Revealing Mental Disorders Through Stylometric Features in Write-Ups

Tamanna Haque Nipa(✉) and A. B. M. Alim Al Islam

Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh
tamanna.haque8@gmail.com, alim_razi@cse.buet.ac.bd

Abstract. Mental disorders present one of the leading causes of worldwide disability and have become a major social concern, as the symptoms behind mental disorders are almost hidden. Most of the conventional approaches used for diagnosing and identifying mental disorders rely on oral conversations (through interviews) having a limited focus on write-ups. Therefore, in this study, we attempt to explore identifying different types of mental disorders among people through their write-ups. To do so, we collect a total of 6893 posts and discussions that appeared in different problem-specific Internet forums and utilize them to identify different types of mental disorders. Leveraging appropriate machine learning algorithms over the collected write-ups, our study can categorize Depression, Schizophrenia, Suicidal Intention, Anxiety, Post Traumatic Stress Disorder (PTSD), Borderline Personality Disorder (BPD), and Eating Disorder (ED). To achieve a balanced dataset in the process of our study, we apply a combined sampling approach and achieve up to 89% accuracy in the identification task. We perform varied exploration tasks in our study covering 5-fold cross-validation, 5-times repetition on the used dataset, etc. We explain our findings in terms of precision, recall, specificity, and Matthews correlation coefficient to demonstrate the capability of our proposed approach in identifying mental disorders based on write-ups.

Keywords: Stylometric Marker · Imbalanced Dataset · Personal Pronoun

1 Introduction

The term "mental disorder" refers to illnesses characterized by abnormal thoughts, perceptions, emotions, behavior, relationships with others, and difficulties in dealings with daily grief and making healthy decisions [1]. This includes depression, bipolar disorder, schizophrenia, other psychoses, dementia, and developmental disorders such as autism. It needs proper attention, diagnosis, psychotherapy, medicines, and follow-up, as we do for physical problems. However, mental health issues are often ignored due to the stigma associated

S. Longfei and P. Bodhi (Eds.): MobiQuitous 2022, LNICST 492, pp. 248–265, 2023.
https://doi.org/10.1007/978-3-031-34776-4_14

with seeking help and fear of being perceived as a burden to society as well as the lack of provision of mental services, proper identification, social awareness, and human resources [1,2].

Different mental disorders are associated with different presentations. They are generally understood or predicted through our conversations and behaviors. One of the potential warning signs in this regard is talking or writing about hopeless feelings. Researchers have found that 50% to 75% of people with suicidal ideation gave a warning sign through sharing their thoughts with a friend or relative [3].

Nowadays, the widespread use of the Internet has allowed various communication platforms to flourish, where users can share their thoughts via posts or comments. To address the complex issue of diagnosing mental disorders through leveraging this reality, one vital technique can be exploiting the stylometric markers of an individual to reflect the person. Even some usual features of writing can link with abnormalities in mental states [4,5]. Research studies have already been conducted in this area. Still, the topic remains open, as the characteristics of human behavior depend on the surrounding conditions resulting in a yet-to-arrive of efficient approach for revealing mental disorders through stylometric features in write-ups.

To focus on this gap in the literature, we conduct a thorough study over a dataset of 6893 messages posted in 63 forums [6]. We identified seven different mental disorder groups and a control group through mining the text messages of normal and physically ill users. To deal with each user group separately from the rest, we applied binary categorization. As the nature of this dataset is imbalanced, we employed a combined sampling method (oversampling-under sampling) to get a better-balanced dataset. Then, we applied different classification models such as Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) successively together to reckon mental disorders from the personal pronouns used in the texts. Next, we utilized two ensemble models namely Random Forest (RF) and Stacking Classifier (SC) combining the baseline models for better prediction. We also investigated combining similar groups in a class. Among all the applied models and explorations, we achieved the best performance with Stacking Classifier while the Random Forest classifier also performed well. Our contributions made in this paper are summarized as follows.

- We proposed a new method of detecting different mental diseases from personal pronouns used in text data.
- We explored different approaches for feature extraction, machine learning-based model build-up, and evaluations over a benchmark dataset to demonstrate the efficiency of our proposed method.

The rest of this paper is organized as follows. Section 2 describes the previous work and Sect. 3 shows our method. Section 4 explains the training and testing processes. Section 5 and Sect. 6 provide the results and performance analysis. Section 7 concludes the paper by summarizing it with a recommended idea and notations to implement in the future.

2 Related Work

Existing studies show that many researchers have conducted their studies on author-ship attribution [7], authorship verification [8], authorship profiling [9], and classified text properties into five categories namely lexical, syntactic, structural, content-specific, and idiosyncratic [10,11]. Different text analyses considering lexical and syntactic features have been done based on word density [5,12], sentence length [13], frequency of nouns or pronouns [12], functional words [14], and part-of-speech n-grams [15,16]. Besides, structural attributes include attributes relating to text organization and layout [17]. Content-specific features follow word n-grams and depend on important keywords [17,18]. Idiosyncratic features show cultural differences in word formation such as incorrect spelling, misuse of words, and inaccurate verb forms [18,19].

Recent research studies have shown that pronouns can be used to identify gender [12]. It has been also demonstrated that Alzheimer's disease can be detected early through personal pronouns and sensory words [5,6]. For gender detection, 25,000 words have been aggregated across 30 selected articles by word frequency count based on a total of 29 personal pronouns and tested the possibility to be male or female. According to the study, 90% accuracy has been achieved using three personal pronouns namely my, her, and its. Another study has investigated the importance of stylometric markers for depression and Alzheimer's detection, which was based on 45 novels by Iris Murdoch and PD James covering a period from 1954 to 1995. RPAS visualization was applied based on richness, personal pronoun, activity power, and sensory adjectives and it was concluded that Alzheimer's can be detected from written text 12 years earlier.

Another investigation [6] has suggested that absolutist thinking is a vulnerable factor and related to suicidal ideation, borderline personality disorder, and an eating disorder. This study showed that suicidal ideation more relates to absolutist words than psychological distress through a statical analysis and a list of absolutist words was published. Besides, as investigated in [20], dark traits of human beings describe the causes of abuses and crimes, which can be measured by some words identified as negative or positive based on the Russian language.

Machine learning techniques with Natural Language Processing (NPL) can potentially offer new routes for learning mental health conditions and risk factors that have been surveyed in [21] and [22]. Another study presented a systematic review based on space disease diagnosis, psychological disorders, and classification techniques to address ADHD, Alzheimer's, Parkinson's, insomnia, schizophrenia, and mood disorder [23]. A similar review was done based on 565 relevant types of research from 2015 to 2020 [24]. Another research [25] applied unsupervised ML techniques based on 826,961 unique user posts during COVID-19 from 2018 to 2020 on 90 text-driven features. This study presented that BPD and PTSD are significantly associated with the suicidality cluster.

Some other studies reported a positive correlation between some specific words used, which frequently can help to figure out the intention of dying [14,26,27]. Several surprising facts also observed that suicidal tendency increases due

to the death of any famous person [14], both depression and suicidal ideation episodes contain 14 days [26], females and young adults aged from 15–24 years are the main victims of depression [26,27]. Besides, the study in [28] predicted the depression level of the writers based on a computational marker from the written text addressed by DASS-21 (Depression, Anxiety, and Stress Scale). This analysis covered 172 people and summarized that informal text is more suitable to detect depression. Additionally, an algorithm namely SAIPA was proposed using ML for predicting the future risk of suicide from 512,526 tweets followed by the countrywide death rate [29]. PSPO, a Chinese online suicide prevention system has been designed to detect people at risk through suicidal thoughts and behaviors. A total of 27,007 comments were analyzed and realized that suicide ideation is more related to future-oriented words than death-oriented words. All the reported classification accuracies varied around 56%–83% using different statistical measures, ML algorithms, and Deep Learning algorithms. Here, investigations confirmed that language plays an important role in life.

3 Our Proposed Methodology

We provided an overview of our solution in this section. As per our study process depicted in Fig. 1, we (1) collected written messages from different online group members related to their diseases or difficulties, (2) screened them for potential causes with proper data preprocessing, feature extraction, and data balancing method, (3) machine learning algorithms have applied to predict the disorders, and (4) finally we explored and compared the findings with a fitness assessment. This section is divided into different subsections that illustrate the workflow of the proposed experimental procedures in detail.

3.1 Data Collection and Arrangement

The dataset used in this study has been collected from https://figshare.com/ and the source of all the messages or posts are different English language-based internet forums [6,30]. For our study, we divided our study into two segments and rearranged the group formation according to our research aim. For the first portion, the test group is formed with any one group among the concerned mental illnesses and the control group consists of the rest of the mental diseases, general members, asthma, diabetes, cancer patient, students (young people), mums' group, elderly peoples with pension problem and people with job problems. This source dataset also consists of some members who have recovered from high depression and mentioned their improvements. However, this recovery group has not been included in our rearranged groups due to the possibility of noise and outliers. We kept the other groups remain the same. Therefore, this part of the analysis is based on 7 different datasets with different mental problem specifications. We have also cross-checked them with the excel file provided with the reference paper [6] and removed 350 indirect messages that address the problem of family members or friends.

For the second part of our study, we have prepared the four major classes combining the related groups to find out whether mental disorders can be separated from other classes or not. At this point, we have organized the analysis considering the mental disorder group vs. normal + physical health problem + recovery group, normal vs. mental disorder + physical health problem + recovery group, physical health problem vs. mental disorder + normal + recovery group, and recovery group vs mental disorder + physical health problem + normal group. Table 1 shows the number of messages that we consider for this experiment.

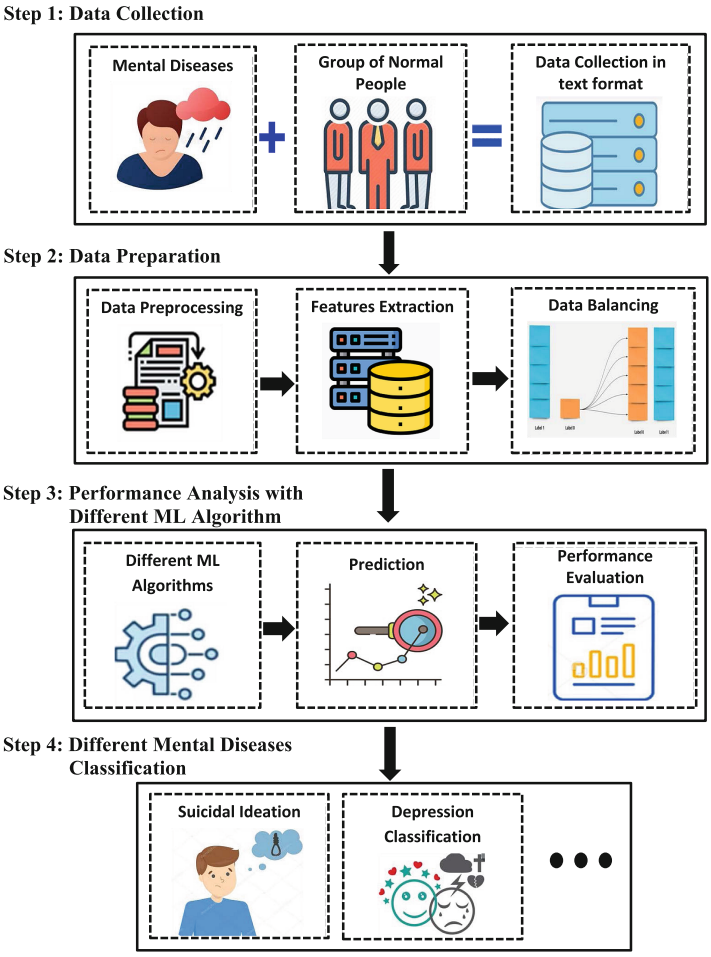

Fig. 1. Schematic diagram of our study to explore different mental diseases through Stylometric features in texts.

Four criteria have been maintained in the procedure of data collection: (a) all the messages must be in the English language (b) the minimum length of the post should be 100 words (b) the messages be directly represented by the members, and (c) the written text must be in continuous prose. Here only the first post of the members has been considered and for some cases, multiple posts have been combined into a single post as the signal post was very poor in length for consideration.

Table 1. Statistics of the dataset were collected from different internet forums [8].

Groups	No of messages	Sample percentage
Anxiety	596	8.6%
BPD	325	4.7%
Depression	531	7.7%
ED	546	8.0%
PTSD	535	7.7%
Schizophrenia	592	8.5%
Suicidal Ideation	366	5.3%
Normal (General, Mums, Pension, Student, Job Problem)	1373 (959, 157, 16, 132, 109)	20.0%
Physical Diseases (Asthma, Diabetes, Breast Cancer, Bowel Cancer, Lung Cancer, Prostate Cancer)	1461 (418, 590, 141, 96, 123, 93)	21.1%
Recovery	568	8.2%
Total	6893	100%

3.2 Data Preparation

After the arrangement of the groups, all the text files have been preprocessed for the avoidance of unwanted noise or redundancy. We have added the group type and index with each file. Each text file is presented as a tuple in the dataset with the index number.

Internet Acronyms Replacement. Non-standard words and phrases of language are mostly observed in the messages of internet users, such as U, GR8, MSG, B4, etc. These internet acronyms are used for easy interpretation by others and to save time in typing. It has become very important to detect, translate and replace the short form with the actual word or group of words. We have collected 2174 short forms of words from different websites [31,32]. While replacing them we found that the same acronyms can have multiple meanings. To avoid conflicts, we tried to pick the most potential one. For example, 'yr' can be 'your' or 'yeah right' or 'you'. In this case, we selected 'your' to substitute 'yr'.

Feature Extraction and Data Balancing. Our main effort is based on pronouns and used as an alternative to nouns to avoid repetition. English literature

contains more than 100 pronouns which can be grouped into 10 categories: personal, possessive, reflexive, intensive, demonstrative, interrogative, relative, indefinite, reciprocal, and archaic. Among these, personal pronouns, progressive pronouns, reflexive pronouns, and intensive pronouns are used for a per-son or thing in a sentence specifically. Personal pronouns are words to highlight the people or things in our sentences which can be both subjective and objective. The personal pronouns for subjects are I, you, he, she, it, we, and they. For objects, personal pronouns are me, us, you, her, him, it, and them. A possessive pronoun (mine, ours, yours, hers, his, theirs) designates ownership and can substitute for noun phrases. Reflexive pronouns are pronouns that are used to show that the subject of the sentence is receiving the action of the verb (myself, yourself, himself, herself, itself, ourselves, themselves). Intensive Pronouns are pronouns that are used only to emphasize the subject. Intensive and reflexive pronouns are the same words; however, they act differently in the sentence. With these features, we also considered the possessive adjectives (my, our, your, her, his, their) as they resemble a pronoun [33]. A total of 29 words have been considered for the analysis.

According to our observation, all the groups are dissimilar to each other in respect of the total number of group members shown in Table 1. We have addressed 7 different mental problems to analyze every group separately from the others. We also considered 4 classes and explored every class separately from the other classes. Therefore, every test group and class is found highly imbalanced concerning others. To handle the issues of imbalance problem, we have applied SMOTEENN. Synthetic Minority Over-sampling Technique (SMOTE) and its variations were always found to be the better choice for oversampling. To avoid overfitting, we choose SMOTEENN, a hybridization of the oversampling and under-sampling process. It is a combination of SMOTE and, the Edited Nearest Neighbor (ENN) technique [34].

As a part of preprocessing, every sample has been normalized with the standard scaling procedure. Every instance of the concerned group or class is marked as 1 and other instances from the rest of the groups or classes are marked as 0. To keep the words in the messages, remain the same, we avoided the stemming process. Only the frequency of the identified words has been measured.

4 Evaluation Procedure

This section discusses the experimental techniques and classification models used for this research in detail. Aside from that, it has also shown the performance measures that have been used to evaluate the outcomes. Python 3.6 with scikit-learn is used as a machine learning tool. We have also investigated our prepared datasets using Weka 3.9.4 and chose the best-suggested algorithms for this processing.

4.1 Classification Methods

We employed the 3 most used Machine Learning algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) to

examine our experiments. The ratio of the training and testing data is 8:2. 5-fold cross-validation with 5 times repetition is used and the results are summarized by taking the mean of the outcomes. Further refinement is made by following the ensemble process, Stacking Classifier (SC), and Random Forest (RF) which shows a noteworthy upgrading.

4.2 Assessing Performances

Various assessing metrics are available in Machine Learning to analyze model performance. For the comparative performance analysis of our study, we have presented the consequence of the testing set using balanced accuracy, precision, recall, specificity, f1-score, and Matthews Correlation Coefficient (MCC). Instead of accuracy as a valuation tool, we used balanced accuracy as the dataset is imbalanced. We also included recall and specificity to illustrate how well a model can predict both majority and minority classes. Matthews correlation coefficient (MCC) is another powerful and informative measure that finds the correlation of true classes with the predicted labels. Likewise, the f1-score exhibits more significant performance measures than accuracy in the case of the imbalanced dataset considering precision and recall. Precision is the rate of correctly classified positive predictions made over all the positive predicted samples. Similarly, specificity measures the proportion of correctly identified negatives over the total negative prediction made by the model. However, recall measures the correctly identified positive samples from all actual positives. High precision with low recall indicates that very few results have been predicted and most of the predicted labels are correct. A model with low precision and high recall behave oppositely. High precision with high recall represents a better classification model [35,36]. We have also shown the tradeoff between precision and recall using the precision-recall curve.

$$BalancedAccuracy = \frac{Sensitivity + Specificity}{2} \tag{1}$$

$$RecallorSensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

Here, TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative respectively. Moreover, the experimental result has been presented using the necessary graphical representation to compare different ML models and assorted datasets. We have also applied the Chi-square test to identify the best combination of pronouns following the feature selection method for better identification of mental disorders among all types of combined groups.

5 Experimental Results

In this section, we presented and compared the models used in our study on the prepared datasets as mentioned before. First, we showed our testing to identify Anxiety, BPD, Depression, ED, PTSD, Schizophrenia, and Suicidal ideation. Next, we explored our experiment to differentiate among the classes of normal people, recovery groups, people with mental disorders, and physically ill people.

5.1 Mental Disorder Detection and Performance Measure

First, we explore 5 machine learning algorithms and find the best fit for our purpose. It is observed that both Stacking Classifier and Random Forest outperform on K-Nearest Neighbors, Support Vector Machine, and Logistic Regression approach in all classification scenarios. While implementing Stacking Classifier, we applied different model combinations and achieved the best result while the base models were K-Nearest Neighbors, and the final estimator to combine the predicted values of the base models was Logistic Regression. Stacking Classifier is slightly better than Random Forest concerning the performance metrics. However, Random Forest performs much faster than Stacking Classifier. The performance measures of the assigned models for the identification of seven mental disorder groups are listed in Table 2. It exhibits the mean values and standard deviation of balanced accuracy, specificity, f1-score, and MCC representing all the groups respectively.

Our first dataset is based on the samples labeled as Anxiety. Among the implemented models, Stacking Classifier shows the best performance with highly balanced accuracy, f1-score, and MCC (B.Ac. 86%, F1. 82%, MCC 73%). Precision and recall values are both found promising. However, precision is a little higher in respect of its recall.

The second, third, and fourth datasets are based on BPD, Depression, and ED. The similarity is found with Anxiety for BPD (B.Ac. 87%, F1. 84%, MCC 76%), Depression (B.Ac. 85%, F1. 80%, MCC 72%), and ED (B.Ac. 89%, F1. 87%, MCC 80%) detection using the Stacking Classifier model with higher precision and recall values. Random Forest and K-Nearest Neighbors are likewise showing satisfactory performances. On the other hand, Support Vector Machine, and Logistic Regression are found dis-satisfactory with lower evaluation metrics compared to others.

The last three datasets are about PTSD (B.Ac. 59%–83%, F1. 77%–90%, MCC. 25%–69%), Schizophrenia (B.Ac. 58%–83%, F1. 76%–90%, MCC. 20%–69%), and Suicide Ideation (B.Ac. 69%–88%, F1. 78%–89%, MCC. 42%–78%)

Table 2. Class-wise different performance metrics (mean values) of ML models (SC = Stacking Classifier, RF = Random Forest, KNN = K-Nearest Neighbors, SVM = Support Vector Machine, LR = Logistic Regression, B.Ac. = Balanced Accuracy, MCC = Matthews Correlation Coefficient, F1 = F1-score, Pre. = Precision, Rec. = Recall, and Spe. = Specificity)

Group	Algo.	B.Ac.		MCC		F1		Spe.	
		AVG	STD	AVG	STD	AVG	STD	AVG	STD
Anxiety	LR	0.68	0.01	0.38	0.02	0.58	0.02	0.84	0.00
	RF	0.84	0.12	0.69	0.23	0.80	0.16	0.91	0.05
	SVM	0.80	0.13	0.61	0.24	0.74	0.17	0.91	0.05
	KNN	0.82	0.12	0.66	0.24	0.77	0.17	0.92	0.05
	SC	0.86	0.12	0.73	0.23	0.82	0.16	0.94	0.05
BPD	LR	0.66	0.01	0.33	0.01	0.57	0.01	0.82	0.01
	RF	0.86	0.14	0.72	0.28	0.82	0.18	0.91	0.07
	SVM	0.80	0.16	0.62	0.30	0.75	0.20	0.90	0.06
	KNN	0.83	0.15	0.68	0.29	0.79	0.20	0.92	0.07
	SC	0.87	0.14	0.76	0.27	0.84	0.19	0.94	0.07
Depression	LR	0.65	0.01	0.33	0.02	0.53	0.01	0.86	0.01
	RF	0.83	0.13	0.68	0.25	0.78	0.18	0.92	0.04
	SVM	0.78	0.15	0.58	0.28	0.69	0.23	0.92	0.04
	KNN	0.81	0.15	0.64	0.27	0.74	0.22	0.93	0.04
	SC	0.85	0.14	0.72	0.27	0.80	0.21	0.95	0.05
ED	LR	0.76	0.01	0.55	0.02	0.70	0.01	0.90	0.01
	RF	0.88	0.09	0.77	0.16	0.85	0.11	0.93	0.03
	SVM	0.85	0.10	0.71	0.17	0.81	0.12	0.93	0.03
	KNN	0.86	0.09	0.74	0.17	0.83	0.12	0.94	0.04
	SC	0.89	0.09	0.80	0.17	0.87	0.11	0.96	0.04
PTSD	LR	0.59	0.01	0.25	0.02	0.77	0.00	0.27	0.01
	RF	0.82	0.16	0.65	0.29	0.88	0.08	0.70	0.31
	SVM	0.75	0.18	0.53	0.33	0.85	0.09	0.55	0.37
	KNN	0.78	0.18	0.60	0.32	0.87	0.09	0.61	0.35
	SC	0.83	0.17	0.69	0.31	0.90	0.09	0.70	0.33
Schizophrenia	LR	0.58	0.01	0.20	0.02	0.76	0.00	0.28	0.02
	RF	0.80	0.16	0.62	0.30	0.88	0.09	0.69	0.30
	SVM	0.74	0.18	0.51	0.33	0.85	0.09	0.54	0.37
	KNN	0.77	0.18	0.58	0.33	0.87	0.09	0.60	0.35
	SC	0.83	0.17	0.68	0.32	0.90	0.09	0.69	0.34
Suicide	LR	0.69	0.01	0.42	0.02	0.78	0.01	0.54	0.02
	RF	0.86	0.12	0.74	0.23	0.90	0.09	0.80	0.19
	SVM	0.82	0.13	0.66	0.24	0.87	0.09	0.72	0.22
	KNN	0.84	0.13	0.71	0.24	0.89	0.09	0.76	0.20
	SC	0.88	0.12	0.78	0.23	0.92	0.09	0.82	0.20

respectively. Among the 5 ML models with five-fold cross-validation and five times repetition, Stacking Classifier has shown higher balance accuracy with a higher f1-score.

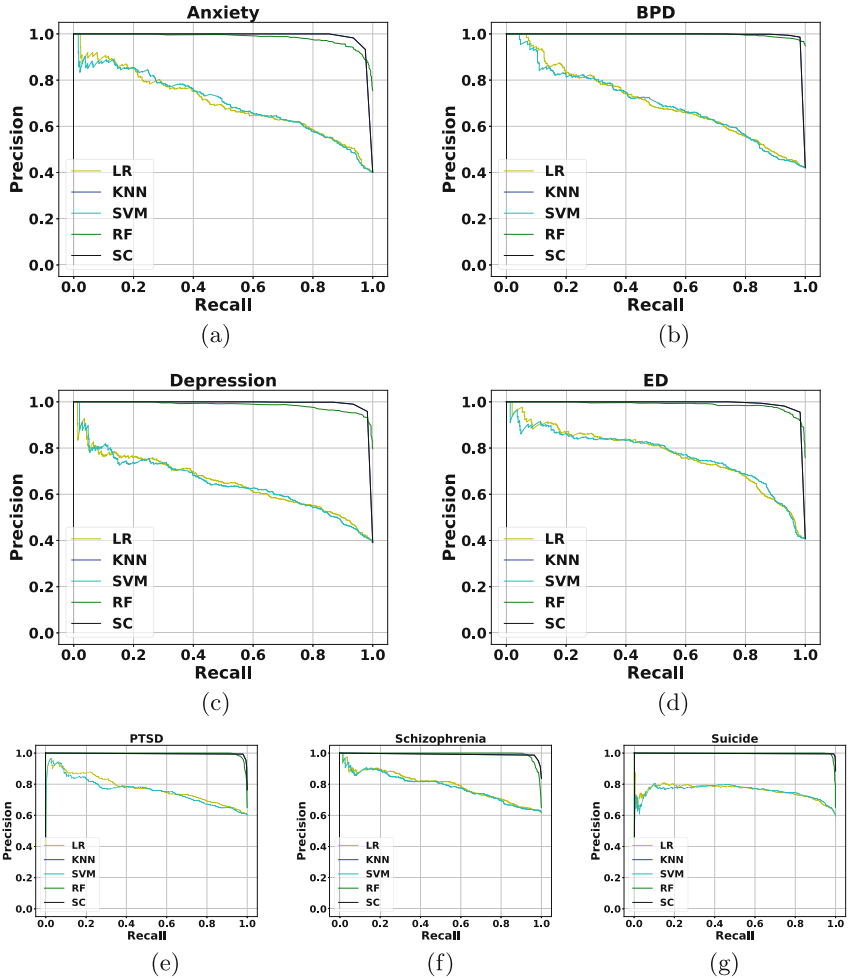

Fig. 2. Precision-recall curve for various mental disorder groups for different algorithms (a) Anxiety, (b) BPD (c) Depression (d) ED (e) PTSD, (f) Schizophrenia, and (g) Suicide Ideation

Figure 2 delineates a comparison among 5 ML algorithms using a precision-recall curve for the seven mental disorder groups. Our method has higher recall with higher precision in detecting specific mental problems from the other groups. The performance of the Stacking Classifier, Random Forest, and K-Nearest Neighbors for all the disorder detection showed significant achievement

compared to Logistic Regression and Support Vector Machine. Both Stacking Classifier and Random Forest approach outperforms over baseline approach in all classification scenario. However, Stacking Classifier is slightly better than Random Forest in most cases, but Random Forest is much faster.

5.2 Class-Specific Performance Measure

In this study, we have also addressed specific class identification as groups in the same class can act similarly. We aim to differentiate people with mental suffering among a variety of people. We have mentioned above the class formation. Therefore, 5 ML algorithms are used for classification with these 4 mentioned classes. All the group statistics and the outcome of all models are revealed in Table 3.

Table 3. Class-wise different performance metrics (mean values) of ML models (SC = Stacking Classifier, RF = Random Forest, KNN = K-Nearest Neighbors, SVM = Support Vector Machine, LR = Logistic Regression, B.Ac. = Balanced Accuracy, MCC = Matthews Correlation Coefficient, F1 = F1-score, Pre. = Precision, Rec. = Recall, and Spe. = Specificity).

Groups	Mod	B.Ac	MCC	F1	Pre	Rec	Spe
Mental disorder vs. Normal +	LR	0.86	0.73	0.88	0.85	0.91	0.82
Physical health problem +	RF	0.88	0.75	0.88	0.88	0.88	0.87
Recovery group	SVM	0.87	0.74	0.88	0.87	0.89	0.85
	KNN	0.88	0.76	0.89	0.88	0.90	0.86
	SC	0.89	0.79	0.90	0.90	0.90	0.88
Normal vs. Recovery+	LR	0.77	0.57	0.71	0.81	0.64	0.90
Physical health problem+	RF	0.87	0.74	0.84	0.87	0.81	0.92
Mental disorder group	SVM	0.84	0.70	0.80	0.85	0.76	0.92
	KNN	0.86	0.73	0.82	0.88	0.78	0.93
	SC	0.88	0.78	0.86	0.90	0.82	0.94
Recovery vs. General+	LR	0.70	0.40	0.77	0.74	0.80	0.60
Physical health problem+	RF	0.85	0.71	0.89	0.88	0.89	0.81
Mental disorder group	SVM	0.81	0.63	0.86	0.84	0.89	0.74
	KNN	0.83	0.68	0.88	0.85	0.91	0.76
	SC	0.87	0.76	0.91	0.89	0.93	0.82
Physical health problem vs General+	LR	0.79	0.60	0.85	0.81	0.90	0.67
Mental disorder+	RF	0.86	0.73	0.90	0.88	0.92	0.80
Recovery group	SVM	0.84	0.70	0.89	0.86	0.92	0.76
	KNN	0.85	0.72	0.90	0.87	0.93	0.77
	SC	0.87	0.76	0.91	0.89	0.94	0.81

Findings are depicted in respect of the bar diagram representing balanced accuracy, precision, recall, specificity, f1-score, and MCC in Fig. 3. According to our observation, all the classes have good balanced accuracy and precision. Moreover, our method has shown a significant f1-score in detecting mental disorders (F1. 88–90%) and recovery groups (F1. 77–91%). It might be for high

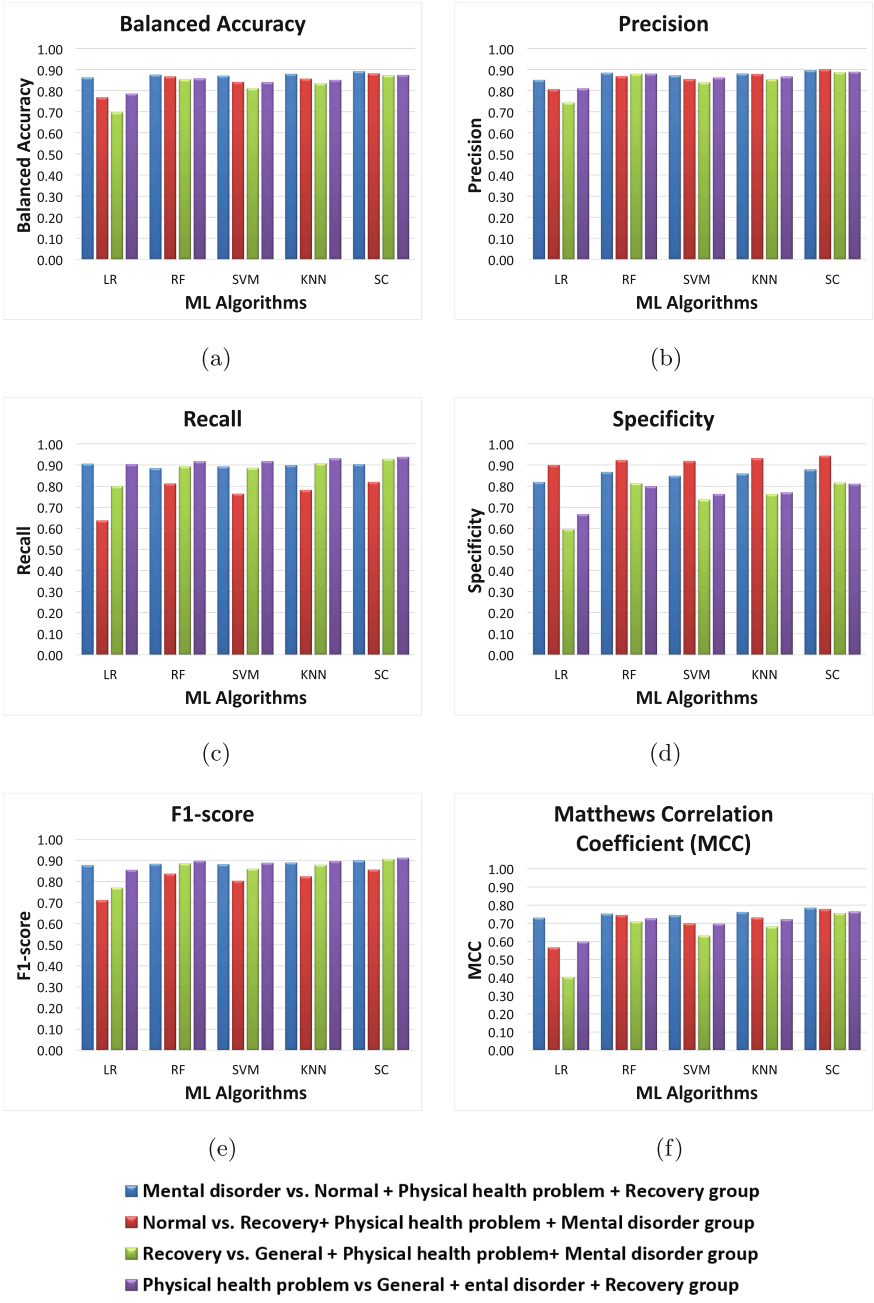

Fig. 3. Different performance metrics for comparison among the 4 classes with similar groups (a) Balanced Accuracy, (b) Precision (c) Recall (d) Specificity (e) F1-score, and (f) MCC

recall and precision indicating to be appropriate for mental problem classification. Specificity calculated for the physical health problem (Spe. 67–81%) and normal class (Spe. 92–94%) is much higher than the other classes which also indicates that opposite classes have been detected properly.

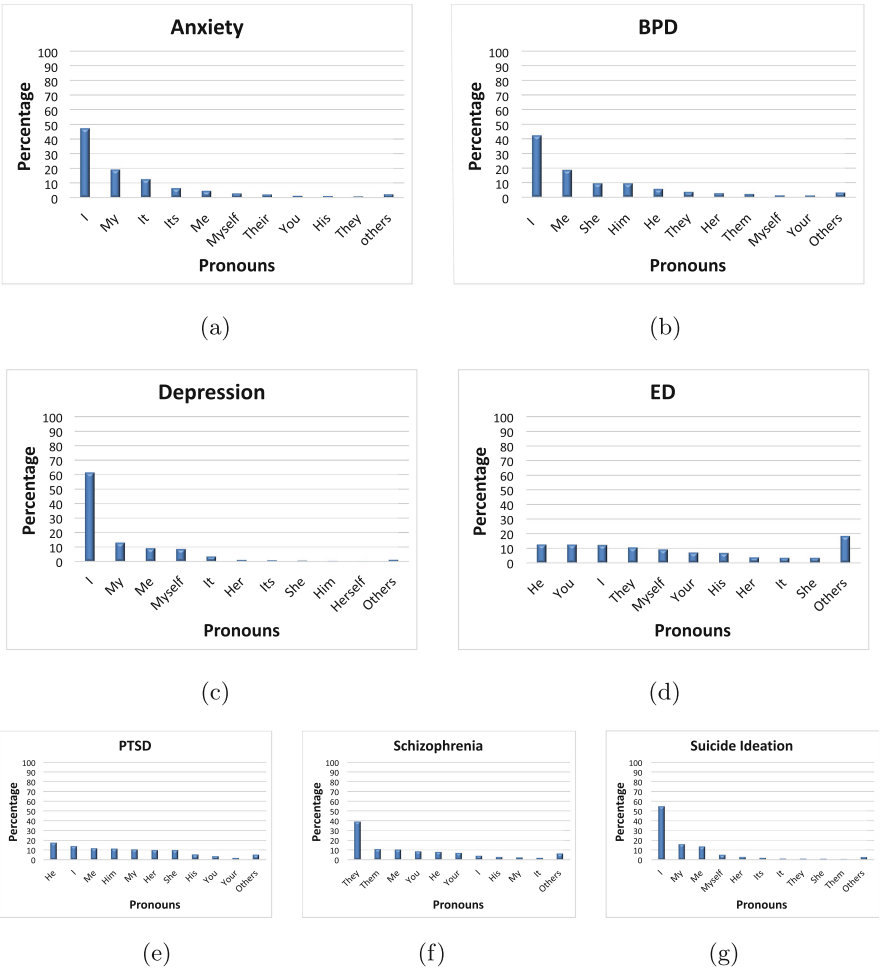

Fig. 4. Bar diagram for various mental disorder groups and 4 classes showing feature scoring for specific pronouns (a) Anxiety, (b) BPD (c) Depression (d) ED (e) PTSD, (f) Schizophrenia, and (g) Suicide Ideation

5.3 Performance for Specific Pronouns

We have tested the chi-squared feature selection method for 29 pronouns described before and observed some specific pronouns have significant scores compared to the other pronouns. The combination of pronouns is different for

different groups. It is noticeable that every group and class has the pronoun 'I'. Surprisingly, both the mental disorder class and the normal class have 'I' with the highest score with almost 60% value. Similarity has been found for another two pronouns 'My' and 'Me' for Anxiety, BPD, Depression, and Suicide Ideation including 'I'. Therefore, combining several pronouns for disorder detection is required. Figure 4 and Fig. 5 show 10 pronouns with the largest score from the others for all the mental disorder groups and four classes respectively. The scores are presented in percentages.

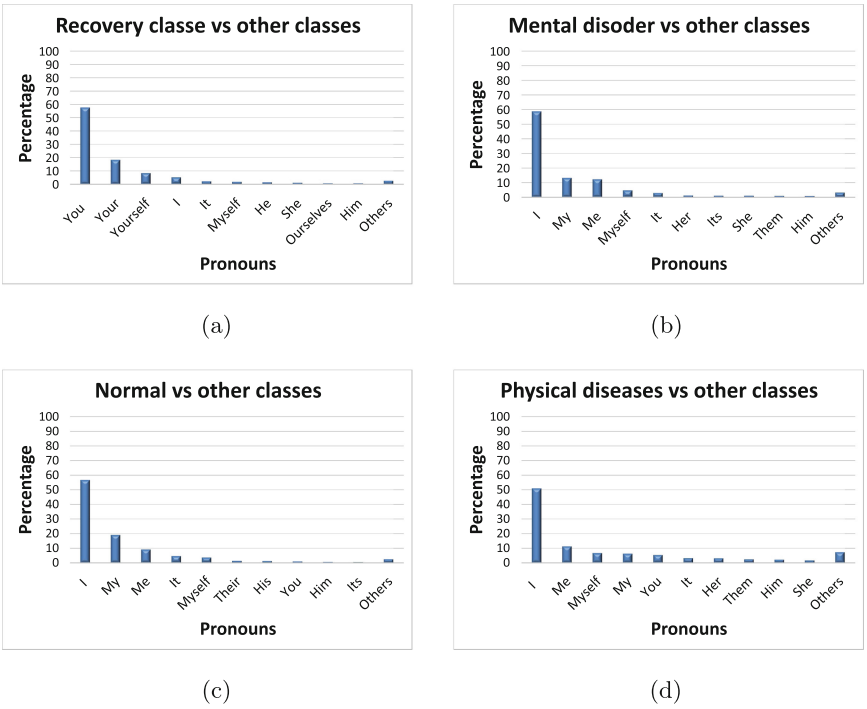

Fig. 5. Bar diagram for different classes showing feature scoring for specific pronouns (a) Recovery Class vs. others (b) Mental disorder vs. others (c) Normal class vs. others, and (d) Physical diseases vs. others

6 Discussion

Among the seven mental disorder groups, Suicide Ideation reported the highest score for correct prediction. PTSD and Schizophrenia also achieved higher scores respectively. Among the algorithms used to measure the implications, it is viewed that Stacking Classifier has performed better than the other algorithms with the highest balanced accuracy and high f1-scores. Random Forest and K-Nearest Neighbors are also investigated with better outputs. Similarly, a good result has

been achieved for class-specific models' implementation and a sharp distance has been observed among the classes considering recall and f1-score. A surprising fact is found that the recovery group still has high similarity with the class mental disorder. Overall, this indicates that there are significant effects of functional words in screening mental diseases.

7 Conclusion

In this paper, we have developed ML-based models to differentiate between mental disease persons and normal individuals from their text messages written in English. Here, we have conducted our study over a benchmark dataset. We have rearranged the group and class formation in the dataset according to our requirements and ob-served the significance of pronouns in defining the status of individuals. Different types of pronouns show promising consequences for mental disease detection through presenting a simple, cost-effective, and less time-consuming approach for diagnosing. Here, as the benchmark dataset was highly imbalanced, the dataset presented a common limitation in data analysis. We have overcome the limitation through applying the necessary combined sampling methods.

To enhance this research further in the future, we need to expand our dataset through including more variety and using different classification processes. Besides, to perform our study more precisely in the future, we need to collect write-ups of clinically diagnosed and classified persons with proper pieces of evidence, experiences, and details that can enhance the diagnosis process. Still, with the success of prediction that can avert future discomforts and support people who are at risk, this study can be a big turn for the future.

Acknowledgement. The work has been conducted at and supported by the Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh.

References

1. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/mental-disorders. Accessed 10 July 2022
2. Mental health: lessons learned in 2020 for 2021 and forward. https://blogs.worldbank.org/health/mental-health-lessons-learned-2020-2021-and-forward. Accessed 10 July 2022
3. Recognizing Suicidal Behavior. https://www.webmd.com/mental-health/recognizing-suicidal-behavior. Accessed 10 July 2022
4. How heavy use of social media is linked to mental illness. https://www.economist.com/graphic-detail/2018/05/18/how-heavy-use-of-social-media-is-linked-to-mental-illness. Accessed 10 July 2022
5. Kernot, D., Bossomaier, T., Bradbury, R.: The stylometric impacts of ageing and life events on identity. J. Quant. Linguist. **26**, 1–21 (2017)
6. Al-Mosaiwi, M., Johnstone, T.: In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clin. Psychol. Sci. **6**, 529–542 (2018)

7. Jockers, M.L., Witten, D.M.: A comparative study of machine-learning methods for authorship attribution. Literary Linguist. Comput. **25**(2), 215–223 (2010)
8. Halvani, O., Winter, C., Pflug, A.: Authorship verification for different languages, genres and topics. Digit. Investig. **16**, 33–43 (2016)
9. Mir, E., Novas, C., Seymour, M.: Social Media and Adolescents' and Young Adults' Mental Health. National Center for Health Research. http://www.center4research.org/social-media-affects-mental-health/. Accessed 10 July 2022
10. Roffo, G., Cristani, M., Bazzani, L., Minh, H.Q., Murino, V.: Trusting skype: learning the way people chat for fast user recognition and verification. In: IEEE International Conference on Computer Vision Workshops (ICCVW), Sydney, Australia, pp. 748–754 (2013)
11. Brocardo, M.L., Traore, I.: Continuous authentication using micro-messages. In: 12th Annual International Conference on Privacy, Security, and Trust, Toronto, Canada (2014)
12. Kernot, D.: Can three pronouns discriminate identity in writing? In: Sarker, R., Abbass, H.A., Dunstall, S., Kilby, P., Davis, R., Young, L. (eds.) Data and Decision Sciences in Action. LNMIE, pp. 397–411. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-55914-8_29
13. López-Escobedo, F., Méndez-Cruz, C.F., Sierra, G., Solórzano-Soto, J.: Analysis of stylometric variables in long and short texts. In: International Conference on Corpus Linguistics (CILC 2013), pp. 604–611 (2013)
14. Burnap, P., Colombo, G., Amery, R., Hodorog, A., Scourfield, J.: Multi-class machine classification of suicide-related communication on Twitter. Online Soc. Netw. Media **2**, 32–44 (2017)
15. Argamon, S., Koppel, M., Avneri, G.: Routing documents according to style. In: Proceedings of the 1st International Workshop on Innovative Information (1998)
16. Baayen, H., Halteren, H.V., Tweedie, F.: Outside the cave of shadows: using syntactic an-notation to enhance authorship attribution. Literary Linguist. Comput. **2**, 110–120 (1996)
17. Hayne, S.C., Pollard, C.E., Rice, R.E.: Identification of comment authorship in anonymous group support systems. J. Manag. Inf. Syst. **20**, 301–329 (2003)
18. Koppel, M., Schler, J.: Exploiting stylistic idiosyncrasies for authorship attribution. In: Proceedings of the IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis, pp. 69–72 (2003)
19. Abbasi, A., Chen, H.: Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. **26**(7), 1–29 (2008)
20. Panicheva, P., Ledovaya, Y., Bogolyubova, O.: Morphological and semantic correlates of the dark triad personality traits in Russian Fackbook tests. In: Proceedings of the IEEE Artificial Intelligence and Natural Language Conference (AINL) Fruct Conference, Saint-Petersburg, Russia (2016)
21. Thieme, A., Belgrave, D., Doherty, G.: Machine learning in mental health: a systematic review of HCI literature to support the development of effective and implementable ML systems. ACM Trans. Comput.-Hum. Interact. **27**(5), 1–53 (2020)
22. Calvo, R.A., Milne, D.N., Hussain, S., Christensen, H.: Natural language processing in mental health application using non-clinical tests. Nat. Lang. Eng. **23**(5), 649–685 (2017)
23. Kaur, P., Sharma, M.: Diagnosis of human psychological disorders using supervised learning and nature-inspired counting techniques: a meta-analysis. J. Med. Syst. **43**(7), 1–30 (2019)

24. Kim, J., Lee, D., Park, E.: Machine learning for mental health in social media: biblio-metric study. J. Med. Internet Res. **23**(3), e24870 (2021)
25. Low, D.M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., Ghosh, S.: Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: an observational study. J. Med. Internet Res. **22**(10), e22635 (2020)
26. Nobles, A.L., Glenn, J.J., Kowsari, K., Teachman, B.A., Barnes, L.E.: Identification of imminent suicide risk among young adults using text messages. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 413–435 (2018). https://doi.org/10.1145/3173574.3173987
27. Du, J., et al.: Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med. Inform. Decis. Making **18**(43), 77–87 (2018)
28. Havigerová, J.M., Haviger, J., Kučera, D., Hoffmannová, P.: Text-based detection of the risk of depression. Front. Psychol. **10**, 513 (2019)
29. Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., Kaminsky, J.A.: A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit. Med. **3**(1), 78 (2020)
30. Supplemental Material (2022). https://doi.org/10.6084/m9.figshare.4743547.v1. Accessed 10 July 2022
31. More Than 2000 of The Most Common Text Abbreviations. https://dexatel.com/blog/text-abbreviations/. Accessed 10 July 2022
32. The Complete List of 1697 Common Text Abbreviations & Acronyms. https://www.webopedia.com/reference/text-abbreviations/. Accessed 10 July 2022
33. The Free Dictionary by Farlex. https://www.thefreedictionary.com/List-of-pronouns.html. Accessed 10 July 2022
34. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Imbalanced classification for big data. In: Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F. (eds.) Learning from Imbalanced Data Sets, pp. 327–349. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98074-4_13
35. Shamsudin, H., Yusof, U.K., Jayalakshmi, A., Khalid, M.N.A.: Combining oversampling and undersampling techniques for imbalanced classification: a comparative study using credit card fraudulent transaction dataset. In: 6th International Conference on Control & Automation (ICCA), pp. 803–808 (2020)
36. Precision-Recall. https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall. Accessed 10 July 2022