Depression Classification with Imbalanced Data Problems: Literature Survey

Artitayaporn Rojarath[1], Wararat Songpan[2] and Olarik Surinta[1,*]

[1] Multi-agent Intelligent Simulation Laboratory (MISL) Research Unit
Department of Information Technology, Faculty of Informatics, Mahasarakham University
Khamriang Sub-District, Kantarawichai District, Mahasarakham 44150, Thailand

[2] Department of Computer Science, College of Computing, Khon Kaen University
Nai Muang sub-District, Muang District
Khon Kaen, THAILAND, 40002

*Corresponding Email : olarik.s@msu.ac.th

Received June 19, 2024, Revised September 21, 2024, Accepted October 15, 2024, Published June 27, 2025

Abstract. *Depression is an increasingly serious global mental health concern, with the number of affected individuals rising steadily. In Thailand, more than 70% of the working-age population is at risk of developing depressive conditions, as reported by the Thai Depression Center. A significant challenge in depression research is the issue of imbalanced datasets, where the number of depressive cases (minority class) is significantly lower than non-depressive cases (majority class). This imbalance often results in biased classification models that favor the majority class, thereby reducing the accuracy and effectiveness of depression classification. This literature survey addresses critical gaps in the field by focusing on the imbalanced data problem in depression classification. While previous studies have primarily relied on traditional oversampling and undersampling techniques, these approaches often intensify the problem of overfitting and lead to the loss of valuable information. Our research explores these issues by reviewing various resampling methods, with a particular emphasis on advanced oversampling techniques that aim to preserve data integrity while mitigating overfitting. The survey also presents a comparative analysis of evaluation metrics, including accuracy, precision, recall, F1-score, and AUC, to provide a more nuanced understanding of classifier performance in the context of imbalanced data. Our findings indicate that while oversampling methods are generally effective, careful implementation is essential to avoid overfitting, which can distort the predictive accuracy of the model.*

Keywords:

Depressive classification, Imbalanced data, Resampling method, Oversampling technique, Machine learning

1. Introduction

Currently, depression is a frequent and increasingly serious mental health problem worldwide. The many factors that can make people suffer from depression, such as psychological health, genetic disorders, and even a negative opinion of their surroundings, are increasing rapidly. These factors may be the reason that depression has become a serious health condition [1]. The World Health Organization (WHO) has revealed that more than 264 million people worldwide suffer from depression, equivalent to almost 4% of the population. In Thailand, the Department of Mental Health has reported that more than 1.5 million Thai people aged 15years or more suffer from depression, and according to World Health Statistics [2], almost 70% of Thai people die early. In-depth medical research of Zhang et al. [3] identified depression risk factors which included gender, age, sleep, exercise, stress, medical conditions, and negative life occurrences.

Furthermore, several studies have identified the causes of depression, including the study of Yazdaver et al. [4] in which the researcher mentioned that social media behavior might directly cause depression. When people use social media improperly, it might cause an increase in depression, resulting in stress and violence through both speech and actions. Depression patients sometimes compare their lives to others and then share negative comments through social media platforms (i.e., Facebook, Instagram, and Twitter). Haand and Shuwang [5] mentioned that people would be able to openly recognize the lives of others as their access to social media increases, which is why they are comparing their lives to others and thinking negatively about themselves and their surroundings. As a result, many negative feelings and opinions are expressed on social media. For this reason, comments and posts published on social media by depressed people could be analyzed to predict distress or a trend of suffering from depression. Therefore, the data from social media was collected so as to analyze the associated factors that are derived from data classification which were identified with depression. Further, the operations for data collection pertaining to depressive disorder must be managed before constructing a model based on the CRISP-

DM process, including data understanding, data preparation, and data transformation. These processes must be accomplished before initiating the model creation process in order to achieve a diagnosis of depression precision [6], [7].

In recent years, machine learning algorithms have been proposed to address mental illness problems. [8], [9], [10], [11]. Hence, precision models can assist in classifying depressed patients [1], [12]. However, Sanchez-Carro et al. [12] were interested in solving the imbalanced data problem because depression data is generally inadequate and affects the precision of the machine learning models. Many studies have employed two classification methods. individual machine learning and ensemble classification, to classify the depression data into two main classes: normal (majority class) and depressive (minority class) patients [13], [14]. Classification using individual machine learning might not succeed in the precision result. Further, the ensemble learning algorithm combines the results of each machine learning model and classifies them as the final results in order to enhance classification performance [15]. However, the classification results when using machine learning and ensemble learning methods still achieved only a low precision value regarding the imbalanced data of the relevant clinical data [16], [17]. In practice, if the patient is depressed, but the predictive result of the classification model classifies them as not being depressives, then, the diagnosis results will have a negative effect on the patient, who will not be treated properly and in a timely manner.

The primary objective of this survey research is to explore methods or processes that can enhance the precision of classifying between normal and depressive patient classes in a depression dataset. The depression dataset of interest in this study is characterized as an imbalanced dataset. A review of previous research indicates that optimizing the dataset significantly impacts the accuracy of classifying outcome classes. Especially, research into mental health data sets within medicinal databases. However, many studies have not placed significant emphasis on dataset optimization, instead focusing on parameter tuning, experimenting with various methods to find the most effective approach for the data, or primarily on data cleaning. This research aims to fill this gap by focusing on addressing the issue of dataset imbalance to improve the accuracy of outcome class classification. The research employed resampling techniques, specifically focusing on the use of oversampling methods, to address this data imbalance problem.

The main contributions of the survey research are as follows.

- The authors are interested in medical data, which frequently suffers from imbalanced problems within the medical data group. If the dataset is imbalanced, there will be difficulties classifying incorrect data. Therefore, the authors were interested in resampling techniques as a solution for imbalanced datasets.
- In this survey, the authors intended to employ an oversampling technique, which is one of the resampling methods, to address the problems of imbalanced data. The oversampling techniques randomly increase in a minority class to an equal amount as the majority class. The authors are interested in methods to increase minority classes to reduce the risk of losing crucial data when employing techniques that require reducing the number of majority classes.

This research is structured as follows: Section 2 describes resolving imbalanced data for depressive disorder, including undersampling and oversampling techniques. Section 3 includes discussing relevant research, particularly on managing imbalanced datasets, and examining various techniques used to address this issue. Section 4 provides examples of datasets relevant to the topic of interest, specifically datasets related to depression. It then presents the techniques used by each method for adjusting imbalanced datasets. The data preprocessing process, the algorithm for resolving imbalanced data, data construction, and model evaluation are described in Section 5. Section 6 provides a comparison of experimental results obtained from various datasets, based on several metrics. In Section 7, the discussion focuses on classification using depression-related data from various studies, aiming to reach a clear conclusion on the research topic.

2. Imbalanced Data Problems

Data imbalance is a frequent problem encountered in medical datasets. In many cases, the proportion of individuals diagnosed with a disease is significantly lower than that of the general population [18], [19]. When comparing the two outcomes of disease diagnosis, the majority of medical data tends to cluster around the group of non-diseased individuals. Most of the diagnostic results are classified into two categories: disease and non-disease. The category of diagnostic data is known as ‘class labels’ including such designations as Patient or Healthy, Depression or Non-depression, and Yes or No, which have different numerical data distributions. The analysis of numerous studies revealed significant differences between classes, in which a minor fraction of the population was diagnosed with the disease. Consequently, this might result in data misclassification problems or imbalanced data. A recent study that utilized diagnostic data for testing to develop models employed machine learning techniques to improve imbalanced data [20], [21].

This survey is focused on imbalanced data, which can yield biased classification. Consequently, machine learning models may only prioritize the majority class [22]. Hence, imbalanced data classification is proposed to solve abnormal data distribution, where one class has a significantly larger number than the other [23]. Based on clinical data collected from many patients receiving treatment for disease risk factors, it appears that only a small number of patients were diagnosed with depression. As a result, the clinical data emerging was frequently imbalanced [24]. When evaluating clinical data related to depression, the machine learning model requires avoiding bias and considering only the majority class. The model should also take into account the minority class and provide high precision for both classes.

Shi et al. [15] addressed the problem of imbalanced data and explained that classifying imbalanced data often leads to results biased toward majority classifications, which makes it challenging to predict class results accurately. Gao et al. [25] also explained that minority classes are important classes requiring accurate classification. Incorrectly predicting a minority class has a more severe impact than predicting a majority class. If patients with depression are predicted not to have depression, they may not receive the necessary treatment on time and could suffer severe consequences [16].

A proposed method to address imbalanced data is called the resampling method, which adjusts the number of instances in every class to be comparable. Pereira, Costa, and Silla Jr. [26] mentioned that resampling methods are the most popular and commonly employed for solving imbalanced data. The resampling methods consist of both undersampling and oversampling. Undersampling reduces the number of instances in the majority class. Further, the undersampling technique is suitable for handling large amounts of data. However, the oversampling technique creates new instances in the minority class [27]. Furthermore, increasing the data by repeating random instances from the minority class might lead to overfitting.

3. Literature Review on Imbalanced Data Classification

Numerous research studies have proposed classification methods for depression, which is often characterized by imbalanced data [24], [28], [29]. Addressing data imbalance is crucial, particularly in cases where the objective is to identify rare but highly significant events, such as diagnosing rare diseases or serious conditions that require timely treatment. This research study focuses on depression data that reveals characteristics of imbalanced data. In the context of imbalanced datasets, the classification involves identifying the minority class of interest [30]. In medical data, this minority class is referred to as the positive class, which represents the group of patients diagnosed with a particular condition, such as the patient class, depression class, or heart failure class. Imbalanced data refers to a dataset where the distribution of result classes is highly unequal. In such datasets, one class, known as the majority class, contains a significantly larger number of samples compared to the other class. This majority class typically represents individuals who are not diagnosed with the condition [31]. This results in the model's inability to effectively learn the characteristics of the minority class.

Typically, conventional classification methods are designed to maximize overall accuracy, making them appropriate for balanced datasets where the result classes are equal or nearly equal [32], [33]. Evaluating model performance using only accuracy can provide a comprehensive assessment of overall effectiveness. Generating a model utilizing datasets with balanced or nearly balanced class distribution decreases bias resulting from imbalanced data and enhances prediction accuracy. Consequently, the performance of the model is frequently decreased when generated using imbalanced datasets. Adjusting imbalanced datasets requires the appropriate analysis and techniques to ensure that the model effectively classifies the minority class [23], [30]. Therefore, selecting appropriate performance metrics is crucial for evaluating the effectiveness of models generated from imbalanced datasets. Relying on general metrics such as accuracy value can be misleading, as the value may not accurately represent the actual performance of the model. This characteristic of the model involves an abnormal class classification pattern, where the model primarily learns to predict the majority class due to its higher accuracy in that class [34].

This research focuses on classifying depression using both binary and multi-class classification methods. In addition, we review research studies related to the problems of generating classification models from imbalanced data. This problem leads the model to favor predicting the majority class, resulting in inaccurate predictions for the minority class. In addition to reviewing research on the characteristics of imbalanced datasets, we also examined methods for balancing the data using resampling techniques. We focus on resampling methods to resolve the problem of imbalanced class distribution. Two widely used techniques are oversampling and undersampling, both of which are popular for adjusting imbalanced dataset problems Li D et al. [18] as demonstrated in the following research studies.

Xin and Rashid [24] studied the prediction of depression in Malaysian women using the random forest (RF) approach and then addressed the imbalanced data problems using the oversampling technique. The research demonstrated a comparison of the imbalanced ratio of the Minor class, which is the Positive class, with an imbalanced ratio of 1:10. Due to the tiny size of the used dataset, the SMOTE technique (Randomizing technique of adding data to each class) was utilized to achieve data balance. It was found that the RF model obtained predictions that were 95% more accurate than the imbalanced dataset. By using the oversampling technique, it was possible to predict depression in women in Malaysia accurately. After adjusting the dataset, the distribution of the imbalanced ratio decreased to 1:2.3.

Asare et al. [35] experimented with emotional ranking using digital biomarkers, and the dataset used was raw patient data from the study by Moshe et al. [36]. The researchers employed an oversampling technique to augment data on individuals with depressive symptoms by having them wear intelligent wearables. So, the wearable device was connected to a smartphone and kept track of sleeping patterns and physical activity and also utilized the mobile device to acquire GPS location data. The emotional ranking proposed to investigate the distinctions between 8 people with depressive and 46 people with non-depressive groups based on imbalanced clinical data. The SMOTE method was employed to balance the data. The experimental results showed that the machine learning model achieved high accuracy when the data was balanced. The result indicated that the XGBoost model achieved the highest accuracy of 81.43%.

Karima and Anggraeni [37] conducted research on risk classification for hypertension using an imbalanced dataset. They employed the Adaboost method and utilized the relief feature selection algorithm to identify the top five most relevant attributes. The dataset was a binary class, consisting of Normal and Hypertension, with an imbalanced class distribution for hypertension patients. The data originated from hypertension cases at Bumi Makmur Community Health Centers, Tanah Laut Regency, South Kalimantan, Indonesia. In this study, the imbalanced dataset was addressed using resampling technique with class weights to enhance the effectiveness of data adjustment. A random resampling technique was employed, using the oversampling method to increase the Hypertension class to balance with the Normal class. The results from the Adaboost method, which were compared with the performance of the Naive Bayes (NB) method, indicate that feature classification was stable and highly effective. The model achieved a Precision of 0.826, a Recall of 0.820, an F1 score of 0.819, and an AUC of 0.891.

Zuo et al. [38] examined stroke datasets, which are classified as imbalanced data. The stroke screening data reveals the problem of imbalance, as the proportion of stroke patients is significantly smaller compared to the total population screened. This research outlines a data preprocessing approach, introducing the MICFS algorithm for feature selection to identify the primary risk factors for stroke based on stroke screening data. As stroke is a statistically rare event, the data from stroke screening has features of class imbalance. The research employed the MRF-SMOTE algorithm, an oversampling technique that integrates MAHAKIL Random, a method within semi-supervised learning, with SMOTE. This approach aims to enhance the quality of the synthetic minority class samples. Subsequently, a classification model is constructed using Deep Reinforcement Learning based on the Dueling Deep Q-Network (DQN) algorithm for stroke classification. It was found that the accuracy, AUC, precision, and F1-measure values were 0.8982, 0.96, 0.899, and 0.8981, respectively.

Dengao Li et al. [18] investigated the diagnosis of heart failure using image data, specifically chest X-ray images (CXR images). The massive number of CXR images imposes a significant burden on physicians and contributes to the problems associated with data imbalance. In their research, they utilized the publicly accessible CheXpert dataset, which is intended for multi-class classification. The researchers employed a random under-sampling method to address the problem of excessive CXR images. However, under-sampling can lead to the loss of important and valuable data, limiting the ability to fully learn the distinctive characteristics of each class. To address the problem of losing critical data, their research proposed a method that combines under-sampling techniques with instance selection to maintain the completeness data distribution. They also employed an end-to-end multi-level classification approach to assist physicians in diagnosing the specific causes of heart failure. The experiments showed that combining both techniques for data balancing enhanced the average accuracy by 3.78% compared to the traditional random under-sampling method, achieving an accuracy rate of 84.44%.

The technique of imbalanced data classification can be categorized into two techniques, which are undersampling and oversampling, as explained below.

A. Undersampling Techniques.

Undersampling techniques aim to balance the number of instances by focusing on the majority class. They select all the instances that belong to the minority class and randomly select instances from the majority class. As a result, instances from the majority class are eliminated to balance the distribution of each class [26], [39]. These same researchers also discussed imbalanced datasets, in which improvement techniques divided data into three levels: data level, algorithm level, and cost-sensitive level. The problem of class distribution was resolved at the data level. Resampling techniques were used to address imbalanced class distribution to achieve the data balance at the data level. Pereira, Costa, and Silla Jr. [26], and Seng, Kareem, and Varathan [40] proposed a novel undersampling technique to yield the undersampling problem called neighborhood undersampling (NUS). Further, Hoyos-Osorio et al. [41] proposed a relevant information-based undersampling (RIUS) method to select the most relevant instances from the majority class; the RIUS method effectively minimized lost data by selecting the most relevant instances. The problem of lost data can be solved with the undersampling technique, and although this method usually decreases the quantity of majority class data, it potentially results in the loss of critical data.

Moreover, Ren et al. [42] mentioned the large-scale problem of imbalanced data using the equalization ensemble method (EASE) to reduce the majority class samples by minimizing the essential data loss of the majority class. They created a new dataset by adjusting the larger-scale data in the majority class sample into equal parts, achieved through bin-based equalization using the undersampling method. The data was adjusted to reduce the possibility of selecting unusual sample data, such as noise or an outlier. Then, the base model was aggregated through the weighted averaging procedure to combine the models. The model aggregation process used G-mean scores obtained from base classifiers evaluated on the initial imbalanced data as weighted values. The experimental findings indicated that the EASE method performed better than baseline models.

B. Oversampling Techniques.

The key concept underlying oversampling is to address the problems of minority classes by creating additional instances that are equivalent to the majority class. In this survey research, we addressed the issue of oversampling, which can create new noisy instances that overlap with those in the majority class. Liu et al. [43] proposed a noise-robust oversampling algorithm called a noise-robust oversampling algorithm for mixed-type and multi-class imbalanced data (NROMM) to address the problem of noisy instances. NROMM technique effectively eliminates the noise that arises while generating new instances of the minority class. The noise-robust oversampling algorithm is designed based on grouping-based oversampling to generate new instances from the overlapping portion. Most oversampling methods operate inefficiently when the data contains noise and unusual distribution data. In order to solve this problem, Wei et al. [44] proposed an Improved and random synthetic minority oversampling technique (IR-SMOTE) to reduce noisy instances and balance the distribution of each class.

The oversampling technique is an approach to considering a minority class. The approach adopts the principle of increasing the quantity of data in the minority class to equal that of the majority class. As a method for resolving issues with imbalanced data, oversampling generates additional samples of minority data to obtain a more balanced data distribution. The problem of class imbalance is caused by the unequal distribution of data, as mentioned by Liu et al. [43]. In addition, the authors also discussed oversampling problems that often overlap with the majority class, resulting in noisy data from new minority samples. To solve this problem, the researchers proposed a noise-robust oversampling algorithm for imbalanced data that eliminates the noise occurring in minority samples, and the experimental results show that the proposed approach reduces noisy data and effectively overlaps with majority classes. According to Ren et al. [42], the solution to overlap results from the oversampling technique can be used to resolve the distribution of imbalanced data classes. The common oversampling techniques frequently result in class overlaps due to improper sample selection. Therefore, a new overlapping technique based on grouping-based oversampling was employed to generate new samples away from the overlapping region in which the grouping for the minority class sample was performed first. Moreover, the research of Wei et al. [44] discussed the problem of oversampling when trying to solve the imbalanced data; to handle this problem, they proposed a new oversampling technique known as IR-SMOTE to reduce the noisy data and construct a balanced distribution of data classes.

4. Depressive Datasets

Depression datasets collect data from several sources, such as the Children's Depression Inventory (CDI), the Hospital Anxiety and Depression Scale (HADS), the Patient Health Questionnaire (PHQ-9), data from expressions of emotions and sentiments via social media, and signal measurement using wearable sensors. The data collected usually includes symptoms, environment, emotional state, and chronic illness. The following represents an example of a depression-related dataset:

A. DAIC-Woz Dataset.

The dataset consists of image file data for diagnosing mental distress conditions such as depression, anxiety, and post-traumatic stress disorder. The image files containing these data show the facial patterns of depressed people, with the data distribution being characteristic of an imbalanced data set. The dataset consists of two classes: non-depression class and depression class. Due to the DAIC-Woz dataset being imbalanced data, the number of non-depressed subjects is three times greater than the number of depressed [45].

The dataset is downloadable from https://github.com/Jackustc/Question-Level-Feature-Extraction-on-DAIC-WOZ-dataset

B. SMS Spam Collection Dataset.

The research of Priya and Karthika [46] utilized the SMS Spam Collection dataset in an experiment, which pertained to social media. The data from social media was characterized by short posts that convey emotions called short text messages, including tweets, posts, headlines, short messages, product descriptions, and searching. Social media users construct short text messages, such as tweets and Facebook posts. These messages have become necessary for analyzing social media data to detect the symptoms of depression in individuals. The short text data set consists of binary classes, with 1002 correct messages and 322 spam messages.

The dataset is available for download at https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

C. Depresjon Dataset.

The Depresjon Dataset comprises sensor data that is measured from Actigraph watches collected from patients with depression. The dataset comprises recordings of body movement performed by patients diagnosed with depression, divided into two types: Unipolar and Bipolar. The dataset comprises motor activity recordings from 32 healthy controls and 23 depressed patients with unipolar and bipolar depression [47].

The dataset is available for download at https://datasets.simula.no/depresjon/#dataset-details.

5. Methods

Through the research survey, the resampling method was applied to address the issue of imbalanced datasets, serving as a process to achieve data balance. By increasing or decreasing the number of samples in classes with varying sample sizes. The selection of an appropriate method depends on the characteristics of the data and the specific problem to be addressed. These methods are categorized into two main techniques: undersampling and oversampling. The procedures for these two techniques can be summarized as follows:

A. Undersampling Methods

This survey study explores popular undersampling methods, comprising the following methods: 1) Neighborhood undersampling (NUS), 2) Relevant information-based undersampling (RIUS), 3) Equalization ensemble (EASE), and 4) Undersampling method based on majority class data distribution (UMCDD). Each method involves the following four approaches.

1) Neighborhood Undersampling based Stacked Ensemble (NUS-SE)

The NUS-SE approach is proposed to address the performance decrease in imbalanced classification problems. Traditional methods struggle with issues involves class-imbalance, class-overlapping, and class-noise. While heterogeneous stacked ensembles have shown more favorable trends, the proposed NUS-SE integrates undersampling directly within the ensemble framework rather than as a pre-processing step. This method improves sample diversity by allowing all instances a chance to be used in training, unlike pre-processing which removes part of the data before training [40]. The feasibility of this approach is examined through the integration of undersampling within a stacked ensemble framework that uses cross-validation prediction as a metadata generation method. The resulting NUS-SE approach addresses potential issues with incomplete metadata [31].

2) Relevant Information-based Undersampling (RIUS)

The RIUS approach, also known as Relevant Information-based Undersampling, is used to select the most pertinent samples from the majority class in order to improve the classification performance on imbalanced datasets. The RIUS effectively represents the majority class distribution by selecting the most informative instances, minimizing information loss during undersampling [41].

3) Equalization Ensemble Method (EASE)

EASE is an ensemble method that employs an undersampling technique to adjust imbalanced data. EASE is a technique that assists in decreasing the quantity of majority class samples in an imbalanced dataset. This technique carries the risk of inappropriate sampling, which may result in the loss of valuable data from the majority class. Increased data imbalance provides a greater challenge for proper classification and increases complexity to the learning process for models, especially when dealing with large and highly imbalanced datasets. The EASE method primarily consists of a learning module and a combination module. The implementation follows two main approaches: 1) Perform equalization under-sampling to create a balanced dataset for each base classifier, which decreases the impact of class imbalance on the performance of the base classifiers. 2) Combine the results by applying weights based on the G-mean values from base classifiers on the imbalanced dataset. These weights are used in the final decision-making classification and can be adjusted for different imbalanced datasets [42].

4) Undersampling Method based on Majority Class Data Distribution (UMCDD)

UMCDD balances the dataset by creating multiple balanced training subsets [48]. The k-means algorithm is employed to manage the distribution of the majority class by computing Euclidean distances and iterating until the centroids converge. After clustering the majority class, a new distribution of data is obtained, with samples being allocated to different clusters. Subsequently, samples are randomly selected from each cluster in varying proportions and combined to approximate the number of samples in the minority class. When undersampling the majority class with consideration of class distribution, the randomly selected samples may not be representative, which can adversely affect the classification performance of the base classifier [30].

B. Oversampling Methods

This study explores popular Oversampling methods, which consist of the following approaces: 1) Noise-robust oversampling algorithm for mixed-type and multi-class imbalanced data (NROMM), 2) Synthetic minority over-sampling technique (SMOTE), 3) Improved and random synthetic minority oversampling technique (IR-SMOTE), and 4) The synthetic minority oversampling technique with natural neighbors (NaNSMOTE). Each method involves the four procedures detailed as follows.

1) Noise-robust Oversampling Algorithm for Mixed-type and Multi-class Imbalanced Data (NROMM)

The noise-robust oversampling technique for addressing imbalanced data classification in multiclass and mixed-type data is also known as NORMM [43]. The NROMM technique effectively decreases the creation of noisy samples and is particularly well-suited for addressing the challenges of multiclass classification. NROMM comprises an algorithm designed to eliminate noise within each class of samples and effectively manage the appropriate boundaries of each class. NROMM consists of two main steps: 1) performing oversampling on the minority class and 2) cleaning the samples within the majority class [49].

2) Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is an oversampling technique used to address the problem of imbalanced data. The SMOTE approach produces synthetic samples in the minority class that are not duplicates of the current samples [50], [51]. SMOTE can be extended into various techniques that adjust the weight of the minority class, such as ADASYN, MWMOTE, and k-means SMOTE [48], [52], [53]. These techniques increase the weight of samples near the boundary to create synthetic samples. There are also techniques developed from SMOTE aimed at adjusting noisy data, such as SMOTE-ENN and SMOTE-TL, which use Tomek links (TL) to remove noise from the boundary between classes during the SMOTE process [54], [55].

3) Improved and Random Synthetic Minority Oversampling Technique (IR-SMOTE)

Oversampling methods often underperform when dealing with noisy data and complex, irregular distributions. To address these problems, the author introduces a novel oversampling technique named the improved and random synthetic minority oversampling technique (IR-SMOTE), designed to address problems associated with imbalanced data. The IR-SMOTE technique is an advanced development of the SMOTE method. The process starts by employing the K-means algorithm to cluster samples from both the majority and minority classes [56]. A distance metric is used to eliminate noise from the minority class, thereby improving the quality of the synthetic samples. Subsequently, the mean of each cluster is computed to establish new centroids, resulting in updated centroids for both the minority and majority classes. IR-SMOTE utilizes kernel density estimation technique to determine the number of synthetic samples for each cluster in the minority class and assigns weights to the new synthetic samples. At this stage, a synthetic approach is employed to select suitable attributes, thereby ensuring diversity among the synthetic samples [57].

4) The Synthetic Minority Oversampling Technique with Natural Neighbors (NaNSMOTE)

The NaNSMOTE technique was developed from the SMOTE and natural neighbor techniques to address the problem of imbalanced data. The NaNSMOTE technique involves selecting the parameter k and determining the number of neighbors for each sample [58]. The NaNSMOTE approach operates by randomly selecting samples with varying differences, and then generating synthetic samples by selecting one of the natural neighbors. This approach ensures that samples closer to the class centroid have more neighbors, thereby enhancing the creation of new synthetic samples. Conversely, samples near the class boundary have fewer neighbors, which contributes to the reduction of errors in synthetic samples and improves the removal of outliers [59].

C. The Procedure of Imbalanced Data Classification.

In the research on data classification, we investigated relevant factors that caused the depressive disorder, including the algorithm that is used to construct the classification model. A variety of research studies have revealed depression data, including medical data that was frequently imbalanced data. Therefore, it was interesting to examine approaches to solving the distribution of imbalanced data. The literature survey studied the resampling technique to improve the distribution of data classes that are appropriately balanced. Then, metrics and performance evaluations were evaluated, not just the evaluation employing accuracy values. The construction of classification models involved the following two stages.

1) Data Preprocessing

Data preprocessing is the process of preparing data for another step of processing; it is a crucial stage in the knowledge-based data discovery (KDD) procedure. According to Benhar, Idri, and Fernández-Alemán [32], the process is a technique for preparing data to enhance the performance of prediction systems in research related to the classification of heart disease. Ridzuan and Zainon [60] mention that data cleaning is the first step in the data preparation process; this is a procedure that includes the detection and correction of data to provide the most accurate and comprehensive result. Examples from the research of Maghraby and Ali [61] relevant to cleaning up the Arabic depression datasets from Twitter found that most of the problems were caused by missing values. For this reason, handling the missing values involves removing data values from each column, such as irrelevant columns, duplicate data, empty data, and other symbols. The next step is data transformation, AL-Alimi et al. [62] mentioned that sometimes the data was obtained from multiple sources, and each source may have different storage. During data preparation before processing the data, the most common data transformation process, such as transforming the numeric representation of the “Gender” that is the gender characteristic, was studied by Ojokoh et al. [63]. Male characteristics were indicated by 1 and females by 2. In several research studies, the emphasis was focused on preprocessing data as a result of the need for complete data to construct an effective model. Also, the classification is effective and provides quality results.

2) Processing of Imbalanced Data

According to Farshidvard, Hooshmand, and MirHassani [64] solving the imbalanced data involves balancing each class that is significantly different in order to achieve a balance or have result classes that are equal. Without effective data management during the modeling process, there was an overlap between the data of the majority and the minority classes. As a result, the data was biased toward the majority class, resulting in the inability to effectively classify the minority class. Therefore, in order to improve the effectiveness of data classification, it is essential to balance the data as the first process. Huang et al. [65] indicated that, according to relevant research studies, two techniques are frequently employed to solve imbalanced data. These methods are described below.

- The resampling approach is a random sampling method that is divided into two techniques. – (1) undersampling which randomly decreases instances of majority class data to be similar to the minority class and (2) oversampling techniques which randomly increase the minority class. Mohammed, Rawashdeh, and Abdullah [66] in summarizing the study of the two sampling techniques, found that randomly increasing the number of minority classes was more efficient than randomly decreasing the number of majority classes that employed several classifiers, the model evaluation of which provided higher accuracy than that of the initial model.
- For SMOTE, Zhang et al. [28] stated that SMOTE is a technique for solving imbalanced data. Due to the improper distribution of data, the number of instances in each class differs significantly, resulting from the classification occurring in the majority class. Therefore, the SMOTE technique focused on increasing the number of minority class data. Li et al [67]. proposed the SMOTE technique known as NaNSMOTE to solve the selection of the parameter *k*, which the value of *k* represents the number of neighbors surrounding the centroid and determine the number of neighbors. As a result, the new SMOTE technique provided the best technique when compared with other methods consisting of the existing SMOTE technique or SMOTE's extension method.

D. Model Construction and Evaluation.

In the survey research, we studied the data set from Kaggle open dataset that is relevant to the depression dataset. We investigated how to deal with imbalanced data in the dataset.

1) Model Construction

In the study of classification research, there are two types of classification methods for imbalanced data. –

- *Individual classification* combined with several techniques for balancing the imbalanced data were used to measure and evaluate the performance of the model in order to find the model with the best precision. According to various research studies, the most popular individual classification models consist of decision trees (DT) and support vector machines (SVM), according to research conducted by Han et al. [68].
- *Ensemble classification* combined with individual classification for balancing the imbalanced data which is popular in the ensemble model as the RF, is utilized to balance the binary classification described by Ding et al. [69]. From the two classification types, it was found that the most popular models used to evaluate the performance of classification models on the imbalanced data and provide the best accuracy included RF, SVM, and XGBoost. The new model was then compared with others that are the same or different types in order to achieve the best performance of the classification model. In addition to establishing the appropriate models, it is also essential to utilize effective techniques to solve the problem of imbalanced data, such as the SMOTE algorithm, oversampling techniques, or undersampling techniques. This achieves effectiveness for classifying results and is an appropriate approach for imbalanced data.

2) Evaluation Metrics

Various evaluation metrics have been proposed to measure the performance of depression, including precision (PR), recall (RC), accuracy (ACC), F1-score, and a receiver operating characteristic (ROC) curve [15], [25], [70], [71], [72]. For example, Jere et al. [73] used evaluation metrics (PR, RC, ACC, and F1-score) to evaluate a depression-related model that was trained from the patient risk factor. Chiong et al. [74] studied the detection of depressive disorder by collecting data from social media platforms that were classified using several machine learning methods, and the model effectiveness was measured using PR, RC, ACC, and F1-score. Additionally, Fang et al.[75], and Ahmed et al. [76] proposed machine learning techniques to classify depressive data when the data set is high-dimensional noisy data with small data sets, and using evaluation metrics that included PR, RC, ACC, and AUC ROC.

Several appropriate metrics for evaluating the performance of classification models include accuracy, F1-score, precision, and recall. The primary objective of this research study is to examine clinical data that reveals imbalanced characteristics. Therefore, it is crucial to select appropriate metrics to enhance the efficiency of classifying depression conditions. The analysis of studies on depression datasets reveals that the datasets utilized to develop machine learning models were frequently imbalanced. The data of critical importance for consideration are those within the minority class or diagnosing rare diseases, specifically referring to patients diagnosed with depression. These minority classes typically involved notably less numerous data than those representing the normal classes. Even though the dataset is balanced using data balancing techniques prior to model creation, the metrics and evaluation methods must also be appropriate for imbalanced datasets. Research study indicates that the precision-recall curve (PR-curve) and area under the precision-recall curve (AUC-PR) metrics [77] are suitable for evaluating model performance with imbalanced datasets.

Both measures are based on the following general concepts. First, PR-curve is a technique for assessing the performance of machine learning models, particularly in classification tasks [78]. This technique illustrates the relationship between precision and recall, with an optimal model providing both high precision and recall. Given the specific performance evaluation characteristics of each class, this technique assists imbalanced datasets effectively. Typically, the metric is employed to evaluate the performance of a model in classifying data across various classes, with a particular emphasis on the minority class. This technique demonstrates the effectiveness of the model in accurately identifying true positives while decreasing false positives to the lowest feasible.

Another relevant metric, AUC-PR is appropriate for evaluating the performance of models on highly imbalanced datasets, as it effectively assesses the ability to classify between different classes [77]. The evaluation concentrates on measuring the performance of the positive case (minority class), with particular emphasis on positive predictive value (PPV) and true positive rate (TPR) as the primary metrics. In general, positive cases constitute a minority class but are of significant importance for predicting the occurrence of the disease. In a medical context, false negatives may imply that a patient is not afflicted with a disease, despite the presence of the disease in the patient. If the prediction results from the model are of this nature, inclusion of the patient in the data can be extremely harmful, as they may lead to patients not receiving timely and appropriate treatment.

PR-curve and AUC-PR are often considered the best metrics for evaluating imbalanced datasets. Both metrics effectively address the impact of class imbalance, thereby offering a more comprehensive evaluation of model performance on imbalanced datasets. Focusing on the accurate classification of the minority class in relation to the associated factors within the dataset, where the minority class often holds greater significance than the majority class in imbalanced datasets. Although the ROC AUC metric is commonly used for imbalanced datasets, this metric may not be suitable for datasets with a significant disparity in data distribution between classes. The ROC AUC value can appear high because it primarily considers the true negative rate (negative class is normal patients). However, this metric may not be particularly relevant when evaluating performance for predicting disease occurrence, as it focuses mainly on the majority class. In contrast, the AUC-PR metric specifically focuses on the minority class (positive class is depressive patients), offering a more accurate and relevant assessment [77], [78].

6. Comparative Study on the Depression Data

This section of the survey research aims to demonstrate research related to classification using machine learning methods applied to both imbalanced depressive data (see Table 1) and balanced depressive data (see Table 2). The study is divided into two parts: 1) The analysis of imbalanced depressive data and 2) The analysis of balanced depressive data.

A. The Study on Imbalanced Depressive Data.

In the evolving landscape of psychiatric diagnostics, the utilization of machine learning algorithms such as SVM, DT, XGBoost and other ensemble models to navigate the complexities of depressive disorders presents both immense potential and significant challenges for both binary classes and multi-classes as shown in Table 1.

Data utilized for the experiment regarding the varied depressive symptoms, such as Beck's cognitive triad, major depression (MD) and subclinical depression (SD), questionnaire, comment on social media, and electroencephalogram (EEG, which is a test that measures electrical activity in the brain) and eye movements (Ems) are shown in Table 1.

In most cases, machine learning was employed to establish the models, which were SVM, DT, NB, RF, XGBoost, and stacking ensemble. The efficiency measurement results indicate that social media datasets yielded experimental results with a high accuracy of 99.8% [74]. The accuracy of brain activity and eye movement data types was 82.5% and 92.65%, respectively. However, Jere et al. [73] tested the cognitive triad datasets on Twitter and achieved an accuracy of just 60.54 % using the SVM method, the most efficient method for this dataset.

B. The Study on Balanced Depressive Data.

To solve the problem of imbalanced datasets, especially medical data sets. The research study employed experimental methodologies adjusted for imbalanced depressive data. The literature survey was interested in studying resampling techniques, notably SMOTE, and other oversampling techniques to equilibrate the dataset. Therefore, using this newly improved technique enhances the ability of the model to classify depressive classes within a dataset that contains many non-depressive classes. This experimental approach is pivotal in elucidating the effectiveness of balancing strategies, setting a foundational precedent for future endeavors aimed at refining depression classification models for improved clinical outcomes and patient care as shown Table 2.

Table 2 demonstrates that medical datasets were frequently imbalanced data that usually had two classes (i.e., healthy and depressed) and also evaluated model performance with several metrics. For example, the research of Emre et al. [33], and Begum et al. [17] was applied to depressive disorder data using the oversampling technique to rectify the imbalanced dataset. This table demonstrates 76.2% the evaluation metrics for evaluating the performance model, such as accuracy, precision, recall, and F1-score.

Table 1 The performance of depression classification using various machine learning models on various depression datasets

References	Year	Depression Datasets	Machine Learning Models	Evaluation Metrics (%)			
				F1-score	Precision (PR)	Recall (RC)	Accuracy (ACC)
Zhang et al. [28]	2022	MD and SD dataset (2 classes)	SVM	-	-	-	83.09
Jere et al. [73]	2021	Cognitive triad dataset (6 classes)	DT	52.62	53.31	52.65	52.65
			NB	42.17	46.75	44.15	44.65
			SVM	60.58	61.69	60.35	60.54
			RF	58.26	59.05	58.52	58.64
Ojokoh et al. [63]	2021	Questionnaire -depress dataset (2 classes)	NB	-	92.0	-	96.0
Chiong et al. [74]	2021	Twitter dataset (2 classes)	DT	-	83.4	84.2	82.2
			SVM	-	99.9	99.6	99.8
Zhu et al. [79]	2020	EEG dataset (2 classes)	Content-based ensemble method (CBEM)	-	-	-	82.5
		EMs dataset (2 classes)	CBEM	-	-	-	92.65
Fang et al. [75]	2022	Reward positivity (RewP) and late positive potential (LPP) datasets (2 classes)	SVM with RBF Kernel	-	-	93.8	64.4
			Stacking ensemble method	-	-	93.8	64.4

In addition, Table 2 makes a comparison of accuracy between imbalanced and balanced data, finding that after the data sets were adjusted to balance data, both datasets increased the accuracy to 84% and 98.33% from 76.2% and 83.33%, respectively. The above suggests that the balanced data set is more significant in precision.

The research study of data from depressive disorder datasets reveals that the data is collected from various sources, such as questionnaires, opinions gathered from social media platforms, and physiological signal monitoring, among others. The author aims to demonstrate that there are numerous studies using depressive disorder datasets in various experiments that do not prioritize data balancing, as illustrated in Table 1. Which illustrates that researchers tend to focus more on developing algorithms for classifying result classes rather than on adjusting the dataset. The evaluation and performance measurement typically report only the accuracy value. However, measuring model performance just based on accuracy is insufficient to evaluate models constructed on imbalanced datasets. Therefore, relying on a single performance metric is insufficient to determine whether the model developed from the dataset demonstrates good classification performance. This is evident in several studies, including the works of Zhang et al. [45], Zhu et al. [51], Ojokoh et al. [41], Fang et al. [53], as presented in Table 1. This research study utilizes a dataset characterized by imbalanced data, reporting a relatively high accuracy, which is derived from the overall performance evaluation of the model. However, when analyzing the data in-depth regarding the classification of each class, the results do not reveal whether the developed model truly demonstrates high accuracy. This is because the model's performance is evaluated using only the accuracy value. In Table 2, we present a list of research study on depression datasets where data balancing techniques are applied. The focus in the present study is on adjusting imbalanced data to enhance the effectiveness of classifying class results.

Table 2 The methods used for solving the imbalanced data using the oversampling techniques

References	Year	Depression Datasets	Resampling Methods	Evaluation Metrics (%)				
				Accuracy (before applying resampling method)	Accuracy (after applying resampling method)	Precision	Recall	F1-score
Othmani and Zeghina [45]	2022	DAIC-Woz dataset (2 classes)	SMOTE	78.97	82.55	-	-	79.9
			Undersampling	71.18	80.99	-	-	75.9
Suen et al. [29]	2021	Depressed and bipolar patient dataset (2 classes)	XGBoost + Oversampling	76.0	78.0	81	75	78
			Elastic net + Oversampling	72.0	76.0	79	74	76.52
Emre et al. [33]	2023	EEG dataset (8 classes)	Oversampling	76.2	84	85	84	82
Xiao et al. [17]	2021	BRCA dataset (2 classes)	Oversampling	83.33	98.33	100	96.67	98.31
Sharma and Verbike [80]	2020	Dutch citizen dataset (2 classes)	Oversampling	90.35	97.29	95.48	99.87	97.62
			Undersampling	90.35	51.64	50.17	57.37	51.83
			Over-Under sampling	90.35	94.42	91.07	98.35	94.57
			ROSE sampling	90.35	66.18	65.65	65.38	65.51

We are interested in studying research that employs resampling methods to adjust imbalanced datasets, using various techniques to balance the class distributions of the result classes to achieve similarity.

Table 2 demonstrates that the resampling method includes oversampling and undersampling techniques which effectively balance the data and achieves superior classification performance is the oversampling method. Comparing the accuracy metrics using the same dataset between the imbalanced and balanced datasets, it is observed that accuracy increases with data balancing. For example, in the BRCA dataset, which contains two classes (Non-depression and Depression), applying the oversampling technique to balance the data results in a 15% increase in accuracy compared to the original dataset that remains unadjusted. According to the research by Sharma and Verbike [52], which utilized the Dutch citizen dataset with two classes (Depression cases and Healthy cases) and various techniques were utilized for data balancing. Performance evaluation of the model indicates a decrease in accuracy of 38.71% when using the undersampling technique to adjust the dataset. Conversely, applying the oversampling technique resulted in 6.94% increase in accuracy compared to the original dataset.

The comparison table demonstrates that the oversampling technique, which involves increasing the data for the minority class to balance it with the majority class, yields higher accuracy in the adjusted dataset. The oversampling technique, when applied to adjust datasets, results in higher accuracy compared to the undersampling technique. The utilization of the identical dataset as exemplified by the Dutch citizen dataset provides a more obvious evaluation of performance. An examination of several studies related to depression datasets reveals that oversampling techniques have been widely employed to adjust imbalanced data. Therefore, the comparison table, based on our review of various research studies, demonstrates that increasing the instances of the minority class (depression class) results in better accuracy than reducing the instances of the majority class.

7. Discussion

To discuss depression as a growing global health concern, this work highlighted its prevalence and the multifactorial nature of risk factors, including social media behavior, lifestyle, and genetic predispositions. It also delves into different approaches to depression classification, including the use of ensemble learning methods and individual machine learning models, and the necessity for balanced data to enhance model accuracy and reliability. In terms of methodology, the paper reviews various datasets related to depression, detailing their characteristics and how imbalanced data within these datasets is addressed, which covers both undersampling and oversampling techniques, and presents a comparison of model performances across different datasets to demonstrate the effectiveness of modeling on a balanced dataset.

To construct an effective classification model, selecting an appropriate performance metric is also crucial. Comprehending evaluation metrics is essential for the creation of effective machine learning models, especially when dealing with imbalanced data. Even with data preparation or balancing techniques such as cluster-based oversampling, SMOTE, IR-SMOTE, and ADASYN, understanding evaluation metrics remains essential. When increasing or decreasing data in each class, using inappropriate evaluation metrics can lead to errors in assessing model performance and possibly induce bias in the results.

Apart from reviewing relevant research studies, identifying suitable methodologies and performance metrics are also important. This study aims to demonstrate how the findings can be utilized to enhance the classification of patients with depression. The imbalanced data classification utilizes research-based knowledge to enhance the classification of patients with depression by incorporating the analysis of relevant factors and causes, thereby enabling more accurate identification of depressive conditions. This preliminary data is used to inform the general public about the potential appearance of depression and the various risk factors that may contribute to different levels of depressive conditions. Since depression is not a physical illness that can be easily observed externally or through changes in physical condition, Rather the condition of depression is an emotional disorder that affects feelings, thoughts, and expressed behaviors. Living a normal life proves challenging; therefore, individuals with depression require compassionate and continuous treatment. Applying research knowledge to identify individuals with depression involves developing preliminary assessment criteria for diagnosing depressive disorders based on research findings. In addition to using the data to assess and screen patients for depression, behavioral data are analyzed alongside research findings to identify high-risk groups for depression. This approach aims to develop personalized treatment plans for each patient.

The primary motivation for this research is aimed to study various methods that improve the accuracy of depression diagnosis using the resampling techniques to address the imbalanced data problem, particularly the oversampling technique is of primary interest here. While previous studies have focused on applying machine learning algorithms to mental health data, they have not adequately addressed the impact of data imbalance on model performance. This research study aims to fill this gap by systematically evaluating the effectiveness of oversampling techniques in enhancing the classification accuracy of depression-related data

Therefore, this report stresses the criticality of employing resampling methods to counteract the imbalanced data problem in depression classification. The superiority of oversampling in preserving valuable data and enhancing model predictions by ensuring a balanced representation of both minority and majority classes is emphasized. This work suggests future research directions, including the exploration of novel resampling approaches that mitigate the risk of overfitting while effectively addressing data imbalance. Moreover, the study provides a comprehensive overview of current practices and challenges in the field of depression classification with imbalanced datasets.

8. Conclusion

We found various existing research reports that emphasizes the effectiveness of resampling techniques in addressing imbalanced data. The research indicated that resampling techniques enhance the performance of classification models when applied to imbalanced datasets, contributing to a more balanced distribution of data across classes. This technique also improves the capacity of the model to learn and achieve higher accuracy in data classification, particularly in cases where the minority class is critical for analysis. Balancing the data adjusts decrease classification bias arising from one class having a greater number of samples than another. This data is common in medical datasets, where the minority class represents patients with the condition, and the majority class consists of individuals without the condition, typically with a significantly larger number of samples. Therefore, once the data is balanced to achieve similar sample sizes across classes, the model can learn the characteristics of each result class more equally. In terms of evaluating model performance, the utilization of resampling techniques leads to a more balanced data distribution, which enhances both accuracy and precision in the assessment. Based on a review of research studies utilizing medical datasets with imbalanced data, many of the studies we studied applied resampling techniques to address the data imbalance. These research findings demonstrated improved classification performance compared to using the original, imbalanced datasets.

This study focuses on the importance of balancing datasets, especially within the context of medical data. The researchers have concentrated especially on datasets related to depression. Models trained on balanced datasets tend to demonstrate superior performance in predicting rare or minority classes. This research also explored increasing data by generating new synthetic samples from existing ones to balance datasets, specifically utilizing oversampling techniques. A major problem with data enlargement through the generation of new samples from existing ones is the risk of overfitting. This technique may cause the model to learn excessively specific details from the existing data, which could impair its predictive performance on new datasets. Sometimes, generating samples that are inconsistent to the actual data distribution may increase irregularities in the learning process of the model.

In this study, we focused primarily on examining oversampling methods. Given the interest in medical datasets, where the minority class representing patients is already extremely few in size. Employing undersampling methods, which involve reducing the number of majority class instances to approximate the size of the minority class, may lead to the loss of valuable data. The issue of overfitting may impact a model's ability to learn the characteristics of data from the majority class, which is essential for improving machine learning models utilized in mental health diagnostics. From our findings, we have identified methods for addressing the issue of imbalanced datasets, particularly focusing on medical data. Addressing the problem of imbalanced data by using resampling techniques to balance the dataset aims to serve as initial research leading to the development of diagnostic tools for patients with depression. Additionally, it can provide preliminary data for screening individuals who may be experiencing depression, enabling timely intervention and treatment. By improving the data and using appropriate performance metrics, this research study could lead to the development of more accurate and reliable classification models. Ultimately, improving the dataset for developing classification models will enhance the accuracy of patient classification outcomes.

Furthermore, future research will continue to study, test, and refine these methods. We will focus on oversampling techniques and emphasize addressing the issue of overfitting in the dataset. To develop machine learning models that are both effective and adaptable to suitable medical datasets.

Acknowledgements

This research project was financially supported by Mahasarakham University. (Grant number 6707163/2024).

References

[1] Z. Liu *et al.*, "Classification of major depressive disorder using machine learning on brain structure and functional connectivity," *J Affect Disord Rep*, vol. 10, pp. 1–11, 2022.

[2] World Health Organization's, *World Health Statistics 2022*. World Health Organization 2022, 2022.

[3] X. Zhang *et al.*, "Prevalence and risk factors of depression and anxiety among Chinese adults who received SARS-CoV-2 vaccine — A cross-sectional survey," *J Affect Disord*, vol. 324, pp. 53–60, 2023.

[4] H. A. Yazdavar *et al.*, "Multimodal mental health analysis in social media," *Public Library of Science*, vol. 15, no. 4, p. e0226248, 2020.

[5] R. Haand and Z. Shuwang, "The relationship between social media addiction and depression: A quantitative study among university students in Khost, Afghanistan," *Int J Adolesc Youth*, vol. 25, no. 1, pp. 780–786, 2020.

[6] Md. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, and H. Wang, "Depression detection from social network data using machine learning techniques," *Health Inf Sci Syst*, vol. 6, no. 1, pp. 1–13, 2018.

[7] R. S. Begum and Y. S. Sait, "Effective techniques for depression detection on social media: A comprehensive review," in *International Conference on Computer Communication and Informatics (ICCCI)*, India: IEEE, 2022, pp. 1–9.

[8] D. Geng, Q. An, Z. Fu, C. Wang, and H. An, "Identification of major depression patients using machine learning models based on heart rate variability during sleep stages for pre-hospital screening," *Comput Biol Med*, vol. 162, p. 107060, 2023.

[9] K.-I. Jang, S. Kim, J.-H. Chae, and C. Lee, "Machine learning-based classification using electroencephalographic multi-paradigms between drug-naïve patients with depression and healthy controls," *J Affect Disord*, vol. 338, pp. 270–277, 2023.

[10] Sofia, A. Malik, M. Shabaz, and E. Asenso, "Machine learning based model for detecting depression during Covid-19 crisis," *Sci Afr*, vol. 20, p. e01716, 2023.

[11] Y. Chen, W. , J. Stewart, J. Ge, B. Cheng, A. Chekroud, and J. , D. Hellerstein, "Personalized symptom clusters that predict depression treatment outcomes: A replication of machine learning methods," *J Affect Disord Rep*, vol. 11, p. 100470, 2023.

[12] Y. Sánchez-Carro *et al.*, "Importance of immunometabolic markers for the classification of patients with major depressive disorder using machine learning," *Prog Neuropsychopharmacol Biol Psychiatry*, vol. 121, p. 110674, 2023.

[13] A. Occhipinti, L. Rogers, and C. Angione, "A pipeline and comparative study of 12 machine learning models for text classification," *Expert Syst Appl*, vol. 201, p. 117193, 2022.

[14] J. Zhai, J. Qi, and C. Shen, "Binary imbalanced data classification based on diversity oversampling by generative models," *Inf Sci (N Y)*, vol. 585, pp. 313–343, 2022.

[15] S. Shi, J. Li, D. Zhu, F. Yang, and Y. Xu, "A hybrid imbalanced classification model based on data density," *Inf Sci (N Y)*, vol. 624, pp. 50–67, 2023.

[16] ao, Z. Huang, Y. Sang, Y. Sun, and J. Lv, "A neural network learning algorithm for highly imbalanced data classification," *Inf Sci (N Y)*, vol. 612, pp. 496–513, 2022.

[17] Y. Xiao, J. Wu, and Z. Lin, "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data," *Comput Biol Med*, vol. 135, p. 104540, 2021.

[18] D. Li, C. Zheng, J. Zhao, and Y. Liu, "Diagnosis of heart failure from imbalance datasets using multi-level classification," *Biomed Signal Process Control*, vol. 81, p. 104538, 2023.

[19] Inamullah, S. Hassan, S. B. Belhaouari, and I. Amin, "Deciphering the impact of diversity in CNN-based ensembles on overcoming data imbalance and scarcity in medical datasets:

A case study on diabetic retinopathy," *Inform Med Unlocked*, vol. 49, p. 101557, 2024, doi: 10.1016/j.imu.2024.101557.

[20] L. Bai, T. Ju, H. Wang, M. Lei, and X. Pan, "Two-step ensemble under-sampling algorithm for massive imbalanced data classification," *Inf Sci (N Y)*, vol. 665, p. 120351, 2024.

[21] F. Wang, M. Zheng, X. Hu, H. Li, T. Wang, and F. Chen, "FIAO: Feature information aggregation oversampling for imbalanced data classification," *Appl Soft Comput*, vol. 161, p. 111774, 2024.

[22] J. Guo, H. Wu, X. Chen, and W. Lin, "Adaptive SV-borderline SMOTE-SVM algorithm for imbalanced data classification," *Appl Soft Comput*, vol. 150, p. 110986, 2024.

[23] A. , S. Alex, V. J. J. Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl Soft Comput*, vol. 156, p. 111491, 2024.

[24] K. L. Xin and A. binti, N. Rashid, "Prediction of depression among women using random oversampling and random forest," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, Taif, Saudi Arabia: IEEE, 2021, pp. 1–5.

[25] X. Gao *et al.*, "An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling," *Expert Syst Appl*, vol. 160, p. 113660, 2020.

[26] R. M. Pereira, Y. M. G. Costa, and S. C. N. Jr., "Toward hierarchical classification of imbalanced data using random resampling algorithms," *Inf Sci (N Y)*, vol. 578, pp. 344–363, 2021.

[27] H. Li, X. Dong, W. Shen, F. Ge, and H. Li, "Resampling-based cost loss attention network for explainable imbalanced diabetic retinopathy grading," *Comput Biol Med*, vol. 149, p. 105970, 2022.

[28] B. Zhang *et al.*, "Discriminating subclinical depression from major depression using multi-scale brain functional features: A radiomics analysis," *J Affect Disord*, vol. 297, pp. 542–552, 2022.

[29] J. C. P. Suen, S. Goerigk, B. L. Razza, F. Padberg, C. I. Passos, and R. A. Brunoni, "Classification of unipolar and bipolar depression using machine learning techniques," *Psychiatry Res*, vol. 295, p. 113624, 2021.

[30] Z. Sun, W. Ying, W. Zhang, and S. Gong, "Undersampling method based on minority class density for imbalanced data," *Expert Syst Appl*, vol. 249, p. 123328, 2024.

[31] P. Vuttipittayamongkol and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Inf Sci (N Y)*, vol. 509, pp. 47–70, 2020.

[32] H. Benhar, A. Idri, and J.L. Fernández-Alemán, "Data preprocessing for heart disease classification: A systematic literature review," *Comput Methods Programs Biomed*, vol. 195, pp. 1–30, 2020.

[33] E. I. Emre, Ç. Erol, C. Tas¸, and N. Tarhan, "Multi-class classification model for psychiatric disorder discrimination," *Int J Med Inform*, vol. 170, p. 104926, 2023.

[34] Chenxi Huang *et al.*, "Sample imbalance disease classification model based on association rule feature selection," *Pattern Recognit Lett*, vol. 133, pp. 280–286, 2020.

[35] O. K. Asare *et al.*, "Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis," *Pervasive Mob Comput*, vol. 83, p. 101621, 2022.

[36] I. Moshe *et al.*, "Predicting symptoms of depression and anxiety using smartphone and wearable data," *Front Psychiatry*, vol. 12, p. 625247, 2021.

[37] C. Karima and W. Anggraeni, "Performance analysis of the Ada-Boost algorithm for classification of hypertension risk with clinical imbalanced dataset," *Procedia Comput Sci*, vol. 234, pp. 645–653, 2024.

[38] T. Zuo, F. Li, X. Zhang, F. Hu, L. Huang, and W. Jia, "Stroke classification based on deep reinforcement learning over stroke screening imbalanced data," *Computers and Electrical Engineering*, vol. 114, p. 109069, 2024.

[39] K. Niu, Z. Zhang, Y. Liu, and R. Li, "Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending," *Inf Sci (N Y)*, vol. 536, pp. 120–134, 2020.

[40] Z. Seng, A. S. Kareem, and D. K. Varathan, "A neighborhood undersampling stacked ensemble (NUS-SE) in imbalanced classification," *Expert Syst Appl*, vol. 168, p. 114246, 2021.

[41] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, and G. Castellanos-Dominguez, "Relevant information undersampling to support imbalanced data classification," *Neurocomputing*, vol. 436, pp. 136–146, 2021.

[42] J. Ren, Y. Wang, M. Mao, and Y. Cheung, "Equalization ensemble for large scale highly imbalanced data classification," *Knowl Based Syst*, vol. 242, p. 108295, 2022.

[43] Y. Liu, Y. Liu, X. B. , B. Yu, S. Zhong, and Z. Hu, "Noise-robust oversampling for imbalanced data classification," *Pattern Recognit*, vol. 133, p. 109008, 2023.

[44] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowl Based Syst*, vol. 248, p. 108839, 2022.

[45] A. Othmani and O. A. Zeghina, "A multimodal computer-aided diagnostic system for depression relapse prediction using audiovisual cues: A proof of concept," *Healthcare Analytics*, vol. 2, p. 100090, 2022.

[46] K. Priya S. and P. Karthika K., "An embedded feature selection approach for depression classification using short text sequences," *Appl Soft Comput*, vol. 147, p. 110828, 2023.

[47] E. Garcia-Ceja *et al.*, "Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients," in *The 9th ACM International Conference on Multimedia Systems (MMsys 2018)*, Amsterdam, 2018, pp. 472–477.

[48] X. Yuan, C. Sun, and S. Chen, "A clustering-based adaptive undersampling ensemble method for highly unbalanced data classification," *Appl Soft Comput*, vol. 159, p. 111659, 2024.

[49] S. Shen, Z. Li, Z. Huan, F. Shang, Y. Wang, and Y. Chen, "Neighborhood repartition-based oversampling algorithm for multiclass imbalanced data with label noise," *Neurocomputing*, vol. 600, p. 128090, 2024.

[50] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *Int J Min Sci Technol*, vol. 32, pp. 309–322, 2022.

[51] S. Maldonado, C. Vairetti, A. Fernandez, and F. Herrera, "FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification," *Pattern Recognit*, vol. 124, p. 108511, 2022.

[52] C. Rao, Y. Xu, X. Xiao, F. Hu, and M. Goh, "Imbalanced customer churn classification using a new multi-strategy collaborative processing method," *Expert Syst Appl*, vol. 247, p. 123251, 2024.

[53] J. Wei, H. Huang, L. Yao, Y. Hu, Q. Fan, and D. Huang, "New imbalanced fault diagnosis framework based on Cluster-

MWMOTE and MFO-optimized LS-SVM using limited and complex bearing data," *Eng Appl Artif Intell*, vol. 96, p. 103966, 2020.

[54] L. Han *et al.*, "An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images," *Comput Electron Agric*, vol. 194, p. 106804, 2022.

[55] K. G. R. Narayan *et al.*, "Attenuating majority attack class bias using hybrid deep learning based IDS framework," *Journal of Network and Computer Applications*, vol. 230, p. 103954, 2024.

[56] F. Soleymani, S. Zhu, and X. Hu, "An unsupervised k-means machine learning algorithm via overlapping to improve the nodes selection for solving elliptic problems," *Eng Anal Bound Elem*, vol. 168, p. 105919, 2024.

[57] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowl Based Syst*, vol. 248, p. 108839, 2022.

[58] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf Sci (N Y)*, vol. 565, pp. 438–455, 2021.

[59] W. Wang, L. Yang, J. Zhang, J. Yang, D. Tang, and T. Liu, "Natural local density-based adaptive oversampling algorithm for imbalanced classification," *Knowl Based Syst*, vol. 295, p. 111845, 2024.

[60] F. Ridzuan and W. N. M. W. Zainon, "A review on data cleansing methods for big data," in *The Fifth Information Systems International Conference 2019*, Surabaya, 2019, pp. 731–738.

[61] A. Maghraby and H. Ali, "Modern standard Arabic mood changing and depression dataset," *Data Brief*, vol. 41, p. 107999, 2022.

[62] D. AL-Alimi, Z. Cai, A. A. , M. Al-qaness, and A. E. Alawamy, "ETR: Enhancing transformation reduction for reducing dimensionality and classification complexity in hyperspectral images," *Expert Syst Appl*, vol. 213, Part B, p. 118971, 2023.

[63] A. , B. Ojokoh, A. , O. Olaku, A. , O. Sarumi, and I. , S. Olotu, "Predictive analytics for economic crisis triggered depression risk level identification among some adults in Nigeria," *Sci Afr*, vol. 14, p. e01056, 2021.

[64] A. Farshidvard, F. F. Hooshmand, and S. A. S.A. MirHassani, "A novel two-phase clustering-based under-sampling method for imbalanced classification problems," *Expert Syst Appl*, vol. 213, Part B, p. 119003, 2023.

[65] Y. Huang, B. Giledereli, A. Köksal, A. Özgür, and E. Ozkirimli, "Balancing methods for multi-label text classification with long-tailed class distribution," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Dominican Republic, 2021, pp. 8153–8161.

[66] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, Jordan, 2020, pp. 243–248.

[67] J. Li, Q. Zhu, Q. Wu, and Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," *Inf Sci (N Y)*, vol. 565, pp. 438–455, 2021.

[68] T. T. Han *et al.*, "Machine learning based classification model for screening of infected patients using vital signs," *Inform Med Unlocked*, vol. 24, p. 100592, 2021.

[69] H. Ding, Y. Sun, Z. Wang, N. Huang, Z. Shen, and X. Cui, "RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification," *Inf Process Manag*, vol. 60, p. 103235, 2023.

[70] L. Cañete-Sifuentes, R. Monroy, and A. M. Medina-Pérez, "FT4cip: A new functional tree for classification in class imbalance problems," *Knowl Based Syst*, vol. 252, p. 109294, 2022.

[71] L.-H. Yang, T.-Y. Ren, F.-F. Ye, P. Nicholl, Y.-M. Wang, and H. Lu, "An ensemble extended belief rule base decision model for imbalanced classification problems," *Knowl Based Syst*, vol. 242, p. 108410, 2022.

[72] C. Morris and J. , J. Yang, "Effectiveness of resampling methods in coping with imbalanced crash data: Crash type analysis and predictive modeling," *Accid Anal Prev*, vol. 159, p. 106240, 2021.

[73] S. Jere, P. , A. Patil, I. , G. Shidaganti, S. , S. Aladakatti, and L. Jayannavar, "Dataset for modeling Beck's cognitive triad to understand depression," *Data Brief*, vol. 38, p. 107431, 2021.

[74] R. Chiong, S. G. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput Biol Med*, vol. 135, p. 104499, 2021.

[75] X. Fang *et al.*, "Accurate classification of depression through optimized machine learning models on high-dimensional noisy data," *Biomed Signal Process Control*, vol. 71, Part B, p. 103237, 2022.

[76] A. Ahmed *et al.*, "Machine learning models to detect anxiety and depression through social media: A scoping review," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100066, 2022.

[77] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[78] S. A. Khan and Z. Ali Rana, "Evaluating performance of software defect prediction models using area under precision-recall curve (AUC-PR)," in *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)*, IEEE, Feb. 2019, pp. 1–6. doi: 10.23919/ICACS.2019.8689135.

[79] J. Zhu *et al.*, "An improved classification model for depression detection using EGG and eye tracking data," *IEEE Trans Nanobioscience*, vol. 19, no. 3, pp. 527–537, 2020.

[80] A. Sharma and J. M. I. , W. Verbeke, "Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch dataset (n = 11,081)," *Front Big Data*, vol. 3, p. 15, 2020.