Explainable hybrid tabular Variational Autoencoder and feature Tokenizer Transformer for depression prediction

Article *in* Expert Systems with Applications · March 2025

DOI: 10.1016/j.eswa.2024.126084

2 authors, including:

Haewon Byeon
Korea University of Technology and Education
571 PUBLICATIONS **2,660** CITATIONS

Explainable Hybrid Tabular Variational Autoencoder and Feature Tokenizer Transformer for Depression Prediction

Vinh Quang Tran[1,2], Haewon Byeon[1,3,*]

[1]Department of Digital Anti-Aging Healthcare (BK21), Inje University, Gimhae 50834, Republic of Korea.

[2]trnvinh5799@oasis.inje.ac.kr, ORCID iD: https://orcid.org/0009-0003-5044-0688

[3] byeon@inje.ac.kr, ORCID iD: https://orcid.org/0000-0002-3363-390X.

Abstract

Recent advancements in machine learning (ML) and deep learning (DL) have significantly improved the diagnosis, detection, prediction, and prognosis of depressive disorders. Nonetheless, these methodologies often face challenges related to generalizability, transparency, data scarcity, privacy concerns, and class imbalance. This study develops a robust and interpretable model for predicting depression in South Korea, addressing these limitations. We employed a hybrid deep learning approach that integrates the Feature Tokenizer Transformer (FT-Transformer) model, specifically designed for tabular data, to effectively tokenize and process both categorical and numerical features, alongside synthetic data generated by the Tabular Variational Autoencoder (TVAE). TVAE, an adaptation of Variational Autoencoders with a specialized loss function for tabular data, generated high-quality synthetic data from the Korea National Health and Nutrition Examination Survey (KNHANES) dataset. The efficacy of the TVAE-generated data was validated using non-parametric statistical tests, achieving 86.30% in the Kolmogorov-Smirnov (KS) test and 76.65% in the Chi-squared (CS) test. Performance evaluation metrics, including accuracy, recall, F1-score, and AUC, demonstrated our model's effectiveness, yielding an accuracy of 0.7783, a recall of 0.5310, an F1-score of 0.4657, and an AUC of 0.6822, outperforming state-of-the-art models. Additionally, SHapley Additive exPlanations (SHAP) analysis was incorporated to elucidate feature importance, offering valuable insights for healthcare professionals. This research highlights the potential of deep learning and synthetic data techniques to enhance depression prediction, addressing critical challenges such as generalizability, class imbalance, data privacy, and interpretability over existing models.

Key words: variational autoencoder; feature tokenizer; transformer; depression; explainable.

[*]Corresponding author at: Department of Digital Anti-Aging Healthcare (BK21), Inje University, Gimhae 50834, Republic of Korea. Email: bhwpuma@naver.com

1 Introduction

Depressive disorder (depression), a pervasive mental health condition, has reached epidemic proportions globally, imposing a substantial burden on individuals, societies, and economies. Depression, defined by persistent loss of interest, sadness, and changes in sleep, appetite, and energy, can have a significant negative impact on daily functioning (Otte et al., 2016). It marked by a sustained period of sadness or a loss of interest in activities, which is clearly different from normal mood swings and everyday life experiences. Depression affects around 3.8% of the population, with 5% of adults affected (4% of men and 6% of women). Among adults older than 60, the prevalence of depression is estimated at 5.7%. Globally, an estimated 280 million people suffer from depression (Bhatt et al., 2023). Depression is also a major risk factor for suicide, with South Korea experiencing the highest suicide rates among Organization for Economic Cooperation and Development (OECD) countries (G. E. Kim et al., 2020; Shin et al., 2017). This concerning statistic has brought depression and mental health to the forefront of public health research in South Korea, prompting calls for more effective diagnostic and predictive tools.

Traditional diagnostic methods for depression rely heavily on clinical interviews and self-reported questionnaires. However, these techniques are subject to significant limitations, including subjectivity, time-consuming procedures, and limited accessibility. Such limitations can hinder accurate diagnosis, particularly when administered by non-specialist clinicians, potentially leading to delayed or missed treatment opportunities (McGorry, 2015; McGorry et al., 2006). Electronic health records (EHRs) offer a promising avenue for addressing these challenges. Beyond storing patient information and performing administrative tasks, EHRs can be utilized for diagnostic purposes (Shickel et al., 2018). By analyzing patient data within EHRs, clinicians can potentially identify patterns and symptoms indicative of depression, facilitating earlier detection and intervention. However, effective disease tracking and surveillance through EHRs require professional expertise and experience. Excessive time spent on EHR tasks can also contribute to clinician burnout (Budd, 2023). Additionally, patient data privacy and ethical concerns related to EHR usage must be carefully considered (Sulmasy et al., 2017).

With recent advancements in technology and data science have created opportunities for developing automated and efficient depression screening tools by EHR. As information and technology continue to grow, the use of deep learning (DL) and machine learning (ML) algorithms to extract meaningful patterns from EHR data across various sectors is becoming increasingly prevalent. While ML algorithms have been widely adopted in the medical and health sectors, their application in psychological analysis remains relatively limited (Orrù et al., 2020). Due to the challenges associated with statistical inference, researchers are increasingly turning to ML and DL for psychological analysis. However, applying deep neural networks to tabular data presents challenges such as missing values, the absence of locality, mixed feature types, and limited understanding of the dataset's organization (Hwang & Song, 2023; Shwartz-Ziv & Armon, 2022). Furthermore, existing ML and DL models frequently face challenges in generalizing to varied populations and clinical environments, leading to inconsistent performance across different populations or datasets. Privacy concerns also complicate the sharing and use of sensitive health data (Thapa & Camtepe, 2021), which can limit the breadth and depth of model training. A notable limitation of current studies is the tendency to treat ML and DL models as black boxes, which obstructs our understanding of the reasoning behind their predictions. This opacity presents challenges in fostering confidence in the predictions made by these models among both patients and healthcare professionals (Confalonieri et al., 2021).

To bridge these technical gaps, this study aims to develop a robust and interpretable model for predicting depression in South Korea, with a focus on addressing data imbalance, privacy concerns, and model interpretability to ensure its applicability in clinical settings. In pursuit of these objectives, we evaluated the model’s performance, generalizability, and interpretability, testing several ML and DL models, including the Feature Tokenizer Transformer (FT-Transformer) (Gorishniy et al., 2023), TabNet (Arik & Pfister, 2021), LightGBM (Ke et al., 2017), Random Forest (RF) (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), Gradient Boosting (Friedman, 2002), and AdaBoost (Freund & Schapire, 1999). To improve generalizability and reduce noise in the data, we implemented Maximum Relevance Minimum Redundancy (mRmR) for feature selection. Additionally, we employed TVAE to generate synthetic data to address data imbalance and missing value issues. The use of synthetically generated EHR data (Goncalves et al., 2020) eliminates patient confidentiality concerns, enabling a comprehensive evaluation of all model aspects without compromising privacy.

The research is characterized by several key contributions:

- The study utilizes a comprehensive dataset from the Korea National Health and Nutrition Examination Survey, which includes data from 10,194 individuals. This dataset is meticulously analyzed to select features that are crucial for predicting depression.
- A systematic evaluation of various ML and DL algorithms is conducted to identify those that perform optimally and suit the specifics of our tabular depression dataset. This research introduces a robust hybrid model that integrates the FT-Transformer with synthetic data generated from TVAE, demonstrating superior performance compared to traditional data augmentation methods, such as oversampling and under sampling.
- By employing synthetic data, we effectively mitigate data imbalance, enhancing the representativeness of the training dataset while addressing privacy concerns associated with the

use of sensitive health information. This allows for a thorough evaluation of the model without risking patient confidentiality, facilitating its implementation in clinical settings.

- The integration of SHAP provides valuable insights into feature importance and local explanations, enabling healthcare professionals and data analysts to interpret model predictions more effectively.

This paper is structured as follows: Section 2 reviews existing literature. Section 3 details our proposed approach, including the datasets, augmentation methods, and algorithms used. Section 4 presents the model's performance and explains our rationale for choosing these methods. Section 5 explores the model's interpretability using SHAP techniques. Finally, Section 6 concludes by summarizing our key findings and outlining potential areas for future research.

2 Literature review

There are three key areas of interest that must be considered for clinical application: generalizability, interpretability, and performance. Generalizability refers to the model's ability to be reused with different populations. In predicting and diagnosing depression, a variety of modalities—such as self-questionnaires, audiovisual markers, and EEG—are commonly used (Yasin et al., 2023). However, studies using these modalities often have limited sample sizes, typically between 40 to 60 individuals (Khadidos et al., 2023; Mahato et al., 2022; Mohanty et al., 2024). By contrast, research on EHRs or tabular data often involves larger cohorts because such data is more accessible in healthcare systems. Consequently, focusing on tabular data enhances generalizability due to the larger, more diverse sample sizes typically available in EHR datasets. This allows for broader applicability of the model to different demographic and clinical profiles. Interpretability is the ease with which a model's predictions can be understood and trusted by clinicians and other stakeholders. Performance is a measure of the model's ability to meet the intended purpose, as measured by metrics such as Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and F1-score. Recent studies provide varying insights and findings relevant to these areas, underscoring both advancements and limitations.

In terms of performance, advancements in ML and DL have led to significant improvements in predicting depression based on EHRs. Several studies have reported AUC-ROC values ranging from 0.55 to 0.94 and F1-score from 0.45 to 0.7 (Kasthurirathne et al., 2019; Półchłopek et al., 2020; Zhang et al., 2021). In term of accuracy, the highest accuracy was in Sau et al research (Sau & Bhakta, 2017) by 0.91. However, the performance of predictors could be affect by EHRs variables such as antidepressants drug codes, or family history of depression, alcohol, drug, physical and sexual abuse, and co-morbidity with other mental health conditions, are strong predictors of depression (Nickson et al., 2023). Overly boosting the performance of predictors may lead to misclassification of cases. Therefore, careful feature selection is crucial for generalizability when predicting depression.

In relation to the KNHANES dataset, Kim et al. (N. H. Kim et al., 2024) conducted a comprehensive analysis using the KNHANES dataset to predict depression in individuals exposed to second-hand smoke. Employing the SHAP method, they identified significant predictors and evaluated the performance of various ML algorithms. Support Vector Machines (SVM) demonstrated superior overall accuracy, achieving an AUC of 0.900 while LightGBM exhibited higher positive predictive value of 0.646, indicating its effectiveness in identifying true cases. However, the relatively low F1-score of 0.444 for LightGBM suggests class imbalance within the dataset, indicating that the model may overemphasize certain predictions at the expense of others. This limitation underscores the importance of managing data imbalance for effective application across broader populations. In a separate study, Lee et al. (Lee et al., 2023) utilized SVM to predict depression in a larger cohort of 3,007 individuals in patient with diabetes. Their model achieved a respectable AUC-ROC of 0.835 (95% CI = 0.730-0.901), demonstrating good overall performance. The F1-score of 0.6510 suggests a reasonable balance between precision and recall. Despite these strengths, the study lacks an in-depth interpretability component, making it challenging to understand the specific contributions of each feature to the model's predictions.

Regarding interpretability, Nemesure et al (Nemesure et al., 2021) used SHAP (Shapley Additive Explanations) scores to aid interpretability. Stacked Random Forest, XGBoost, Support Vector Machine, a neural network, and K-nearest-neighbors got AUC of 0.73. SHAP analysis identified several key features that contributed to the model's predictions, including living conditions, blood pressure, vaccination status, and marijuana use. However, the research did not delve into the model's internal workings to visualize individual predictions or assess the model's confidence in its classifications.

There remains a gap in applying these technologies effectively for clinicians, particularly in user interface integration. Studies in other fields have demonstrated the potential for enhanced user engagement through well-designed interfaces. For example, (Naga Srinivasu et al., 2024) developed a crop recommendation model that uses environmental data—such as rainfall, temperature, and humidity—to help stakeholders, including farmers and government agencies, optimize crop cycles, enhance harvest profitability, and manage food supply and demand effectively. Similarly, (Dharmarathne et al., 2024) introduced a self-explanatory XAI interface aimed at diabetes diagnosis. This tool offers users an interactive explanation of their likelihood of having diabetes, fostering awareness and enabling proactive health decisions. Given the high global incidence of diabetes and the impact of delayed diagnosis, this application serves as a valuable educational resource, helping users understand both their health condition and the factors influencing the model's predictions. Together, these studies highlight both advancements and remaining challenges in creating clinically applicable models for depression prediction, as summarized in Table 1.

Table 1. Recent Research on ML and DL for Depression Diagnosis

Authors	Year	Dataset	Models	Performance	Pros	Cons
Sau et al.	2017	Depression Clinical Dataset	SVM, Decision Trees, Neural Network	Accuracy: 0.91	High accuracy achieved; diverse model comparison	Limited to a single dataset; lacks interpretability
Kasthurirathne et al.	2019	EHRs (general population)	Logistic Regression, Random Forest	AUC-ROC: 0.86–0.94	Wide range of models tested	High variability in AUC-ROC, dependent on dataset
Półchłopek et al.	2020	EHRs (Poland dataset)	Random Forest, Gradient Boosting	AUC-ROC: 0.70–0.82 F1-core: 0.45-0.70	Good generalizability with EHRs	Limited interpretability; data-specific
Nemesure et al.	2021	Mixed clinical and lifestyle data	Stacked Random Forest, XGBoost, SVM, KNN	AUC-ROC: 0.73	Uses SHAP for interpretability	Lacks per-instance classifier visualization
Zhang et al.	2021	Depression clinical dataset	Neural Networks, SVM	AUC-ROC: 0.89	Inclusion of various clinical predictors	Low accuracy in some cases, inconsistent performance
Lee et al.	2023	KNHANES (diabetes patients)	SVM	AUC-ROC: 0.84, F1-score: 0.65	Good balance between precision and recall	Limited to diabetes patients; generalizability uncertain
Kim et al.	2024	KNHANES (second-hand smoke exposure)	SVM, LightGBM XGBoost	AUC-ROC: 0.900 (SVM), F1-score: 0.44	Uses SHAP to identify significant predictors	Limited positive predictive value; focused on specific population, low F1-score indicating data imbalance

EHR: Electronic Health Record; KNHANES: Korea National Health and Nutrition Examination Survey; Random Forest: Random Forest Classifier; LightGBM: Light Gradient Boosting Machine Classifier; XGBoost Classifier: XGBoost; Random Forest Classifier: Random Forest; SVM: Support Vector Machine Classifier;

3 Material and methods

To develop a robust and interpretable model for predicting depression, this study followed a systematic approach as illustrated in Figure 1. The raw data was subjected to a preprocessing phase to enhance its suitability for analysis. Subsequently, a feature selection process was implemented to identify the most significant variables. Next, a variety of augmentation techniques, including conventional and synthetic methods, were applied and evaluated to determine the most effective approach. The model was trained using the augmented data, and the optimal hyperparameters were selected through a tuning process. Finally, the model's interpretability was enhanced by using SHAP analysis to gain insights into its predictions.

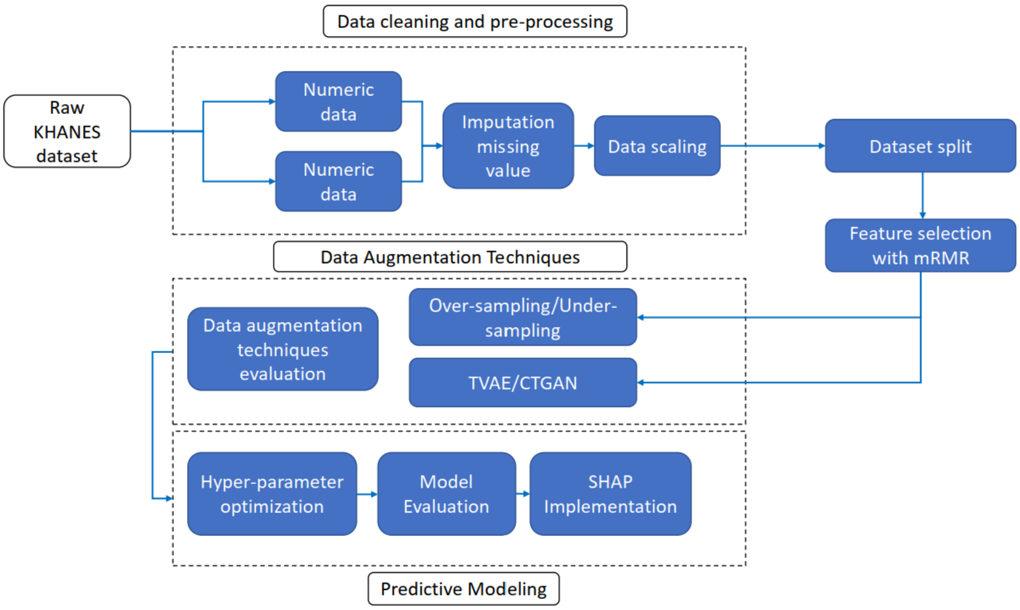
Data cleaning and pre-processing
Raw KHANES dataset
Numeric data
Numeric data
Imputation missing value
Data scaling
Dataset split
Feature selection with mRMR
Data Augmentation Techniques
Data augmentation techniques evaluation
Over-sampling/Under-sampling
TVAE/CTGAN
Hyper-parameter optimization
Model Evaluation
SHAP Implementation
Predictive Modeling

Figure 1. Visual diagram of the research methodology.

3.1 Dataset description

This study utilized data from the Korea National Health and Nutrition Examination Survey (KNHANES), conducted between 2019 and 2022. Detailed information regarding the KNHANES design and data can be found in (Kweon et al., 2014). The KNHANES is a cross-sectional survey conducted by the Korea Centers for Disease Control and Prevention (KCDC), collecting data on various aspects of health through health interviews, health examinations, and a dietary survey. The data covers a wide range of topics, including socioeconomic factors, health-related behaviors, overall well-being, healthcare utilization, physical measurements, health indicators, and dietary habits.

All participants gave written consent before joining the study. This research only used existing data that had been made anonymous. The research followed the ethical guidelines of the Declaration of Helsinki, and the 2016-2018 KNHANES protocol was approved by the Institutional Review Board (IRB) of the Korea Centers for Disease Control and Prevention (IRB approval numbers: 2018-01-03-P-A and 2018-01-03-C-A).

3.2 Data preprocessing

3.2.1 Data cleaning and splitting for model training

Thorough data preprocessing is essential for constructing robust models, especially when working with tabular data (Maharana et al., 2022). This study combined data from four years (2019-2022) of the

KNHANES survey, encompassing 28,824 individuals and resulting in a dataset with 562 features. The target variable was the "MH_PHQ_S" column, representing the Patient Health Questionnaire-9 (PHQ-9) score (*Patient Health Questionnaire-9 (PHQ-9) - Mental Health Screening - National HIV Curriculum*, n.d.).

Given the survey-based nature of the dataset, certain questions were optional, leading to a higher prevalence of missing values in some columns. However, these columns primarily contained non-essential or supplementary information, justifying their exclusion. To address missing data, columns with over 50% null values were removed. Additionally, existing depression and anxiety-related columns, such as PHQ-9, GAD-7, suicide thoughts, attempts, or stress-related questions, were excluded.

This data cleaning process yielded a final dataset of 10,194 samples and 239 variables, focusing on individuals aged 19 and older who responded to depression-related questions. The PHQ-9 scores, proposed treatment actions, and corresponding sample sizes were identified in Table 2 as follows:

Table 2. Distribution of PHQ-9 Score, proposed treatment actions, and sample sizes

PHQ-9 Score	Level of depression	Proposed Treatment Actions	n
1 to 4	None	None	8,339
5 to 9	Mild	Observe carefully; repeat the PHQ-9 questionnaire at subsequent visits	1,347
10 to 14	Moderate	Treatment plan, incorporating counseling, follow-up appointments, and medication if needed	343
15 to 19	Moderately Severe	Begin treatment with medication or therapy immediately	105
20 to 27	Severe	Begin medication immediately. If symptoms are severe or treatment isn't working well, refer to a mental health specialist for therapy or combined care	60

A PHQ-9 score of 5 or higher was considered indicative of depression. The output variables were categorized as either 'presence of depression' (PHQ-9 ≥ 5) or 'absence of depression' (PHQ-9 < 5) (Kroenke et al., 2001). Of the total 10,194 participants, 8,339 individuals were classified as having no depression (MH_PHQ_S = 0). The remaining 1,855 individuals reported depressive symptoms, with 1,347 scoring between 5 and 9,343 scoring between 10 and 14,105 scoring between 15 and 19, and 60 scoring between 20 and 27. These scores represent varying levels of depression severity, ranging from mild to severe.

To enhance model generalizability, all individuals with depressive symptoms (PHQ-9 ≥ 5) were combined into a single category, regardless of the specific severity level. This approach aimed to capture the overall presence of depression while reducing the negative effects of having unequal class sizes on the model's accuracy.

To evaluate model performance effectively, cross-validation is a common technique in ML (Domingos, 2012). However, for complex DL models, cross-validation can be computationally demanding (Bergman et al., 2024). As an alternative, we employed a stratified split, partitioning the dataset into training, validation, and testing sets, with 60%, 20%, and 20% respectively.

The training set was used for data augmentation and model development. Hyperparameter tuning was done using the validation set, and the hold-out test set was reserved for final assessment. To further mitigate overfitting, we utilized a separate hold-out test set during model training. This set was not used for training but served exclusively to assess the model's generalizability on unfamiliar data. The performance metrics on this set helped us evaluate the model's ability to perform well on new data, preventing it from becoming too reliant on the training data.

3.2.2 Feature scaling and missing value handling

To ensure all features contribute equally to the model and enhance its generalizability, feature scaling was applied. This technique transforms feature values to a similar scale, balancing the influence of features with different magnitudes in the learning process. For datasets with features of varying ranges, units, or magnitudes, feature scaling is crucial.

The data was initially separated into numerical (32 columns) and categorical (207 columns) features. Missing values were replaced with the median for numerical variables and the mode for categorical variables. Feature scaling was then applied to standardize numerical features, ensuring each feature has a mean of 0 and a standard deviation of 1. The equations for calculating standardization and mean value are provided in (1) and (2), respectively. The standard deviation, which indicates the variability of the data, is calculated in equation (3) as follows.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \tag{2}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\mathrm{x} - \mu)^2}{N}} \tag{3}$$

3.3 Conventional imbalanced handling techniques

The dataset exhibited a significant class imbalance, with a disproportionate number of non-depression patients (8339) compared to depressed patients (1855). This imbalance poses a challenge for standard classification algorithms, as they may become biased towards the majority class. To address this issue, we employed a variety of techniques, including oversampling, under-sampling, a combination of oversampling and under-sampling, and synthetic data generation. These methods were compared to evaluate their effectiveness in mitigating the effects of class imbalance and improving the model's generalization performance.

3.3.1 Synthetic Minority Oversampling Technique (SMOTE)

One of the traditional methods to handle class imbalance is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE mitigates the limitations of simple class replication by artificially increasing the number of samples in the underrepresented class. This technique involves interpolating between a minority class instance and its k-nearest neighbors within the feature space (Chawla et al., 2002). By creating new synthetic samples in this manner, SMOTE effectively increases the representation of the smaller group in the training data, resulting in a more even distribution and potentially improving model generalizability. Equation (4) details the specific mathematical formula used by SMOTE to generate synthetic samples.

$$x_{smote} = x_i + (\hat{x}_i - x_i) \times \delta, \quad \delta \in [0, 1] \tag{4}$$

Equation (4) demonstrates how SMOTE generates synthetic minority samples. Here, x_i represents a data point belonging to the minority class. Its k-nearest neighbors (k-NNs) are identified, denoted by $\hat{x}_i$.. SMOTE calculates the difference between x_i and each of its k-NNs. Then, this difference is multiplied by a random value between 0 and 1, and added to the original sample x_i to create a new synthetic sample, denoted by x_{smote}. This process continues until the underrepresented class has the same number of samples as the larger group.

3.3.2 Tomek Links

Tomek Links ("Two Modifications of CNN," 1976) is a technique that reduces the overrepresented class to make it more similar to the underrepresented class. A Tomek link is a pair of data points, x and y, from different groups, and $d(x, y)$ being the distance between x and y. They are called a Tomek link if there is no data point z, that is $d(x, z)$< $d(x, y)$or $d(y, z)$< $d(y, x)$.

The Tomek Links procedure involves detecting and removing inaccurate classification boundaries by identifying Tomek links. This helps prevent the model from creating boundaries that are too close to instances of the minority class. Once Tomek links are identified, observations belonging to the majority class are removed, effectively under-sampling the majority class and addressing the class imbalance. This under-sampling technique helps to improve the model's performance on minority class instances

3.3.3 SMOTE-ENN

Batista et al. (2004) (Batista et al., 2004) proposed SMOTE-ENN, a hybrid approach that leverages the strengths of both SMOTE and Edited Nearest Neighbors (ENN) (*Statistical Learning Theory | Wiley*, 2024) to address class imbalance. This technique tackles the issue by first employing SMOTE to oversample the underrepresented minority class. During this stage, SMOTE randomly selects a minority class data point, identifies its k-nearest neighbors, and generates synthetic samples by interpolating between the selected point and its neighbors within the feature space. This process effectively increases the representation of the underrepresented class in the training data.

Following the oversampling step, SMOTE-ENN incorporates ENN for targeted under-sampling. Here, a predefined number of nearest neighbors (k) is determined for each data point, both from the minority and majority classes. The algorithm then analyzes the k-nearest neighbors of each data point and identifies the majority class within that region. If a data point's class label contradicts the majority class among its k-nearest neighbors, both the data point and its k-nearest neighbors are removed from the dataset.

3.4 Synthetic tabular data generation

This section explains the two advanced techniques used to generate synthetic data and enhance the model's performance. A total of 8,000 synthetic samples were generated for each model, with a balanced distribution of 4,000 samples from the non-depression class (class 0) and 4,000 samples from the depression class (class 1). Table A1 presents the hyperparameters used for the two synthesizers.

3.4.1 Tabular Variational Autoencoder (TVAE)

Variational Autoencoders (VAEs) (Kingma & Welling, 2022) are form of generative model that utilize an encoder and decoder to derive a latent space, similar to an autoencoder. VAEs aim to approximate the true distribution of input data, $p(x)$. The VAE architecture consists of a latent space, decoder, and encoder, as depicted in Figure 2. The encoder maps the input into a latent space, approximating the distribution $q(z|x)$. The decoder converts the latent space back into the input space, approximating the true distribution $p(x|z)$.

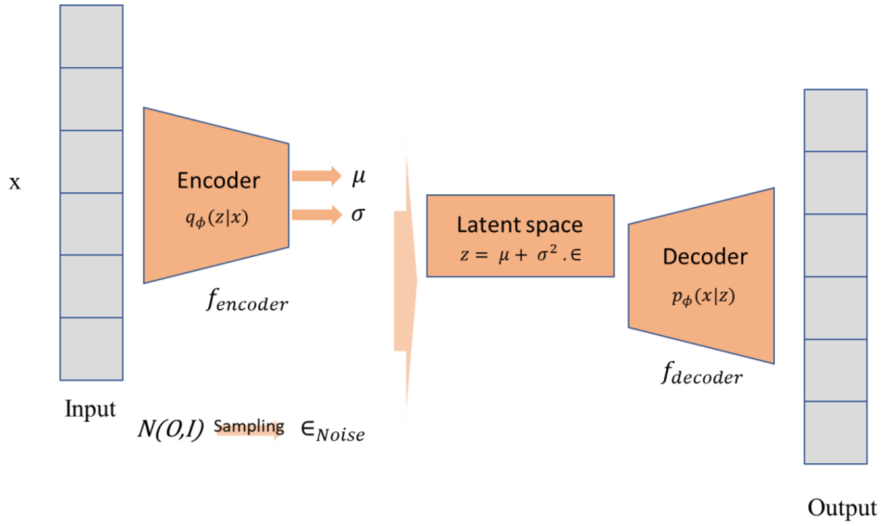

Figure 2. Architecture of a Variational Autoencoder.

The latent space is a hidden vector representation that the decoder can use to generate data. To prevent the latent space from always producing data with the same structure as the input, it samples noise and create a latent vector z. This technique is called the reparameterization trick. The objective when training a VAE is to maximize the log-likelihood $p_{\phi}(x)$. This can be achieved by minimizing the negative log-likelihood, which is also known as the Evidence Lower Bound (ELBO).

Xu et al.'s (Xu et al., 2019) introduced Tabular Variational Autoencoder (TVAE) by adapting the VAE framework by using the same preprocessing steps and modifying the loss function. TVAE employs two neural networks to model $p_{\phi}(r_j|z_j)$ and $q_{\phi}(z_j|r_j)$, and trains them using the ELBO loss. The key part of TVAE is a neural network that accurately captures the probability distribution. This network outputs a combined distribution for multiple variables. The approach assumes that variables $\alpha_{i,j}$ follow a Gaussian distribution with unique means and variances, while others ($\beta_{i,j}$ and $d_{i,j}$) follow a categorical probability distribution. By applying TVAE to our KNHANES depression dataset, we were able to generate high-quality synthetic data.

3.4.2 Generative Adversarial Network

To compare the performance of synthetic data generation methods, we evaluated Conditional Tabular Generative Adversarial Networks (CTGAN), introduced by Xu et al. (Xu et al., 2019). CTGAN (Goodfellow et al., 2014), a popular deep generative model, consists of a generator and a discriminator. The generator generates new data from random information, and the discriminator tries to distinguish between real and fake data. The quality of the new data is enhanced by learning from the discriminator's feedback. CTGAN is an improved version of the basic GAN, specifically designed to handle the difficulties of creating synthetic tabular data, which often has a mix of different data types, non-normal distributions, multiple possible values, and unbalanced categorical columns.

CTGAN uses tanh and softmax activation functions on the output to create a combination of categorical and numerical data. The softmax function is defined in Equation (5) as follows:

$$f(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{5}$$

The tanh activation function is given by equation (6) as follows:

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{6}$$

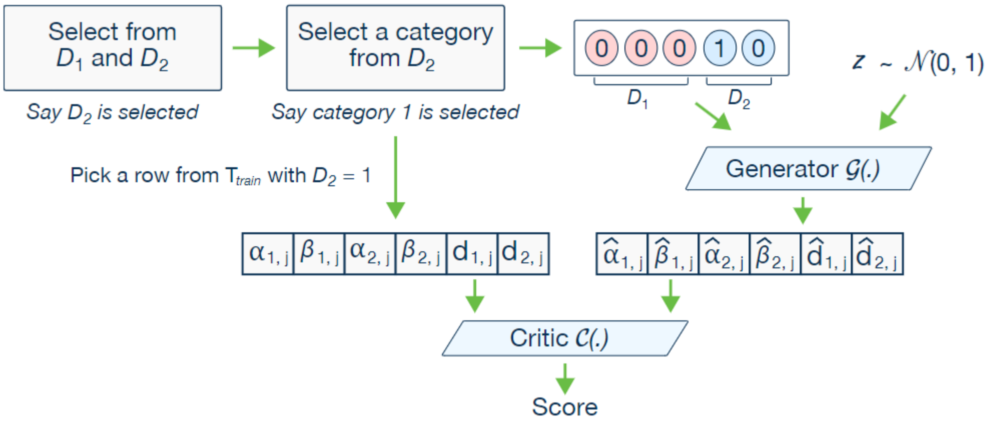
Select from
D_1 and D_2
Say D_2 is selected
Select a category
from D_2
Say category 1 is selected
0 0 0 1 0
D_1
D_2
$z \sim \mathcal{N}(0, 1)$
Pick a row from T_{train} with $D_2 = 1$
Generator $\mathcal{G}(.)$
$\alpha_{1,j}$ $\beta_{1,j}$ $\alpha_{2,j}$ $\beta_{2,j}$ $d_{1,j}$ $d_{2,j}$
$\hat{\alpha}_{1,j}$ $\hat{\beta}_{1,j}$ $\hat{\alpha}_{2,j}$ $\hat{\beta}_{2,j}$ $\hat{d}_{1,j}$ $\hat{d}_{2,j}$
Critic $\mathcal{C}(.)$
Score

Figure 3. Concept of CTGAN illustrated by Xu et al.

Here, $f(x)_i$ and $f(x)$ represents the activation function, where x is the input of the function. Figure 3 illustrates the concept of CTGAN, as introduced by Xu et al. Table 3

Table 3. Training datasets before and after different imbalance handling techniques

Imbalanced Handling Techniques	Number in Class 0 (Non-Depression)	Number in Class 1 (Depression)
None	5003	1113
SMOTE	5003	5003
Tomek Links	4777	1113
SMOTE-ENN	2188	4807
TVAE	4000	4000
CTGAN	4000	4000

SMOTE: Synthetic Minority Oversampling Technique; SMOTE-ENN: SMOTE and Edited Nearest Neighbors; TVAE: Tabular Variational Autoencoder; CTGAN: Conditional Tabular Generative Adversarial Network

3.5 Feature selection

High dimensionality, characterized by a large number of features, can lead to overfitting in machine learning models (Ying, 2019). Overfitting happens when a model becomes too complicated and learns unimportant details instead of the main patterns in the data. This can result in poor generalization to unseen data. Several ML models possess inherent feature selection mechanisms. Random Forest, XGBoost, LightGBM, and Random Forest fall into this category. Conversely, TabNet and FT-Transformer models utilize attention layers to assess feature importance. However, these approaches have limitations, as they may not be able to generalize and can decrease the performance of other models. Additionally, solely relying on accuracy for feature inclusion or exclusion might be inadequate (Ying, 2019). Therefore, a feature selection method that identifies all relevant features, rather than solely focusing on minimal optimal ones, is desirable. Our goal is to gain a complete understanding of the underlying phenomenon, encompassing all contributing factors, not just non-redundant ones. This approach not only helps us to mitigates the risk of overfitting that can arise from a high number of features.

To address the well-documented challenges associated with high-dimensional data, various feature selection techniques have been established (Bolón-Canedo et al., 2013). These methods aim to achieve dimensionality reduction by selecting a more manageable subset of features that effectively capture the essential information within the dataset. These methods generally fall into three categories: wrapper methods, filter methods, and embedded methods. In our study, we aim to employ a feature selection approach that emphasizes a balance between relevance and redundancy across features. Instead of focusing solely on a minimal optimal feature set, we seek a comprehensive selection that includes all relevant contributors to the underlying phenomena. Such an approach provides a holistic understanding of the data's key dimensions, which is essential for interpretability and robustness in clinical applications. This research focuses on implementing a filter method – mRMR (Zhao et al., 2019), (Peng et al., 2005).

The mRMR algorithm operates iteratively, selecting features that are both highly relevant to the target class and exhibit minimal redundancy with previously chosen features. In each iteration, it calculates relevance using metrics like Pearson correlation and redundancy using distance measures. These are then combined into a Max-Relevance Max-Distance (MRMD) score. The feature with the highest MRMD is selected – maximizing relevance and minimizing redundancy – and removed for the next iteration. This process continues until the desired number of features is reached, resulting in a reduced set that retains key information within the dataset. By applying mRMR in this iterative manner, we effectively reduce the dimensionality of our dataset, selecting only those features that balance high relevance and low redundancy. This reduces the risk of overfitting by excluding noisy or redundant features and enhances the model's generalizability across diverse datasets.

Our initial step involved establishing a performance benchmark. We trained and validated all machine learning models using their default hyperparameters and the complete dataset encompassing all 239 features. This process identified the model with the strongest overall performance.

Following model selection, we implemented the mRMR algorithm for feature selection. mRMR iteratively evaluates different feature set sizes. The choice of k, the number of selected features, is a critical aspect in feature selection. Selecting too few features may omit relevant information, while selecting too many can reintroduce redundancy and risk overfitting. To determine the optimal k, we implemented an iterative loop that evaluated various feature set sizes ranging from 10 to 80 features. After selecting the optimal k with mRMR, we retrained and validated our models on the reduced feature sets. By reducing the feature space, we also observed enhanced interpretability, as the final model relied on a more concise set of relevant features that could be further analyzed to understand the clinical predictors of depression in our dataset.

Algorithm 1: Feature selection with mRMR

```
For $k$ in range (20,70)
selected_features = mRMR (features, k) # Select top k features
model = train_model (top_performing_model, selected_features) # Train best default model with selected features
performance = evaluate_model (model, validation_data) # Evaluate performance
record_metrics (k, performance) # Record performance metrics for each k
```

3.6 Classification models

To address the challenges posed by the depression dataset, we employed a variety of advanced ML and DL algorithms. These included state-of-the-art ensemble methods like Random Forest (RF), XGBoost, LightGBM, Gradient Boosting Machine (GBM), and AdaBoost. Additionally, we explored specialized models tailored for tabular data, such as TabNet and the FT-Transformer. A brief overview of these algorithms is provided in the following section.

- RF: Random Forest (Breiman, 2001), an ensemble method that consists of a group of decision trees. Known for its robustness and accuracy, this technique combines the predictions of diverse trees to deliver reliable results, demonstrating resilience against noise and overfitting.
- Gradient Boosting is a boosting ensemble technique, constructs a sequence of improved models. Unlike bagging's focus on diversity, gradient boosting follows a stage-wise approach. Each new model tackles the errors of its predecessor by leveraging gradients, which indicate areas of weakness in the current ensemble's predictions. This iterative refinement is further enhanced in stochastic gradient boosting by using random subsets of training data, leading to improved accuracy and efficiency.
- XGBoost (Chen & Guestrin, 2016) or eXtreme Gradient Boosting, extends gradient boosting by employing a unique regularization term (e.g., L1/L2) and parallel computing to achieve superior accuracy across a diverse range of tasks, including regression, classification, and ranking.
- LightGBM (Ke et al., 2017) builds upon the Gradient Boosting Decision Tree (GBDT) with innovative techniques like Gradient-based One-Side Sampling and the Histogram-based Algorithm. These methods accelerate training time, reduce memory usage, and ultimately enhance the precision of its GBDT model.
- AdaBoost, a breakthrough in boosting algorithms introduced by (Freund & Schapire, 1999), tackles limitations of earlier methods. It excels at combining weak learners (simple models or estimator) into a powerful final classifier. The key lies in strategically weighting data points during training.

Initially, all points have equal weight. As the algorithm progresses, it focuses on previously misclassified data by increasing their weight, ultimately leading to a more robust classifier.

- TabNet (Arik & Pfister, 2021), a robust deep learning model, has demonstrated commendable performance across diverse datasets. The architecture encompasses an encoder module that capitalizes on sequential decision steps to encode features. It employs sparse learned masks to select pertinent features for each row through an attention mechanism. A distinct characteristic of TabNet is its utilization of sparsemax layers, compelling the selection of a compact set of features. This approach deviates from traditional all-or-nothing feature selection, allowing for nuanced decisions via learnable masks. This not only circumvents the rigidity of hard feature thresholds but also enables a soft, differentiable approach to feature selection.
- The FT-Transformer model (Gorishniy et al., 2023) is an adaptation of the Transformer architecture (Vaswani et al., 2023) specifically designed for tabular data. It simplifies the process by transforming all features (both categorical and numerical) into tokens. These tokens are then fed into a series of stacked Transformer layers, where each layer analyzes the features of individual data points. Within the Transformer component, a special [CLS] token is added and processed alongside the other tokens through multiple layers. This pre-normalization step improves optimization and overall performance. Finally, the model uses the final representation of the [CLS] token for prediction. The FT-Transformer, as described earlier and depicted in Figure 4.

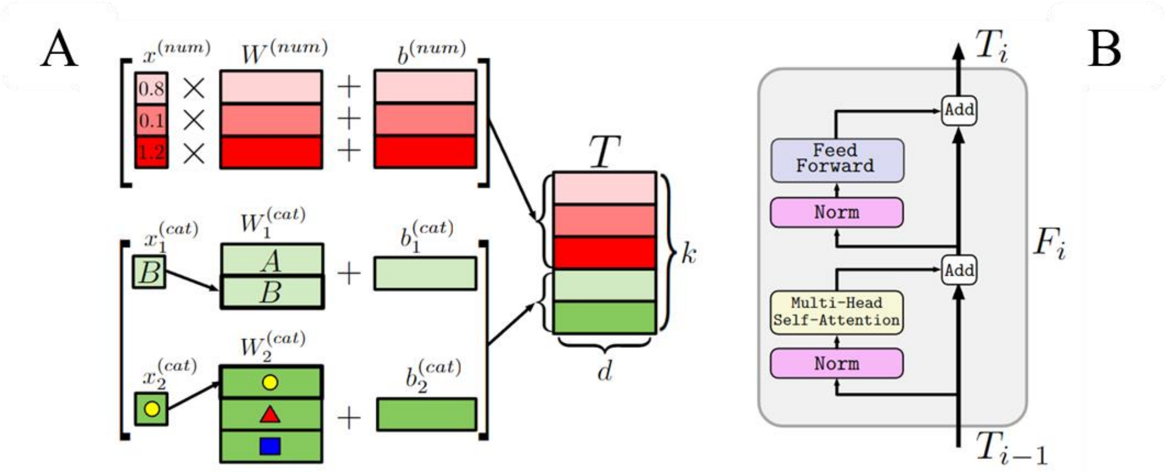

Figure 4. Concept of the FT-Transformer Model. (A) Feature Tokenizer: This part transforms raw features (in this case, two categorical and three numerical) into tokens for the Transformer to process (B) Transformer Layer: This layer analyzes the relationships between the tokens, ultimately using them to make predictions

For comparison, we utilized a baseline model to benchmark the performance of the advanced ML/DL algorithms. Typically, the baseline model for tabular datasets might consist of a simple logistic regression or a decision tree classifier, both of which can serve as reference points for assessing the performance gains of more complex models. The baseline model provides a useful measure of how much improvement is gained by incorporating more sophisticated techniques, such as ensemble methods or deep learning models. Given that the depression dataset likely involves complex interactions between variables and potential non-linear relationships, we expected substantial improvements over these simpler models when utilizing more advanced techniques like AdaBoost, XGBoost, TabNet, and FT-Transformer.

3.7 Fine-tunning Hyperparameters

To optimize the predictive capabilities of each model, we carried out hyperparameter optimization after rebalancing the training subset. This process involved adjusting the hyperparameters to deliver peak performance within the validation set, with the assurance that the test set was not engaged during optimization. The Optuna library (Akiba et al., 2019) facilitated this process, leveraging the efficiency of Bayesian optimization through the Tree-Structured Parzen Estimator (TPE) method. This technique is well-documented for its effectiveness in surpassing the outcomes of random search methods (Turner et al., 2021). A total of 100 iterations were dedicated to each model in pursuit of the best hyperparameter values, with the optimization's objective being the maximization of the F1-score. The hyperparameter search ranges for the respective models are systematically presented in Table 4.

Table 4. Hyperparameter search spaces for each model.

Model	Hyperparameter spaces
LightGBM	num_leaves: [0, 1024], max_depth: [1, 10], learning_rate: [0.0001, 0.1], n_estimators: [8, 1024], class_weight: ['balanced', None], min_child_samples: [10, 50], subsample: [0.7, 1.0], colsample_bytree :[0.7, 1.0], reg_alpha: [0.0, 1.0], reg_lambda: [0.0, 10.0]
XGBoost	num_leaves: [0, 1024], max_depth: [1, 10], learning_rate: [0.0001, 0.1], n_estimators: [8, 1024], class_weight: ['balanced', None], min_child_samples: [10, 50], subsample: [0.7, 1.0], colsample_bytree: [0.7, 1.0], reg_alpha:[0.0, 1.0], reg_lambda: [0.0, 10.0]
AdaBoost	n_estimators: [50, 100], learning_rate: [0.1, 50]
RandomForest	n_estimators: [1, 500], max_depth: [1, 500], min_samples_split: [2, 10], min_samples_leaf: [1, 5], max_features: ['auto', 'sqrt', 'log2'], bootstrap: [True, False], ccp_alphe: [0.01 to 1.0]
GradientBoosting	n_estimators: [100, 5000], learning_rate: [1e-4, 0.3], max_depth: [3, 9], subsample: [0.5, 0.9]
TabNet	n_d: [8, 64], gamma: [1.0, 2.0], cat_emb_dim: [1, 4], n_step: [3, 10], lr: [1e-5, 1e-2]
FT-Transformer	d_token: [64, 128, 256, 512], n_blocks: [1, 4], attention_dropout: [0, 0.5], ffn_d_hidden: [64, 1028], ffn_dropout: [0, 0.5], residual_dropout: [0, 0.2], lr: [1e-5, 1e-2], weigth_decay: [1e-6, 1e-3]

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer

3.8 Performance metrics

The quality of the synthetic depression numerical values was evaluated using a non-parametric Kolmogorov-Smirnov test (KSTest). The KSTest compares the empirical Cumulative Distribution Functions (CDFs) of real and generated samples to assess their similarity. The output for each column is 1 minus the KS Test D statistic, which represents the largest difference between the predicted and observed cumulative distribution function values.

To compare the distributions of two categorical columns, a Chi-Squared test (CS test) was employed. The CS test p-value indicates the likelihood that the two columns were drawn from the same distribution. This metric validates the quality of the synthetically generated categorical columns compared to the real values.

To evaluate the effectiveness of the prediction models, we evaluated the model using a set of performance metrics, including accuracy, precision, recall, and F1-score. These metrics are computed based on the following formulas:

$$Accuracy : \frac{(True_{positive} + True_{negative})}{(True_{positive} + True_{negative} + False_{positive} + False_{negative})} \tag{7}$$

$$Precision : \frac{True_{positive}}{(True_{positive} + False_{positive})} \tag{8}$$

$$Recall : \frac{True_{positive}}{(True_{positive} + False_{negative})} \tag{9}$$

$$F1\ score : 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

where $True_{negative}$ and $True_{positive}$ represent correct predictions for non-depression and depression patients, respectively. $False_{negative}$ and $False_{positive}$ indicate incorrect predictions for these groups.

In addition to the previously discussed performance metrics, we utilized the Area Under the Receiver Operating Characteristic Curve (AUC) as a key metric for evaluating the performance of our prediction models. The AUC score is a robust performance indicator that is not influenced by specific classification thresholds (Sokolova & Lapalme, 2009). In our research, due to the significant class imbalance, the F1-score was prioritized as a primary performance metric. To enhance the F1-score, we will meticulously optimize the model's hyperparameters and explore various feature engineering techniques and augmentation strategies. While AUC will also be considered, it may be less critical due to the potential for overfitting in highly imbalanced datasets.

3.9 Experimental setup

In our research environment, we used Jupyter Lab with Python 3.11.5. The 'sklearn' library was employed to design various ML models such as AdaBoost, LightGBM, XGBoost, and RF, while the 'pytorch' package was instrumental in implementing the TabNet and FT-Transformer models. In addition to these libraries, various tools for hyperparameter optimization, model evaluation, and visualization were used, including matplotlib, seaborn, and pandas for data processing and visualization. All experiments were conducted on a machine CPU (Ryzen 7 7800X3D), 32GB of RAM, and a GPU (NVIDIA 4070Ti) for the DL models, which helped accelerate the training process, especially for models like TabNet and FT-Transformer that benefit from parallel computation.

3.10 SHapley Additive exPlanations

We employed SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017) to determine the significance of various features and contribution to model predictions for depression classification. SHAP considers every possible combination of features to assess how much each feature contributes to the final prediction (Aumann & Hart, 1992). SHAP determines the impact of each feature value on the model's prediction by evaluating all possible combinations of features. This contribution is then weighted and summed up to arrive at a Shapley value:

$$\phi_j(val)\colon \textstyle\sum_{S\subset\{1,\ldots,p\}\{j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{j\} - val(S)) \qquad (11)$$

This calculation considers all possible combinations of features (represented by S) in the model. It then focuses on a specific data point (instance to be explained) with its own set of feature values (represented by the vector x). There are p total features in the model. The part represents the model's prediction when only considering the features in set S, but accounting for the average effect of all other features:

$$val_x(S) : \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x\notin S} - E_X(\hat{f}(X)) \qquad (12)$$

To understand the nuanced effects of features on individual predictions, a SHAP Beeswarm Plot was employed. Each data point was represented by a single point on the plot, positioned along the x-axis according to its SHAP value for a specific feature. The density of points in each feature row indicated the strength of that feature's influence on the model's prediction for that specific feature. Finally, we utilized SHAP Local Waterfall Plots to deconstruct the model's prediction for individual data points. This visualization commenced with the baseline prediction (average prediction on the training set) and sequentially displayed how each feature value in that data point either increased (red) or decreased (blue) the prediction. The SHAP explainer functions were implemented from the SHAP Python module by Slundberg et al. available at https://github.com/slundberg/shap (*Shap/Shap*, 2016/2024).

4 Results

4.1 Results of feature selection process

To establish a baseline performance benchmark, all models were initially trained and validated using their default hyperparameters on the complete dataset of 239 variables. Given the significant class imbalance in the dataset, the F1-score was used as the primary evaluation metric instead of AUC to mitigate the potential for model bias and overfitting. F1-score is a more robust metric for imbalanced datasets as it considers both precision and recall, providing a balanced assessment of model performance. As summarized in Table 5, this preliminary evaluation, conducted prior to feature selection, revealed that the AdaBoost model achieved the highest F1-score of 0.3897 using its default hyperparameters. Consequently, the AdaBoost model was selected as the foundation for further exploration of the optimal number of features (k) using the minimum mRMR feature selection method.

Table 5. Model's performance with default parameter

ML/DL architecture	Accuracy	AUC	Recall	Precision	F1-score
GradientBoosting	**0.8345**	**0.7822**	0.2177	0.6326	0.3228
LightGBM	0.8309	0.7765	0.2304	0.5923	0.3314
AdaBoost	0.8303	0.7744	**0.2978**	0.5663	**0.3897**
RandomForest	0.8260	0.7522	0.0579	**0.8058**	0.1079
XGBoost	0.8250	0.7520	0.2460	0.5421	0.3383
Tabnet	0.8210	0.5137	0.0303	0.6923	0.0581
FT-Transformer	0.8265	0.6113	0.2727	0.5473	0.3640

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer

The relationship between the F1-score and the number of selected features is illustrated in Figure 5. As evident from the figure, the F1-score reaches its peak value of 0.3985 at 36 features. For further exploration, a comprehensive description and list of the 36 selected features are provided in Appendix Table S1.

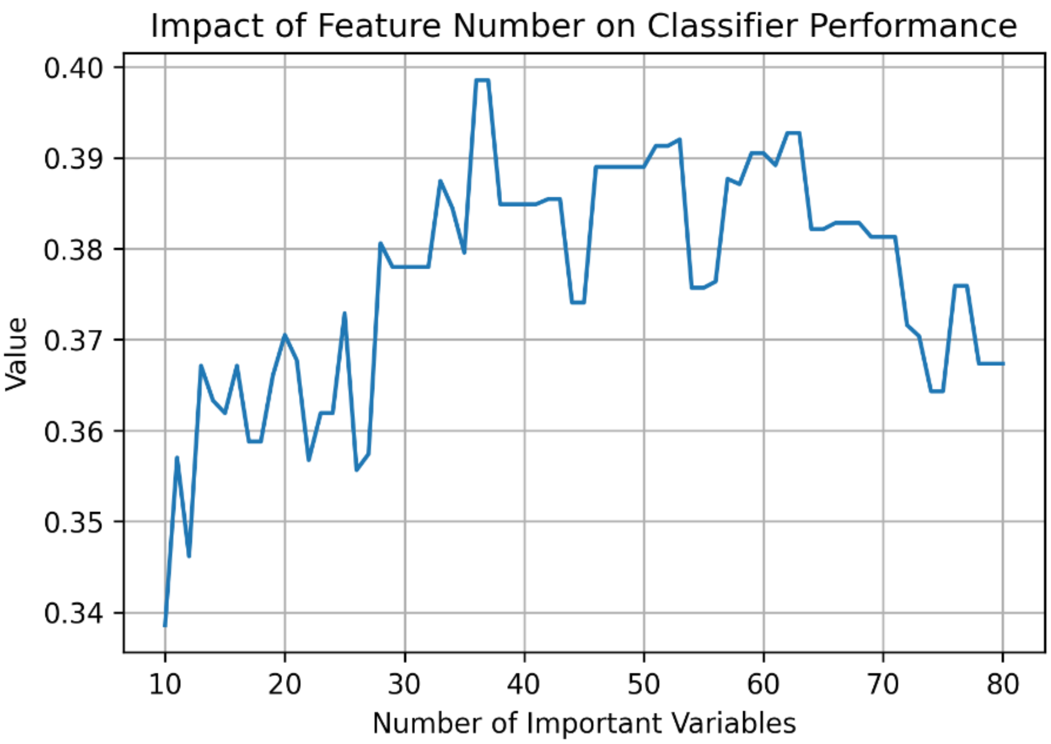

Figure 5. Evolution of F1-Score and Feature Subset Size Using AdaBoost.

Table 6. Model's performance with 36 selected features

ML/DL architecture	Accuracy	AUC	Recall	Precision	F1-score
GradientBoosting	0.8341	0.7840	0.2298	0.6205	0.3343
LightGBM	0.8330	0.7684	0.2803	0.5860	0.3789
AdaBoost	**0.8358**	**0.7876**	0.2857	0.6056	**0.3876**
RandomForest	0.8347	0.7695	0.1772	**0.6793**	0.2792
XGBoost	0.8221	0.7395	**0.2945**	0.5205	0.3758
Tabnet	0.8146	0.5377	0.1024	0.4578	0.1674
FT-Transformer	0.8216	0.5808	0.2020	0.5263	0.2920

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer

Table 6 presents the models' performance after feature selection using the 36 most informative features. The results reveal contrasting trends. Notably, Random Forest exhibited improvements across all evaluation metrics.

Conversely, the AdaBoost model experienced a slight performance decline, with its F1-score dropping from 0.3897 to 0.3876, while the FT-Transformer model experienced a more significant decline, with its F1-score dropping from 0.3640 to 0.2920. However, LightGBM demonstrated a notable increase in F1-score (from 0.3314 to 0.3789) and potentially other metrics. Moreover, LightGBM, Gradient Boosting, XGBoost, and TabNet all exhibited performance gains after feature selection. The observed performance improvements for certain models after feature selection suggest the potential utility of this technique in alleviating overfitting issues commonly encountered in high-dimensional settings, resulting in a more generalizable model.

4.2 Results of data augmentation

Following feature selection, the training data was utilized for data augmentation using various methods, ranging from conventional techniques to synthetic data generation. The quality of the synthetic data generated by TVAE and CTGAN is illustrated in Figure 6, which presents the KS test scores and CS test scores. As shown in Figure 6, the TVAE model demonstrated slightly higher test scores than CTGAN, achieving 86.30% and 76.65% for the KS test compared to CTGAN's 85.54% and 73.54%, respectively.

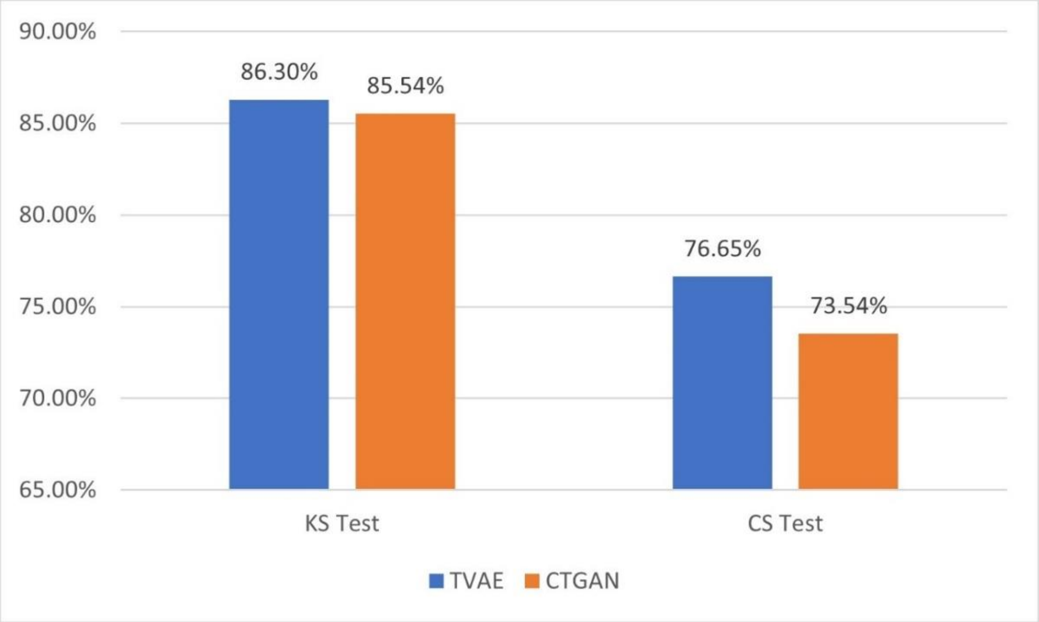

Figure 6. Evaluation of Synthetic Data Quality.

To address the significant class imbalance in our dataset, we employed and evaluated various data augmentation techniques for the machine learning and deep learning models. The effectiveness of these methods is presented in Tables 5 and 6. Regarding conventional data augmentation methods, the ML and DL models were trained with data augmented using different techniques. The classification effectiveness on the validation set is summarized in Table 8.

Table 8. Performance Evaluation of Conventional Imbalance Handling Methods

Imbalance Handling Method	ML/DL architecture	Accuracy	AUC	Recall	Precision	F1-score
	GradientBoosting	**0.8357**	**0.7861**	0.2776	0.6074	0.3808
	LightGBM	0.8318	0.7722	0.2830	0.5773	0.3795
	AdaBoost	0.8239	0.7682	0.3551	0.5251	0.4234
SMOTE	RandomForest	0.8310	0.7737	0.2210	0.5967	0.3224
	XGBoost	0.8207	0.7419	0.2958	0.5137	0.3748
	Tabnet	0.6851	0.5948	0.4528	0.2768	0.3436
	FT-Transformer	0.5506	0.6284	0.7508	0.2528	0.3783
	GradientBoosting	0.8320	0.7837	0.2527	0.5909	0.3534
	LightGBM	0.8283	0.7708	0.2971	0.5531	0.3861
	AdaBoost	0.8354	0.7849	0.3234	0.5879	0.4170
Tomek Links	RandomForest	0.8316	0.7703	0.1860	**0.6276**	0.2862
	XGBoost	0.8164	0.7456	0.3086	0.4930	0.3794
	Tabnet	0.8136	0.5235	0.0674	0.4237	0.1163
	FT-Transformer	0.8234	0.6277	0.3199	0.5249	0.3975
	GradientBoosting	0.6974	0.6903	0.6792	0.3360	0.4496
	LightGBM	0.6925	0.6884	0.6819	0.3320	0.4466
SMOTE-ENN	AdaBoost	0.7008	0.7029	0.7062	0.3434	0.4621
	RandomForest	0.7013	0.6854	0.6604	0.3365	0.4459
	XGBoost	0.6989	0.6839	0.6604	0.3342	**0.4636**
	Tabnet	0.3026	0.5255	0.5255	0.1911	0.3026
	FT-Transformer	0.1821	0.5000	**1.0000**	0.1821	0.3081

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer; SMOTE: Synthetic Minority Oversampling Technique; SMOTE-ENN: SMOTE and Edited Nearest Neighbors; TVAE: Tabular Variational Autoencoder; CTGAN: Conditional Tabular Generative Adversarial Network

The application of SMOTE-ENN for addressing class imbalance within the XGBoost model yielded promising results. It achieved the highest F1-score of 0.4636 among all methods and classifiers evaluated. Furthermore, SMOTE-ENN exhibited superior performance on other machine learning classifiers, including Gradient Boosting, LightGBM, AdaBoost, and Random Forest, compared to other class imbalance handling methods. However, when applied to deep learning classifiers such as TabNet and FT-Transformer, SMOTE-ENN's performance was less effective. Both TabNet and FT-Transformer achieved very low precision scores of 0.1911 and 0.1821, respectively, indicating their inability to accurately predict the minority class. This suggests that the complexity and non-linearity of deep learning models may require alternative approaches or further optimization to effectively handle class imbalance in combination with SMOTE-ENN.

Table 7. Comparison of Classification Performance on CTGAN vs. TVAE Generated Data

Synthesizer	ML/DL architecture	Accuracy	AUC	Recall	Precision	F1-score

CTGAN	GradientBoosting	0.7509	0.66125	0.5067	0.3665	0.4253
	LightGBM	0.7106	0.6345	0.5148	0.3178	0.3930
	AdaBoost	0.7626	0.4932	0.4340	0.3701	0.3995
	RandomForest	0.7420	0.6118	0.4070	0.3304	0.3647
	XGBoost	0.7180	0.6495	0.5418	0.3317	0.4115
	Tabnet	0.5875	0.5572	0.5094	0.2229	0.3101
	FT-Transformer	0.7509	0.6350	0.4528	0.3552	0.3981
TVAE	GradientBoosting	0.7950	**0.6977**	0.5202	0.4457	0.4801
	LightGBM	0.7921	0.6937	0.5499	0.4425	**0.4904**
	AdaBoost	**0.8053**	0.6527	0.3989	**0.4596**	0.4271
	RandomForest	0.7739	0.6952	0.5714	0.4125	0.4791
	XGBoost	0.7876	0.6807	0.5526	0.4343	0.4864
	Tabnet	0.7244	0.5957	0.3935	0.3023	0.3419
	FT-Transformer	0.7249	0.6851	**0.6226**	0.3543	0.4516

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer

The application of SMOTE-ENN for addressing class imbalance within the XGBoost model yielded promising results. It achieved the highest F1-score of 0.4636 among all methods and classifiers evaluated. Furthermore, SMOTE-ENN exhibited superior performance on other machine learning classifiers, including Gradient Boosting, LightGBM, AdaBoost, and Random Forest, compared to other class imbalance handling methods. However, when applied to deep learning classifiers such as TabNet and FT-Transformer, SMOTE-ENN's performance was less effective. Both TabNet and FT-Transformer achieved very low precision scores of 0.1911 and 0.1821, respectively, indicating their inability to accurately predict the minority class. This suggests that the complexity and non-linearity of deep learning models may require alternative approaches or further optimization to effectively handle class imbalance in combination with SMOTE-ENN.

Using a dataset of 8,000 synthetic samples, we evaluated the performance of LightGBM, TabNet, and FT-Transformer models. Our results demonstrate that LightGBM achieved the highest F1-score (0.4904) on the validation set when trained on TVAE-generated data. Notably, deep learning models, including TabNet and FT-Transformer, exhibited significant improvements when trained on synthetic data generated by both TVAE and CTGAN. These models achieved F1-scores of 0.3419 and 0.4516 on TVAE-generated data, respectively, and 0.3101 and 0.3981 on CTGAN-generated data. This performance surpasses the performance of SMOTE-ENN, particularly in terms of precision. Given the positive results obtained with TVAE-generated data, it was used for hyperparameter optimization to further enhance model performance. The hyperparameter details of each classifier are provided in Appendix Table A2 after fine-tuning with Optuna.

4.3 Performance comparison among optimized models

The baseline model, with optimized hyperparameters, was then trained on different synthesizers for evaluation. Figures 7 and 8 present the training and validation accuracy and loss for both TabNet and FT-Transformer. Analyzing the effectiveness of fine-tuning on our base models for predicting individuals with depression reveals valuable insights (Table 9). FT-Transformer achieved the highest F1-score (0.4657) with a recall of 0.5310 and an F1-score of 0.6821 when trained on the TVAE-generated dataset.

Compared to the validation data in Table 7, the hold-out test set results demonstrate an increase in the F1-score for deep learning models but a decrease for other machine learning models. XGBoost, for example, decreased from 0.4864 on the validation set to 0.4549 on the hold-out test set. Similar trends were observed for Random Forest, LightGBM, AdaBoost, and Gradient Boosting, indicating that these models may not perform as well on unseen data. Figure 9 illustrates the performance of the classifiers included in this study, using scores of precisions, recall, F1-score, and accuracy on the depression dataset.

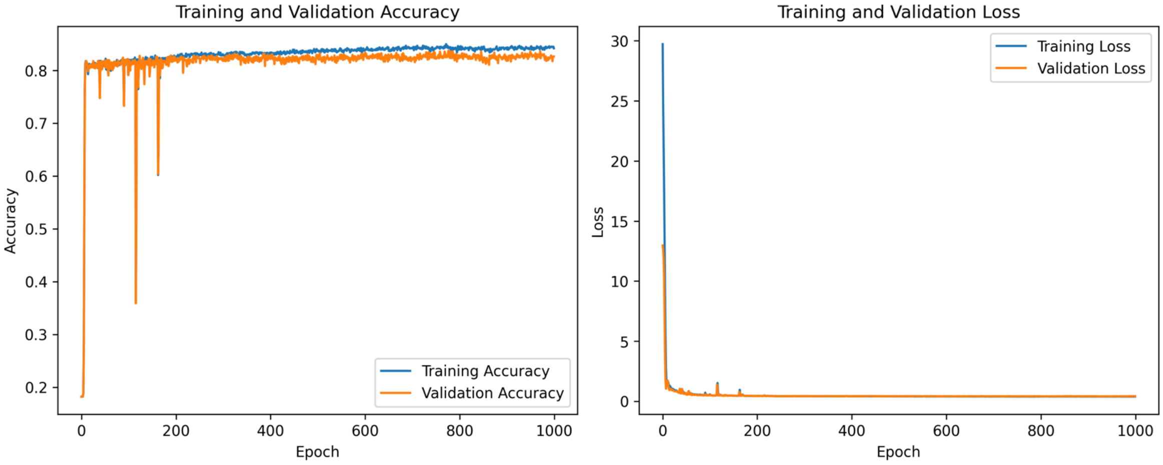

Figure 7. Training and Validation Accuracy, and Loss for TabNet.

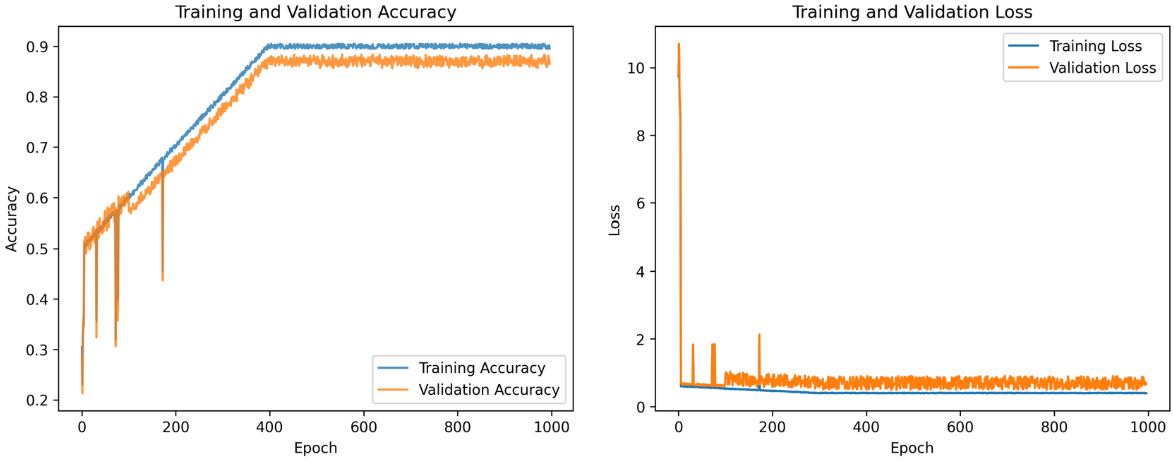

Figure 8. Training and Validation Accuracy, and Loss for FT-Tranformer.

Table 9. Comparison performance of optimized base models.

Synthesizer	ML/DL architecture	Accuracy	AUC	Recall	Precision	F1-score
CTGAN	GradientBoosting	0.7862	0.6367	0.4016	0.4105	0.4060
	LightGBM	0.7675	0.6347	0.4259	0.3771	0.4000
	AdaBoost	0.7818	0.6549	0.4555	0.4102	0.4317
	RandomForest	0.7670	0.6218	0.3935	0.3687	0.3807
	XGBoost	0.7376	0.6437	0.4960	0.3459	0.4075
	Tabnet	**0.8161**	0.5575	0.1509	0.4828	0.2300
	FT-Transformer	0.7749	0.5920	0.3046	0.3599	0.3299
TVAE	GradientBoosting	0.7764	0.6621	0.4825	0.4041	0.4398
	LightGBM	0.7700	0.6666	0.5040	0.3962	0.4437
	AdaBoost	0.8073	0.6412	0.3801	**0.4638**	0.4178
	RandomForest	0.7582	0.6793	0.5553	0.3858	0.4552
	XGBoost	0.7719	0.6751	0.4025	0.5229	0.4549
	Tabnet	0.6827	0.6122	0.5013	0.2870	0.3651
	FT-Transformer	0.7783	**0.6822**	**0.5310**	0.4147	**0.4657**

GradientBoosting: Gradient Boosting Classifier;LightGBM: Light Gradient Boosting Machine;AdaBoost: AdaBoost Classifier; RandomForest: Random Forest Classifier; XGBoost: XGBoost Classifier; FT-Transformer: Feature Tokenizer Transformer

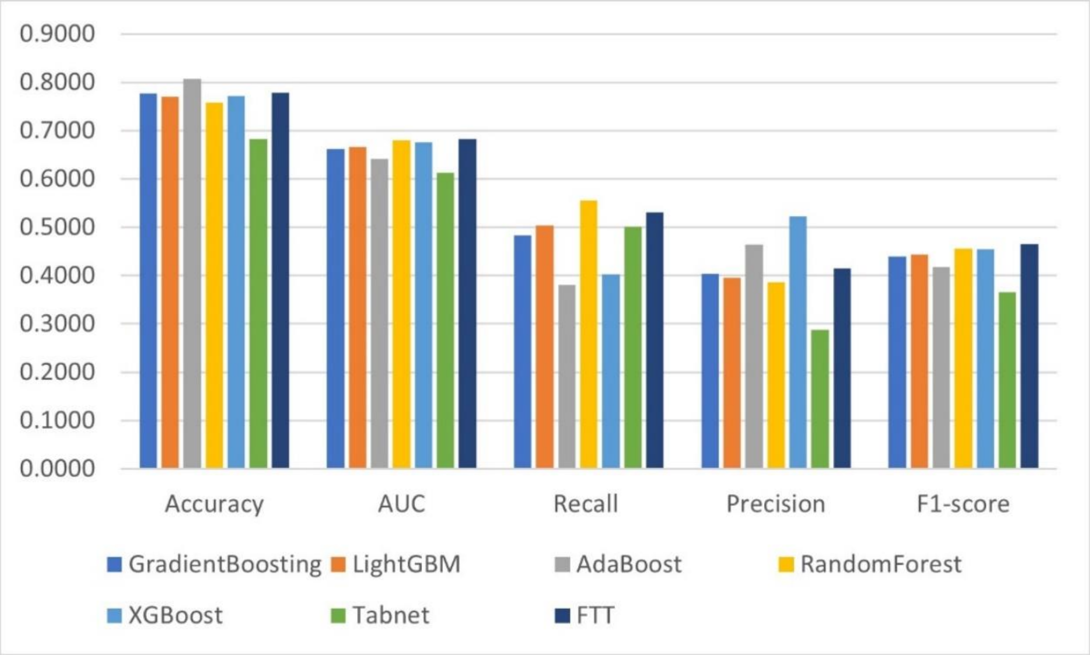

Figure 9. Comparison of evaluation metrics for depression predicting with TVAE generated synthetic data.

Based on the metrics, the FT-Transformer model trained on the TVAE dataset was selected as the best-performing model. The confusion matrix (Figure 10) shows that the model correctly identified 1,390 instances of depression (True Positive, TP) and 197 instances of non-depression (True Negative, TN). However, it also incorrectly classified 278 instances as depression (False Positive, FP) and 174 instances as non-depression (False Negative, FN). While the model achieved a relatively high true positive rate, the significant number of false positives indicates that it may be overly sensitive, potentially leading to unnecessary interventions.

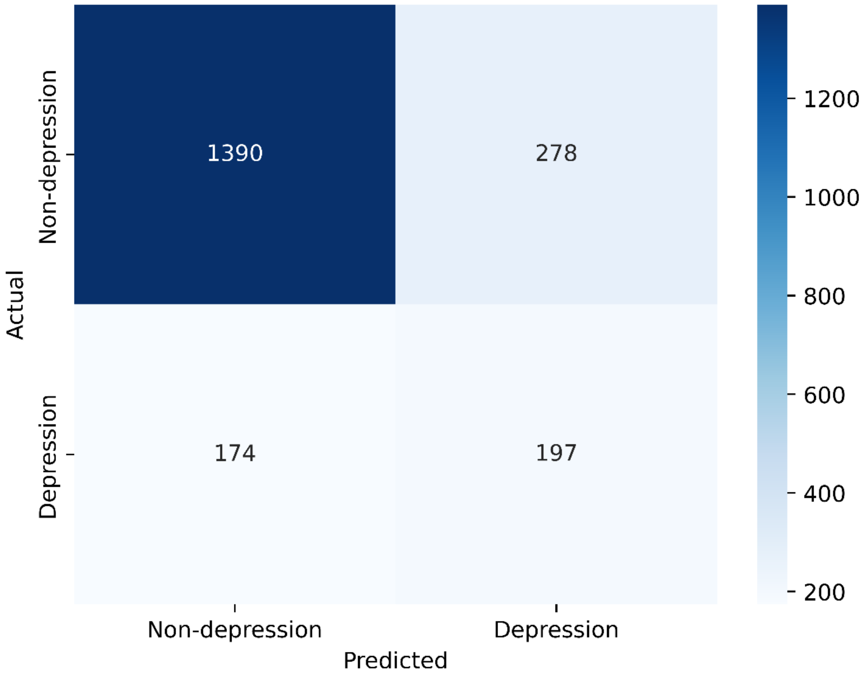

Figure 10. Comparison of evaluation metrics for depression predicting with TVAE generated synthetic data.

4.4 Interpretation of the FT-Transformer Model Using SHAP

To comprehensively illuminate the interpretability of our FT-Transformer model for predicting depression, Figure 11 presents the significance ranking of features employed in this predictive endeavor and their role in influencing the resulting predictions. Figure 12 highlights the key attributes that exert a substantial influence on the predictions concerning suicidal thoughts. Notably, the top 3 features with the highest contribution are "D_1_1" (Subjective health awareness), "BP16_1" (Average sleep time per day during the week or on workdays), and "B01_1" (Weight change over the past year).

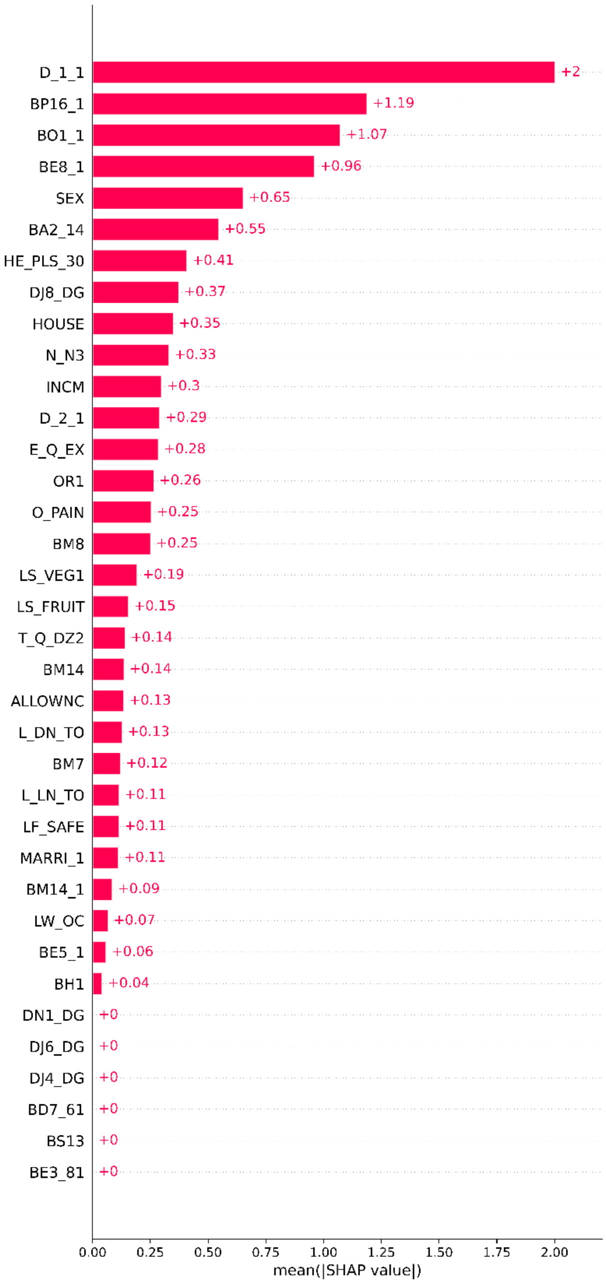

Figure 11. SHAP's global explanation: Global bar plot showing the order of feature importance.

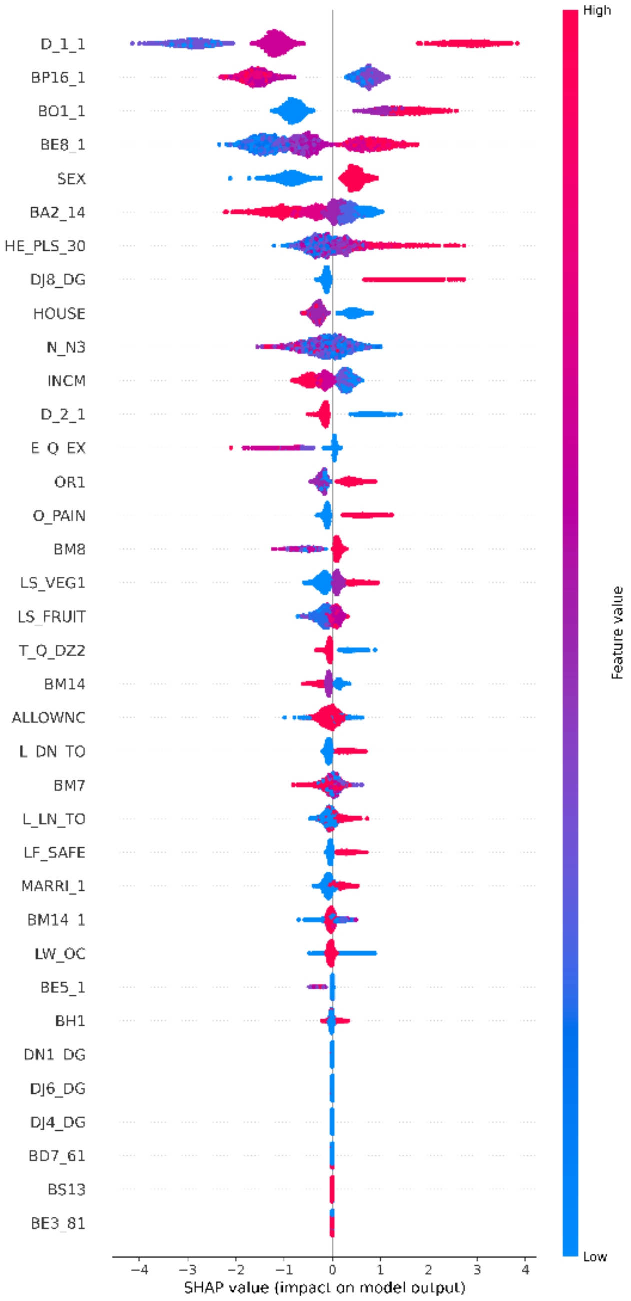

Figure 12. SHAP’s global explanation: beeswarm plot showing how relevant each feature is.

To gain a deeper understanding of the model's decision-making process, we examined specific individuals who were incorrectly classified, as shown in Figure 13. We utilized SHAP's color-coded visualizations to identify the features that significantly influenced the model's prediction for each person. Red hues highlight features that strongly support a Class 1 prediction (indicating depression), while blue hues indicate features aligned with a Class 0 prediction (non-depression). For these individuals, the model predicted class 0 (non-depression), but their actual class was 1 (depression). Among the features, subjective health awareness (D_1_1) appeared to have the most significant influence on the model's incorrect predictions, particularly when rated as "very good".

- Subjective health awareness (D_1_1): Very good (0)
- Average sleep time per day during the week (or on work days) (BP16_1) : 6
- Weight change over the past year (BO1_1): Weight loss (1)
- The amount of time you normally spend sitting down (hours) (BE8_1) : 4
- Gender (SEX): Male (0)
- Daily n-3 fatty acid intake (g) (N_N3): 0.7214 (g)
- Income Quaternary (individual) (INCM): Low (0)

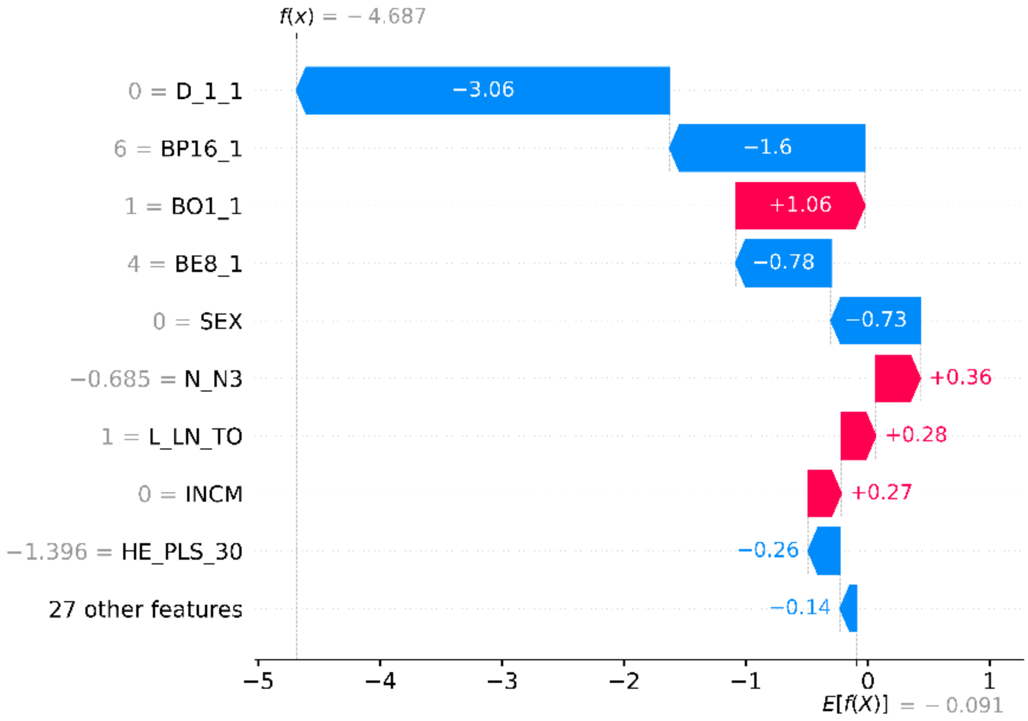

Figure 13. SHAP's local explanation: The initial instance's prediction explained by force plot.

5 Discussion

This study aimed to identify an effective model for predicting depression in South Korea, specifically focusing on improving mental health prediction outcomes. We meticulously evaluated seven individual models: FT-Transformer, TabNet, LightGBM, Random Forest, XGBoost, Gradient Boosting, and AdaBoost. Our findings aligned with previous research, demonstrating that the FT-Transformer model trained on TVAE-generated synthetic data achieved superior performance compared to conventional data augmentation methods like oversampling and under-sampling. Furthermore, we integrated SHAP with the FT-Transformer to gain valuable insights into feature importance. This approach offers a beneficial tool for healthcare professionals experiencing burnout with traditional EHR (Budd, 2023) and data analyst specialists. SHAP provides two-fold benefits: global explanations for analyzing the overall importance of each feature and individual explanations for understanding the rationale behind specific model predictions.

Our research aimed to predict depression using a DL model that incorporated feature tokenization and synthetic data augmentation. Employing the KNHANES dataset (2019-2022), we achieved promising results, with an accuracy of 0.7783, recall of 0.5310, F1-score of 0.4657, and AUC of 0.6822. To situate our findings within the broader research context, we compared them to existing studies.

Kim et al. (N. H. Kim et al., 2024) investigated depression prediction in second-hand smokers using SHAP and the KNHANES dataset (2014-2018). Their study identified quality of life, stress perception, and subjective health status as the most influential factors for depression prediction using a LightGBM model. While the SVM model achieved the highest AUC (0.900), the LGBM model demonstrated the highest positive predictive value of 0.646. However, the F1-score of 0.444 raises concerns regarding overfitting and generalization across groups.

In a study by Lee et al. (Lee et al., 2023), a support vector machine (SVM) model demonstrated superior performance in predicting depression within a cohort of 3,007 individuals. The model exhibited a cross-validated AUC-ROC of 0.835 (95% CI: 0.730-0.901). Key features influencing the model included subjective health status, general stress awareness, and stress recognition rate. While the F1-score of 0.6510 was obtained, incorporating the features "BP1" and "mh_stress" could potentially enhance model performance (Fuchs & Flügge, 2004; van Praag, 2004).

In our research, these features were initially excluded to ensure generalizability. However, compared to these studies, our study may have limitations in terms of performance due to the smaller dataset, lack of diversity, and class imbalance. While TVAE can help balance precision, recall, and F1-score, further research with a larger and more diverse dataset could potentially improve performance. However, compared to these studies, our study has the advantage of excluding biased features that could have boosted algorithm performance. This approach can enhance generalizability across multiple groups of individuals.

Our research underscores the potential of DL techniques with transformer layers, particularly deep neural networks like the FT-Transformer, which is specifically designed to handle tabular data and incorporates a tokenization layer for encoding categorical and numerical features. The FT-Transformer transforms all features (categorical and numerical) into tokens and processes them using a series of stacked Transformer layers. However, when dealing with imbalanced datasets, the performance of our model was suboptimal when trained on conventional augmented data (Table 7). This highlights the necessity of synthetic data generation methods such as CTGAN and TVAE. To our knowledge, the use of synthetic data for training depression prediction models has not been extensively explored in recent research, suggesting a promising avenue for future investigations.

Our research indicates that the TVAE model generates synthetic data of high quality for the KNHANES dataset, as confirmed by rigorous CS and KS tests. Extensive evaluations at various stages demonstrated the effectiveness of TVAE-generated synthetic data. Notably, all ML and DL classifiers achieved significantly improved F1-scores when trained on TVAE-generated data compared to traditional methods.

The integration of TVAE demonstrates its potential to significantly enhance the performance and generalizability of both ML and DL models in medical applications. By utilizing synthetic data generated by TVAE, researchers and organizations can train and evaluate models without compromising patient privacy, addressing critical data privacy concerns (Jordon et al., 2022; Raghunathan, 2021). Additionally, TVAE can address the challenge of small datasets. Our research encourages future studies to focus on improving model performance and developing an interpretable EHR system for depression diagnosis and prediction.

6 Conclusions

This study aimed to identify an effective model for predicting depression in South Korea, with a particular emphasis on improving mental health prediction outcomes. Through a comprehensive evaluation of seven machine learning models—FT-Transformer, TabNet, LightGBM, Random Forest, XGBoost, Gradient Boosting, and AdaBoost—we drew several key conclusions:

1. The FT-Transformer model, trained on TVAE-generated synthetic data, outperformed traditional data augmentation methods such as oversampling and under-sampling.
2. Feature selection techniques, such as mRMR, were found to be effective in reducing noise and improving model generalization. However, overuse of such methods may lead to overly optimistic results on training data, with the model performing poorly on external datasets.
3. Our research highlights the potential of DL techniques, particularly the FT-Transformer, for handling tabular data by effectively encoding both categorical and numerical features through tokenization.
4. Despite challenges faced when training on traditionally augmented data, our study emphasizes the importance of synthetic data generation methods like CTGAN and TVAE for improving the performance of both ML and DL models in medical applications. Additionally, synthetic data addresses privacy concerns, enabling model training without compromising patient confidentiality.

5. The integration of SHAP, both globally and locally, helps to provide a comprehensive understanding of model predictions, offering insights into the overall feature importance as well as the rationale behind individual predictions.

To the best of our knowledge, this is the first study to use TVAE-generated data for depression prediction. Although limitations such as a smaller dataset and class imbalance exist, we suggest that future research should focus on utilizing synthetic datasets, as well as expanding the dataset, to improve predictive accuracy and generalizability. Furthermore, the development of a user-friendly interface for depression prediction will be essential for facilitating practical implementation and enhancing accessibility for healthcare professionals.

Abbreviations

Table 10. List of abbreviations

Abbreviation	Definition
AdaBoost	Adaptive Boosting
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CPU	Central Processing Unit
CS Test	Chi-Squared Test
CTGAN	Conditional Tabular Generative Adversarial Network
DL	Deep Learning
EHR	Electronic Health Record
ELBO	Evidence Lower Bound
ENN	Edited Nearest Neighbors
F1-score	F1 Score (Harmonic Mean of Precision and Recall)
FT-Transformer	Feature Tokenizer Transformer
FT-Transformer	Feature Tokenizer Transformer
GAD-7	Generalized Anxiety Disorder-7
GBM	Gradient Boosting Machine
GPU	Graphics Processing Unit
KCDC	Korea Centers for Disease Control and Prevention
KNHANES	Korea National Health and Nutrition Examination Survey
KS	Kolmogorov-Smirnov (Test)
KSTest	Kolmogorov-Smirnov Test
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
MRMD	Max-Relevance Max-Distance
mRmR	Maximum Relevance Minimum Redundancy
Optuna	Hyperparameter Optimization Framework
PHQ-9	Patient Health Questionnaire-9
RF	Random Forest
SHAP	SHapley Additive exPlanations

Author Contributions

Conceptualization, V.Q. Tran and H. Byeon; software, V.Q. Tran; methodology, V.Q. Tran and H. Byeon; validation V.Q. Tran and H. Byeon; investigation, V.Q. Tran; writing—original draft preparation, V.Q. Tran; formal analysis, B.H; writing—review and editing, V.Q. Tran and H. Byeon; visualization, H. Byeon; supervision, B.H; project administration, B.H; funding acquisition, B.H. All authors have read and agreed to the published version of the manuscript.).

Funding Information

The Basic Science Research Program under the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526), and regional innovation projects based on university-local government cooperation (2021RIS-003) are the sources of support for this research.

Data Availability Statement

This study leveraged data from the Korea National Health and Nutrition Examination Survey (KNHANES) conducted between 2019 and 2022. Data are publicly available through the KNHANES website. For more information, please visit: http://knhanes.cdc.go.kr.

Institutional Review Board Statement

The study was carried out in accordance with the Helsinki Declaration and was approved by the Korea Workers' Compensation and Welfare Service's Institutional Review Board (or Ethics Committee) (protocol code 0439001, date of approval 31 January 2018).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), Article 8. https://doi.org/10.1609/aaai.v35i8.16826

Aumann, R. J., & Hart, S. (1992). Handbook of Game Theory with Economic Applications. Elsevier.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), Article 1. https://doi.org/10.1145/1007730.1007735

Bergman, E., Purucker, L., & Hutter, F. (2024). Don’t Waste Your Time: Early Stopping Cross-Validation (arXiv:2405.03389). arXiv. http://arxiv.org/abs/2405.03389

Bhatt, S., Devadoss, T., Jha, N. K., Baidya, M., Gupta, G., Chellappan, D. K., Singh, S. K., & Dua, K. (2023). Targeting inflammation: A potential approach for the treatment of depression. Metabolic Brain Disease, 38(1), 45–59. https://doi.org/10.1007/s11011-022-01095-1

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and Information Systems, 34(3), 483–519. https://doi.org/10.1007/s10115-012-0487-8

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), Article 1. https://doi.org/10.1023/A:1010933404324

Budd, J. (2023). Burnout Related to Electronic Health Record Use in Primary Care. Journal of Primary Care & Community Health, 14, 21501319231166921. https://doi.org/10.1177/21501319231166921

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), Article 1.

Dharmarathne, G., Jayasinghe, T. N., Bogahawaththa, M., Meddage, D. P. P., & Rathnayake, U. (2024). A novel machine learning approach for diagnosing diabetes with a self-explainable interface. Healthcare Analytics, 5, 100301. https://doi.org/10.1016/j.health.2024.100301

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755

Freund, Y., & Schapire, R. E. (n.d.-a). A Short Introduction to Boosting.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), Article 4. https://doi.org/10.1016/S0167-9473(01)00065-2

Fuchs, E., & Flügge, G. (2004). Cellular consequences of stress and depression. Dialogues in Clinical Neuroscience, 6(2), 171–183.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. BMC Medical Research Methodology, 20(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks (arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2023). Revisiting Deep Learning Models for Tabular Data (arXiv:2106.11959). arXiv. http://arxiv.org/abs/2106.11959

Hwang, Y., & Song, J. (2023). Recent deep learning methods for tabular data. Communications for Statistical Applications and Methods, 30(2), 215–226. https://doi.org/10.29220/CSAM.2023.30.2.215

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022, May 6). Synthetic Data—What, why and how? arXiv.Org. https://arxiv.org/abs/2205.03257v1

Kasthurirathne, S. N., Biondich, P. G., Grannis, S. J., Purkayastha, S., Vest, J. R., & Jones, J. F. (2019). Identification of Patients in Need of Advanced Care for Depression Using Data Extracted From a Statewide Health Information Exchange: A Machine Learning Approach. Journal of Medical Internet Research, 21(7), e13809. https://doi.org/10.2196/13809

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khadidos, A. O., Alyoubi, K. H., Mahato, S., Khadidos, A. O., & Nandan Mohanty, S. (2023). Machine Learning and Electroencephalogram Signal based Diagnosis of Dipression. Neuroscience Letters, 809, 137313. https://doi.org/10.1016/j.neulet.2023.137313

Kim, G. E., Jo, M.-W., & Shin, Y.-W. (2020). Increased prevalence of depression in South Korea from 2002 to 2013. Scientific Reports, 10, 16979. https://doi.org/10.1038/s41598-020-74119-4

Kim, N. H., Kim, M., Han, J. S., Sohn, H., Oh, B., Lee, J. W., & Ahn, S. (2024). Machine-learning model for predicting depression in second-hand smokers in cross-sectional data using the Korea National Health and Nutrition Examination Survey. DIGITAL HEALTH, 10, 20552076241257046. https://doi.org/10.1177/20552076241257046

Kingma, D. P., & Welling, M. (2022). Auto-Encoding Variational Bayes (arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. Journal of General Internal Medicine, 16(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kweon, S., Kim, Y., Jang, M., Kim, Y., Kim, K., Choi, S., Chun, C., Khang, Y.-H., & Oh, K. (2014). Data resource profile: The Korea National Health and Nutrition Examination Survey (KNHANES). International Journal of Epidemiology, 43(1), 69–77. https://doi.org/10.1093/ije/dyt228

Lee, J.-Y., Won, D., & Lee, K. (2023). Machine learning-based identification and related features of depression in patients with diabetes mellitus based on the Korea National Health and Nutrition Examination Survey: A cross-sectional study. PloS One, 18(7), e0288648. https://doi.org/10.1371/journal.pone.0288648

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91–99. https://doi.org/10.1016/j.gltp.2022.04.020

Mahato, S., Paul, S., Goyal, N., Mohanty, S., & Jain, S. (2022). 3EDANFIS: Three Channel EEG-Based Depression Detection Technique with Hybrid Adaptive Neuro Fuzzy Inference System. Recent Patents on Engineering, 17. https://doi.org/10.2174/1872212117666220801105612

McGorry, P. D. (2015). Early Intervention in Psychosis: Obvious, Effective, Overdue. The Journal of Nervous and Mental Disease, 203(5), 310. https://doi.org/10.1097/NMD.0000000000000284

McGorry, P. D., Hickie, I. B., Yung, A. R., Pantelis, C., & Jackson, H. J. (2006). Clinical Staging of Psychiatric Disorders: A Heuristic Framework for Choosing Earlier, Safer and more Effective Interventions. Australian & New Zealand Journal of Psychiatry, 40(8), 616–622. https://doi.org/10.1080/j.1440-1614.2006.01860.x

Mohanty, S. N., Satpathy, S., Chopra, R., & Mahato, S. (2024). Investigating the impact of Mahā Mantra chanting on anxiety and depression: An EEG Rhythm Analysis Approach. Advances in Integrative Medicine, 11(2), 74–83. https://doi.org/10.1016/j.aimed.2024.04.003

Naga Srinivasu, P., Ijaz, M. F., & Woźniak, M. (2024). XAI-driven model for crop recommender system for use in precision agriculture. Computational Intelligence, 40(1), e12629. https://doi.org/10.1111/coin.12629

Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Scientific Reports, 11(1), 1980. https://doi.org/10.1038/s41598-021-81368-4

Nickson, D., Meyer, C., Walasek, L., & Toro, C. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: A systematic review. BMC Medical Informatics and Decision Making, 23(1), 271. https://doi.org/10.1186/s12911-023-02341-x

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. Frontiers in Psychology, 10. https://doi.org/10.3389/fpsyg.2019.02970

Otte, C., Gold, S. M., Penninx, B. W., Pariante, C. M., Etkin, A., Fava, M., Mohr, D. C., & Schatzberg, A. F. (2016). Major depressive disorder. Nature Reviews Disease Primers, 2(1), 1–20. https://doi.org/10.1038/nrdp.2016.65

Patient Health Questionnaire-9 (PHQ-9)—Mental Health Screening—National HIV Curriculum. (n.d.). Retrieved September 4, 2024, from https://www.hiv.uw.edu/page/mental-health-screening/phq-9

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Półchłopek, O., Koning, N. R., Büchner, F. L., Crone, M. R., Numans, M. E., & Hoogendoorn, M. (2020). Quantitative and temporal approach to utilising electronic medical records from general practices in mental health prediction. Computers in Biology and Medicine, 125, 103973. https://doi.org/10.1016/j.compbiomed.2020.103973

Raghunathan, T. E. (2021). Synthetic Data. Annual Review of Statistics and Its Application, 8(Volume 8, 2021), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Sau, A., & Bhakta, I. (2017). Predicting anxiety and depression in elderly patients using machine learning technology. Healthcare Technology Letters, 4. https://doi.org/10.1049/htl.2016.0096

Shap/shap. (2024). [Jupyter Notebook]. shap. https://github.com/shap/shap (Original work published 2016)

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2017.2767063

Shin, C., Kim, Y., Park, S., Yoon, S., Ko, Y.-H., Kim, Y.-K., Kim, S.-H., Jeon, S. W., & Han, C. (2017). Prevalence and Associated Factors of Depression in General Population of Korea: Results from the Korea National Health and Nutrition Examination Survey, 2014. Journal of Korean Medical Science, 32(11), 1861–1869. https://doi.org/10.3346/jkms.2017.32.11.1861

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Statistical Learning Theory | Wiley. (2024, June 14). Wiley.Com. https://www.wiley.com/en-us/Statistical+Learning+Theory-p-9780471030034

Sulmasy, L. S., López, A. M., Horwitch, C. A., & American College of Physicians Ethics, Professionalism and Human Rights Committee. (2017). Ethical Implications of the Electronic Health Record: In the Service of the Patient. Journal of General Internal Medicine, 32(8), 935–939. https://doi.org/10.1007/s11606-017-4030-1

Thapa, C., & Camtepe, S. (2021). Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in Biology and Medicine, 129, 104130. https://doi.org/10.1016/j.compbiomed.2020.104130

Two Modifications of CNN. (1976). IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(11), 769–772. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/TSMC.1976.4309452

van Praag, H. M. (2004). Can stress cause depression? Progress in Neuro-Psychopharmacology and Biological Psychiatry, 28(5), 891–907. https://doi.org/10.1016/j.pnpbp.2004.05.031

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN (arXiv:1907.00503). arXiv. http://arxiv.org/abs/1907.00503

Yasin, S., Othmani, A., Raza, I., & Hussain, S. A. (2023). Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and EEG modalities: A comprehensive review. Computers in Biology and Medicine, 159, 106741. https://doi.org/10.1016/j.compbiomed.2023.106741

Ying, X. (2019). An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series, 1168(2), Article 2. https://doi.org/10.1088/1742-6596/1168/2/022022

Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021). Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. Journal of Affective Disorders, 279, 1–8. https://doi.org/10.1016/j.jad.2020.09.113

Zhao, Z., Anand, R., & Wang, M. (2019). Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform (arXiv:1908.05376). arXiv. http://arxiv.org/abs/1908.05376