Advantage***	***Disadvantage***
Supervised learning	Algorithm learns attributes from the labeled dataset	Ease of training	Overstrain might occur if the training dataset does not have an example of a desired class
Unsupervised learning	Algorithm learns by its own observations without human intervention	Useful in finding hidden patterns in data and no need for hand-labeling of data	The obtained results are not always useful since there is no label to confirm their accuracy
Reinforcement Learning	Algorithm is given a goal and rewards and punishment are defined then the algorithm is self-directed to reach the endgame	Similar to the learning behavior of human beings in which a perfect model could be developed to achieve long-term results	Data-hungry so it is highly reliance on the exploration of the environment
Semi-supervised Learning	Algorithm learns from labeled dataset and extrapolates to the unlabeled data	Overcome the drawback of supervised and unsupervised learning	Underperforms if the labeled data is not indicating the whole data distribution

In short, supervised learning requires human supervision and is easy to train but suffers from overstrain if lack of examples of desired class. Unsupervised learning does not involve human intervention but the acquired results might not be useful due to the absence of labels to ensure accuracy. On the other hand, reinforcement learning is similar to human learning behavior but highly relies on exploration of the environment. Semi-supervised Learning overcomes the

downside of supervised and unsupervised learning but needs a sufficient combination of labeled data.

B. Related Work

The study [9] explored the data from the 2019 Open Source Mental Disorders (OSMI) mental health survey, which included information on employees from both technology and non-technology organizations, to identify seven out of 70 qualities as those that most contribute to mental health illness. Machine Learning algorithms such as Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Decision Tree, and Random Forest were applied to the chosen features. Decision Tree showed the highest accuracy and precision. Furthermore, it has been discovered that a person's mental health history and family background were the two important factors that influence one's mental health the most.

In [10], the author studied eight approaches: Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, Extreme Gradient Boosting (XGB), Gradient Boost Machine, and Artificial Neural Network to predict whether an individual suffered from depression. Sensitivity, specificity, precision, Negative Predictive Value (NPV), F1 Score, False Negative Rate (FNR), False Positive Rate (FPR), False Discovery Rate (FDR), False Omission Rate (FOR), and accuracy are the evaluation metrics utilized. The finding showed that Naive Bayes underperformed with the lowest accuracy of 21.67% while Support Vector Machine showcased a high accuracy of 7.38%.

The author from [11] implemented eleven approaches, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, Adaptive Boost, SDG Classifier, Gradient Boosting, XGB, and Light Gradient Boosting to examine the external factors influencing an employee's mental health using the data from the OSMI mental health survey. To emphasize, Light Gradient Boosting is a gradient boost methodology based on Decision Tree that splits the data by leaves rather than by tree depth, producing improved accuracy with shorter training times. The finding showed that the XGB model surpassed the other Machine Learning models in terms of overall accuracy, precision, recall, and F1 score.

The author of [12] utilized Decision Tree, Support Vector Machine, Neural Network, Naive Bayes, and Logistic Regression using SPSS Modeler to discover the determinants in mental health problems among students in higher education institutions and classify the individuals into distinct categories such as stress, depression, and anxiety. Decision Tree outperformed all other models in terms of stress prediction. While Logistic Regression had the highest accuracy, specificity, and precision in anxiety prediction, it unexpectedly showcased a low sensitivity value that is almost 20% lower than Neural Network (the best sensitivity performer) .Support Vector Machine led the other models in depression prediction with an accuracy of 88.15%. The study showed that a lack of social support, financial difficulties, and the learning environment contributed to mental health problems.

The study [13] highlighted the prediction of post-traumatic stress disorder (PTSD), by examining whether supervised Machine Learning algorithms could identify associations between the occurrence of PTSD symptoms in a patient and at one month after the trauma event. In this study, classifiers such as Logistic Regression, Naive Bayes, and Random Forest were employed. Support Vector Machine ensembled with linear, Gaussian, and polynomial kernels were also employed to increase prediction robustness. There were two methods of voting: hard voting in which the winning class label was the class with the most votes, and soft voting in which the class with the highest probability was chosen by summing up the probabilities predicted by each classifier for each class label. With an Operating Characteristics Curve (ROC) of 0.8465576, Support Vector Machine with a Gaussian kernel transcended the other classifiers.

In study [14], the Average One Dependence Estimator, Multilayer Perceptron, Radial Basis Function network, Instance-based Learning IB1, K-star algorithm, Multi-class Classifier, Functional Trees, and Logical Analysis of Data (LAD) tree were used to predict mental health issues in children. The authors spoke with a psychologist to determine the challenges encountered during mental health diagnosis to better comprehend the clinical flow in mental health disorder diagnosis. Measures of the accuracy of the classifiers include Kappa statistics, accuracy, and area under the receiver operating characteristics curve (AUROC). It is observed that Multilayer Perceptron, Multiclass Classifier, and LAD Tree exhibited greater accuracy and Kappa values.

The study [15] aimed to investigate the impact of the Coronavirus disease (COVID-19) lockdown on various aspects of young Indian students' lives, including their social life, general mood, and thoughts about the lockdown. Association rules were learned and visualized using the Apriori algorithm in R, and the analysis was based on metrics such as lift, confidence, and support. The dataset for the study was collected through an online survey distributed via email and WhatsApp. The findings indicated that 37.9% of students felt calm and hopeful during the lockdown, while 42% of students felt dissatisfied, anxious, and despondent.

The researchers proposed the Improved Global Chaos Bat Back Propagation Neural Network (IGCBA-BPNN) as a prediction model for detecting mental health issues among medical workers during COVID-19 [16]. The model combined an optimization algorithm and a neural network, incorporating Stepwise Logistic Regression, Binary Bat Algorithm, and a hybrid improved dragonfly algorithm. The Global Chaos Bat Algorithm (IGCBA) was introduced to address the limitations of the Bat Algorithm and was further improved to create the IGCBA with better performance. The IGCBA-BPNN optimized the feature variables to enhance prediction accuracy. Experimental results showed that using algorithms in conjunction with BPNN improved prediction

accuracy by an average of 2.46%. The IGCBA-BPNN-4 model demonstrated the best overall performance with a prediction accuracy of 92.55% and reduced redundant characteristics.

The study [17] predicted an individual's treatment response to a digital mental health intervention for treating depression and anxiety. Using clinical characteristics such as past suicide attempts, trauma history, medication use, etc., the capability of the algorithms such as Logistic Regression, Support Vector Machine, Naive Bayes, and Random Forest in predicting reductions in the symptoms of depression and anxiety were measured. It is observed that Random Forest excels in cross-validation with an AUROC of 0.64.

The study [18] utilized six Machine Learning algorithms including Boruta Random Forest, Lasso regression, Elastic-net regression, Bayesian Additive Regression Trees (BART), and Logistic Regression to predict self-harm in young people within six months. Evaluation metrics such as Brier scores, sensitivity, specificity, positive predictive value (PPV), NPV, Area under the Precision-Recall Curve (AUPRC), and AUROC are employed. Important predictors include history of self-harm, age, social and occupational functioning, sex, bipolar disorder, psychosis-like experiences, treatment with antipsychotics, and history of suicide ideation. It is observed that the Boruta Random Forest model demonstrated the lowest Brier scores and highest AUPRC, PPV, and specificity. BART achieved the highest mean AUROC, while Lasso regression models had the highest mean NPV and sensitivity.

The related works reviewed are summarized in a high-level manner in Table II.

TABLE II. SUMMARY OF RELATED WORKS

Reference	*Evaluation Matrix*	*Finding*
[9]	Accuracy, Precision, Recall, F1 Score	Multiple models were used to determine the contribution of personal and professional factors to mental health problems. Decision Tree demonstrated the highest accuracy and precision. The most influential factors were personal and family history of mental health problems.
[10]	Sensitivity, specificity, precision, NPV, F1 Score, FNR, FPR, FDR, FOR	Eight models were utilized to predict depression using attributes such as earning and spending patterns, household conditions, and family members. The Support Vector Machine model outperformed all others, achieving an accuracy of 87.38%, while Naive Bayes had the lowest accuracy at 21.67%.
[11]	Accuracy, Precision, Recall, F1 Score	The causes of mental illness in the workplace are studied by analyzing external factors such as company location, company size, leaves provided, and company wellness programs. The XGB model performed the best across all evaluation metrics.
[12]	Accuracy, Sensitivity, Specificity, Precision	The study aimed to classify higher education students with mental health problems into categories of stress, depression, and anxiety. Decision Tree performed best for stress, Support Vector Machine for depression, and Neural Network for anxiety. The contributing factors included a lack of social support, financial difficulties, and the learning environment.
[13]	Accuracy, ROC, Confusion Matrix	Three Machine Learning algorithms and Support Vector Machine ensembled with linear, Gaussian, and polynomial kernels are deployed to find correlations between the occurrence of PTSD symptoms in a PTSD patient after trauma within a month. With a ROC of 0.8465576, Support Vector Machine with a Gaussian kernel outperformed the other eight classifiers.
[14]	Kappa statistics, AUROC	Eight Machine Learning techniques are utilized. to predict mental health issues in children. The results proved that Multilayer Perceptron, Multiclass Classifier, and LAD Tree achieved more significant performance.
[15]	Support, Confidence, Lift	The study examined the impact of the COVID-19 lockdown on young Indian students. Association rules were generated using the Apriori algorithm in R. It revealed that 37.9% of students felt calm during the lockdown, while 42% felt anxious.
[16]	Accuracy	The IGCBA-BPNN model was proposed to predict mental health issues among medical workers during COVID-19. The finding showed that the best-performing model, IGCBA-BPNN-4, achieved a prediction accuracy of 92.55% and reduced redundant characteristics.
[17]	ROC	The study predicted an individual's treatment response such as past suicide attempts, trauma history, and medication use to a digital mental health intervention in order to treat depression and anxiety. It is observed that Random Forest excelled at cross-validation with an AUROC of 0.64.
[18]	Brier scores, sensitivity, specificity, PPV, NPV, net benefit, AUPRC, AUROC	The study aimed to predict self-harm in young individuals. The Boruta Random Forest model showed the lowest Brier scores and highest AUPRC, PPV, and specificity. BART had the highest mean AUROC, and Lasso regression models had the highest mean NPV and sensitivity. Important predictors included history of self-harm, psychosis-like experiences, treatment with antipsychotics, history of suicide ideation, etc.

III. MACHINE LEARNING TECHNIQUES USED

As mentioned in Section I, Machine Learning methods are categorized into supervised Machine Learning, unsupervised Machine Learning, reinforcement learning, and semi-supervised learning. Supervised Machine Learning is primarily used for classification and prediction modeling using structured training datasets while unsupervised Machine Learning involves data handling without supervision. In this study, only supervised learning algorithms are implemented as the dataset utilized is labeled.

A. Logistic Regression (LR)

The fundamental principle underlying Logistic Regression is to build a model that can estimate the probability of a binary outcome or a categorical value by leveraging the relationship between a dependent variable and one or more independent variables. As a result, the value of the dependent variable falls between 0 and 1 since Logistic Regression predicts the result in probability. The logistic function, commonly referred to as the sigmoid function, is the equation that the logistic regression model applies. The sigmoid function is an S-shaped curve, as illustrated in Figure 1, that converts any real number to a value between 0 and 1, which can then be used to predict the class of the dependent variable as a probability. The sigmoid function can be represented as

$$p = \frac{1}{1+e^{-(b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n)}} \quad (1)$$

Where p is the probability that the dependent variable belongs to a particular class; e is the natural logarithm base; and b0, b1, b2, ..., bn are the beta parameters, or regression coefficients of the independent variables x1, x2, x3, ..., xn. The most common method for estimating the beta parameter is by evaluating many beta values to find the one that best fits log odds [19]. Once the optimal beta parameter is determined, the conditional probabilities can be calculated to generate a predicted probability. The sigmoid function produces outputs in the range of 0 to 1, with the midpoint serving as a threshold to distinguish between class 1 and class 0. An input that results in an outcome greater than 0.5 in a binary classification is regarded to belong to class 1. In contrast, the corresponding input is categorized as falling into class 0 if the output is smaller than 0.5 [20].

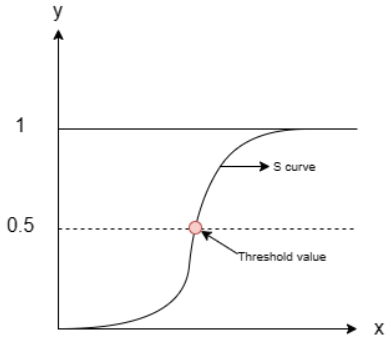

Figure 1. Sigmoid Function

There are three types of Logistic Regression models: binary, multinomial, and ordinal. The dependent variable we want to analyze is dichotomous, having only two output classes (yes or no) as it is employed in this project to predict whether or not an individual has mental health problems. Logistic Regression is chosen as the proposed model because it can manage non-linear correlations between predictors and dependent variables by applying a non-linear transformation of the predictors. Additionally, it is simple to reconstruct to avoid overfitting, making it more resilient to datasets with small sample sizes or high noise levels. It is also simple to comprehend, which makes it simple to interpret and put into practice.

B. K-Nearest Neighbors (KNN)

K-Nearest Neighbors uses proximity to predict or classify how a single data point will be categorized. Figure 2 shows that it locates the k-number of training data points closest to a new data point and then determines the class or value of that new data point based on the majority class or average value of those k-nearest neighbors.

K-Nearest Neighbors works in the subsequent steps:

1. Initialize the k value to choose the number of neighbors
2. Calculate the distance using a distance metric between the new observation and the training data
3. Decide the k-number of training observations that are closest to the new observation
4. Assign the mode of the K label with the majority vote to the new observation in the classification situation. The average value of its k-nearest neighbors is used to predict the regression problem
5. Repeat for each new observation

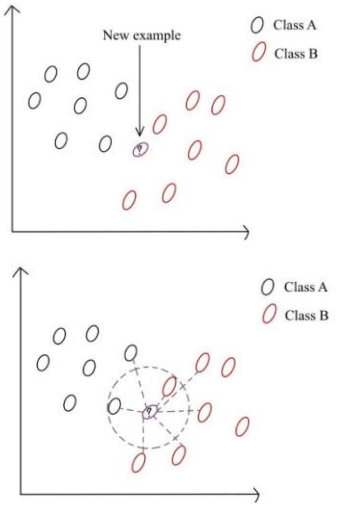

Figure 2. K-Nearest Neighbors

To emphasize, the k value determines the number of neighbors to establish classification; hence, the selection of k has a major influence on the performance of the K-Nearest Neighbors. With a lower k, the model will be more susceptible to data noise and more likely to base predictions on outliers and errors in the training data, leading to high variance and overfitting. A large k, on the other hand, will reduce the model's sensitivity to capturing the underlying patterns in the data, leading to underfitting. In other words, data with more noise or outliers will probably perform better when k is higher.

It is important to note that K-Nearest Neighbors is a lazy learning method that retains the training dataset in the memory rather than learning a distinction-making function from it. It then classifies new points using a similarity metric that compares the value to be classified with the remembered values. It is necessary to specify the distance before classification. User-defined distance metrics can be used to determine the nearest observations. The Euclidean Distance is a popular distance metric that determines the straight-line distance in n-dimensional space between any two points. It is frequently used when the data comprises continuous variables and is in a multi-dimensional space. It is also utilized in applications such as clustering to group similar data points together by measuring the distance between data points. The Euclidean distance can be obtained by calculating the square root of the sum of the squares of the coordinate difference between the two points, which is represented as

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

Manhattan Distance is the distance between two points without taking the diagonal or shortest route. It is commonly used in robotics as a navigation in an environment that resembles a grid. The Manhattan Distance is computed using the sum of the Cartesian coordinates' absolute differences of two points, as in

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

Minkowski Distance is the generalized version of the Manhattan distance and the Euclidean distance. The distance is referred to as the Manhattan distance if the p-value is set to 1 and the Euclidean distance if the p-value is set to 2. The equation is represented as

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \quad (4)$$

The Euclidean, Manhattan, and Minkowski distances work well for continuous and numerical data. Different distance metrics will be tested during hyperparameter fine-tuning to determine which one performs best because the choice of distance metric can significantly impact the K-Nearest Neighbors' performance. To summarize, KNN is simple to implement and is tolerant and resistant to noise present in the training dataset [21], thus, it is picked as one of the proposed models.

C. Random Forest (RF)

Random forest is an ensemble learning method that constructs numerous decision trees during training, which are widely used for classification and regression applications due to their simplicity and adaptability.

The Decision Tree is the building block of the random forest model. It is a tree structure that resembles a flowchart and is used to depict a set of decisions and their potential outcomes as illustrated in Figure 3. Starting with a single node, known as the root node, the Decision Tree is constructed by splitting it into two or more child nodes according to the value of a feature in the dataset. The splitting process repeats until a stopping standard is satisfied. Each leaf node in the tree represents a specific class, while each internal node represents a test on a specific feature. The test results are represented by each branch of the tree. Despite being simple to comprehend and interpret, Decision Tree is susceptible to overfitting when the sample size of the dataset is limited [22] and the trees are deep. However, when numerous Decision Trees are combined into an ensemble using the RF algorithm, they are capable of predicting outcomes more precisely, especially when the individual trees are uncorrelated with each other.

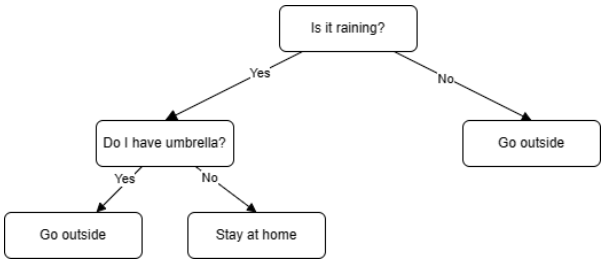

Figure 3. Decision Tree

The creation of a Random Forest begins with selecting a random data sample from the original dataset, commonly referred to as bootstrapping. With this bootstrapped sample of data as a foundation, a Decision Tree is then developed with a portion of features randomly picked at each split. A forest of Decision Trees is generated by repeatedly executing this strategy. The Random Forest casts a majority vote among the predictions of the different trees when predicting a new observation. In a classification problem, the predicted class is the one with the most votes, whereas, in a regression problem, the final prediction is the average of the predictions.

Random Forest comes with a plethora of hyperparameters such as the number of Decision Tree, the number of features, etc. It is important to note that the ideal values of these hyperparameters vary on the dataset, hence methods like fine-tuning and cross-validation will be used to determine the best hyperparameter values. To summarize, in contrast to single decision trees, Random Forests provide several advantages. They frequently have higher prediction accuracy and are less likely to overfit.

IV. RESEARCH METHODOLOGY

Figure 4 illustrates the steps of research methodology used in this project which include background research, problem statement and research objectives formulation, literature review, design of research methodology, data collection, data pre-processing, models design and implementation, models training and testing, model selection, model evaluation, model prediction, and finding discussion.

Firstly, a background study is conducted by studying the theory and practice concerning the topic. This step is then extended to formulating problem statements and

research objectives to identify the issue that is a concern. Then, the relevant researches are reviewed to identify the feasible ways in prior research and discover the respective limitations. Data pre-processing, which entails activities such as data quality assessment, data cleaning, discretization, and data encoding, is carried out to convert the raw data into a usable form after the dataset has been gathered. For example, irrelevant columns such as "Timestamp," "comments," and "state" are dropped from the dataset. The column "country" is also excluded to prevent any potential bias, as Figure 5 illustrates a majority of respondents originating from the United States.

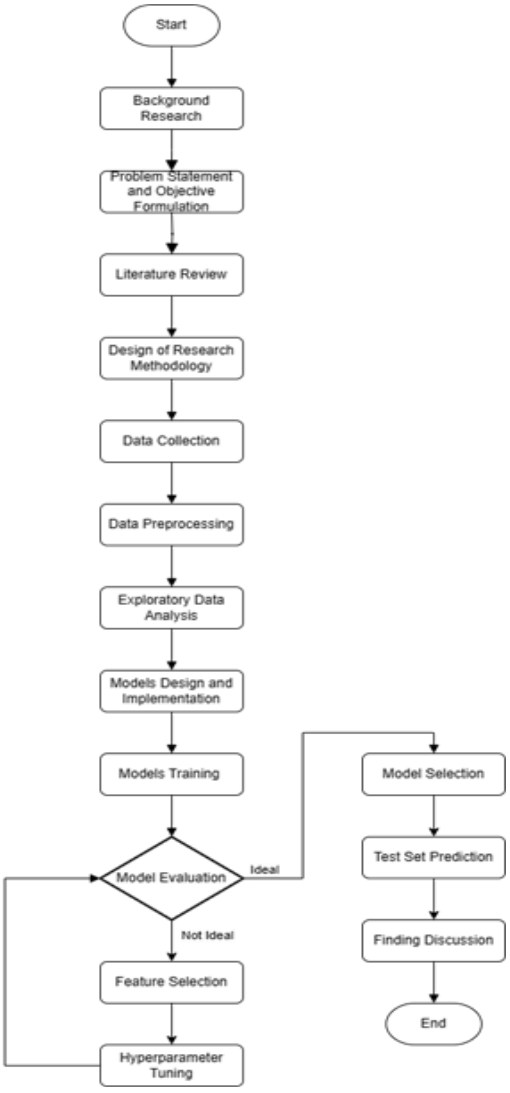

Figure 4. Research Methodology Flowchart

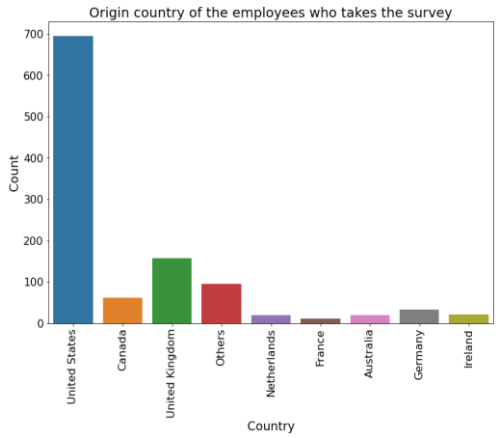

Figure 5. Origin Country of the Respondents

Furthermore, the meaningless data entries are corrected, particularly within the “gender” column, which contains nonsensical entries such as “'something kinda male?'. To address this, the data values are categorized into 'Female', 'Male', and 'Other'. Moreover, as depicted in Figure 6, the “age” column exhibits outliers such as -29 years old, -1726 years old, 5 years old, and 329 years old, which are illogical as human age cannot be negative and it is unlikely for a 5-year-old to be employed. The mentioned outliers in the “age” column are replaced with the median of the column. To avoid bias towards specific ages, the “age” column is discretized into age ranges such as 0-20, 21-30, 31-40, 41-50, 51-60, and 61-80 years old. Then, data encoding is performed using Label Encoding to convert categorical columns into numerical columns as most of the columns exhibit a certain level of sequencing in their data values such as “No”, “Not sure” and “Yes”.

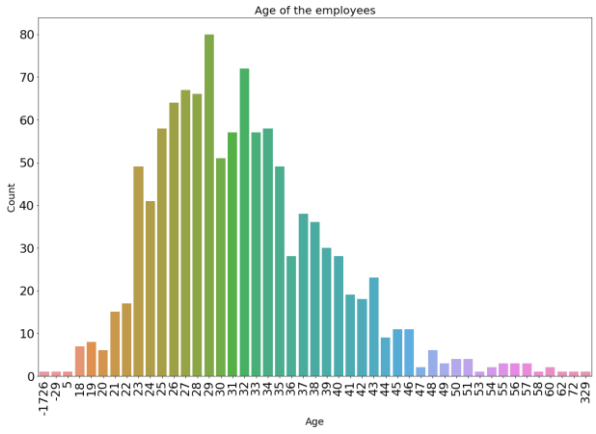

Figure 6. Age distribution of the Respondents

The dataset is then analyzed and investigated using data visualization as part of the exploratory data analysis process. Next, the dataset is divided into the training, validation, and test sets at 70%–15%–15%. The baseline model is trained and evaluated to analyze the baseline models' results as a reference point. In continuation, feature selection using Recursive Feature

Elimination, Cross-validated (RFECV) with GradientBoostingClassifier, and hyperparameter fine-tuning using RandomizedSearchCV and GridSearchCV are carried out. The best-performing model in cross-validation with 10-folds is chosen to perform prediction on the test set. Finally, the finding outcomes will then be discussed.

The configuration of the hyperparameters setting is represented in Table III. When applying the identical range of hyperparameters to fine-tune the Random Forest model using GridSearchCV, it is observed that the fine-tuning process takes more than two days to complete due to a large number of combinations and the exhaustive search nature of GridSearchCV, which explores all possible combinations of hyperparameters in the parameter grid. [23]. Hence, to reduce the lengthy duration, the range of hyperparameters implemented to fine-tune the Random Forest model using GridSearchCV is scaled down, as depicted in Table IV.

TABLE III. CONFIGURATION OF HYPERPARAMETER SETTING

Hyperparameter Setting (RandomizedSearchCV)		
Model	***Hyperparameter***	***Range of Value***
LR	solver	'liblinear'
	penalty	l2
	C	10
KNN	leaf_size	40
	n_neighbors	11
	p	1
	weights	'uniform'
	metric	'minkowski'
RF	n_estimators	282
	max_features	'auto'
	max_depth	7
	min_samples_split	6
	min_samples_leaf	2
	bootstrap	True
	criterion	'gini'
	oob_score	False
	max_leaf_nodes	9

TABLE IV. GRIDSEARCHCV HYPERPARAMMETERS' RANGE OF VALUE FOR RANDOM FOREST MODEL

Hyperparameter Setting (GridSearchCV)		
Model	***Hyperparameter***	***Range of Value***
RF	n_estimators	[100, 200, 300, 400]
	max_features	['auto', 'sqrt']
	max_depth	[3, 6, 9]
	min_samples_split	[3, 6, 9]
	min_samples_leaf	[3, 6, 9]
	bootstrap	[True, False]
	criterion	['gini', 'entropy']
	oob_score	[True, False]
	max_leaf_nodes	[3, 6, 9]

A. Dataset

The dataset that is utilized for this study [24] is a publicly available secondary dataset in CSV format, published on Kaggle, and originated from Open Sourcing Mental Health. It is collected from a survey answered by 1260 respondents working at companies from the digital technology sector to interpret the individuals' opinions on mental health and also study the number of occurrences of mental health problems at work. The dataset contains 1260 rows and 27 columns with assorted attributes related to the corresponding workplace. Table V shows the distribution of the dataset at a 0.15 test split ratio.

TABLE V. DISTRIBUTION OF TRAIN, VALIDATION, AND TEST SET

Data	***Number of rows of data***
Training	779
Validation	167
Testing	167
Overall	1113

B. Evaluation Metrics

Several popular evaluation metrics are employed in this study to assess the effectiveness of the proposed models, including the confusion matrix, precision, recall, accuracy, F1 score, and AUROC. The commonly used terms for evaluation metrics are as follows:

- True Positive (TP): correctly predicted the positive class
- False Positive (FP): predicted as a positive class but it is actually a negative class
- False Negative (FN): predicted as a negative class but it is actually a positive class
- True Negative (TN): correctly predicted the negative class

1) Confusion Matrix

Figure 7 illustrates the confusion matrix table. It measures the classification model's performance by showing possible outputs such as TP, FP, FN, and TN in a table. These are combinations of predicted values and actual values. Each value's rate can be calculated, and TP and TN should have the highest rate possible; meanwhile, the rate of FP and FN should be as low as possible. With these four values, other performance metrics such as precision, accuracy, recall, and accuracy, can then be calculated.

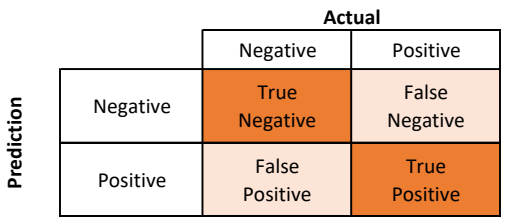

Figure 7. Confusion Matrix

2) Accuracy

The ratio of true predictions to all predictions is calculated using (5). It is acknowledged that accuracy is favorable for well-balanced classes and might not be appropriate for classes with uneven distribution. The predictable variable is balanced

distributed in the dataset; hence, accuracy acts as a reference point in this study.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Predictions}} \quad (5)$$

3) Recall

Recall is also known as sensitivity, indicating how accurate the positive predictions are compared to the ground truth. In other words, out of the total positives, what is the rate of predicted positives? In this case, out of all the people with mental health problems, how many get positive test results? To highlight, recall is considered more critical than precision in this research. Similar to healthcare applications such as cancer screening, prioritizing high recall is common practice to ensure early detection [25]. Hence, it is crucial not to miss any patients in mental health prediction. Equation (6) represents the calculation of recall.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (6)$$

4) Precision

Equation (7) measures the rate of how many of the predicted positive cases are truly positive. It measures the number of TP over the number of total positives predicted by the model. In this study, it measures how many patients predicted to have mental health problems actually have a corresponding ground-truth annotation confirming the predictions.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (7)$$

5) F1 score

The F1 score is a measure that balances precision and recall by considering the two most critical values: FP and FN. The maximum value is 1 when recall and precision are equal. As such, it works effectively on imbalanced datasets having uneven class distribution. Equation (8) depicts the F1 score calculation.

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (8)$$

6) Area Under the Receiver Operating Characteristics Curve (AUROC)

The Receiver operating characteristic (ROC) is a curve that compares the TP rate on the y-axis to the FP rate on the x-axis. The Area Under the Curve (AUC) assesses the capability of a classification model to distinguish between the classes at a threshold point. The closer the AUROC is to the value of 1, the better the model is at differentiating patients with mental health illness and no illness. However, if the AUROC is equal to 0, it implies that the model predicts positive as negative and vice versa, which is the worst-case scenario. Meanwhile, if the AUROC is equivalent to 0.5, it indicates that the model is incapable of differentiating between positive and negative classes. Figure 8 illustrates the area under the ROC.

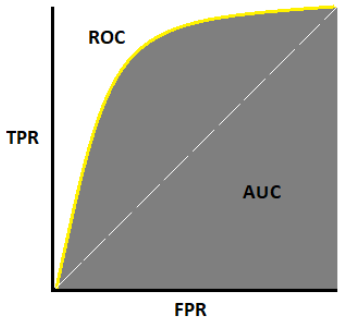

Figure 8. AUROC.

V. Results and Performance Analysis of Different Methods

A. Data Analysis

1) Family History and Treatment

Figure 9 presents a grouped bar chart that shows the relationship between having a family history of mental health problems and actively seeking mental health care. 321 responders are both seeking treatment and have a family history of mental health issues.

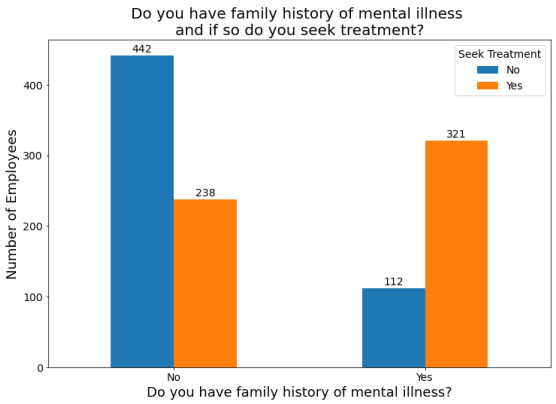

Figure 9. Number of Respondents with or without Family History of Mental Illness and do they Seek Treatment for Mental Illness.

2) Concern of Mental Health vs Physical Health

Figure 10 depicts an intriguing insight into respondents' attitudes toward disclosing one's health condition to a potential employer during an interview. The bar chart on the left concerns mental health, while the bar chart on the right concerns physical health. It is startling to learn that while 41% of respondents will not disclose a physical health issue, 81% of respondents have decided to keep their mental health issues a secret from a potential employer. This disparity of 40% could indicate that employees are concerned that disclosing a mental health issue could unintentionally impact their careers.

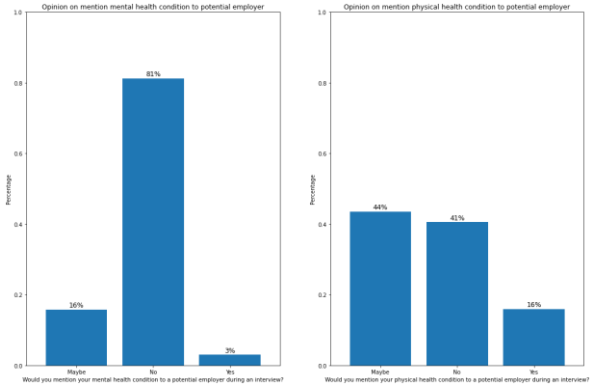

Figure 10. (i) Opinion on Mention Mental Health Condition to Potential Employer, (ii) Opinion on Mention Physical Health Condition to Potential Employer

3) *Company Healthcare Benefits and Treatment*

Figure 11 depicts the association between company benefits and the percentage of employees seeking mental healthcare. The number of employees seeking for mental health treatment increases by 34% for those who receive any form of benefits from their company. There is a difference of 12% in treatments as compared to employees without company benefits.

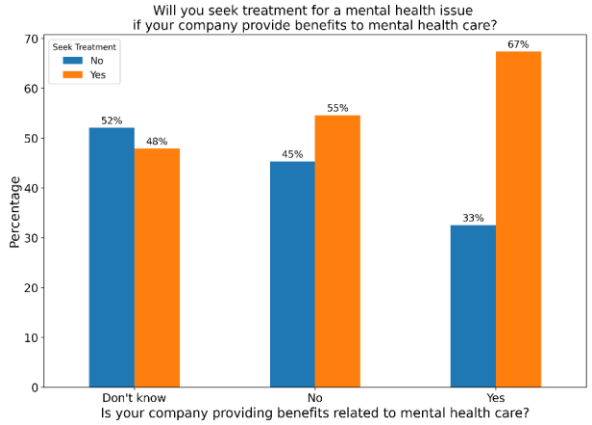

Figure 11. Number of Respondents who will or will not Seek Treatment for a Mental Illness if Healthcare Benefit is or is not provided by Employer

4) *Company Healthcare Benefits and Treatment*

Figure 12 depicts the number of employees who will or will not tell every employer about their mental health condition or only to certain employers and if they are receiving mental health treatment. Each group's response is very evenly distributed.

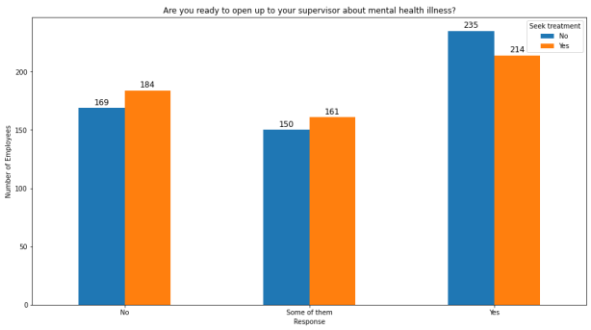

Figure 12. Number of Respondent Ready or Not Ready to inform their Supervisor about their Mental Illness

B. *Baseline Models*

The baseline models are trained on the training set prior to feature selection and fine-tuning. Table VI illustrates the preliminary result of the baseline models.

TABLE VI. PERFORMANCE OF THE BASELINE MODELS

Model	*Accuracy*	*Recall*	*Precision*	*F1 Score*	*AUROC*
LR	81.78%	85.1%	79.89%	82.35%	88.44%
KNN	77.54%	79.95%	76.27%	77.99%	82.27%
RF	81.52%	85.87%	80.80%	83.17%	88.91%

Both Logistic Regression and Random Forest baseline models demonstrate strong performance, achieving accuracy slightly above 80%, with a recall of 85% and precision near or at 80%. In contrast, K-Nearest Neighbors baseline model performs relatively poorer, with an accuracy of 77.54%, and a precision of only 76.27%. The F1 score of Logistic Regression and Random Forest is above 80%, while the K-Nearest Neighbors' F1 score can be further improved. It is important to note that the baseline model is a simple and less sophisticated approach involving fewer features and minimal preprocessing, which serves as a reference for future comparisons as more advanced models are developed.

C. *Feature Selection*

This study employs Recursive Feature Elimination, Cross-validated (RFECV) to perform feature selection, and to avoid biasness, the GradientBoostingClassifier is used to calculate the feature importance. Figure 13 illustrates the features and their feature importance score.

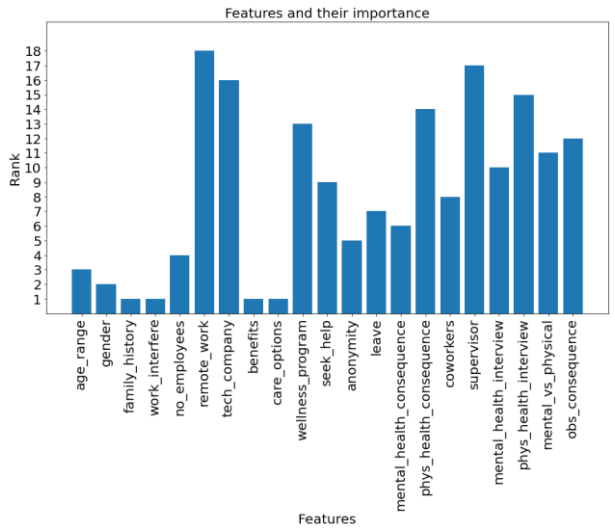

Figure 13. Features and Their Importance Score

It is observed that out of the initial set of 21 features, the RFECV algorithm selects four features:

- family_history: the presence of mental illness in the family
- work_intefere: the extent to which mental health problems affect work performance
- care_options: the level of awareness and availability of mental healthcare options provided by the company
- benefits: whether the employer offers mental healthcare benefits to employees

One of the selected features, family history, can be supported by the findings of [26] which suggest that the presence of mental illness is often attributed to the influence of genetic factors. On the other hand, [27] has established a connection between workplace environmental quality and certain mental health indicators, such as concentration and stress. Additionally, [28] highlights that mental disorders will affect an individual's productivity at work. Both findings reinforce the relevance of the selected feature: "work_interfere". All four selected features hold the highest rank of importance, indicating that they have the highest influence on mental health problems; hence, they will be utilized to train the models.

D. Hyperparameters Fine-tuning

To further refine the model, the pre-trained baseline models undergo fine-tuned using GridSearchCV and RandomizedSearchCV.on the data of the validation set with the features selected. The optimal hyperparameters obtained through fine-tuning using RandomizedSearchCV and GridSearchCV approaches are illustrated in Table VII and Table VIII respectively. The performance of various models utilizing the optimal hyperparameters on the validation dataset is evaluated to assess the performance improvement. Table IX depicts the performance of the proposed models after hyperparameter fine-tuning.

TABLE VII. OPTIMAL HYPERPARAMETERS OBTAINED (RANDOMIZEDSEARCHCV)

Optimal Hyperparameter Value Obtained (RandomizedSearchCV)		
Model	***Hyperparameter***	***Range of Value***
LR	solver	'liblinear'
	penalty	l2
	C	10
KNN	leaf_size	40
	n_neighbors	11
	p	1
	weights	'uniform'
	metric	'minkowski'
RF	n_estimators	282
	max_features	'auto'
	max_depth	7
	min_samples_split	6
	min_samples_leaf	2
	bootstrap	True
	criterion	'gini'
	oob_score	False
	max_leaf_nodes	9

TABLE VIII. OPTIMAL HYPERPARAMETERS OBTAINED (GRIDSEARCHCV)

Optimal Hyperparameter Value Obtained (GridSearchCV)		
Model	***Hyperparameter***	***Range of Value***
LR	solver	'liblinear'
	penalty	l2
	C	0.01
KNN	leaf_size	22
	n_neighbors	16
	p	1
	weights	'uniform'
	metric	'minkowski'
RF	n_estimators	200
	max_features	'sqrt'
	max_depth	6
	min_samples_split	6
	min_samples_leaf	9
	bootstrap	True
	criterion	'entropy'
	oob_score	True
	max_leaf_nodes	9

TABLE IX. PERFORMANCE OF THE FINE-TUNED MODELS

Performance of the RandomizedSearchCV-tuned Models					
Model	***Accuracy***	***Recall***	***Precision***	***F1 Score***	***AUROC***
LR	78.44%	82.42%	78.95%	80.65%	78.05%
KNN	82.63%	92.31%	79.25%	85.28%	81.68%
RF	83.83%	92.31%	80.77%	86.15%	83.0%
Performance of the GridSearchCV-tuned Models					
LR	83.23%	98.9%	76.92%	86.54%	81.69%
KNN	82.04%	89.01%	80.2%	84.38%	81.35%
RF	83.23%	94.51%	78.9%	86.0%	82.12%

Inarguably, the RandomizedSearchCV-tuned Random Forest model, Logistic Regression model, and GridSearchCV-tuned K-Nearest Neighbors model exhibit higher consistency in performance across all evaluation metrics. Furthermore, the RandomizedSearchCV-tuned K-Nearest Neighbors model, GridSearchCV-tuned Random Forest model, and Logistic Regression model demonstrate high recall rates exceeding 90%, while their precision remains below 80%, indicating a lack of consistency and demands further examination to identify potential overfitting. Notably, it is found that the use of GridSearchCV increases the time required for hyperparameter optimization while RandomizedSearchCV may not always identify the optimal hyperparameter combination when exploring a larger hyperparameter space, as supported by the findings of [29].

E. Model Selection using K-folds Cross-validation

The fine-tuned models undergo cross-validation on data of the validation set, aiming to select the best-performing model for test set prediction. Table X recapitulates the performance of the RandomizedSearchCV-tuned and GridSearchCV-tuned models during cross-validation.

TABLE X. PERFORMANCE OF THE FINE-TUNED MODELS DURING CROSS-VALIDATION

Performance of the RandomizedSearchCV-tuned Models during Cross-Validation					
Model	***Accuracy***	***Recall***	***Precision***	***F1 Score***	***AUROC***
LR	83.23%	89.87%	78.02%	83.53%	83.57%
KNN	78.38%	87.00%	76.38%	79.56%	85.25%
RF	81.99%	88.17%	81.90%	84.87%	90.47%
Performance of the GridSearchCV-tuned Models during Cross-Validation					
LR	70.59%	100.00 %	64.51%	78.92%	86.91%
KNN	80.88%	86.41%	81.27%	81.05%	89.84%
RF	81.36%	89.39%	79.30%	84.31%	89.97%

It is noteworthy to highlight that the Logistic Regression model utilizing the hyperparameters generated by GridSearchCV exhibits clear signs of overfitting during cross-validation. To clarify, both GridSearchCV-tuned and RandomizedSeachCV-tuned Logistic Regression models have the penalty parameter set to 'l2', and 'solver' parameter is set to 'liblinear'. The distinction lies in the use of the parameter C: the GridSearchCV-tuned Logistic Regression model has a C value of 0.01, while the RandomizedSearchCV-tuned LR model has a higher C value of 10., suggesting that C value of 0.01 is insufficient in preventing overfitting in the multiple folds of the validation data.

For model selection, as illustrated in Table X, the RandomizedSearchCV-tuned Random Forest model outperforms other models, achieving over 80% in all evaluation metrics and excelling in recall with 88.17% and AUC with 90.47%. Therefore, it is chosen as the best model for test set prediction. Figure 14 and Figure 15 illustrates the performance of the fined-tuned models' during cross-validation.

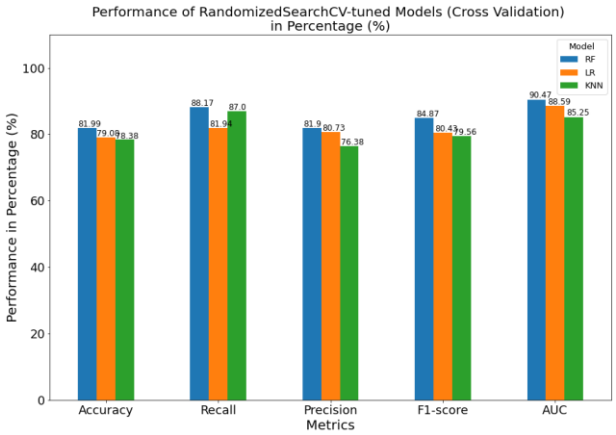

Figure 14. Performance of various RandomizedSearchCV-tuned Models on Cross-validation

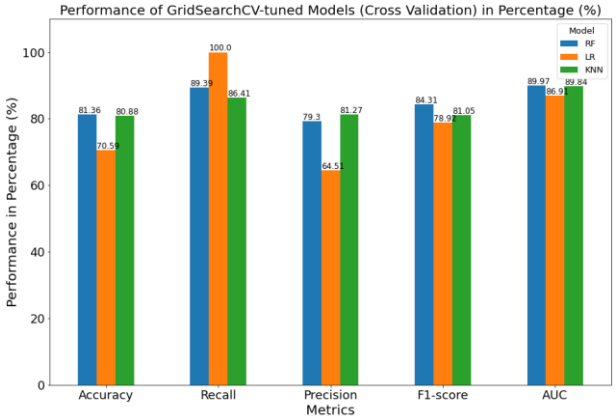

Figure 15. Performance of various GridSearchCV-tuned Models on Cross-validation

F. Best Model on Test Set Prediction

The Random Forest model utilizing hyperparameters generated through RandomizedSearchCV (which will be referred to as the best model from this point onwards) is deployed on the data of the test set for prediction. Figure 16 illustrates the confusion matrix generated by the best model, showing that out of 167 rows of test data, the best model is able to distinguish 68 cases of true negative and 71 cases of true positive. On the other hand, the model predicts 20 false positives and 8 false negatives.

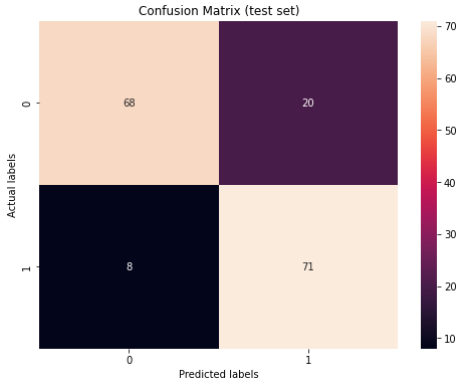

Figure 16. Confusion Matrix by the Best Model

The best model exhibits a good generalization when applied to the test set. Table XI shows that it achieves 83.23% accuracy and 89.87% recall, indicating a low false negative rate, but the precision drops to 78.02%. This indicates that out of all the predicted positive cases in the test set, only 78.02% are actually true positives This suggests a tendency to predict more positive instances, leading to an increase in false positives and resulting in an AUROC of 83.57%, recalling that AUROC is a metric that measures a model's ability to distinguish positive and negative classes. Considering the trade-off between recall and precision, the best model gives a good F1 score of 83.53%. Nevertheless, an AUROC of 83.57% indicates a reasonably good ability to accurately rank positive and negative instances instead of random guessing. The AUROC being closer to 100% than 50% further supports this notion.

TABLE XI. RESULTS OF THE BEST MODEL

Accuracy	*Recall*	*Precision*	*F1 Score*	*AUROC*
83.23%	89.87%	78.02%	83.53%	83.57%

The comparison between the Random Forest baseline model on the training set and the best model on the test set prediction is visualized in Figure 17. Compared to the baseline model, the best model shows an increase in accuracy by 2.10%; a significant improvement in recall by 4.66%; and a slight improvement in the F1 score by 0.43%. However, the precision of the best model is lower than that of the baseline model by -3.44%, and its AUROC is 6.01% lower compared to the baseline model. Despite a drop in precision and AUROC, the best model outperforms the baseline model. As aforementioned, this study emphasizes the recall to ensure the identification of potential patients with mental health problems.

In short, the combination of feature selection and hyperparameter fine-tuning are proven effective in developing a mental health problem prediction model, albeit with potential for improvement.

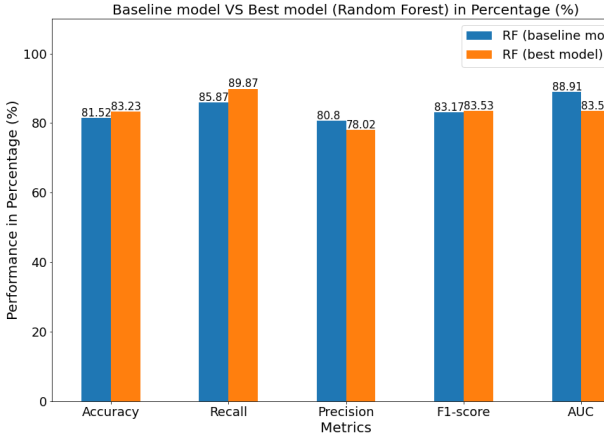

Figure 17. Performance of Baseline Model compared to the Best Model

VI. CONCLUSION AND FUTURE WORK

From our study, it is observed that family history of mental illness, the extent to which mental health problems affect work performance, the level of awareness and availability of mental healthcare options provided by the employer, and whether the employer offers mental healthcare benefits to employees have the highest feature importance in determining whether an individual has mental health problems. The Random Forest model utilizing hyperparameters derived from the RandomizedSearchCV method exemplifies its robustness and reliability during cross-validation. It yields highly significant results and achieves balance across all evaluation metrics, making it the prime choice for test set prediction. During test set prediction, it achieves an accuracy of 83.23% and an impressive recall of 89.87%. However, its precision is slightly low at 78.02%. Considering the trade-off between recall and precision, the model gives a good F1 score of 83.53% and AUROC of 83.57%.

This study has certain limitations such as the current mental health problem prediction model design is more inclined to produce higher recall than higher precision. For instance, the best model has higher recall than precision. Therefore, using a larger volume of data for training might help to improve the precision by providing the model with more information to learn from, especially since the current dataset is relatively small.

For future study, deep learning and hybrid classifiers resulting from ensemble methods such as Bagging and Gradient Boosting can be deployed to study the models' performance. Moreover, consulting a professional in the field of mental health is an ideal approach that will contribute to identifying the features that are directly related to the accuracy of the prediction precisely. The features selected will give the employer and employee an insight on how to build a mental health-friendly workplace environment to help curb the growth of mental health problems.

ACKNOWLEDGMENT

There is no financial support from any agencies funding this research work.

REFERENCES

[1] Zhang, X., Ren, H., Gao, L., Shia, B.-C., Chen, M.-C., Ye, L., Wang, R., & Qin, L. (2023). Identifying the predictors of severe psychological distress by auto-machine learning methods. *Informatics in Medicine Unlocked, 39*, 101258. https://doi.org/10.1016/j.imu.2023.101258

[2] World Health Organization. (2022, June 8). *Mental disorders*. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[3] Minister of Health Malaysia. (2016, September 28). *Mental Health Problems in Malaysia.* https://www.moh.gov.my/moh/modules_resources/english/database_stores/96/337_451.pdf

[4] Henderson, C., Evans-Lacko, S., & Thornicroft, G. (2013). Mental illness stigma, help seeking, and public health programs. *American Journal of Public Health, 103*(5), 777–780. https://doi.org/10.2105/ajph.2012.301056

[5] Arora, A., Bojko, L., Kumar, S., Lillington, J., Panesar, S., & Petrungaro, B. (2023). Assessment of machine learning algorithms in national data to classify the risk of self-harm among young adults in hospital: A retrospective study. *International Journal of Medical Informatics, 177*, 105164. https://doi.org/10.1016/j.ijmedinf.2023.105164

[6] Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: A meta-analysis. *The Lancet, 374*(9690), 609–619. https://doi.org/10.1016/s0140-6736(09)60879-5

[7] Colizzi, M., Lasalvia, A., & Ruggeri, M. (2020). Prevention and early intervention in youth mental health: Is it time for a multidisciplinary and trans-diagnostic model for care? *International Journal of Mental Health Systems, 14*(1). https://doi.org/10.1186/s13033-020-00356-9

[8] Cho, G., Yim, J., Choi, Y., Ko, J., & Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation, 16*(4), 262–269. https://doi.org/10.30773/pi.2018.12.21.2

[9] Katarya, R., Maan, S. (2020). Predicting mental health disorders using machine learning for employees in technical and non-technical companies. *2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE)*, 1–5. https://doi.org/10.1109/icadee51157.2020.9368923.

[10] Jain, T., Jain, A., Hada, P. S., Kumar, H., Verma, V. K., & Patni, A. (2021). Machine learning techniques for prediction of Mental Health. 2021 *Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 1606–1613. https://doi.org/10.1109/icirca51532.2021.9545061

[11] Sujal, B. H., Neelima, K., Deepanjali, C., Bhuvanashree, P., Duraipandian, K., Rajan, S., & Sathiyanarayanan, M. (2022). Mental health analysis of employees using Machine Learning Techniques. *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 1–6. https://doi.org/10.1109/comsnets53615.2022.9668526

[12] Mutalib, S., Shafiee, N. S. M., & Rahman, S. A. (2021). Mental Health Prediction Models Using Machine Learning in Higher Education Institution. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(5), 1782-1792. 10.17762/turcomat.v12i5.2181.

[13] Wshah, S., Skalka, C., & Price, M. (2019). Predicting posttraumatic stress disorder risk: A machine learning approach. *JMIR Mental Health, 6*(7). https://doi.org/10.2196/13946

[14] Sumathi, M. R., & Poorna, B. (2016). Prediction of mental health problems among children using machine learning techniques. *International Journal of Advanced Computer Science and Applications, 7*(1). https://doi.org/10.14569/ijacsa.2016.070176

[15] Khattar, A., Jain, P. R., & Quadri, S. M. K. (2020) Effects of the Disastrous Pandemic COVID 19 on Learning Styles, Activities and Mental Health of Young Indian Students - A Machine Learning Approach, *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1190-1195, doi: 10.1109/ICICCS48265.2020.9120955.

[16] Wang, X., Li, H., Sun, C., Zhang, X., Wang, T., Dong, C., & Guo, D. (2021). Prediction of mental health in medical workers during COVID-19 based on machine learning. *Frontiers in Public Health, 9*. https://doi.org/10.3389/fpubh.2021.697850

[17] Hornstein, S., Hoffman, V. F., Nazander, A., Ranta, K., & Hilbert, K. (2021). Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *DIGITAL HEALTH, 7*, 1–11. https://doi.org/10.1177/20552076211060659

[18] Iorfino, F., Ho, N., Carpenter, J. S., Cross, S. P., Davenport, T. A., Hermens, D. F., Yee, H., Nichles, A., Zmicerevska, N., Guastella, A., Scott, E., & Hickie, I. B. (2020). Predicting self-harm within six months after initial presentation to youth mental health services: A machine learning study. *PLOS ONE, 15*(12). https://doi.org/10.1371/journal.pone.0243467

[19] Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random Forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

[20] Urso, A., Fiannaca, A., La Rosa, M., Ravì, V., & Rizzo, R. (2019). Data Mining: Prediction Methods. *Encyclopedia of Bioinformatics and Computational Biology, 1*, 413–430. https://doi.org/10.1016/b978-0-12-809633-8.20462-7

[21] Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal, 3*, 100071. canohttps://doi.org/10.1016/j.dajour.2022.100071

[22] Amro, A., Al-Akhras, M., Hindi, K., Habib, M. & Shawar, B. (2021). Instance Reduction for Avoiding Overfitting in Decision Trees. *Journal of Intelligent Systems, 30*(1), 438-459. https://doi.org/10.1515/jisys-2020-0061

[23] G, S. G., & Sumathi, B. (2020). Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction. *International Journal of Advanced Computer Science and Applications, 11*(9). https://doi.org/10.14569/ijacsa.2020.0110920

[24] Open Sourcing Mental Illness, LTD. (2014). *Mental Health in Tech Survey, Version 3*. Retrieved February 13, 2023 from https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey

[25] Kourou, K. et al. (2015) 'Machine learning applications in cancer prognosis and prediction', *Computational and Structural Biotechnology Journal*, 13, pp. 8–17. doi:10.1016/j.csbj.2014.11.005.

[26] Grant, J. E., & Chamberlain, S. R. (2020). Family history of substance use disorders: Significance for mental health in young adults who gamble. *Journal of Behavioral Addictions, 9*(2), 289–297. https://doi.org/10.1556/2006.2020.00017

[27] Bergefurt, L., Weijs-Perrée, M., Appel-Meulenbroek, R., & Arentze, T. (2022). The physical office workplace as a resource for mental health – A systematic scoping review. *Building and Environment, 207*, 108505. https://doi.org/10.1016/j.buildenv.2021.108505

[28] World Health Organization. (2023). *Mental health in the Workplace*. https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/mental-health-in-the-workplace#:~:text=Without%20effective%20support%2C%20mental%20disorders,to%20retain%20or%20gain%20work.

[29] Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks, 4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005