Identifying the Signature of Suicidality : A Machine Learning Approach

Md Naimur Rahaman[1], Sudipto Chaki[1], Md Shovon Biswas[1], Milon Biswas[1], Shamim Ahmed[1], Md. Julkar Nayeen Mahi[2], and Nuruzzaman Faruqui[2]
{naimurrahaman248@gmail.com[1], sudiptochakibd@gmail.com[1], shovonbiswas.cse@gmail.com[1], milon@ieee.org[1], shamim.6feb@gmail.com[1], mahi.1992@gmail.com[2], faruquizaman27@gmail.com[2]}

Bangladesh University of Business and Technology, Dhaka, Bangladesh[1],
Daffodil International University, Dhaka, Bangladesh[2]

Abstract. Suicide is a serious problem in today's world. It would be useful to detect it sooner to lower the death rate. This research proposes a machine learning-based prediction model in this respect. This research uses four traditional machine learning techniques to look at suicidality in various groups of people. To achieve higher levels of accuracy, we have included feature engineering in our suggested model. To accomplish so, a Select-K-Best feature scoring system has been developed. The developed procedure assists in selecting the top 12 scoring characteristics while improving the performance accuracy of the suggested model. In this experimental setup, Random Forest, XGBoost, SVM, and LightGBM prediction algorithms has been used for accuracy prediction. Finally, accuracy levels are measured by fitting our model to the test dataset and found that the XGBoost technique has an accuracy of 89%, which is greater than other implemented ML models.

Keywords: Suicide Prediction, Machine Learning, SVM, Feature Engineering.

1 Introduction

Suicide is a serious issue in today's society. It is one of the top causes of mortality in the world, and it's getting worse day by day. According to a WHO estimate [1], around 700,000 suicides are committed every year. Suicide is the fourth greatest cause of mortality among those aged 15 to 19. Every country is seeing an increase in the number of suicides. According to statistics, over 77% of suicides occur in low- and middle-income countries. Our neighboring countries like India have a suicide rate of 16.5 and Pakistan has a suicide rate of 3.1, per 100,000 people. Suicide has been a serious problem in Bangladesh. As per the statistics, Bangladesh has a suicide mortality rate of 6.1 per 100,000 people, and it's increasing every year. In most cases, depression plays the main role behind suicide attempts. The available traditional methods are not much efficient to prevent this. Based on the past 49 years of suicide study, it is apparent that the discipline has not improved its

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318160

prediction capacity. To explore new approaches that might help prevent young suicides, this paper outlines a novel approach to early detection and intervention.

Machine learning (ML) paves the path for guiding risk prediction, explain about the prediction[2] and improving suicide prevention systems. Machine Learning outperforms traditional approaches in terms of suicide prediction because it is more successful when utilizing many variables in basic regression models, which are insufficient to represent multi-factorial complicated behavior like suicide. Traditional statistical approaches are not appropriate for such studies, thus machine learning (ML) techniques are particularly well suited. Machine Learning methods analyze complicated connections between wide ranges of parameters to enhance predictions. We are working on four distinct algorithms for comparison and prediction in our study. In our research, we used the Random Forest method, the XGBoost algorithm, the Support Vector Machine (SVM) algorithm, and the LightGBM algorithms. The key contributions of this paper are featured in the followings:

1. We have enhanced the accuracy level using Select-K-Best feature engineering algorithm.
2. We have proposed machine learning-based approach in terms of suicide attempt prediction instead of traditional methods.

The rest of the part of this paper is organized in the following manner. Section **2** describes the background or previous works that are related to our proposed system. Section **3** provides an overall description of our proposed methodology. Section **4** explains the experimental setup and implementation of our proposed system. Finally, we discuss the conclusion and future works in section **5**.

2 Related Research

Some researchers have contributed to this field intending to analyze suicidal cases while considering different aspects. Some of those notable and recent works are discussed in this section while mentioning their achievements as well as the current limitations of their proposed methods. Birjali et al. [3] suggested a semantic analysis method for tweets in both the training and validation data sets. Their validation data set was WordNet-based, while their training data set was Twitter4J-based. They were able to obtain 82.9 % accuracy for suspected tweets with a risk of suicide and 67.8 % accuracy for suspected tweets without a risk of suicide using the Naive Bayes algorithm. Linthicum et al. [4] provided a comprehensive review of alternative machine learning algorithms for predicting suicide. In their study, they explored several aspects of machine learning techniques in the identification of suicidal ideation. Yang et al. [5] conducted a detailed review of six different machine learning approaches and published their findings. Logistic regression, decision trees, random forests, gradient boosting regression trees, support vector machines, and multilayer perceptrons are among the approaches used in their study. Bernecker et al. [6] conducted a thorough assessment of different machine learning methods for suicide prediction. They investigated many elements of machine learning approaches in the detection of suicidal thoughts in their study. Hill et al. [7] utilized classification tree analysis to prospectively identify suicide attempters within a large adolescent population

sample to demonstrate the merits and limitations of this technique for risk identification. They developed a method that maximized risk prediction, with 90.6 % sensitivity and 70.9 % specificity, respectively. Pan et al. [8] presented a survey that introduces and discusses clinical methods based on interactions between social workers or experts and the targeted individuals, as well as machine learning techniques based on feature engineering or deep learning for automatic detection based on online social content categories. They investigated the many domain-specific applications of suicidal ideation detection. Parghi et al [9] created a suicide crisis assessment using machine learning analysis. The data originated from high-risk psychiatric inpatients who were split into two groups based on their short-term suicidal behavior, that is, those who attempted suicide and those who did not between the intake and one-month follow-up dates. We have summarized all the notable existing methods with their limitations in table **1**.

Table 1: Research findings of the existing works

Authors Information	**Research Findings**
Birjali et al. [3]	Semantic analysis method
Linthicum et al. [4]	Comprehensive Machine learning model
Yang et al. [5]	Tested a dataset with different ages of participants
Bernecker et al. [6]	Performed element wise investigation
Hill et al. [7]	Maximized the risk prediction in terms of sensitivity
Pan et al. [8]	Clinical method based interactions
Parghi et al. [9]	Performed investigation upon high-risk psychiatric patients
Tran et al. [10]	Proposed integrated computational technique
Torous et al. [12]	Population based longitudinal data survey

Tran et al. [10] presented an integrated computational technique for predicting suicide risk. Their proposed technique includes an integrated feature selection process, a set of risk classifiers, and a risk calibration mechanism. They had converted the patient's clinical history into a temporal image and used a bank of one-side filters to extract temporal features. Shen et al. [11] created a machine-learning method to predict the likelihood of suicide attempts. They gathered the information for their dataset from various online sites, the majority of which belonged to medical students. Using the Random Forest approach, they achieved 90.1 % accuracy in their trial. Torous et al. [12] used population-based longitudinal data to provide machine learning-based approaches for predicting future suicide behavior. Their information came from the Scottish Well-being Study. The random forest method outperformed the results of other algorithms that they tested. They may be able to enhance accuracy significantly due to less detailed data.

In our paper, we have validated our outcomes through a machine learning model. The key advantage of the machine learning model over classical solution methods is that we can predict suicidal cases earlier if we can train it with the proper dataset. Moreover, a fast and time savings

framework helps to reduce the mortality rate if necessary actions can be taken based on the predicted reports of an ML model.

3 Research Methodology of Our Framework

We intend to investigate the case of suicidality using the machine learning approach in this paper. To do so, we have collected our dataset, selected a list of feature columns based on the Select-K-Best algorithm, and then trained our ML model to predict suicidality from our test dataset. The whole procedure of our proposed machine learning model is outlined in Fig. **1**.

3.1 Dataset Collection

The dataset we utilized includes demographic information such as gender, sexuality, age, race, employability, weight, and so on. This dataset [13] is collected from Kaggle (an open-source dataset repository) website. Participants in this study were given questions on their age, number of friends, income, and other characteristics. This dataset contains information about 500 people and their mental health. This dataset contains a total of 19 different features and we have used 12 features out of them using feature engineering technique.

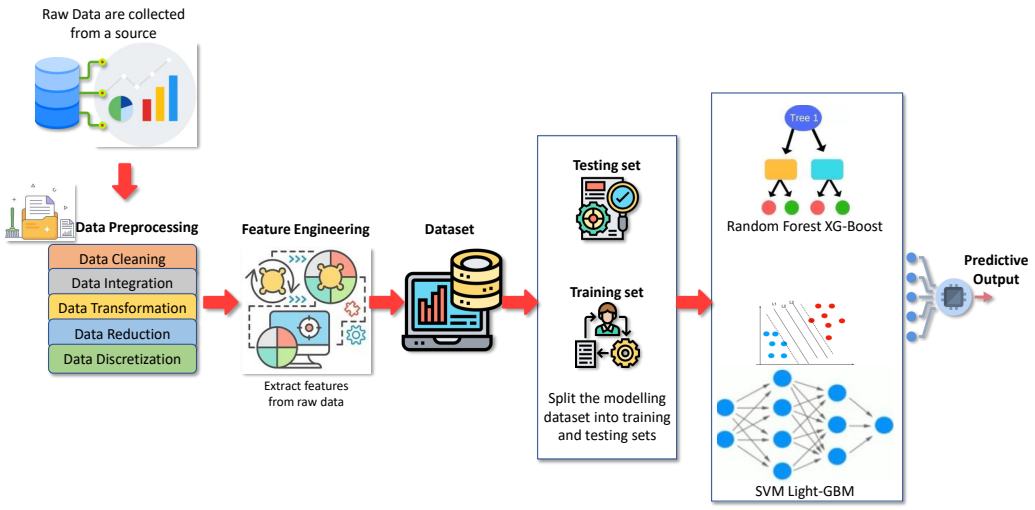

Fig. 1. The developed machine learning based prediction model

3.2 Feature Engineering

The process of converting raw data into features that may be used in machine learning approaches like prediction models is known as feature engineering. Machine learning divides feature engineering into four steps: feature creation, transformations, feature extraction, and feature selection. Feature engineering is the process of creating, modifying, extracting, and selecting features, also known as variables, that are most conducive to the development of an accurate machine learning algorithm. These procedures involve the following processes:

- **Feature Creation:** To create features, one must first determine which variables will be most useful in the prediction model. This is a subjective process that needs human input and creativity. To produce new derived characteristics with better prediction power, existing attributes are merged utilizing addition, subtraction, multiplication, and ratio.
- **Transformations:** Model performance is enhanced by converting predictor variables to guarantee that the model can absorb a wide variety of input, that variables are on the same scale, that the model is simpler to comprehend, that accuracy is raised, and that computational mistakes are avoided.
- **Feature Extraction:** Feature extraction is a technique for producing new variables from raw data automatically. The objective of this step is to reduce the amount of data to a more manageable set for modeling.
- **Feature Selection:** Many characteristics are assessed, appraised, and ranked by feature selection algorithms to decide which are inconsequential and should be eliminated, which are redundant and should be removed, and which are most valuable to the model and should be prioritized.

3.2.1 Steps in Feature Engineering

The steps of feature engineering for most machine learning algorithms are as follows:

- **Data Preparation:** At this stage, raw data from many sources is transformed and aggregated into a common format that can be utilized in a model.
- **Analytical Experiment:** This stage identifies and summarizes the essential aspects of data collection through data analysis and inquiry.
- **Set a benchmark:** Benchmarking is the process of establishing a standard of accuracy against which all other variables are measured. This is done to reduce the rate of error and improve a model's predictability.

3.2.2 Feature Selection Using Select-K-Best

Select-K-Best is a feature selection approach for improving prediction accuracy or performance on high-dimensional data sets. In unilabiate feature selection, the best features are determined using

unilabiate statistical tests. It may be thought of as a stage in the estimator's prepossessing process. Feature selection techniques are exposed by Scikit-learn as objects that implement the transform method: Select-K-Best eliminates everything but the highest-scoring characteristics. We have demonstrated our feature engineering technique in algorithm **1**. Using the scoring methods based on the Select-K-best algorithm, we have measured scores for each feature. Then, we have used 12 features out of 19 based on the score values presented in table **2**.

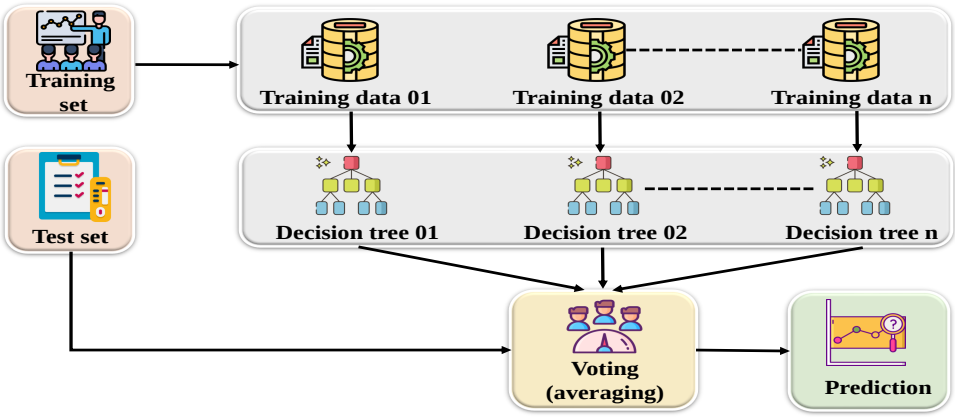

Fig. 2. Working principle model of random forest algorithm

3.3 Random Forest Classifier

Random forest [5] is a flexible, user-friendly machine learning approach that produces great results most of the time even when no hyper-parameters are changed. Random Forest is a supervised learning approach that is used in classification and regression analysis. Given a data sample, Random Forest produces multiple decision trees, compares the outcomes of those trees to the forecast of that result, and then selects the best result. Features variable in the dataset should contain some true values so that the classifier can predict correct outcomes rather than guesses. This approach has the important benefit of being applicable to both classification and regression scenarios. It can avoid over-fitting by averaging or combining the results of different decision trees. It is also applicable to a wide range of data components. Because the random forest uses a large number of trees to forecast the class of the dataset, some decision trees may anticipate the correct output while others may not. However, when all of the trees are merged, they correctly anticipate the outcome. The working principle model of the random forest classifier is shown in Fig **2**.

Algorithm 1: Select-K-Best Feature Engineering

```
Call the class SelectKBest(_BaseFilter)
Load the dataset with n features as X, y =
 load_digits(return_X_y=True)
Iterate over all the features from k :  int or "all";,
 default=12 (select best 12 features)
Call the transformation function as SelectKBest(chi2,
 k=22).fit_transform(X, y)
Set params as _check_params(self, X, y):
if  not (self.k == "all" or 0 <= self.k <= X.shape[1]) then
   Set error value as k should be >=0, <= n_features = %d;
end if
else
   Set error value as k='all' to return all features.
end if
Check score fitting function if self.k == "all" then
   Return np.ones(self.scores_.shape, data_type=bool)
end if
else if self.k == 0 then
   Return np.zeros(self.scores_.shape, data_type=bool)
end if
else
   Set scores as np.zeros(scores.shape, data_type=bool)
end if
Apply merge_sort operation to sort the scores
Return [np.argsort(scores, kind="mergesort")[-self.k :]]  = 1
```

3.4 XGBoost Classifier

XGBoost is an abbreviation for "Extreme Gradient Boosting." XGBoost [6] is a distributed gradient boosting system that has been optimized for efficiency, versatility, and portability. XGBoost is a well-known and widely used machine learning method. The XGBoost ensemble method is based on decision trees. It is utilized in gradient boosted trees, which are intended to be fast and efficient. The library is laser-focused on model performance and computing speed. It builds Machine Learning algorithms using the Gradient Boosting framework. Gradient Boosting, Stochastic Gradient Boosting, and Regularized Gradient Boosting are the three major types of gradient boosting support. It's a lightning-fast algorithm. We have presented the working principle model of the XGBoost classifier in Fig. **3**.

Table 2: The best twelve feature scores after applying the Select-K-Best algorithm

Feature Column	Feature Name	Select-K-Best Score
C-10	Depressed	22.955
C-1	Sexuality	18.948
C-0	Gender	14.151
C-15	Improve-yourself-how	11.985
C-11	What-help-from-others	11.138
C-9	Social-fear	6.669
C-12	Employment	2.297
C-8	Friends	1.154
C-13	Job-title	1.059
C-3	Race	1.008
C-5	Virgin	0.597
C-14	Edu-level	0.532

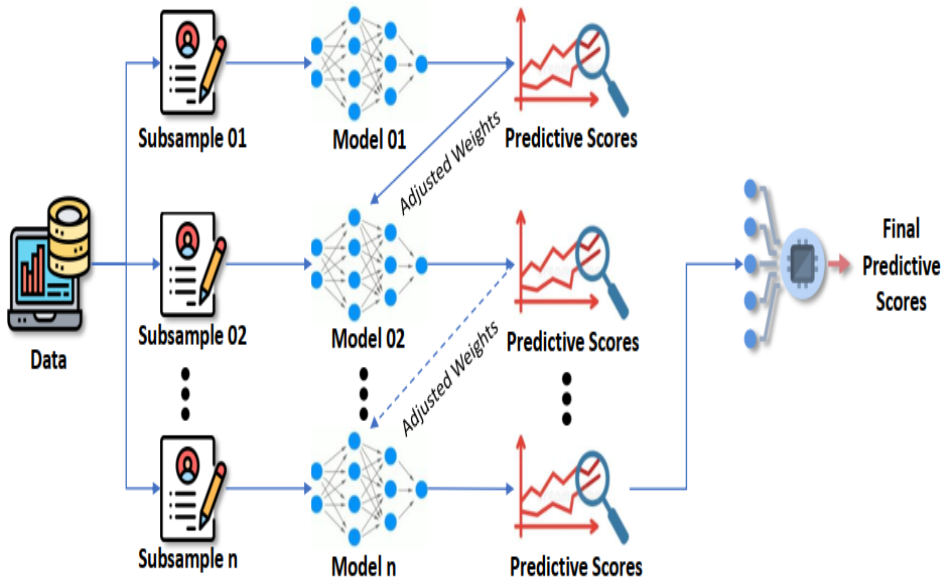

Fig. 3. Working principle model of the XGBoost algorithm

3.5 Support Vector Machine (SVM)

The ”Support Vector Machine” (SVM) [5] is a supervised machine learning method that can be used to solve classification and regression problems. It is, however, primarily utilized in catego-

rization issues. Each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a certain coordinate in the SVM algorithm. Then, we classify by locating the hyperplane that best distinguishes the two classes. SVM will construct the hyperplane repeatedly to reduce error. Individual observation coordinates are used to generate support vectors. The support vector machine is preferred because it achieves great accuracy while using little computing power. SVM (Support Vector Machine) is a type of machine that may be used for both regression and classification. SVM classifiers are very accurate and function well in high-dimensional spaces. Its classifier employs a subset of training points, using relatively minimal memory as a consequence. Because of the long training period, it is not appropriate for huge datasets. SVM is a great tool, but its computation and storage need skyrocket as the number of training vectors increases.

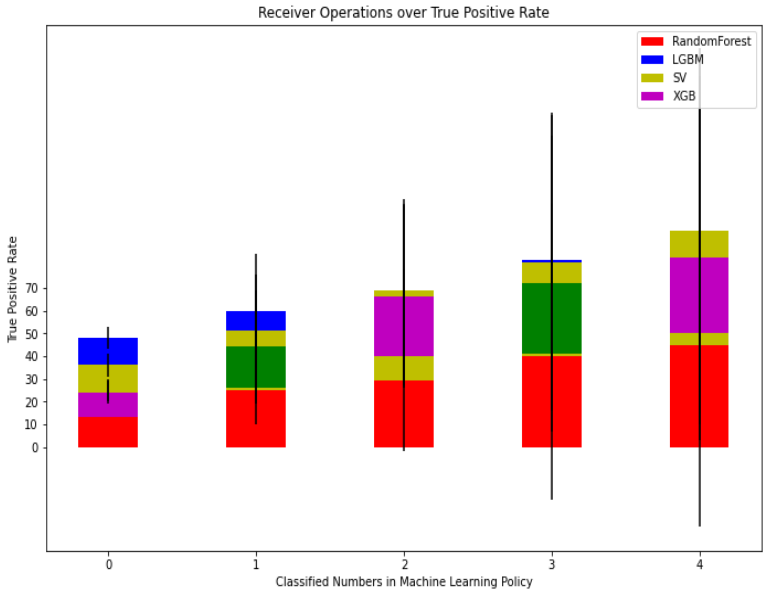

Fig. 4. An increased number of classification result orientation based on Receiver Operations error metrics

3.6 LightGBM Classifier

LightGBM is a fast, distributed, and high-performing decision tree-based gradient boosting system [6]. It's used for several machine learning tasks, including ranking and classification, among others. A decision tree underpins the algorithm. Along the leaf axis, this tree splits. The leaf-wise method can reduce loss more effectively than the level-wise method while simultaneously increasing the accuracy of other boosting techniques. Because of its processing power and capacity to give results rapidly, LightGBM is employed. It uses less memory and is capable of handling large amounts of data. It's a processing algorithm that's lightning fast. It contains almost 100 parameters.

LightGBM has several advantages, including faster training speed and efficiency, reduced memory usage, higher accuracy than any other boosting technique, compatibility with large datasets, and parallel learning support.

4 Implementation and Experimental Results

The dataset [13] comprises of a total of 470 observations (instances) and 19 feature attributes. We have trained our machine learning models with 80% of the total dataset and for validation purposes, 20% of the total dataset were tested. Table **3** displays all the parameter values [14] are kept fixed in our experimental set-up while measuring accuracy level. Fig. **4** shows the increasing classification relationships among the receiving operations based on iterations.

Table 3: Parameter values in our experimental set-up

Maximum Depth	**Subsample**	**Learning Rate**	**Random State**	**Maximum Iteration**
12	0.7	0.01	25	400

We have displayed the generated confusion matrix while testing our machine learning model in Table **4** for the cases of four classifier algorithms. These confusion matrices comprise of four factors and to be more specific, the values of true positive, true negative, false positive, and false-negative rates. The following equations (1), (2), and (3) respectively signify the value of precision, recall, and f1-measure that are able to provide better intuition in terms of prediction results. Here, TP= true positive, FP= false positive, FN= false negative. Table **5** summarizes our experimental set-up values of precision (P), recall (R), and f1-measure with the help of our four different predictive models. From the table, it is evident that the accuracy level of the XGBoost algorithm is comparatively higher than others. We have compared the performance of our four different classifiers in terms of the receiver operating characteristic (ROC) curve shown in Fig. **5**. These comparative ROC curves support our performance measurement parameter values in terms of accuracy level [15], [16].

$$\text{Precision, (P)} = \frac{TP}{TP+FP} \tag{1}$$

$$\text{Recall, (R)} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{F1:measure} = \frac{2PR}{P+R} \tag{3}$$

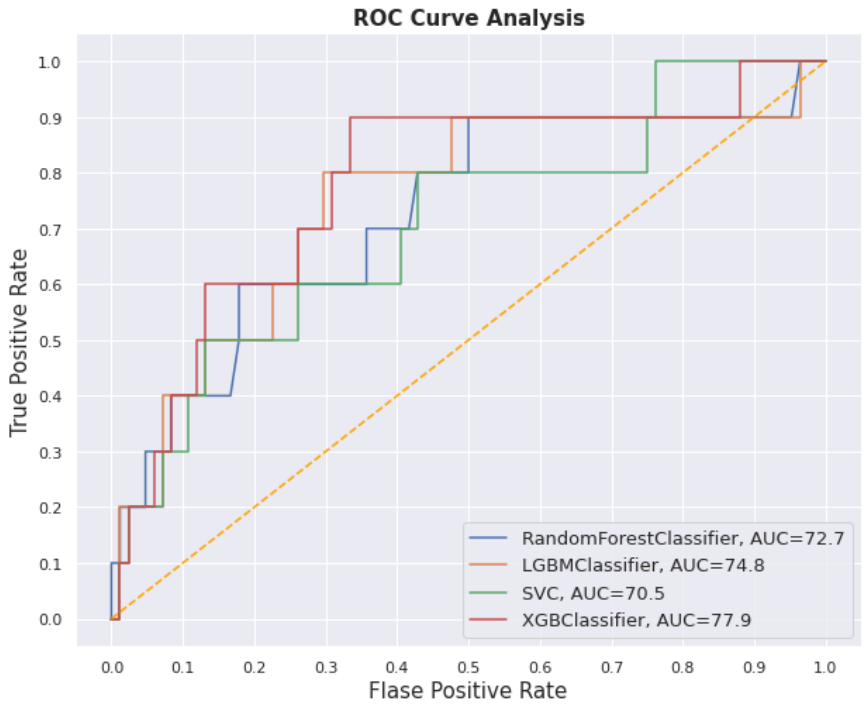

Fig. 5. The comparative receiver operating characteristic curve for four different prediction algorithms

Table 4: Generated confusion matrix for four different ML models XGBoost, Light-GBM, Support Vector Machine (SVM), and Random Forest

Parameters	XGBoost	Light-GBM	SVM	Random Forest
True Pos.	82	75	75	77
False Neg.	2	9	9	7
False Pos.	8	6	6	7
True Neg.	2	4	4	3

5 Conclusion

Suicide is one of the leading causes of death, and its occurrence is on the rise. Machine learning-based early detection of suicide attempts can be immensely valuable to our society because it reduces death rates. It is feasible to extract the text that is judged depressed since people are increasingly expressing their diverse thoughts and views in the form of text on social media. A Red-

Table 5: Performance analysis in terms of Precision, Recall, and F1-score

Parameters	XGBoost	Light-GBM	SVM	Random Forest
Precision	0.91	0.93	0.93	0.92
Recall	0.98	0.92	0.89	0.92
F1-score	0.94	0.92	0.91	0.92
Accuracy	**0.89**	0.86	0.84	0.85

dit survey was used to collect the data, which contained 19 different questions and replies. The top 12 features were retrieved using the Select-K-Best feature engineering approach from these replies, which were employed as features in our experiment. After extracting the features, we analyzed them using four different machine learning techniques. We observed that XGBoost outperformed the other strategies in terms of accuracy, with an accuracy rate of 89%. However, while the findings show a spectacular result in testing, the dataset [13] was very limited, thus there are still a few flaws for real-life prediction. We wish to expand on our existing study by creating our dataset and using a deep learning-based model to get better results.

References

[1] [World Health Organization] https://www.who.int/news-room/fact-sheets/detail/suicide [Accessed on, October 5,2021]

[2] Biswas M, Kaiser MS, Mahmud M, Al Mamun S, Hossain M, Rahman MA. An XAI Based Autism Detection: The Context Behind the Detection. InInternational Conference on Brain Informatics 2021 Sep 17 (pp. 448-459). Springer, Cham.

[3] Birjali M, Beni-Hssane A, Erritali M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. Procedia Computer Science. 2017 Jan 1;113:65-72.

[4] Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behavioral sciences the law. 2019 May;37(3):214-22.

[5] Lin GM, Nagamine M, Yang SN, Tai YM, Lin C, Sato H. Machine learning based suicide ideation prediction for military personnel. IEEE journal of biomedical and health informatics. 2020 Apr 20;24(7):1907-16.

[6] Bernecker SL, Zuromski KL, Gutierrez PM, Joiner TE, King AJ, Liu H, Nock MK, Sampson NA, Zaslavsky AM, Stein MB, Ursano RJ. Predicting suicide attempts among soldiers who deny suicidal ideation in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Behaviour research and therapy. 2019 Sep 1;120:103350.

[7] Hill RM, Oosterhoff B, Do C. Using machine learning to identify suicide risk: a classification tree approach to prospectively identify adolescent suicide attempters. Archives of suicide research. 2020 Apr 2;24(2):218-35.

[8] Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems. 2020 Sep 17;8(1):214-26.

[9] Parghi N, Chennapragada L, Barzilay S, Newkirk S, Ahmedani B, Lok B, Galynker I. Assessing the predictive ability of the Suicide Crisis Inventory for near-term suicidal behavior using machine learning approaches. International journal of methods in psychiatric research. 2021 Mar;30(1):e1863.

[10] Tran T, Phung D, Luo W, Harvey R, Berk M, Venkatesh S. An integrated framework for suicide risk prediction. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013 Aug 11 (pp. 1410-1418).

[11] Shen Y, Zhang W, Chan BS, Zhang Y, Meng F, Kennon EA, Wu HE, Luo X, Zhang X. Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. Journal of affective disorders. 2020 Aug 1;273:18-23.

[12] Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, Firth J. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. Current psychiatry reports. 2018 Jul;20(7):1-6.

[13] [Suicide Attempting Dataset: a Survey performed in Reddit Website] https://www.kaggle.com/kingburrito666/the-demographic-rforeveralone-dataset [Accessed on, October,1,2021]

[14] Biswas M, Tania MH, Kaiser MS, Kabir R, Mahmud M, Kemal AA. ACCU3RATE: A mobile health application rating scale based on user reviews. PloS one. 2021 Dec 16;16(12):e0258050.

[15] Mahi M, Nayeen J, Hossain KM, Biswas M, Whaiduzzaman M. Sentrac: A novel real time sentiment analysis approach through twitter cloud environment. InAdvances in Electrical and Computer Technologies 2020 (pp. 21-32). Springer, Singapore.

[16] Ray B, Saha KK, Biswas M, Rahman MM. User perspective on usages and privacy of ehealth systems in bangladesh: A dhaka based survey. In2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) 2020 Dec 16 (pp. 1-5). IEEE.