COVID-19 and Suicide Tendency: Prediction and Risk Factor Analysis Using Machine Learning and Explainable AI

Khalid Been Badruzzaman Biplob[1], Musabbir Hasan Sammak[1], Abu Kowshir Bitto[1,*] and Imran Mahmud[1]

[1]Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh

Abstract

INTRODUCTION: Pandemics and epidemics have frequently led to a significant increase in the suicide rate in affected regions. However, these unnecessary deaths can be prevented by identifying the risk factors and intervening earlier with those at risk. Numerous empirical studies have exhaustively documented multiple suicide risk factors. In addition, many evidence-based approaches have employed machine learning models to diagnose vulnerable groups, a task that would otherwise be challenging if only human cognition were employed. To date, to the best of our knowledge, no research has been conducted on COVID-19-related suicide prediction.
OBJECTIVES: This research aims to develop a machine-learning model capable of identifying individuals who are contemplating suicide due to COVID-19-related complexities and assessing the potential risk factors.
METHODS: We trained a gradient-boosting model based on tree-based learners on 10067 data consisting of 76 features, which were primarily responses to socio-demographic, behavioural, and psychological questions about COVID-19 and suicidal behaviours.
RESULTS: The final model predicted individuals at risk with an auROC score of 0.77 and a 95% confidence interval of 0.77 to 0.88. The optimal cutoff produced a sensitivity of 31.37 percent and a specificity of 82.35 percent in predicting suicidal tendencies. However, the auPRC was only 0.26, with a 95 percent confidence interval of 0.13 to 0.38, as the class distribution was extremely unbalanced. Consequently, the scores for precision and recall were 0.35 and 0.31, respectively.
CONCLUSION: We investigated the risk factors, the majority of which were associated with sleeping difficulties, fear of COVID-19, social interactions, and other socio-demographic factors. The identified risk factors can be considered when formulating a policy to prevent COVID-19-related suicides, which can impose a long-term economic and health burden on society.

1. Introduction

Pandemics and epidemics have frequently led to a significant increase in the suicide rate in affected regions. However, these unnecessary deaths can be prevented by identifying the risk factors and intervening earlier with those at risk. Numerous empirical studies have exhaustively documented multiple suicide risk factors. In addition, many evidence-based approaches have employed machine learning models to diagnose vulnerable groups [34], a task that would otherwise be challenging if only human cognition were employed. To date, to the best of our knowledge, no research has been conducted on COVID-19-related suicide prediction.

*Corresponding author. Email: abu.kowshir777@gmail.com

Prospective Increased suicide rates in the time of pandemics and epidemics have been observed throughout the history. For example, historical data shows that deaths by suicide increased during the 1918-19 great Influenza epidemic [1]. Among the more recent events, there is evidence that suicide rate increased among the older people in Hong Kong during the severe acute respiratory syndrome (SARS) epidemic [2]. Similarly, the latest Coronavirus Disease 2019 (COVID-19) pandemic has been sweeping across the earth since 2019, and a plethora of mental health effects have already been reported of which depression, anxiety, self-harm, and, most importantly, suicides are on the rise [3]. However, this increase in pandemic induced suicides need not be inescapable if the possible risk factors are identified and the at-risk individuals are intervened earlier [3].

The existing body of empirical research underscores a strong association between social isolation, physical distancing, and loneliness with increased risks of suicide attempts, depression, and anxiety [4]–[7]. Suicide rate is growing day by day [35]. This nexus is further compounded by socioeconomic factors, with unemployment and financial insecurity emerging as significant contributors to the prevalence of suicide attempts [8]–[10]. Beyond these, a myriad of risk factors has been identified, ranging from access to lethal means, such as firearms and pesticides, to the influence of irresponsible media coverage, alcohol consumption, and the intricate dynamics of social relationships [11]–[13].

Despite advancements in evidence-based approaches aimed at enhancing the precision of clinical decision-making, clinicians grapple with the complex task of assessing an individual's suicide risk. The process demands a synthesis of extensive empirical knowledge, clinical intuitions, and personal experience [14]. Complicating matters further are the diverse individual factors at play, including age, religiosity, and relationship status, which often contribute to suboptimal and less accurate decision-making [15]–[17].

While evidence-based approaches have indeed improved the accuracy of clinical decisions, their reliance on regression models has limitations. These models, initially foundational, struggle to unearth the intricate and dynamic relationships among multiple interrelated risk factors [18]–[20]. Enter machine learning-based models, offering a promising avenue to navigate the complexity of these relationships by leveraging a multitude of interrelated variables [14]. Recent studies have showcased the success of various machine learning techniques, such as decision trees, classification tree analysis, random forests, extreme gradient boosting, and artificial neural networks, in identifying at-risk individuals in advance [14], [21]–[24].

Yet, in the context of the unprecedented challenges posed by the COVID-19 pandemic, no study has thus far probed into the potential risk factors associated with COVID-19-induced suicides. Furthermore, the untapped potential of machine learning models in identifying at-risk individuals in the specific context of the pandemic remains unexplored. The aim of this study is twofold. First, we propose a machine learning model that predicts suicide tendency, meaning whether an individual is thinking of committing suicide or not, using information about his or her socio-demographics, attitudes towards lockdowns, knowledge about COVID-19, behaviors, fear of COVID-19, health status, and sleeping difficulties. Next, we identify the most important features or risk factors of the COVID-19 induced suicide tendency from the model. Therefore, our model can be deployed to identify at-risk individuals in advance. In addition, it can also help in making national policies to mitigate the risk factors and prevent further casualties among vulnerable groups.

2. Methodology

2.1. Dataset

The dataset contains 10067 responses to survey questions concerning COVID-19 knowledge, preventive behaviours, psychological behaviours, sleeping difficulties, and suicide tendencies along with other social-demographic information [25]. From the questionnaires, 86 answers that correspond to questions about individuals' social-demographics, attitudes towards lockdowns, knowledge about COVID-19, behaviours, fear of COVID-19, and sleeping difficulties were selected for the modelling purpose. The target variable was a binary variable that answered whether the individual was thinking about committing suicide or not. Data were then divided into training and test sets. The training set contained 90% of the total data, whereas the test set contained the remaining 10%. As a result, the training and test set had 9060 and 1007 observations respectively.

2.2. Model

A gradient boosting model built with decision tree learners was trained to predict individuals with suicide tendencies [26]. Currently, Gradient boosting models are the state-of-the-art for modeling tabular data due to its many advantages over other algorithms [27]. We used the Python library named LightGBM to train the model [28]. Early stopping was utilized using the validation set constructed earlier [29], and area under receiving operating characteristic curve (auROC) score was set as the performance measure to account for the class imbalance [30]. Moreover, missing values were inherently handled by the algorithm itself. Finally, we used a 10-fold cross validation to measure the training performance.

For our second objective, SHapley Additive exPlanations or SHAP values were calculated to find out the features that contributed most to the outcome [31]. SHAP values are suitable for explaining feature importance in complex models like decision trees, neural networks, and so on [32]. We used the Python library named Shap to calculate the SHAP values of each feature of our model [31].

2.3. Evaluation

The final model was evaluated on the test set using the performance metric area under the receiver operating characteristic curve (auROC). Similarly, area under precision-recall curve (auPRC) was also computed as auROC sometimes can mask poor performance if the data is imbalanced [30]. Plots of auROC and auPRC was created using different thresholds. Confidence intervals of various metrics were calculated using bootstrapping with 1000 repetitions [33].

3. Result and Discussion

During training, the model achieved 0.77 auROC with 95% CI: 0.77-0.80 on the validation data as shown in Figure 1. On the other hand, it only achieved 0.17 auPRC with 95% CI: 0.12-0.21 as shown in Figure 2
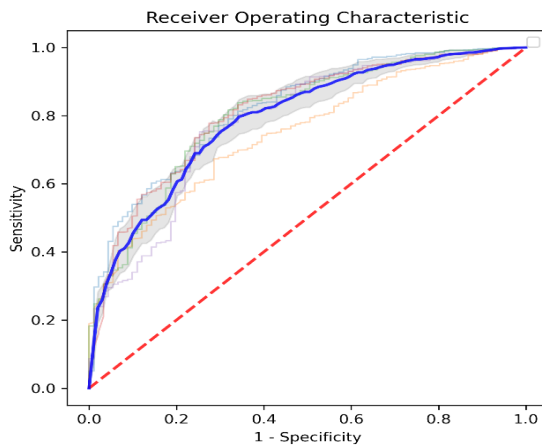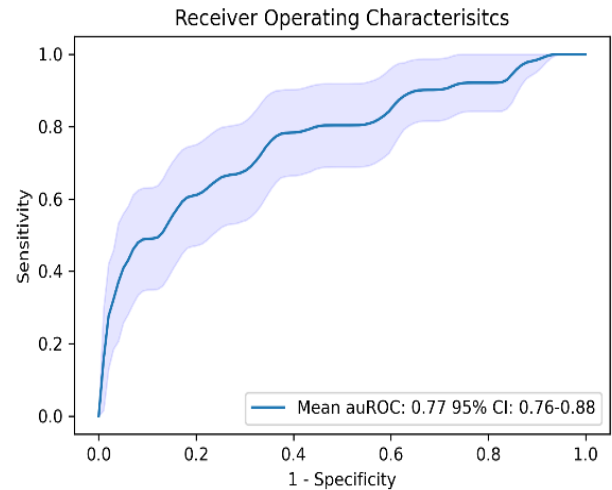

Figure 1. ROC curves of different validation folds. The blue line is the average auROC and the light area around it is the pointwise confidence intervals. The red dotted line is the ROC curve of a model with no skill.
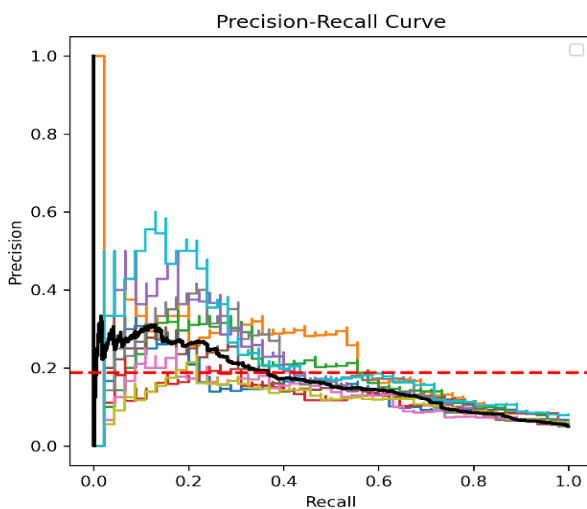

Figure 2. Precision-recall curves of different validation folds. The red dotted line represents average precision

On the test data, it achieved 0.77 auROC with 95% CI: 0.76-0.88 as shown in Figure 3. However, it achieved only 0.26 auPRC with 95% CI: 0.13-0.38 as shown in Figure 4. From the auROC plot of the test data, the best threshold achieved 31.37% sensitivity and 82.35% specificity in predicting suicide tendencies. In contrast, it achieved a precision score of 0.35 and a recall score of 0.31 in classifying suicide prone individuals.

Figure 3. ROC curve on the test data. The blue line is the average auROC and the light area around it is the pointwise confidence intervals.
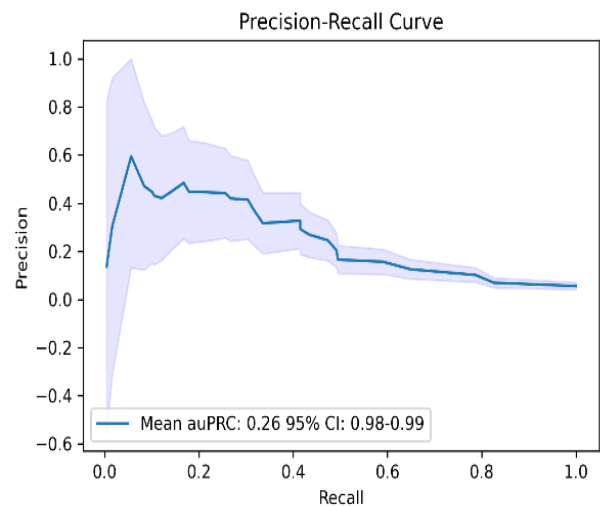
Figure 4. Precision-recall curve of the test data. The blue line is the average precision and the light area around it is the pointwise confidence intervals.

The most important features or risk factors and their contributions to the model outcome is shown in Figure 5. As expected, social and family relationships, interactions, isolation or loneliness, social media usage, gender, age group, fear, and sleeping difficulties had high impact on the model outcome.

Many empirical studies have comprehensively documented multiple risk factors of suicide attempts. However, it was infeasible for the clinicians to identify the at-risk individuals because the information to do so were too complex for human cognition to process. Evidence based approaches then integrated various regression and machine learning models like Decision Tree, Random Forests, Naïve Bayes, Artificial Neural Networks, etc. However, no study, so far, has tried to assess the risk factors of a pandemic induced suicide, neither anyone tried to create a model that can identify the suicide prone individuals. In this study, we tried to construct a machine learning model that can predict suicide prone individuals and assessed the risk factors of the pandemic induced suicide tendencies.

Our first objective, which was to predict suicide prone individuals, comes with many shortcomings. Even though it

scores decent auROC at different thresholds, auPRC reveals the actual problem induced by data imbalance. As the class distribution was highly imbalanced, the precision and recall rates were below average which is evident from the auPRC curve. The skewness in the class distribution is natural as the probability of getting responses of people thinking about suicides is inherently low. Oversampling or under-sampling techniques like SMOTE, random under-sampling, etc. were used to correct class imbalance. However, none of those could improve the model outcome significantly. Other than that, missing values, even if they were imputed inherently by the algorithm, were also a shortcoming of the final model. Therefore, future works should focus on integrating a more balanced and complete dataset for a more accurate model.

Figure 5. Important features or risk factors. The beeswarm plot shows the SHAP values of the most important features of the model. The feature names in the y axis are ranked according to their mean absolute SHAP values. Each point in the plot is an individual. Each position in x axis corresponds to the impact of the feature on the model outcome for a given individual.

For our second objective, we found many risk factors, as expected, which were directly or indirectly associated with loneliness, isolation, physical distancing, etc., as reported by previous studies. For example, there was an impact of social media usage on the suicide tendencies. Those who did not use social media or used them less than several hours a day were more prone to suicide. In addition, the amount of time spent outdoor or amount of time an individual had face to face contact during last 7 days also had an impact on the model. The less they spent time outdoor or had face to face contacts, to more suicide prone they were.

Demographics had some role to play in the model outcome. For example, as shown in Figure 3, being male, married, and aged between 20 and 29 were potential risk factors of suicide attempts. In addition, fear of COVID-19, for example, those who feared constantly about getting COVID-19 were also a predictive marker of possible suicide attempt. Finally, having insomnia or any kind of sleep difficulties were also a red flag against suicide attempt.

4. Conclusion

Suicides are one of the long-term effects of pandemics or epidemics like COVID-19. These events put a considerable amount of economic and mental health burden on the society. However, this unnecessary loss of lives can be prevented if the risk factors are identified, and the at-risk individuals are detected earlier. Empirical studies provide us with plenty of risk factors of suicide attempts, but they are often infeasible to use for a clinician due to our limitation of cognitive abilities. Therefore, evidence-based approaches like regression or machine learning models often come handy in both identifying at-risk individuals and assessing suicide risk factors in a way that neither empirical studies nor clinicians could do. However, no work has been carried out so far to investigate the risk factors of COVID-19 induced suicide tendencies, neither any machine learning based model has been created to predict the at-risk individuals in advance.

We proposed a gradient boosting decision tree model to predict individuals who might be thinking of committing suicides. However, the class distribution of the dataset was highly imbalanced. Therefore, the model performed below average in metrics like precision and recall. Even oversampling and under-sampling techniques could not improve the model performance. On the other hand, we identified several risk factors of COVID-19-induced suicides, of which sleeping difficulties, fear of COVID-19, social-demographics, etc., were highly important.

Future works on this topic should try to focus on compiling a more balanced and complete dataset. Suicides, stemming from events like COVID-19, cast prolonged economic and mental health challenges on society.

Identifying risk factors and intervening early are crucial in preventing these losses. While existing studies offer valuable insights, practical limitations call for innovative approaches. Our proposed machine learning model marks a step forward, yet challenges in class distribution impacted its performance. Future endeavors should explore refined techniques to enhance model accuracy. Additionally, the identified risk factors—sleeping difficulties, COVID-19 fear, social-demographics—lay the foundation for future research, guiding targeted interventions and policy formulations to curb the growing concern of pandemic-induced suicides.

References

[1] Wasserman IM. The impact of epidemic, war, prohibition and media on suicide: United States, 1910–1920. Suicide and Life-Threatening Behavior. 1992 Jun;22(2):240-54.

[2] Cheung YT, Chau PH, Yip PS. A revisit on older adults suicides and Severe Acute Respiratory Syndrome (SARS) epidemic in Hong Kong. International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences. 2008 Dec;23(12):1231-8.

[3] Cheung YT, Chau PH, Yip PS. A revisit on older adults suicides and Severe Acute Respiratory Syndrome (SARS) epidemic in Hong Kong. International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences. 2008 Dec;23(12):1231-8.

[4] Elovainio M, Hakulinen C, Pulkki-Råback L, Virtanen M, Josefsson K, Jokela M, Vahtera J, Kivimäki M. Contribution of risk factors to excess mortality in isolated and lonely individuals: an analysis of data from the UK Biobank cohort study. The Lancet Public Health. 2017 Jun 1;2(6):e260-6.

[5] Matthews T, Danese A, Caspi A, Fisher HL, Goldman-Mellor S, Kepa A, Moffitt TE, Odgers CL, Arseneault L. Lonely young adults in modern Britain: findings from an epidemiological cohort study. Psychological medicine. 2019 Jan;49(2):268-77.

[6] O'Connor RC, Kirtley OJ. The integrated motivational–volitional model of suicidal behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018 Sep 5;373(1754):20170268.

[7] Yao H, Chen JH, Xu YF. Patients with mental health disorders in the COVID-19 epidemic.

[8] Barr B, Taylor-Robinson D, Scott-Samuel A, McKee M, Stuckler D. Suicides associated with the 2008-10 economic recession in England: time trend analysis. Bmj. 2012 Aug 14;345:e5142.

[9] Barr B, Taylor-Robinson D, Scott-Samuel A, McKee M, Stuckler D. Suicides associated with the 2008-10 economic recession in England: time trend analysis. Bmj. 2012 Aug 14;345:e5142.

[10] Barr B, Taylor-Robinson D, Scott-Samuel A, McKee M, Stuckler D. Suicides associated with the 2008-10 economic recession in England: time trend analysis. Bmj. 2012 Aug 14;345:e5142.

[11] Gunnell D, Appleby L, Arensman E, Hawton K, John A, Kapur N, Khan M, O'Connor RC, Pirkis J, Caine ED, Chan LF. Suicide risk and prevention during the COVID-19 pandemic. The Lancet Psychiatry. 2020 Jun 1;7(6):468-71.

[12] Garfin DR, Silver RC, Holman EA. The novel coronavirus (COVID-2019) outbreak: Amplification of public health consequences by media exposure. Health psychology. 2020 May;39(5):355.

[13] Torok M, Han J, Baker S, Werner-Seidler A, Wong I, Larsen ME, Christensen H. Suicide prevention using self-guided digital interventions: a systematic review and meta-analysis

of randomised controlled trials. The Lancet Digital Health. 2020 Jan 1;2(1):e25-36.

[14] 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.

[15] Berman NC, Stark A, Cooperman A, Wilhelm S, Cohen IG. Effect of patient and therapist factors on suicide risk assessment. Death studies. 2015 Aug 9;39(7):433-41.

[16] Joiner Jr TE, Walker RL, Rudd MD, Jobes DA. Scientizing and routinizing the assessment of suicidality in outpatient practice. Professional psychology: Research and practice. 1999 Oct;30(5):447.

[17] Bryan CJ, Rudd MD. Advances in the assessment of suicide risk. Journal of clinical psychology. 2006 Feb;62(2):185-200.

[18] McCarthy JF, Bossarte RM, Katz IR, Thompson C, Kemp J, Hannemann CM, Nielson C, Schoenbaum M. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. American journal of public health. 2015 Sep;105(9):1935-42.

[19] Bolton JM, Spiwak R, Sareen J. Predicting suicide attempts with the SAD PERSONS scale: a longitudinal analysis. The Journal of clinical psychiatry. 2012 Jun 15;73(6):15009.

[20] Su C, Aseltine R, Doshi R, Chen K, Rogers SC, Wang F. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Translational psychiatry. 2020 Nov 26;10(1):413.

[21] De La Garza ÁG, Blanco C, Olfson M, Wall MM. Identification of suicide attempt risk factors in a national US survey using machine learning. JAMA psychiatry. 2021 Apr 1;78(4):398-406.

[22] Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science. 2017 May;5(3):457-69.

[23] Mann JJ, Ellis SP, Waternaux CM, Liu X, Oquendo MA, Malone KM, Brodsky BS, Haas GL, Currier D. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. Journal of Clinical Psychiatry. 2008 Jan 1;69(1):23.

[24] Kessler RC, Stein MB, Petukhova MV, Bliese P, Bossarte RM, Bromet EJ, Fullerton CS, Gilman SE, Ivany C, Lewandowski-Romps L, Millikan Bell A. Predicting suicides after outpatient mental health visits in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). Molecular psychiatry. 2017 Apr;22(4):544-51.

[25] Pakpour AH, Al Mamun F, Hosen I, Griffiths MD, Mamun MA. A population-based nationwide dataset concerning the COVID-19 pandemic and serious psychological consequences in Bangladesh. Data in brief. 2020 Dec 1;33:106621.

[26] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Jun.

[27] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems?. The journal of machine learning research. 2014 Jan 1;15(1):3133-81.

[28] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30.

[29] Raskutti G, Wainwright MJ, Yu B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. The Journal of Machine Learning Research. 2014 Jan 1;15(1):335-66.

[30] Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data-- recommendations for the use of performance metrics. In2013 Humaine association conference on affective computing and intelligent interaction 2013 Sep 2 (pp. 245-251). IEEE.

[31] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.

[32] Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK, Newman SF, Kim J, Lee SI. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature biomedical engineering. 2018 Oct;2(10):749-60.

[33] Johnson RW. An introduction to the bootstrap. Teaching statistics. 2001 Jun 1;23(2):49-54.

[34] Hasan MM, Knight P, Tania MH, Bitto AK, Das A, Punja H. A Novel Framework for Co-designing of An Artificial Intelligence Based Television-enabled Application to Address Social Isolation and Alleviating Loneliness for Older People. In2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) 2022 Dec 2 (pp. 42-47). IEEE.

[35] Biplob KB, Bijoy MH, Bitto AK, Das A, Chowdhury A, Hossain SM. Suicidal Ratio Prediction Among the Continent of World: A Machine Learning Approach. In2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) 2023 Apr 21 (pp. 1-6). IEEE.