A Study of Machine Learning Techniques for Predictive Analysis of Suicidal Tendency Across Different Age Groups

Md. Thoufiq Zumma
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
thoufiq15-10968@diu.edu.bd

Md. Anikur Rahaman
Department of Computer Science
Washington University Science and Technology
NewYork, USA
anikura.student@wust.edu

Mohammad Abdul Muneem
Department of Computer Science
Washington University Science and Technology
NewYork, USA
moabdul.student@wust.edu

Obyed Ullah Khan
Department of Computer Science
Wilmington University
NewYork, USA
okhan001@my.wilmu.edu

Nuzhat Noor Islam Prova
Department of Seidenberg School of CSIS
Pace University
NewYork, USA
np23474n@pace.edu

Abstract— **Suicide is a serious public health concern, and life-saving early identification and prevention are essential. When it comes to forecasting suicide, risk based on a variety of criteria and signs, machine learning algorithms have shown encouraging results in recent years. This study uses machine learning techniques to predict suicide tendencies in various age groups. The first step in the research is to collect pertinent datasets including sociodemographic, clinical, behavioral, and psychiatric data on people who have attempted or succeeded in suicide. Preprocessing is done on these datasets to guarantee data quality, handle missing values, and normalize characteristics. Using evaluation measures such as accuracy, precision, recall, and F1-score, the best performing models are chosen to serve as the best suicide prediction classifiers. The findings show that suicide risk may be accurately, sensitively, and specifically predicted using machine learning techniques. With the help of the found predictive traits, at-risk people may get tailored treatments and support networks, as well as insights into the risk factors linked to suicide conduct. This study uses five different algorithms, The support vector machine (SVM) emerges as the technique that performs the best, offering the greatest accuracy of 0.89. The decision tree classifier comes as a close second, delivering the same accuracy. Following that, random forest obtains an accuracy of 0.84, KNN comes in second with 0.81, and gaussian naive bayes comes in third with 0.53.**

Keywords—Suicide, Generation, Machine learning, SVM, Decision Tree classifier

I. INTRODUCTION

Mental health practitioners and researchers aim to predict and prevent suicide, which affects millions globally. Mental illness, drug addiction, social isolation, trauma, and availability to deadly methods may increase suicide risk. Around 703000 individuals worldwide commit suicide each year. Over one in 100 fatalities in 2019 were suicides. The global suicide rate of men is double that of women. Many factors may increase or decrease suicide risk. Suicide is linked to injury and violence. Victims of child maltreatment, bullying, or sexual assault are more likely to commit suicide. Maintaining strong family and community support networks and simple access to medical treatment may prevent suicide thoughts and actions. Healthcare providers may act early and give appropriate treatment and support by accurately diagnosing suicide risk. Machine learning-based suicide prediction algorithms and models are being developed to detect at-risk people. Suicide prediction may be done by examining massive datasets of suicide risk variables using machine learning. Machine learning models may be trained using electronic health records, social media, and mobile or wearable behavioral data. In these data sets, machine learning algorithms may find patterns and relationships that may predict suicide risk. These models must be created and utilized ethically to preserve suicide risk persons' privacy and well-being. These approaches and processes will evaluate many complex interactions between solutions and components to find the best prediction algorithms. This approach may be more successful in ML research for predicting suicide. Suicide prediction in various age groups aims to identify those at risk of trying or committing

suicide and intervene before a tragedy. Mental illness, drug misuse, and social isolation may cause suicide. Using machine learning algorithms to examine massive datasets may reveal suicidal behaviors and risk factors. By identifying at-risk patients and providing proper treatment, suicide prediction may save lives. Counseling, medication treatment, or hospitalization may be recommended depending on the severity of symptoms. Machine learning algorithms might also tailor suicide prevention efforts to particular individuals or demographics, improving their effectiveness. Thus, it is crucial to analyze how these models are applied and utilize them with other diagnostic and intervention strategies to prevent suicide more comprehensively. We want to study user data for a depression scale and a suicide ideology scale to detect and reduce suicide risk factors before it's too late to save a life. Because suicide is a global public health concern, suicide prediction research is crucial. Suicide is one of the top causes of mortality worldwide, with 800,000 fatalities annually. Individuals, families, and communities may be devastated by suicide. Accurate suicide risk identification is essential for suicide prevention. Suicide risk prediction is difficult due to complicated interplay between individual, social, and environmental variables. Current suicide prediction approaches use self-reported data and physician judgments, which may be unreliable. Clinicians and mental health professionals may identify high-risk suiciders and give focused treatments to prevent suicide attempts by developing suicide prediction models. Machine learning may provide insights, but it should be applied carefully in real-world situations. Suicide prediction is delicate and requires multidimensional evaluations that account individual and environmental aspects beyond machine learning techniques. These models should help experts, not solo decision-makers, and crucial circumstances need human involvement and interpretation. The predictive modeling method should always include ethics, privacy, and mental health knowledge.

II. Literature Review

In [1], the authors conducted a scoping review using the method outlined by Arksey and O'Malley et al., and then utilized the PRISMA protocol to select the relevant papers. This extensive review of the literature aims to categorize the machine learning techniques currently in use for evaluating online profiles for suicidal tendencies. Supported vector machines were utilized in ten out of the sixteen studies (62.5%) that included statistical techniques. Lastly, 75% of the research that was examined employed Python to develop machine learning-based models. The findings in [2] show that there is a high level of accuracy in risk classification and Area Under the Curve (AUC) in predicting suicidal behaviors. We outline important limitations in the use of AI/ML frameworks to guide future research that may impact suicide globally and highlight important discoveries from our study. In [3], we address the problem of early suicide ideation identification on the social media network Reddit by using deep learning and machine learning-based classification techniques. In order to do this, we compare our hybrid LSTM-CNN model against other models in the classification space. Word embedding techniques combined with a neural network architecture yields the most accurate relevance categorization results. In terms of word embedding, its performance is comparable to that of the CNN model (90.6%) and the LSTM neural network (91.7%). Currently, machine learning approaches using feature engineering or deep learning for automated diagnosis based on social media material, as well as clinical treatments focused on interaction between social workers or experts and those at risk, are employed for SID [4]. A number of tangible tasks and sets of data are offered and condensed in order to provide the groundwork for further research. In [5], psychiatry's growing focus in using machine learning methods with health data may be advantageous for suicide risk studies. We will analyze sex differences in suicide risk profiles using data from the Danish population and machine learning methods. This is the first research to develop suicide prediction models using population-level data. The random forest model's AUC across folds has a 95% confidence interval (CI) of 0.88. Within [6] The writers address the problem of suicide thoughts using user-generated postings on websites such as Reddit, Facebook, Twitter, Suicide Watch, etc. They gain insight by evaluating the content, and this can be a red flag for suicidal thoughts. They generate characteristics using natural language processing (NLP) that can be recognized by a variety of classifiers (like random forest, SVM, naive bayes, etc.) and neural network models (like CNN, LSTM, BERT, etc.) designed to detect suicidal thoughts. Within [7] With an emphasis on Twitter data from the last two years, this research offers a range of methods to comprehend suicide thoughts via online user contents in an attempt to enable early diagnosis using sentiment analysis and supervised learning approaches. Based on the baseline classifier findings, the updated ensemble random forest (RF) algorithm has a greater accuracy of 0.99% for suicide prediction with tweets including suicidal ideas than the current system.

III. Proposed Methodology

Different methods can solve each analysis. We started with internet data collecting. Remove null values and columns after research data processing.

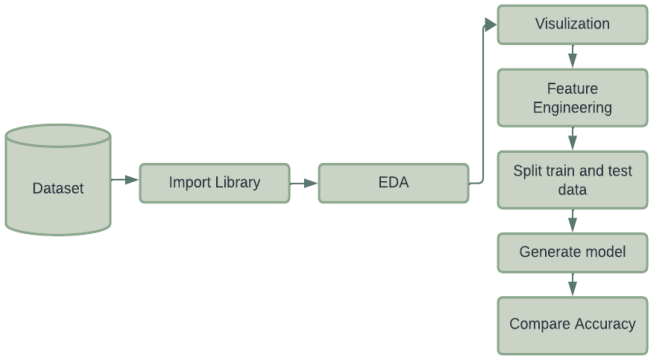

Fig. 1. The proposed methodology of the work

Machine learning algorithm selection. We use five machine learning methods, thus we require data to create the model and fit the algorithm. The model is trained using these data. This selects features. Data formed training and testing sets. Such data is termed test and train data. After fitting several machine learning algorithm models and training using a training data set,We utilized our typical process flow chart for an overview, but we'll discuss numerous techniques utilizing equations and diagrams. This section contains the research methodology and a summary

A. Dataset

We made a few modifications to the data set that Kaggle had provided for us in order to fulfill our criteria. These modifications were rather minor. Furthermore, in order to make use of the dataset, each and every process is necessary. Preprocessing the data is necessary in order to guarantee that it will be included into our model in the suitable manner. It is essential that this be done because, as we have shown in the past, each of the five algorithms that we use acts in a different way. To sum everything up, there are 12 columns and 31756 rows.

B. Data Discription

G.I Generation: The Greatest Generation, or G.I. Generation, was born between 1901 and 1927. The G.I. Generation endured the Great Depression and World War II, which shaped 20th-century history and society. This generation contributed to the battle as troops and citizens. Wartime experiences shaped their duty, tenacity, and patriotism. The G.I. Generation valued family and community. Family values, local bonds, and social and civic engagement were important to them. Many G.I. Generation members valued marriage and motherhood and were acquainted with the traditional family structure.

Silent: The Silent Generation, also known as the Traditionalists or Silent Era, was born between the mid-1928s and the early 1945s. This generation matured amid social upheaval, war, and economic hardship. Civility is important to the Silent Generation. They value institution loyalty, community involvement, and public service. Many Silent Generation leaders in business, government, and other fields helped preserve and advance civilization. Silent Generation people adapted to major technical advances. They saw manual to mechanized employment, telephones, TVs, and computer technology.

Boomers: The "boomer" generation includes baby boomers, born between 1946 and 1964. Birth rates increased after World War II, earning this generation the term "baby boom." Boomers' large numbers and diverse experiences have shaped society, culture, and the economy. Baby boomers make up a large portion of many countries' populations. They have affected consumer, worker, healthcare, and retirement patterns. Popular culture has been shaped by baby boomers. They saw materialism, rock & roll, and television. Boomers have been active in environmental, feminist, and civil rights causes. Boomers have adapted to massive technological advances. Despite not all baby boomers being tech-savvy, the age has seen technology evolve from computers and the internet to smartphones and social media. Technology usage by baby boomers affects communication and information.

Generation X: Generation X (Gen X) is the 1965–1980 birth cohort. This generation precedes millennials and follows baby boomers. Generation X was reared during major social, economic, and technological change. Gen X was the first to use the internet, video games, and PCs. As technology advanced, they switched to digital. People enjoyed how swiftly information technology was being integrated into business and personal lives as adults. Generation X wanted a better work-life balance after seeing how work impacted previous generations. They value adaptation, freedom, and work-life balance. This generation values entrepreneurship, flexibility, and rewarding work.

Millennials: Millennials, often known as Generation Y, were born between the early 1981s and the mid-1996s. This generation matured and experienced major social, economic, and technological developments around the turn of the century. The millennial generation had the first widespread access to computers, the internet, and mobile devices. They use technology well for socializing, information access, and communication. Millennials have struggled financially due to the 2008 financial crisis and student loan debt. They often worry about financial stability and security. Millennials comprehend sustainability and environmental issues better. They choose organizations that operate ethically and sustainably and make choices accordingly.

Generation Z: The generation born between the mid-1997s and the early 2012s is called "Generation Z". This generation, after Generation Y, is the youngest reaching maturity. Growing up amid rapid technological, global, and social change characterizes Generation Z. Gen Z's creativity and entrepreneurship are highly renowned. They embrace new technologies quickly since they were reared in a fast-paced atmosphere. They often tackle problems creatively. Generation Z often wants to start their own businesses or pursue alternative careers. Generation Z has boosted mental health awareness. They prioritize mental health and have sought to reduce stigma. Gen Z openly discusses their stress, worry, and other mental health difficulties to encourage mental health services.

C. Data Visualization

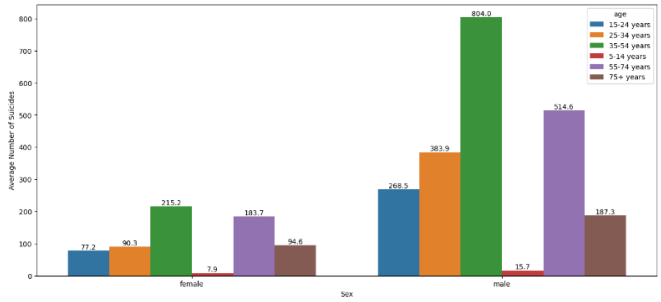

Fig. 2: Average Number of Suicides by age group and gender

There are age groupings in this dataset. Age ranges are given. Teens and young adults have greater suicide rates. Mental health issues, bullying, academic pressure, identity discovery, and

social issues increase vulnerability in this age range. After middle age, suicide risk rises, especially for males. This tendency may be affected by marital issues, money troubles, employment issues, and depression. Many countries show gender-related suicide trends. Because suicide is multifaceted and affected by many factors, it's important to address this topic cautiously. General suicide rate and gender findings. Many countries have higher male suicide rates than female suicide rates.

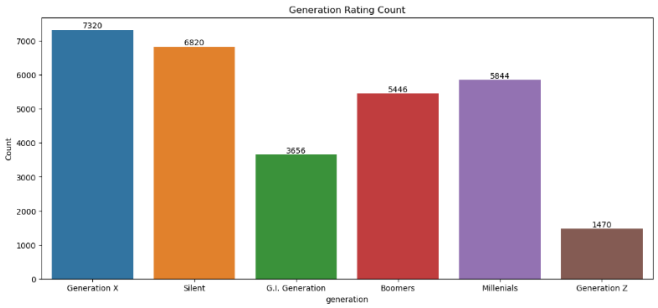

Fig. 3: Bar Chart of generation count

Here is a bar chart of the generation that is seen in figure 3. According to this barchart, how many individuals are there in each generation.

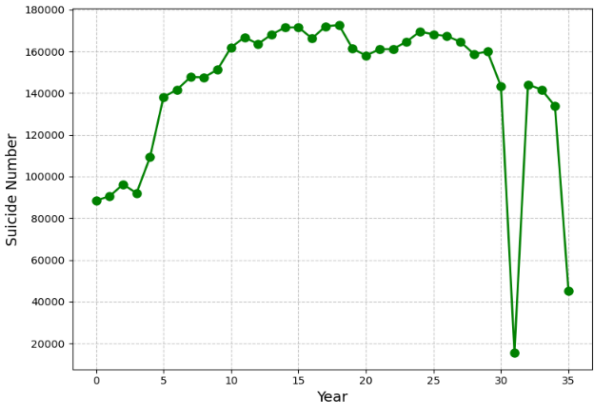

Fig. 4: Total number of suicides from 1985 to 2021

We have included all of the data from 1985 to 2016 in our dataset. In figure 4, a line chart displaying the number of suicides by year is shown. It can now simply comprehend which years had the highest number of suicide attempts.

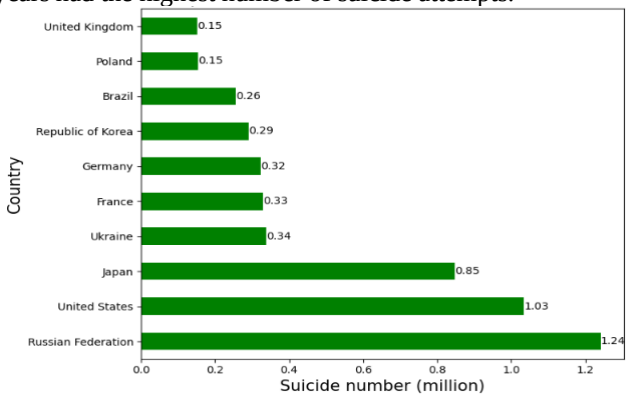

Fig. 5: Top 10 countries for suicide from 1987 to 2021

The top ten nations in which the number of people who commit suicide is highest are shown in Figure 5. There is a higher rate of suicide among persons in the Russian Federation than in other nations between the years 1987 and 2021.

D. Label Encoding

In machine learning and data analysis, label encoding converts categorical variables to numbers. Categories and labels in the variable have various numeric values. Label encoding should be used sparingly, particularly with variables having number names that reflect a logical hierarchy. Label encoding might mislead the model by giving a false sense of ordinality. If a categorical variable lacks order, one-hot encoding or other methods are preferred for machine learning models.

E. Feature Scalling

Scale or normalize dataset properties using machine learning preprocessing techniques like StandardScaler. For each attribute, it sets a mean of 0 and a standard deviation of 1. We use z-score normalization in StandardScaler. Five algorithms are employed.KNN, SVM, decision trees, and Random Forest Classifier.Powerful SVM may be used for classification and regression.A basic yet efficient classification and regression method is KNN. Combining several decision trees, Random Forest ensemble learning makes predictions. Find the best algorithm for your issue and data by comparing numerous solutions.

IV. Results and Discussion

The public health issue of suicide must be identified and prevented early to preserve lives. In recent years, machine learning algorithms have demonstrated encouraging results in predicting suicide risk using a variety of criteria and indications. This study predicts suicide inclinations across age groups using machine learning. Gathering sociodemographic, clinical, behavioral, and psychiatric data on suicide attempters and survivors is the initial step in starting the research. These datasets are preprocessed to assure data quality, handle missing values, and standardize characteristics. The top suicide prediction classifiers are chosen based on accuracy, precision, recall, and F1-grade. Recent studies reveal that machine learning can accurately, sensitively, and precisely predict suicide risk. Predictive characteristics may help at-risk individuals acquire tailored therapy, support networks, and insights into suicide risk factors. Five algorithms are used in this study. The top technique is the support vector machine (SVM), with 0.89 accuracy. Decision tree classifiers are a close second with the same accuracy. Random forest has an accuracy of 0.84, KNN 0.81, and gaussian naive bayes 0.53.

TABLE I. PERFORMANCES OF DIFFERENT CLASSIFIERS

Algorithm	Value name	Precision	Recall	F1-Score	Accuracy
SVM	Boomers	0.89	0.80	0.84	0.89
	G.I Generation	0.98	0.90	0.94	
	Generation X	0.82	0.92	0.87	
	Generation Z	0.95	0.95	0.95	
	Millennials	0.94	0.89	0.92	
	Silent	0.88	0.92	0.90	
Decision Tree	Boomers	0.73	0.92	0.81	0.88
	G.I Generation	1.00	1.00	1.00	
	Generation X	0.83	0.81	0.82	
	Generation Z	1.00	1.00	1.00	
	Millennials	0.92	0.86	0.89	
	Silent	1.00	0.88	0.93	
Random Forest	Boomers	0.77	0.56	0.65	0.83
	G.I Generation	0.99	0.93	0.96	
	Generation X	0.73	0.93	0.82	
	Generation Z	1.00	0.94	0.97	
	Millennials	0.91	0.88	0.89	
	Silent	0.85	0.85	0.85	
KNN	Boomers	0.73	0.74	0.73	0.81
	G.I Generation	0.88	0.87	0.87	
	Generation X	0.78	0.79	0.78	
	Generation Z	0.79	0.91	0.85	
	Millennials	0.83	0.83	0.83	
	Silent	0.86	0.83	0.84	
Gaussian NB	Boomers	0.45	0.29	0.35	0.53
	G.I Generation	0.65	0.74	0.69	
	Generation X	0.50	0.45	0.47	
	Generation Z	0.00	0.00	0.00	
	Millennials	0.44	0.77	0.56	
	Silent	0.69	0.63	0.66	

V. CONCLUSION

The employment of machine learning methods in the prediction of suicidal inclinations across a wide range of age groups shows promising findings in this research. It is possible to make reliable predictions by collecting and preprocessing a wide variety of datasets that include sociodemographic, clinical, behavioral, and psychiatric aspects. Machine learning has been shown to be successful at correctly, sensitively, and particularly identifying persons who are at danger of committing suicide, as shown within these studies. Particularly noteworthy is the fact that the support vector machine (SVM) emerges as the best accurate classifier, with decision tree classifiers, random forest, KNN, and gaussian naive bayes coming in close second and third, respectively. These findings highlight the opportunity for people who have been recognized as being at risk to receive individualized treatments and support networks, as well as insights into the underlying risk factors that are connected with suicide conduct.

REFERENCES

[1] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of suicide ideation in social media forums using Deep Learning," Algorithms, vol. 13, no. 1, p. 7, Dec. 2019. doi:10.3390/a13010007

[2] L. Cao, H. Zhang, and L. Feng, "Building and using personal knowledge graph to improve suicidal ideation detection on social media," IEEE Transactions on Multimedia, vol. 24, pp. 87–102, 2022. doi:10.1109/tmm.2020.3046867

[3] A. A. Choudhury et al., "Predicting depression in Bangladeshi undergraduates using machine learning," 2019 IEEE Region 10 Symposium (TENSYMP), Jun. 2019. doi:10.1109/tensymp46218.2019.8971369

[4] S. T. Rabani, Q. R. Khan, and A. M. Khanday, "Detection of suicidal ideation on Twitter using Machine Learning & Ensemble approaches," Baghdad Science Journal, vol. 17, no. 4, p. 1328, Dec. 2020. doi:10.21123/bsj.2020.17.4.1328

[5] T. Jain et al., "Machine learning techniques for prediction of Mental Health," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Sep. 2021. doi:10.1109/icirca51532.2021.9545061

[6] S. Fodeh et al., "Using machine learning algorithms to detect suicide risk factors on Twitter," 2019 International Conference on Data Mining Workshops (ICDMW), Nov. 2019. doi:10.1109/icdmw.2019.00137

[7] A. Chadha and B. Kaushik, "A survey on prediction of Suicidal Ideation using machine and Ensemble Learning," The Computer Journal, vol. 64, no. 11, pp. 1617–1632, Nov. 2019. doi:10.1093/comjnl/bxz120

[8] S. Jain et al., "A machine learning based depression analysis and suicidal ideation detection system using questionnaires and Twitter," 2019 IEEE Students Conference on Engineering and Systems (SCES), May 2019. doi:10.1109/sces46477.2019.8977211

[9] S. Ji et al., "Suicidal ideation detection: A review of machine learning methods and applications," IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 214–226, Feb. 2021. doi:10.1109/tcss.2020.3021467

[10] G.-M. Lin et al., "Machine learning based suicide ideation prediction for military personnel," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1907–1916, Jul. 2020. doi:10.1109/jbhi.2020.2988393

[11] R. A. Bernert et al., "Artificial Intelligence and Suicide Prevention: A systematic review of Machine Learning Investigations," International Journal of Environmental Research and Public Health, vol. 17, no. 16, p. 5929, Aug. 2020. doi:10.3390/ijerph17165929

[12] Z. Abbass, Z. Ali, M. Ali, B. Akbar, and A. Saleem, "A framework to predict social crime through Twitter tweets by using machine learning," 2020 IEEE 14th International Conference on Semantic Computing (ICSC), Feb. 2020. doi:10.1109/icsc.2020.00073

[13] W. F. Heckler, J. V. de Carvalho, and J. L. Barbosa, “Machine learning for suicidal ideation identification: A systematic literature review,” Computers in Human Behavior, vol. 128, p. 107095, Mar. 2022. doi:10.1016/j.chb.2021.107095

[14] M. Miché et al., “Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning,” Journal of Affective Disorders, vol. 265, pp. 570–578, Mar. 2020. doi:10.1016/j.jad.2019.11.093

[15] R. A. Rahman, K. Omar, S. A. Mohd Noah, M. S. Danuri, and M. A. Al-Garadi, “Application of machine learning methods in mental health detection: A systematic review,” IEEE Access, vol. 8, pp. 183952–183964, 2020. doi:10.1109/access.2020.3029154