

Depression and Suicide Risk Prediction Based on Machine Learning Models

Yunze Li ^{1, 2, *}

¹ Cardiff School of Technologies, Cardiff Metropolitan University

² Western Avenue, Cardiff, CF5 2YB, UK

* Corresponding Author Email: st20202164@outlook.cardiffmet.ac.uk

Abstract. Nowadays depression is one of the most serious diseases in the world. According to incomplete statistics, depression patients in the world are up to tens of millions of people. About 800,000 people committed suicide because of depression, most of them young people aged 15-29. In other words, on average, every 40 seconds someone commits suicide because of depression. In addition, most depressed patients have suicidal tendencies. Research of World Health Organization show that in the 10 years from 2005 to 2015, the total number of people suffering from depression increased by 18.4 percent. Anxiety and stress will be with human over the next 10 years and will become the most common mental illness, according to the World Health Organization. Therefore, pay more attention mental health. Addressing depression can alleviate problems such as healthcare costs that exceed national and corporate budgets. Using machine learning models to predict patients with depression is a key challenge in the field of clinical data analysis and is one of the prevalent techniques for predicting disease. In this paper, a deep network model for anonymous predictive analysis of the data sets provided by "Suicide Watch" and "Depression" on the Reddit platform and finally ensured that the accuracy of each model result reached 95%. Finally, the causative factors of depression were analyzed to help patients and medical staff to prevent depression in time. So that mental health and safety issues are paid attention to reduce the health and safety problems caused by depression.

Keywords: LSTM, machine learning, deep learning, depression, suicide prediction.

1. Introduction

The behavior of depressed patients is like a time bomb erupting without warning, which brings great loss to the family and society. There are many manifestations of depression, the usual manifestations of sadness, fatigue, emotions easily out of control, excessive self-denial, crying, sleep disorders, gain and loss, loss interested in everything, self-doubt, and so on. According to the severity of the disease, depression can be divided into two types. One is patients with heavy depression, which belongs to acute patients. For example, an extreme situation such as the death of a family member suddenly occurs, which makes a normal person receive a huge mental blow in a short period. The other is the widespread depression in the current society. Widespread depression is the generalized depression caused by long-term stress called mild and moderate depressed patient. They are all around people but unnoticed. People with widespread depression may behave crazily under certain circumstances. The most severe manifestations of a heavily depressed patient can lead to suicide [1]. Until now, there has been no accurate clinical characterization of widespread depression, but it is now widely believed that depression can be identified by the daily behavior of depressed people, and even suicidal behaviors. Consider that depression is influenced by more than one non-negligible factor such as health status, education level, social environment, etc., and is widespread in the population and difficult to detect. Therefore, research methods related to depression direction need to use machine learning models for prediction analysis, detection, and prevention. Machine learning models can measure data accurately and extract useful information from large amounts of data. Mining information can accurately measure patients' physical and mental health and predict the probability of depression and the potential risk of depression. In this way, it can help medical staff to analyze the condition of patients more effectively. Section 2 introduces relevant studies on depression, various

factors affecting depression, and various investigations and research conclusions on depression. Section 3 introduces the core content of this research process. The exploration process of data sets (EDA process) and the principle of machine learning model used in the research environment of Anaconda-python. Section 4 introduces the results of the machine learning model used in this study and the information mined from the overall study. Section 5 integrates the general situation of the research, the limitations of the research, and the future work prospects.

2. Research Background

People's understanding of depression is also a process of gradual improvement. Over the last 50 years, there has been a negative bias toward depression, specifically the tendency to process negative self-reasoning and assume the worst. Obviously, depression lacks a proper cognitive process for things. According to research, memory is difficult to escape from negative substances, and these negative substances increase the process of maladaptive emotional mediation and seriously affect adaptive self-emotional regulation. For example, in daily life, many things can cause anxiety, such as the traffic jam that cannot be solved. Generalized depression doesn't happen overnight it's usually living with chronic anxiety. A normal person wouldn't want to kill himself because of a major shock. Generalized depression involves living with years of anxiety. Even a normal person in a very anxious and depressed environment for a long time will become depressed. This external environment includes biological factors, psychological factors, and factors between social variables. Biological stimulation of the external environment will secrete different substances such as hemoglobin, amino acids, hormones, and so on directly affect the occurrence of depression. On the mental side, a good piece of news can make people feel relaxed and happy, which can greatly reduce anxiety and depression. A piece of bad news or a negative reaction to an event can increase anxiety and depression. The influence of social factors and other pressures on inner emotions is immeasurable. The status of the people to contact, the frequency of communication with friends and relatives. These are all processes that can directly affect people's inner state of anxiety. In addition, there are many other factors that can affect depression such as neurogenetic, biological, and environmental factors. There are some studies on depression, such as functional magnetic resonance imaging (fMRI), electroencephalogram (EEG), and genetic or phenomenological data, in which the overall accuracy of ML algorithms for treatment results is about 0.83 [2].

By summarizing the current research on depression, it can be roughly divided into three types according to different data sets. The first is the process of classification, prediction, and recognition of social media text information, obtaining data sources through media information channels and analyzing emotions, to classify and process emotion levels. Through comprehensive analysis, check whether the assessed object is in a state of depression or anxiety or whether there is potential depression [3]. The second, through medical tests, is to examine the brain. The brains of depressed and anxious people compare information differently than the brains of normal people. This difference is used to judge [4]. Third, a comparison is made through objective numerical data, such as human health and life indicators, blood pressure, hormones, heart rate, breathing difficulties, etc., social information indicators, education level, monthly salary, etc. [5]. Although there have been many studies on depression using intelligent models, most of them are based on random sampling analysis of incomplete data sets, and the model's effect on data fitting is not very good. As a result, the overall research results are lack of interpretability, and useful information cannot be mined. Even though a small number of experimental results have excavated some useful information, a comprehensive analysis of the information has not been made, which has no practical research significance. This study took into account the symptoms of depression and divided them into three different categories: mild, moderate, and severe. In addition, it systematically predicted the existence of suicidal behaviors due to depression and proposed relevant solutions for different types of people, thus greatly reducing the number of people who committed suicide due to depression.

3. Methodology

The programming language of this study uses Python, an open-source and free-glue programming language. The Python-compiled code environment is concise and has numerous installation packages for third-party data handling, data visualization, machine learning modeling, and analysis, such as NumPy, pandas, TensorFlow, etc.

3.1. Data Set

The study used data sets from the Reddit platform's "Suicide Watch" data set and "Depression" data set. The data set capacity was 1429*22 numerical category data and 202374 tweets data, respectively. The full data set is publicly available on the Kaggle website. According to the preprocessing of the data set, the number of suicides due to depression in the whole population accounts for 50%. The purpose of this study was to use machine learning models to predict and detect the likelihood of depression and suicidal tendencies. The first findings are that timely treatment for depressed patients can reduce the risk of suicide. Second, it pays attention to the mental health problems of the population and makes a comprehensive analysis of the mental health of the people suffering from depression to ensure the inner health of the population. This could significantly reduce suicide among depressed people and improve overall well-being.

3.2. Data-Processing

Data preprocessing is a key part of machine learning modeling. Data preprocessing is designed to clean data, identifying missing data, noisy data, and erroneous data throughout the process. This can greatly reduce the time consumed by model fitting data after finishing processing the data. In this way, the better-fitting model improves the prediction effect. Considering that the standard data set is used in this study, there are no missing values and noisy data, and only standardized data processing is required. After data preprocessing, it can be concluded that the pathogenic factors of depression can be divided into external environmental factors and psychological factors.

3.3. Model

The machine learning model is an emerging technology that integrates multiple disciplines. Machine learning theory involves mathematics, statistics, data science, computer science, concave-convex theory, and other disciplines. There are four classic models of machine learning, namely regression, classification prediction, collaborative filtering of algorithms, and cluster analysis. According to whether the data type fitted by the model has labels, it can be divided into supervised learning, unsupervised learning, and semi-supervised learning [6]. If the data set has explicit labels for classification, choose the supervised learning model. Unsupervised learning generally provides better performance and results for large data sets that are not explicitly labeled than supervised learning. If the data set is too large, deep learning technology is a very wise choice, because the deep network owners are hungry and naturally adapted to the huge data set [7].

In this study, classification models are used to predict depression and analyze suicide tendencies. The models are as follows: Logistic regression, a machine learning classification model based on linear regression analysis, is used to analyze the influence of one or more independent variables on the results of dependent variables [8]. KNN is a lazy machine learning model in the classification model of machine learning. The model does not need the learning process to complete the calculation of known data samples to predict unknown samples [9]. Decision Tree and Random Forest model based on attribute structure. Based on the attribute structure model, data classification prediction is carried out by displaying a mapping relationship between data attributes and data values [10]. Support vector machine is a binary data analysis model that uses kernel expansion theorem to establish a linear learning machine to solve linear indivisible problems in low-dimensional feature space [11]. DNN, a deep learning model, transmits information through the construction of the model network structure. This network model structure is modeled on the neurons of animal brains to extract and analyze data

features for recognition and prediction. Artificial neural networks can carry out the mathematical model of distributed parallel information processing [12]. The artificial neural network has the ability of self-learning and self-adaptation to achieve the purpose of feature classification, analysis, and prediction by depending on the complexity of the system and by adjusting the inter-connected relations among many internal nodes. Considering that the model used in this study is a kind of LSTM model to solve the long-term dependence of information for the identification and analysis of data sets of tweets type. This model calculates the information provided by the activation function through the loop kernel function and determines whether the information is retained or forgotten in the way of gate structure [12].

4. Result and discussion

In this dissertation. The study aims to draw comprehensive, true and objective conclusions and consider people's anxiety in the context of the impact of COVID-19 [13]. The model research results are shown in Figure 1.

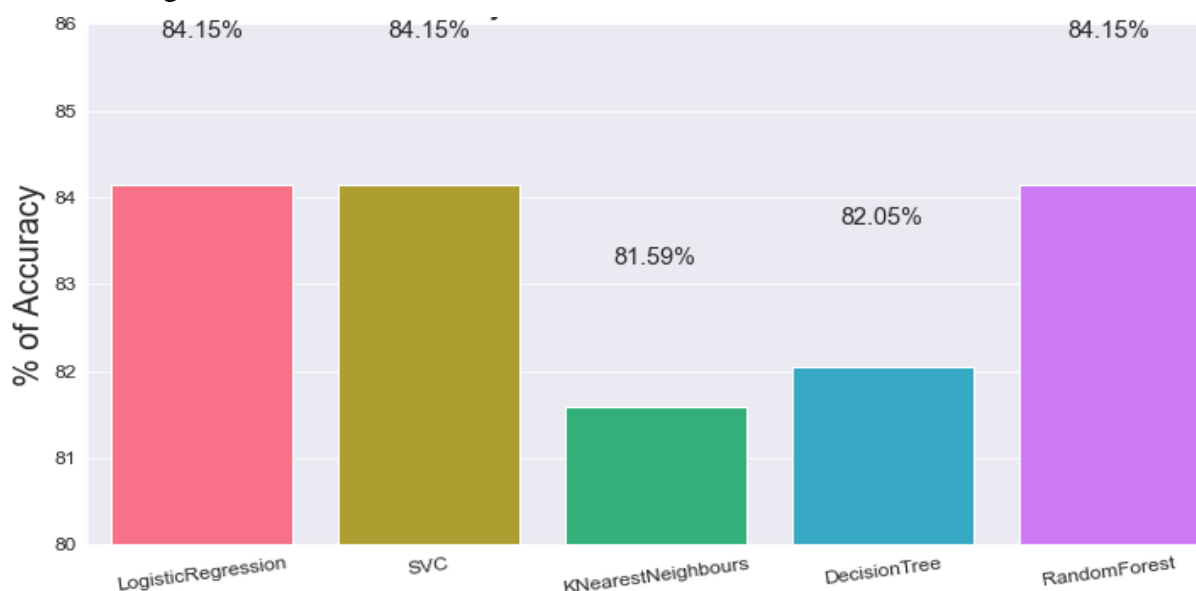


Figure 1. Accuracy of different classifiers.

The scholars point out that according to the symptoms of patients with depression, patients are divided into mild, moderate, and severe depression. The symptoms of mild depression include sadness, avoidance of all activities, diet, sleep, and health fluctuations. People with moderate depression exhibit symptoms such as excessive agitation and mania, excessive energy loss, and frequent fatigue. Symptoms of major depression include a loss of decision-making, a sense of triviality about everything around them, and in certain cases, suicidal tendencies. Healthcare providers need to look at different symptoms to determine how sick a patient is, known as depression. In this study, the accuracy of suicide prediction results for depressed patients reached 95%, as shown in Figure 2.

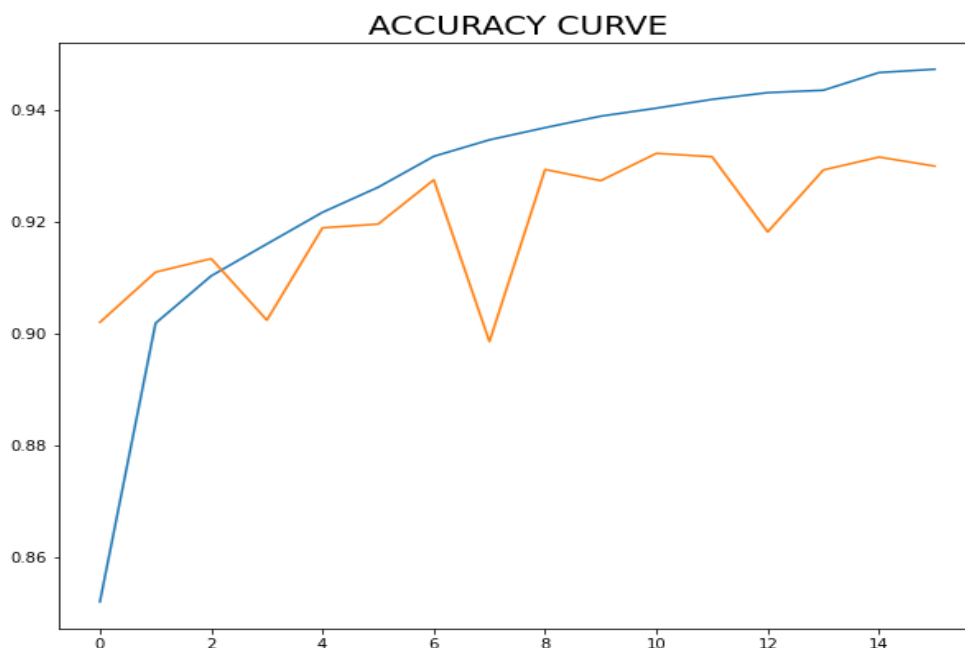


Figure 2. Accuracy curve.

Based on the findings of this study, people with restlessness, indecision, and poor sleep quality are at higher risk for depression. People with mild to moderate depression worry excessively about their social status and their perception and the existence of their worth in themselves. The usual manifestations of depression in this group are a comparison of social status, excessive self-esteem, and excessive negativity. If such patients do not receive timely treatment and after a long period of self-doubt will gradually become severely depressed patients. In addition, studies have shown that women, students, professionals, and people with a history of mental illness are more likely to experience anxiety and depressive symptoms in the general population. People who are restless, indecisive, and sleep poorly are at higher risk of depression. Women need to do housework at home and need to take care of children, there is a lot of outside work pressure, and multiple sources of stress, easy to lead to a high incidence of depression. Students are more likely to have anxiety and depression due to the pressure of college entrance competitions and exam homework. Career personnel due to social pressure, life pressure, and other self-value denial phenomena, which leads to depression. Compared with normal people suffering from a history of psychosis, they are vulnerable groups, due to weak resistance or excessive sensitivity, resulting in depression. These people are likely to have negative emotions, and over time they will engage in extreme behavior.

5. Conclusion

This research involves many fields: biology, psychology, social environment, computer science, data science, etc. Depression studies are subject to multidirectional interactions. Considering that there is often a great deal of variability in the clinical manifestations of people with depression such as post-traumatic stress disorder, schizophrenia, depression, ASD, and bipolar disorder all fall into the realm of mental disorders. In most cases, healthcare workers cannot understand the patient's physical and mental condition conventionally. This study is using a machine learning model to assist healthcare providers in diagnosing mental health conditions. Analyzing data such as social media data, clinical health records, and mobile device sensors to identify problems such as mood disorders has had periodic success and has produced accurate predictions of suicidal behavior in depressed patients. While this study has made some achievements, it also has some limitations. The first is that the data set is a study lasting more than 10 years, and the timeliness of the data set is not very convincing. Second, the correlation between data attributes and the label is low and very sensitive. Depression is affected by the interaction of multiple directions, and many independent variables can affect the dependent

variables. Moreover, the boundary between normal and depressed patients is difficult to distinguish, and each data individual is very sensitive. It is difficult for models to predict with perfect accuracy whether a person is ill or not. It also means that the artificial intelligence model has great potential in the treatment and diagnosis of depression industry, and there is a great space for progress.

References

- [1] S. Aleem, N. ul Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions," *Electronics*, vol. 11, no. 7, pp. 1111, Mar. 2022.
- [2] I. Bhakta and A. Sau, "Prediction of Depression among Senior Citizens using Machine Learning Classifiers," *International Journal of Computer Applications*, vol. 144, no. 7, pp. 11 – 16, Jun. 2016.
- [3] S.-E. Cho, Z. W. Geem, and K.-S. Na, "Predicting Depression in Community Dwellers Using a Machine Learning Algorithm," *Diagnostics*, vol. 11, no. 8, pp. 1429, Aug. 2021.
- [4] S. Gao, V. D. Calhoun, and J. Sui, "Machine learning in major depression: From classification to treatment outcome prediction," *CNS Neuroscience & Therapeutics*, vol. 24, no. 11, pp. 1037 – 1052, Aug. 2018.
- [5] A. Grzenda et al., "Machine Learning Prediction of Treatment Outcome in Late-Life Depression," *Frontiers in Psychiatry*, vol. 12, pp. 738494, Oct. 2021.
- [6] M. T. Hawes, A. K. Szency, D. N. Klein, G. Hajcak, and B. D. Nelson, "Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic," *Psychological Medicine*, vol. 52, no. 14, pp. 1 – 9, Jan. 2021.
- [7] S. Hornstein, V. Forman-Hoffman, A. Nazander, K. Ranta, and K. Hilbert, "Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach," *Digital Health*, vol. 7, pp. 205520762110606, Jan. 2021.
- [8] A. Priya, S. Garg, and N. P. Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms," *Procedia Computer Science*, vol. 167, pp. 1258 – 1267, 2020.
- [9] A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," *Healthcare Technology Letters*, vol. 4, no. 6, pp. 238 – 243, Dec. 2017.
- [10] M. Syed and M. Narvekar, "Early Detection of Depression Indication from Social Media Analysis," *ITM Web of Conferences*, vol. 40, pp. 03029, 2021.
- [11] E. Bailey, J. Robinson, and P. McGorry, "Depression and suicide among medical practitioners in Australia," *Internal Medicine Journal*, vol. 48, no. 3, pp. 254 – 258, Mar. 2018.
- [12] K. C., "Health Care Data Analysis through Machine Learning Meta Heuristic Algorithm," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 7, pp. 196 – 201, Jul. 2020.
- [13] S. K., "Impact of Loss Function Using M-LSTM Classifier for Sequence Data," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 5, pp. 3487 – 3494, Apr. 2020.