REVIEW OF CLASS IMBALANCE DATASET HANDLING TECHNIQUES FOR DEPRESSION PREDICTION AND DETECTION

Simisani Ndaba

Department of Computer Science, Faculty of Science, University of Botswana

ABSTRACT

Depression is a prevailing mental disturbance affecting an individual's thinking and mental development. There have been much research demonstrating effective automated prediction and detection of Depression. Many datasets used suffer from class imbalance where samples of a dominant class outnumber the minority class that is to be detected. This review paper uses the PRISMA review methodology to enlist different class imbalance handling techniques used in Depression prediction and detection research. The articles were taken from information technology databases. The research gap was found that under sampling methods were fewfor predicting and detecting Depression and regression modelling could be considered for future research. Theresults also revealed that the common data level technique is SMOTE as a single method and the common ensemble method is SMOTE, oversampling and under sampling techniques. The model level consisted of various algorithms that can be used to tackle the class imbalance problem.

KEYWORDS

Depression prediction, Depression detection, Class Imbalance, Sampling, Data Level and Model Level

1. INTRODUCTION

[1], [2] and [3] explained how Depression is a severe and well-known public health challenge. Depression is one of the most common psychological problems affecting everyone or through a family member. It causes low moods, a lack of interest in work, guilt, insomnia or disturbed sleep, and feelings of weakness and exhaustion. It is a disorder for which can be a burden to the family, the society and a country. According to [4], the influence of COVID-19 on young people's mental health including Depression was investigated by the University of Surrey between September and November 2019 and May/June 2020. When people are free from Depression, their ability to work increases, they have a healthy brain, their ability to explain things improves, they are able to take on challenges easily, physical unity, family harmony, social harmony and economic development increases.

The major issue is regarding its prediction and detection, which can be solved using machine learning in an efficient manner [5]. Several machine learning methods to automatically classify depression have been proposed. Such machine learning classification methods face a challenging class imbalance problem that arises from the fact that Depression occurs at a low frequency in the general population, yielding imbalanced datasets [6]. Class imbalance situations are pervasive in many fields and applications. Typical examples can be the diagnosis of rare diseases where the number of patients suffering from such diseases is very low in the population, the detection of fraud in card transactions where the number of legitimate transactions is much higher than the number of fraudulent ones [7], breast cancer diagnosis, and bankruptcy prediction [8

David C. Wyld et al. (Eds): DBML, CITE, IOTBC, EDUPAN, NLPAI -2023
pp. 31-45, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120203

1.2. Class Imbalanced Problem

[9] defined that a dataset with skewed class proportions is called imbalanced. [9] continued to saythat classes that make up a large proportion of the dataset are called majority classes. Those that make up a smaller proportion are minority classes. A class imbalanced dataset is when the majority class dominates the minority class. According to [7], the problem of class imbalance datasets occurs when each class does not constitute an equal part of the dataset, but they vary significantly in the number of samples belonging to them. In this situation, the predictive model develops using conventional machine learning algorithms could be biased and inaccurate. This leads to algorithms being biased toward the majority class and their performance becomes unreliable. With few positives relative to negatives, the training model will spend most of its time on negative examples and not learn enough from positive ones. [10] described that most machine learning algorithms in classification operate best when each class's number of instances is approximately equal.

[11] reported that the class imbalance problem is often encountered in the real world, especially in medical data, as there are many patients who are admitted to the hospital. However, the number of patients diagnosed with a certain disease is small compared to the total number of patients who have not been diagnosed with that certain disease. When this data is used to predict outcomes by machine learning and data mining, the learning of the algorithm is affected. It is assumed that the data is drawn from the same distribution as the training data, presenting imbalanced data to the classifierand producing unacceptable results [12].

1.3. Approaches to Class Imbalanced Dataset

[13] identified the ways to solve the class imbalance problem on both the model level and data level. The model level solution includes changing the algorithm of the model. An article by [14] about howto deal with imbalanced classification and regression data, explained that the model level approach concentrates on modifying existing models to alleviate their bias towards the majority groups. This requires good insight into the modified learning algorithm and precise identification of reasons for itsfailure in learning the representations of skewed distributions. The data level includes resampling techniques and data augmentation. The data level approach is usually done in the pre-processing step by altering or modifying the tendency of the class distribution on the dataset. According to [15], resampling or data synthesize is one of the most widely used method that used for the data level approach.

1.4. Objective

This review paper aims to enlist the different class imbalance handling techniques in Depression prediction and detection research. Previous reviews have investigated the machine learning algorithm used in mental health research, and in particularly, Depression. However, very few papers have investigated pre-processing techniques used in Depression prediction and detection.

Firstly, the review paper outlines the search strategies used to find relevant literature. Next, the paper conducts a synthesis of the literature, describing the class imbalance handling techniques used in the research. Finally, the paper summarizes the extant research and the implications for future work.

2. METHOD

2.1. Search Strategy

This review paper followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for reporting in a systematic review. The PRISMA was chosen because it is the recognized standard for reporting evidence in systematic reviews and meta-analyses and the standards are endorsed by organizations and journals. The search strategy used in this paper was adapted from [16] and [17]. As class imbalanced data and Depression prediction and detection fall under machine learning, the search was conducted in only information technology databases such as IEEE Xplore, the ACM Digital Library and Web of Science. Goggle scholar was also used as a source of data collection. The search period for relevant studies was conducted in September 2022. The search terms included variations in the terms for the following:

- (a) imbalanced data (imbalanced data handling*, imbalanced data sampling*, class imbalanceTechnique*)
- (b) Depression prediction (depression prediction*)
- (c) Depression detection (depression detection*)
- (d) Postpartum depression (PPD*)

The search was conducted on titles, keywords, and abstracts with *AND* entered into the database search to link different categories (a, b, c and d) of search terms. Truncation symbols (∗) were used to search for all possible forms of a search term. Forward reference searching, that is, examining the references cited in these articles, and backward reference searching, that is, reviewing the references cited in these articles, were applied to identify further studies that met the inclusion criteria. Table 1 below shows the criteria that were met to include and exclude articles from the review.

Table 1. Inclusion and Exclusion criteria followed that were followed.

Inclusion criteria	**Exclusion criteria**
The article worked on Depression prediction or detection.	The article did not report on class imbalanced handling technique in Depression prediction and detection.
The article reported on a method or application of a class imbalanced handling technique to address Depression prediction and detection only, based on the authors' descriptions of their analyses.	The article did not use a class imbalance dataset.
The article evaluated the performance of the class imbalanced handling technique used to predict and detect Depression.	The full text of the article was not available (eg, conference or abstracts).
The article was available in English.	If articles were commentaries and essays.
The article was published between 2018 and2022.	

2.2. Data extraction and analysis plan

For each article, data was extracted regarding: (i) the aim of research; (ii) area of Depression diagnosis; (iii) dataset; and (iv) class imbalanced data technique. To analyse the data, a narrative review synthesis method was selected to capture the large range of class imbalanced handling techniques for depression prediction and detection. It should be noted that a meta-analysis was

not appropriate for this review given the broad range of machine learning techniques, and types of performance measures used in the studies identified.

3. RESULTS

3.1. Overview of Article Characteristics

The search strategies using a combination of search terms identified 300 articles that included a search term from the category in their abstract or title. The range for publication year of the relevant articles was found to be between 2018 and 2022. The 200 articles were excluded for not using a class imbalance dataset for their experiment. A total of 30 articles were duplicates and removed. Abstracts of the remaining 70 articles were read for an initial screening of eligibility for this scoping review.Of these, 32 were excluded for not focusing on Depression prediction and detection with a class imbalance handling technique. 38 articles were reviewed with the inclusion criteria applied, however, 10 articles were excluded for being essays and commentaries. A total of 28 articles were selected for full text review, but 2 records were excluded for being more than 5 years old except for [18] which met all the criteria. This resulted in a total sample of 26 articles. The selected 26 articles were reviewed in full.

In the subsequent narrative analysis, this review paper focused on the 26 articles that used class imbalanced handling techniques for Depression prediction and detection. Figure 1 below shows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) procedural flowchart for screening class imbalanced handling techniques used in Depression diagnosis research.

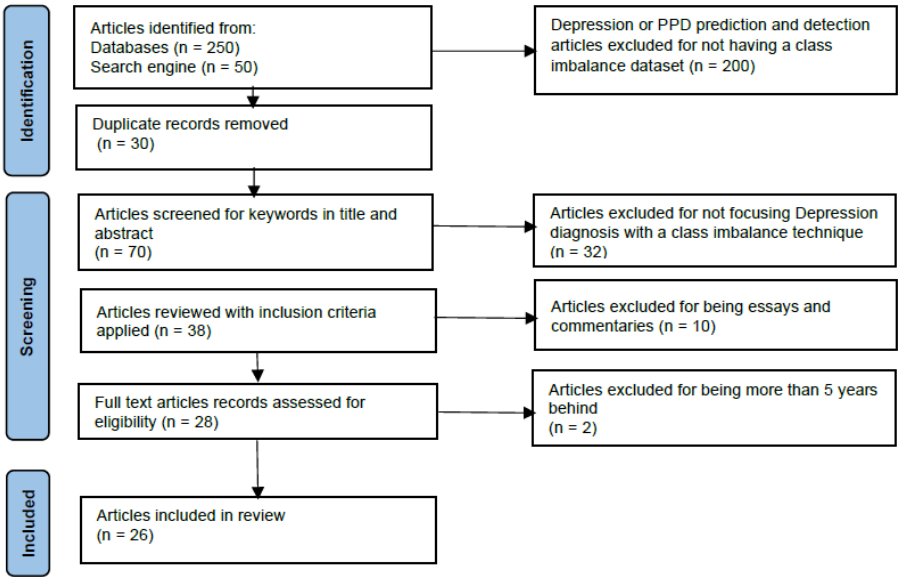

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)procedural flowchart on Depression prediction using class imbalance technique

3.2. Type of Datasets Used

Through synthesis of the articles, seven dataset types used in the Depression prediction and detectionwere identified illustrated in figure 2 below, followed by a detailed description of the datasets in table 2.

Figure 2. Seven dataset types used in Depression prediction and detection

Table 2. Description of data types used in the articles.

Dataset	Description
Distress Analysis Interview Corpus -Wizard of Oz (DAIC-WOZ)	According to [19], the DAIC-WOZ dataset was designed to facilitate research into Depression detection and was released as part of the 2016 Audio-Visual Emotion Challenge (AVEC). In the DAIC-WOZ , the virtual interviewer called Ellie, an animated role controlled by a human interviewer, tries to assess the indicators of distress disorders [20]. Ellie was designed to interact in favor of capturing verbal and non-verbal indicators responsible for the assessment of distress disorders, such as Depression. The corpus consisted of 189 recorded clinical interviews and transcripts aswell as facial features from 189 subjects. The audio recordings were taken from semi- structured interviews between the participants and Ellie. The number of non-depressed subjects was about four times bigger than that of depressed ones in both training and development parts. The articles namely, [3], [21], [20], [22], [13], [23] and [5] used the DAIC-WOZ dataset.
Electronic Health Records (EHR)	The Electronic Health Records used by the articles were taken from different sources. [24] used the Nathan Kline Institute (NKI) dataset which contained a neuroimaging of 420 records which included 76 depressed patients and 344 non-depressed patients. [25] used the MYsore studies of Natal effect on Ageing and Health (MYNAH) which consisted of patient records where patients had undergone a comprehensive assessmentfor cognitive function, mental health and cardio metabolic disorders having a total cohort size of 1321 patient records. [26] used data from the Pregnancy Risk Assessment Monitoring System with 28,755 records (3339 postpartum depression and 25,416 non postpartum depression). [27] used information on patient demographics, diagnoses, and medications available from HER (Electronic Health Records) from Weill Cornell Medicine and NewYork-Presbyterian Hospital as their data source.

Survey	For the past 30 years, clinician administered and self-reported questionnaires remain the gold standard in the assessment and diagnosis of depression. Some of the articles reviewed in this review paper used surveys to collect their own information on risk factors to predict depression. [18] used a survey of 173 responses, and the ratio of postpartum depression (PPD) to non-postpartum depression was 1:3, meaning that there was a 33.33% of postpartum depressed individuals to 66.66% non-postpartum depressed individuals. [11] used data that included 1549 male and female students, with no ratio of non-PPD to those who do have PPD, who filled the Patient Health Questionnaire-9 (PHQ-9) self-administered assessment. The questionnaire was assessed for both mental and physical symptoms. [28] used a Burns Depression Checklist (BDC) survey consisting of 604 respondents. In the training dataset, the percentages of depressed and non-depressed participants were 66.87% and 33.13% respectively. [29] used a nationwide survey data to construct a training and test set. A total of 6,588, 6,067 non-depressed and 521 depressed, participants were included in the study. [30] included 8,628 adults with hypertension, 11.3% with Depression, from the National Health and the 2011–2020 Nutrition Examination Survey. [31] ran a survey that collected data which totaled to 112 respondents with no ratio of non-depressed to depressed individuals.
Social Media	According to [32], over the past few years, social media platforms like Twitter, Facebook, Reddit and Instagram have become a medium for people to share their feelings and experiences. This stands as a significant opportunity for the research community to utilize such data for early diagnosis of psychological issues like Depression. [32] used a dataset that was generated by merging two Kaggle datasets consisting of 13,514 tweets where 76% (10,357) were non-depressive tweets and 24% (3517) were depressive tweets that lead to a class imbalance problem. The dataset [33] used was collected from 843 participants from a cohort of anonymous Android participants worldwide. [34] aimed to detect Depression from social media and the dataset used had 16,632 English comments having a wide range of sentence lengths and was imbalanced. [35] carried out their Depression detection on two datasets that had balanced and imbalanced datasets that consisted of over 1.6 million tweets. [36] had an approach to detect Depression as early as possible using the Reddit social media platform. Their solution for dealing with small and imbalanced data included having 892 subjects with almost 6 time more negative subjects than positive. [37] used 873,524 posts from Facebook from 1,453 students.
Study	Studies have been used to collect their own data sample from a particular population based on geographic location. Research like [38] analyzed the raw data of the 2016 Seoul Panel Study (SEPANS) data. The SEPANS data was conducted for the purpose of estimating the welfare level of Seoul citizens and the actual status situation of socially vulnerable class. The study analyzed 4,085 elderly people who were more than 60 years old living in the community. [39] had a model that evaluated on a large child Depression dataset based on the Longitudinal Study of Australian Children (LSAC) data consisted of multiple bi- annual waves of questionnaire-based interview data of 10,090 children across Australia. [6] utilized the StudentLife study which contained smartphone data continuously gathered from 48 Dartmouth College students over a 10-week semester.

National dataset	[40] is the only article that used a national study that used research based on the Lifelines Cohort study database. Lifelines was a multi-disciplinary prospective population-based cohort study, examining in a unique three-generation design, the health and health-related behaviours of persons living in the north of the Netherlands. [40] classified Depression prognosis from the Lifelines Database that contained biomarkers data and self-reported depression data of Dutch citizens. The dataset had 11,081 subjects with a 5.14% minority class of self-reported depressed symptoms and a 94.86% majority class of the subjects who self-reported no symptoms of Depression.
Workshop Challenge Task	Audio/Emotion Challenge and Workshop (AVEC) 2013 dataset like the the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset are both from challenge workshops but with different tasks. [41] evaluation data came from the Audio/Emotion Challenge and Workshop (AVEC) 2013 which comprised the recordings of 57 speakers with a total of 100 sessions.

3.3. Class Imbalanced Handling Techniques

The articles reviewed used various class imbalanced handling techniques. This section is divided by the articles that used data level and model level approaches.

3.3.1. Data Level Approaches

The data level is further sub-divided into the articles that used a single class imbalanced handling technique and others that used an ensemble class imbalanced handling technique. Figure 3 illustrates the frequency of data level single class handling techniques used in the articles. As explained in section 1.3., a data level approach is done on the pre-processing stage by altering or modifying the tendency of the class distribution on the dataset. The following data level approaches in figure 3were identified to be used in the articles.

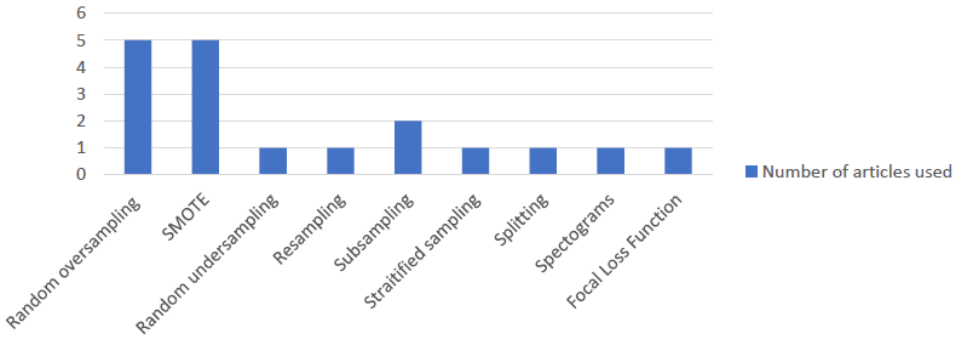

Figure 3. Frequency of Data Level Single Class Handling Techniques identified

Most of the techniques are sampling methods with the exception to Splitting, Spectrogram and Focal loss function. The Sampling technique is one solution to overcome imbalance cases based on data levels. In general, the sampling technique is divided into three, namely: under sampling, oversampling and a combination of under sampling and oversampling methods [42]. Table 3 describes the identified data level class imbalance handling techniques used in the articles.

Table 3. Description of the data level class imbalance handling technique used in the articles.

Class Imbalance HandlingTechnique	Description
Random Oversampling	In the Random Oversampling method, all data points from majority andminority training sets are used. Additionally, instances are randomly picked with replacement from the minority training set till the desired balance is achieved. Adding the same minority samples might result in overfitting, thereby reducing the generalization ability of the classifier.
SMOTE	Synthetic Minority Oversampling Technique (SMOTE) generates newsynthetic data by randomly interpolating pairs of nearest neighbours. SMOTE is a statistical strategy that generates new instances to increase the number of minority samples in the dataset. This approach takes feature space samples for each target class and its nearest neighbours, then generates new samples that blend the features from the target case with the features from its neighbours. The new cases are not exact replicas of extant minority cases [10]
The Random Over-Sampling Examples (ROSE)	The ROSE approach is aimed at oversampling the rare class by creating synthetic data points that were as similar as possible to the real ones with respect to a probability distribution centered on the selected sample [43].
Random Under sampling	All of the training data points from the minority class are used. Instances are randomly removed from the majority training set till the desired balance is achieved. One disadvantage of this approach is that some useful information might be lost from the majority class due to the under sampling. [44] describes that under sampling runs the risk of discarding potentially useful data whilst oversampling, besides increasing the learning time required, it can lead to overfitting, or the generation of a rule specifically for the replicated data. The types of under sampling methods used in the reviewed articles by [11] include Tomek Link and Edited Nearest Neighbour.
Resampling	Re-sampling consists of removing samples from the majority class (under-sampling) and / or adding more examples from the minority class (over-sampling)
Subsampling	[45] defined the sub sample method is randomly drawn from the dominant sample and is then used to form a pooled sample together with the smaller sample. The sub-sampling technique was adopted by [19] from [46] to tackle the class imbalance problem. Sub-sampling equalizes the number of examples from each class in the dataset by randomly selecting a portion of examples from the majority classes.
Stratified sampling	[47] explained that in the Stratified Sampling method, the population is first divided into subgroups, or strata, that all share a similar characteristic. It is used when the measurement of interest to vary between the different subgroups is reasonably expected, and representation from all the subgroupsis ensured.
Splitting	Splitting concatenated subject social media posts into smaller parts, or chunks, was [36] solution to their class imbalance problem. They used the chunks as individual samples and were labeled according to the label of the respective chunk's author. Two ways of voting were proposed for aggregating the classification of individual chunks. The first is majority voting based on predicted chunk classifications, and the second one is the average of predict
	probabilities over each chunk, with a threshold. Chunk size and the threshold value were hyper parameters set to 250 words and 0.5, respectively. To balance out the classes when training, they decided to always include all postsof positive subjects and a fixed number of posts for a negative subject.
Spectrograms	[22] overcame the issue of class imbalance by cropping the spectrograms of each participant into 4 second slices. Then, participants were randomly sampled in equal proportion from each class depressed and not depressed.

Focal Loss Function	The focal loss function was introduced for the classification tasks of imbalanced datasets, where it can concentrate on the hardly predicted class (minority class). [13] said it is a transformed form of the cross-entropy loss function. In 2017, the ICCV award-winning paper by [48] presented a reshaped cross-entropy loss function named Focal Loss, which decreased the weights for the samples in the majority class while focused on the samples of the minority-class.

3.3.2. Data level Ensemble Class Imbalanced Handling Technique

Combining class imbalance techniques were also common in the articles. [42] reported that the hybrid of under sampling and oversampling method is a method of balancing data by combining under sampling and oversampling methods. The number of major class data is reduced by using the concept of under sampling method and so does the amount of minor class data added using the concept of oversampling. Figure 4 below illustrates the percentage frequency of the ensemble class imbalance technique used in the articles with the ensemble Oversampling, Undersampling and SMOTE technique as the most used followed by SMOTE Down sampling and Undersampling, and The Under sampling, Over sampling and ROSE technique as the least used.

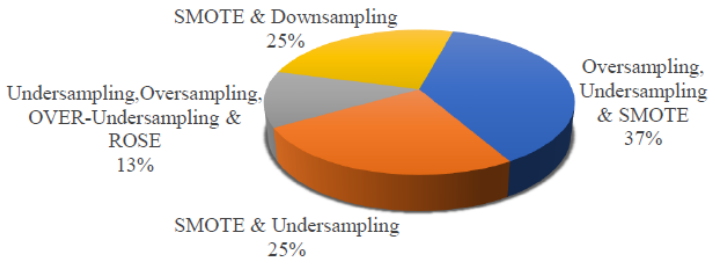

Figure 4. The ensemble class imbalance handling techniques used in the articles

3.3.2. Model Level Approaches

Table 4 describes model (algorithmic) level methods conducted by some of the reviewed articles which section 1.3 explained that the model approach concentrates on modifying existing models to alleviate their bias towards the majority groups.

Table 4. Description of Model Level approach used by some articles

Model Level Approach	Description
MTNet Algorithm	[39] used the MTNet algorithm to generate sample batches with balanced class distribution to achieve more effective optimization. This shares the same spirit as oversampling in imbalanced learning.
AdaBoost and collaborative representation (AdaBoost-CRC)	[41] combined collaborative representation classifier (CRC) and AdaBoost ensemble algorithm named as AdaBoost-CRC to detect Severe Major Depression Disorders (SMDD). Aiming at the data imbalance issue, AdaBoost-CRC classifier structure was created in which AdaBoost was used to discriminate the result of each weak classifier according to its weight.
Autoencoder	[6] investigated an alternate anomaly detection approach to mitigate severe class imbalance in depression datasets. They trained an auto encoder using the location traces of non-depressed users, majority class, which is then able to detect depressed subjects as anomalies. Using the Student Life dataset, [6] established that Depression can be detected from mobility traces and also extracted various mobility features. They then used an auto encoder to project these high dimensional location features into a lower dimensional space.

3.4. Performance with the Imbalanced Dataset Technique

The articles applied machine learning algorithms after the pre-processing application of class imbalance handling methods. Ideally, the F1 metric is used to measure class imbalance techniques performance, however, the articles used a range of performance metrics to measure their model performance not the class imbalance technique. Some of the articles in Table 5 show their experimental results before and after implementation of the class imbalance handling techniques as well as the various model performance measures used.

Although most of the articles do not show the results before implementation, articles such as [28] demonstrate that SMOTE has been used to remove their class imbalance problem by changing their 66% to 33% imbalanced data to 50-50%. [38] has also stated that their study confirmed the effectiveness of SMOTE using an imbalanced binary dataset. [31] has stated that their implementation of SMOTE has improved the result of overall accuracy, sensitivity compared to classification without balancing the data.

Table 5. Experimental results before and after implementation of the class imbalance handlingtechniques as well as the various model performance measures used

Article	Algorithm	Before Score (Average)	Class Imbalance Handling Technique	After score (Average)
[18]	Support Vector Machine (SVM), Naïve Bayes, DecisionTree, Logistic regression, AdaBoostand Bagging	n/a	Undersampling, Oversampling and SMOTE	AUC = 0.75
[11]	Random Forest (RF)	Accuracy=0.91	Random oversampling Undersampling	Accuracy=0.9
[3]	Bidirectional Long ShotMemory and Time Convolution NeuralNetwork (CNN)	Accuracy = 0.78	SMOTE	Accuracy =0.90

[28]	K-Nearest Neighbour, AdaBoost, Bagging, GradientBoosting and XGBoost	n/a	SMOTE	Accuracy =0.92
[40]	n/a	n/a	Oversampling, Undersampling, Over-Under Sampling and ROSE	Accuracy =0.93
[38]	RF	n/a	SMOTE, Under sampling and Oversampling	Accuracy =0.68
[35]	n/a	Accuracy = 0.53	SMOTE	Accuracy =0.72
[24]	RF and SVM	n/a	Oversampling	Accuracy=0.82
[39]	Logistic Regression, SVM and Multiple Layer Perception	n/a	MTNet algorithm	AUC = 0.81
[13]	CNN	n/a	Focal Loss Function	F1= 0.80
[19]	DepAudioNet	F1= 0.62	sub-sampling	F1=0.58
[41]	AdaBoost and collaborative representation (AdaBoost-CRC)	Accuracy = 0.60	AdaBoost-CRC	Accuracy=0.66
[6]	SVM, RF, MLP, XGBoost	n/a	Autoencoder	F1= 0.91
[23]	CNN–LSTM	Accuracy=0.83	Oversampling	Accuracy=0.89
[31]	Random Forest	n/a	SMOTE	Accuracy=0.90

4. DISCUSSION

Different techniques have been used to overcome the class imbalanced dataset problem when predicting and detecting depression. This problem is prevalent in health care records which may show that the majority class are not depressed, and the minority class are depressed. The articles that have been reviewed have used numerous data level and model level approaches to handle the class imbalance problem. The articles have shown how their class imbalance handling techniques have improved their model performance and have a good depression prediction and detection recall. It was found that seven of the articles, used the Distress Analysis Interview Corpus - Wizard of Oz (DAIC- WOZ) dataset which may be because of its development for depression diagnosis. Electronic Health Records were also mostly used with six studies. Surprisingly, social media as a data source for depression was only considered by six of the studies.

Majority of the articles reviewed have used more single sampling techniques, particularly the most popular method SMOTE used by [33], [32] and [31]. The SMOTE implementation has shown to improve classification result on class imbalanced distribution as demonstrated by [28] who achieved 92.56% Accuracy after the applying SMOTE. [35] results illustrated that the SMOTE approach to handle class imbalance gives better results for Depression detection. To ensure that they do not overfit the models, [26] used a cross-validation approach to model building. Another popular methodis Random oversampling which has also been used by majority of the studies for handling class imbalance data like [3] and [24]. However, it has been seen that the oversampling method used to handle the imbalance data may have contributed to overfitting and impacted model performance in most studies [27].

The model level techniques used by [39], [41] and [6] demonstrate thorough understanding of the algorithms to tackle the class imbalance problem. For instance, [39] used the MTNet algorithm to generate sample batches with balanced class distribution to achieve more effective optimization. This shares the same spirit as oversampling in imbalanced learning. This demonstrates that algorithms canalso be used for class imbalance handling.

An ensemble class imbalance technique has mostly been with SMOTE, under sampling and oversampling. One disadvantage of the under-sampling approach is that some useful information might be lost from the majority class due to the under sampling. [44] describes that under sampling runs the risk of discarding potentially useful data whilst oversampling, besides increasing the learning time required, can lead to overfitting, or the generation of a rule specifically for the replicated data. However, [34] and [35] find that the combination of SMOTE and under-sampling performs better than only under-sampling.

4.1. Limitations

The articles that used datasets from various sources, that is from Social Media, Surveys and the DAIZ-WOZ dataset varied in sample size. This may have affected the results accuracy from the classimbalance handling technique used. The small dataset size used by [18] had 173 sample size, [6] had 48 individuals and [41] had 57 samples which may not be enough to generate a general balance sample. Most of the research that have used class imbalanced datasets concentrate on the results after class imbalance handling techniques have been applied and do not document results before they have been applied. Although these techniques do improve Depression classification tasks, it is hard to measure from research just how much they do improve model performance.

This review paper only used research that applied class imbalanced techniques for Depression prediction and detection and did not consider the research that did not apply any technique. The comparison of research that applied the techniques and those that did not, could have assisted in measuring the difference in model performance. This review paper also did not focus on the dataset ratio of class imbalance of every sample size of the reviewed studies because some were not provided.

4.2. Future Work

As much as research has used oversampling and ensemble sampling techniques, to the best of the authors' knowledge, there is few research that applied under sampling techniques on its own forclass imbalanced handling. [11] has proved that under sampling methods with Random Forest can achieve a high Accuracy of 0.93. Due to SMOTE not working well with high dimensional variables, further research is needed to compare the accuracy of SMOTE, under sampling, and oversampling for class imbalanced datasets with high dimensional variables. There are many causes and associations of Depression as evident in [24] where the NKI-Enhanced dataset contained 90 features. Due to the variety of Depression features and the relationship between them, there is a gap in research focused on Depression regression tasks. Data resampling for regression has not been well researched and experimented. Research like [49] plan to explore this in future work.

5. CONCLUSION

In summary, this paper set out to synthesis research conducted using class imbalance data handling techniques for Depression prediction and detection using a PRISMA methodology for systematic review. It was found that in the research, different data sources for datasets came from Task Challenge Workshop forums like Audio/Emotion challenge for 2013, National studies on health- related behaviors of persons living in the north of the Netherlands and Social Media. These datasets can be imbalanced with the majority class are not depressed, and the minority class are depressed.

Research has shown that class imbalance handling methods such as Random oversampling methods, SMOTE, under sampling and model level methods can improve the performance of prediction models after their application. A combination of SMOTE and oversampling techniques showed to bethe most common ensemble technique for class imbalance handling. This paper did not focus on the dataset class imbalance ratio, although, the statistics were provided by the articles. The comparison of model performance between the articles that used different dataset sample sizes may have produced an inaccurate measure of the applied class imbalance techniques. There is few research that show the effectiveness of under sampling techniques used only for handling class imbalance datasets with the exception of [11]. Future research can consider regression modelling as a diagnosis of Depression. Further research is needed to compare the Accuracy of SMOTE, under sampling, and oversampling for imbalanced data with high dimensional variables.

REFERENCES

[1] Rana MS, Kabir MR. “Determining Clinical Depression From The Analysis of Socio- Economic Attributes.” In2020 23rd International Conference on Computer and Information Technology (ICCIT) 2020 Dec 19 (pp. 1-6). IEEE.

[2] Mali, Dixita, Kritika Kumawat, Gaurav Kumawat, Prasun Chakrabarti, Sandeep Poddar, Tulika Chakrabarti, Jemal Hussaine et al. "A Machine Learning Technique to Analyze Depressive Disorders. “2021.

[3] Mao K, Zhang W, Wang DB, Li A, Jiao R, Zhu Y, Wu B, Zheng T, Qian L, Lyu W, Ye M. “Prediction of Depression Severity Based on the Prosodic and Semantic Features with Bidirectional LSTM and Time Distributed CNN.” IEEE Transactions on Affective Computing. 2022 Feb 24.

[4] Keya MS, Han A. “A Performance Analysis of Depression Ratio using Machine Learning Approaches.” In2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS) 2022 Feb 23 (pp. 215-219). IEEE.

[5] Churi, Himanshu, Parul Keshri, Simran Khamkar, and Amruta Sankhe. "A Deep Learning Approach For Depression Classification Using Audio Features." 2021.

[6] Gerych W, Agu E, Rundensteiner E. “Classifying depression in imbalanced datasets using an autoencoder-based anomaly detection approach.” In2019 IEEE 13th International Conference on Semantic Computing (ICSC) 2019 Jan 1 (pp. 124-127). IEEE.

[7] Bach M, Werner A, Palt M. “The proposal of undersampling method for learning from imbalanced datasets.” Procedia Computer Science. 2019 Jan 1; pp. 159:125-34.

[8] Cao L, Shen H. “CSS: Handling imbalanced data by improved clustering with stratified sampling. Concurrency and Computation:” Practice and Experience. 2022 Jan 25;34(2):e6071.

[9] Google. “Imbalanced Data” 2002, [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

[10] Kotb MH, Ming R. “Comparing SMOTE family techniques in predicting insurance premium defaulting using machine learning models.” International Journal of Advanced Computer Science and Applications. 2021;12(9).

[11] Sawangarreerak S, Thanathamathee P. “Random forest with sampling techniques for handling imbalanced prediction of university student depression.” Information. 2020 Nov 5;11(11):519.

[12] Khalilia, M., Chakraborty, S., & Popescu, M. (2011). "Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1), pp. 1-13.
[13] Solieman H, Pustozerov EA. "The detection of depression using multimodal models based on text and voice quality features." In2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus) 2021 Jan 26 (pp. 1843-1848). IEEE.
[14] Canuma. P, "How to Deal With Imbalanced Classification and Regression Data" 2022, [Online]. Available: https://neptune.ai/blog/how-to-deal-with-imbalanced-classification-and-regression-data.
[15] Pristyanto Y, Pratama I, Nugraha AF. "Data level approach for imbalanced class handling on educational data mining multiclass classification." In2018 International Conference on Information and Communications Technology (ICOIACT) 2018 Mar 6. pp. 310-314. IEEE.
[16] Shatte AB, Hutchinson DM, Teague SJ. "Machine learning in mental health: a scoping review of methods and applications." Psychological medicine. 2019 Jul;49(9):1426-48.
[17] Saqib K, Khan AF, Butt ZA. "Machine learning methods for predicting postpartum depression: Scoping review." JMIR mental health. 2021 Nov 24;8(11):e29838.
[18] Natarajan S, Prabhakar A, Ramanan N, Bagilone A, Siek K, Connelly K. "Boosting for postpartum depression prediction." In2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) 2017 Jul 17. pp. 232-240. IEEE.
[19] Bailey A, Plumbley MD. "Gender Bias in Depression Detection Using Audio Features." In2021 29th European Signal Processing Conference (EUSIPCO) 2021 Aug 23 (pp. 596-600). IEEE.
[20] DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, Georgila K, Gratch J, Hartholt A, Lhommet M, Lucas G. SimSensei Kiosk: "A virtual human interviewer for healthcare decision support." InProceedings of the 2014 international conference on Autonomous agents and multi-agent systems 2014 May 5 (pp. 1061-1068).
[21] Shinde SG, Tambe AC, Vishwakarma A, Mhatre SN. "Automated Depression Detection using Audio Features." International Research Journal of Engineering and Technol-ogy (IRJET). 2020 May;7(05).
[22] Saidi A, Othman SB, Saoud SB. "Hybrid CNN-SVM classifier for efficient depression detection system." In2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET) 2020 Dec 15 (pp. 229-234). IEEE.
[23] Prabhu S, Mittal H, Varagani R, Jha S, Singh S. Harnessing emotions for depression detection. Pattern Analysis and Applications. 2022 Aug;25(3): pp. 537-47.
[24] Mousavian M, Chen J, Greening S. "Feature selection and imbalanced data handling for depression detection." InInternational Conference on Brain Informatics 2018 Dec 7. pp. 349- 358. Springer, Cham.
[25] Arun V, Prajwal V, Krishna M, Arunkumar BV, Padma SK, Shyam V. "A boosted machine learning approach for detection of depression." In2018 IEEE Symposium Series on Computational Intelligence (SSCI) 2018 Nov 18 (pp. 41-47). IEEE.
[26] Shin D, Lee KJ, Adeluwa T, Hur J. "Machine learning-based predictive modeling of postpartum depression." Journal of clinical medicine. 2020 Sep 8;9(9):2899.
[27] Wang S, Pathak J, Zhang Y. "Using electronic health records and machine learning to predict postpartum depression." InMEDINFO 2019: Health and Wellbeing e-Networks for All 2019 pp. 888-892. IOS Press.
[28] Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS. "An in-depth analysis of machine learning approaches to predict depression." Current research in behavioral sciences. 2021 Nov 1;2:100044.
[29] Na KS, Cho SE, Geem ZW, Kim YK. "Predicting future onset of depression among Community dwelling adults in the Republic of Korea using a machine learning algorithm." Neuroscience Letters. 2020 Mar 16;721:134804.
[30] Lee C, Kim H. "Machine learning-based predictive modeling of depression in hypertensive populations. "PloS one. 2022 Jul 29;17(7):e0272330.
[31] Xin LK, "Prediction of depression among women using random oversampling and random forest." In2021 International Conference of Women in Data Science at Taif University (WiDSTaif) 2021 Mar 30. pp. 1-5. IEEE.
[32] Nandanwar H, Nallamolu S. "Depression Prediction on Twitter using Machine Learning Algorithms." In2021 2nd Global Conference for Advancement in Technology (GCAT) 2021 Oct 1 (pp. 1-7). IEEE.

[33] Asare KO, Terhorst Y, Vega J, Peltonen E, Lagerspetz E, Ferreira D. "Predicting depression from smartphone behavioural markers using machine learning methods, hyperparameter optimization, and feature importance analysis: exploratory study." JMIR mHealth and uHealth. 2021 Jul 12;9(7):e26540.
[34] Dowlagar S, Mamidi R. "DepressionOne@ LT-EDI-ACL2022: Using Machine Learning with SMOTE and Random UnderSampling to Detect Signs of Depression on Social Media Text." nProceedings of the Second Workshop on Language Technology for Equality, Diversity and inclusion 2022 May (pp. 301-305).
[35] Gupta S, Goel L, Singh A, Prasad A, Ullah MA. "Psychological Analysis for Depression Detection from Social Networking Sites." Computational Intelligence and Neuroscience. 2022 Apr 6;2022.
[36] Banovic L, Fatoric V, Rakovac D. "How Soon Can We Detect Depression?." Text Analysis and Retrieval 2019 Course Project Reports. 2019:1.
[37] Wu MY, Shen CY, Wang ET, Chen AL. A deep architecture for depression detection using posting, behaviour, and living environment data. Journal of Intelligent Information Systems. 2020 Apr;54(2):225- 44.
[38] Byeon H. "Predicting the depression of the South Korean elderly using SMOTE and an imbalanced binary dataset. " International Journal of Advanced Computer Science and Applications. 2021;12(1).
[39] Pang G, Pham NT, Baker E, Bentley R, van den Hengel A. "Deep Depression Prediction on Longitudinal Data via Joint Anomaly Ranking and Classification." InPacific-Asia Conference in Knowledge Discovery and Data Mining 2022 (pp. 236-248). Springer, Cham
[40] Sharma A, Verbeke WJ. "Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch dataset (n= 11,081)." Frontiers in big Data. 2020:15.
[41] Zhang J, Yin H, Wang J, Luan S, Liu C. "Severe major depression disorders detection using adaboost-collaborative representation classification method." In2018 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC) 2018 Aug 15 (pp. 584-588). IEEE.
[42] Choirunnisa S, Lianto J. "Hybrid method of undersampling and oversampling for handling imbalanced data." In2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) 2018 Nov 21 (pp. 276-280). IEEE.
[43] Demir S, Şahin EK. "Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes." Avrupa Bilim ve Teknoloji Dergisi. 2022(34):pp. 142-7.
[44] Colton D, Hofmann M. "Sampling techniques to overcome class imbalance in a cyberbullying context." Journal of Computer-Assisted Linguistic Research. 2019 Jul 16;3(3):pp. 21-40.
[45] Chen L, Dou WW, Qiao Z. "Ensemble Subsampling for Imbalanced Multivariate Two-Sample Tests," Chen, L., Dou, WW, and Qiao, Z.(2013), Journal of the American Statistical Association, 108, 1308–1323. Journal of the American Statistical Association. 2014 Apr 3;109(506):871-.
[46] Krawczyk B. "Learning from imbalanced data: open challenges and future directions." Progress in Artificial Intelligence. 2016 Nov;5(4):221-32.
[47] Faculty of Public Health, "Methods of sampling from a population" 2017 [Online], Available: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a epidemiology/methods-of-sampling-population
[48] Lin TY, Goyal P, Girshick R, He K, Dollár P. "Focal loss for dense object detection." InProceedings of the IEEE international conference on computer vision 2017. pp. 2980-2988.
[49] Qureshi SA, Saha S, Hasanuzzaman M, Dias G. "Multitask representation learning for multimodal estimation of depression level." IEEE Intelligent Systems. 2019 Nov 22;34(5):45-52.

AUTHOR

Simisani Ndaba is a Teaching Assistant in the Department of Computer Science at the University of Botswana. She holds an Master of Science in Computer Information Systems and a Bachelors in Business Information systems. Her research interests are in Machine Learning and Data Science.