Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Stress Detection Through Text in social media Using Machine Learning Techniques

Fauziah Kasmin[1,*], Nur Afeeqah Irsaleena Razali[1], Sharifah Sakinah Syed Ahmad[1], Zuraini Othman[1], Dian Sa'adilah Maylawati[2]

[1] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
[2] Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia

ABSTRACT

In today's digital era, the prevalence of stress-related discussions on social media platforms such as Twitter, Facebook, Instagram, and Reddit has garnered considerable attention. Human stress causes mental and financial problems, impairs one's ability to think clearly at work, strains relationships with co-workers, depresses oneself, and, in extreme circumstances, can result in suicide. Therefore, identifying stress is crucial to reducing its effects. Stress detection and measurement in the large world of social media data is a difficult and time-consuming task. Hence, this comprehensive review explores the crucial realm of detecting and quantifying stress through user behaviour analysis, leveraging the capabilities of machine learning approach. Our primary goals encompass developing a binary stress detection model, conducting a thorough comparative analysis of machine learning models, and designing an intuitive stress detection dashboard for visualizing data. The study utilizes three distinct datasets: the Reddit dataset containing 3,532 records, the Twitter dataset with 1,228 records, and an integrated dataset combining data from both sources, total 4,760 records. Key techniques for feature extraction, particularly Term Frequency-Inverse Document Frequency (TF-IDF), are employed to extract valuable insights from textual data. The study's findings demonstrate how well some machine learning models perform with various datasets and training/testing splits. Interestingly, the Logistic Regression model performs admirably, with an astounding 73% accuracy on the Reddit dataset. All models perform well on the Twitter dataset, however under certain conditions, the Support Vector Machine model outperforms the others with an amazing 81% accuracy. With an accuracy rating of 74% in the combined dataset, the Support Vector Machine likewise shows up as the best performer. The findings contribute significantly to ongoing efforts in enhancing stress detection, early intervention strategies, and health research within the sphere of social media.

Keywords:
Stress detection; machine learning; social media; text

1. Introduction

Life is a journey that has its share of happy, sad, learning, and growing moments. Meaning is woven into this intricate fabric by relationships, experiences, and human development. Life is

[*] *Corresponding author.*
E-mail address: fauziah@utem.edu.my

https://doi.org/10.37934/araset.63.2.162176

undoubtedly beautiful, but it also comes with a lot of difficulties, like grief, uncertainty, and personal troubles. The difficulties in life often lead to stress, which can manifest emotionally, mentally, and physically. Stress is a state of emotional or physical tension arising from any event or thought that causes feelings of frustration, anger, or nervousness [1]. Human stress results in mental and socio-economic issues, a lack of clarity at work, poor workplace relationships, depression, and, in severe cases, can lead to suicide [2]. Thus, stress do affect the quality of life of an individual. According to a mental health and wellness survey conducted by Rakuten Insight in Malaysia in May 2022, 59 percent of respondents aged 16 to 24 reported experiencing higher levels of stress or anxiety in the past 12 months. In contrast, 34 percent of respondents aged 25 to 34 indicated that their stress and anxiety levels remained the same over the past year [3]. As global population growth continues, stress has become increasingly prevalent, and social media platforms have emerged as a space where individuals openly express their experiences with stress. Chronic stress results in a weakened immune system [1], cancer, cardiovascular disease, depression, diabetes and substance addiction [1-4]. Despite improvements in the previous few decades in the diagnosis of mental illness, many cases go undiagnosed [5]. Hence, detection of stress is very important to mitigate the consequences.

There are Some researchers have used wearable sensors. There are researchers using wearable sensors [6-8] that detect the changes in bio physiological and biochemical of a certain individual to detect stress. There are also researchers who are using social media text [9-11] to recognize stress. Detecting and quantifying stress within the vast realm of social media data poses a challenging and time-consuming task. Stress detection using social media involves analysing users' online behaviour, language, and interactions to identify signs of stress. Symptoms related to mental illness are visible on platforms like Twitter, Facebook, web forums, and automated methods are increasingly capable of detecting depression and other mental illnesses [5]. Text analytics in natural language processing (NLP) is used to assess text for indicators like negative sentiment, excessive use of first-person pronouns, and specific keywords associated with stress. This necessitates providing counselling to help stressed individuals cope with their stress [12]. Text analytics have been used in diverse domain such as education [13], healthcare [14], finance [15], retail, marketing, legal [16], customer service and many more.

Through the extensive passive monitoring of social media, automated detection techniques may identify depressed or otherwise at-risk individuals. In the future, they may even supplement current screening protocols [5]. Machine learning offers an efficient and effective approach to address this issue [11]. Machine learning algorithms can classify and predict stress levels based on these textual patterns. There is many research have been done in detecting stress by using machine learning approach. Some of the researches have used deep neural networks [17], decision tree [18], random forest [18], naïve bayes [19], support vector machine [20] and many more.

Recent social media platforms have created a public space where people can express their thoughts and experiences, including those related to stress. This openness has sparked increasing interest in stress research across both academic and practical fields. Researchers have examined various datasets and platforms, including Reddit, Twitter, Facebook, and the PhysioBank dataset. Some studies have also utilized questionnaires and specific sources like Dreaddit [17]. Several studies, including those by Geetha *et al.,* [18] and Rastogi *et al.,* [17], have collected data from Reddit. Rastogi *et al.,* [17] developed four datasets, two from Reddit and two from Twitter, for stress detection with binary labels (0 for no stress and 1 for stress). The "Reddit Title" dataset comprises post titles from stress and non-stress sub Reddits, ensuring class balance from September 2019 to September 2021. The "Reddit Combi" dataset combines titles and body text from relevant sub Reddits, featuring longer, more descriptive text but an unbalanced distribution due to non-text data. The "Twitter Full" dataset includes tweets from stress and non-stress hashtags, providing real-world data from

September 2019 to September 2021. The "Twitter Non-Advert" dataset, derived from "Twitter Full," uses a de noising method to produce cleaner data for stress classification. Preprocessing steps are consistent, involving the removal of HTML tags, links, special characters, punctuation, extra spaces, and converting emoji to text. Extremely short examples and specific Twitter symbols, except for hashtags, are also removed. They utilized a fine-tuned DistilBERT model with student-teacher training for data de noising. This model, fine-tuned on separate datasets for clickbait/advertisements and clean data, was applied to create the "Twitter Non-Advert" dataset. Notably, the Reddit data contained minimal noise. In their evaluation, Rastogi *et al.,* [17] used accuracy and the F1 score as key metrics, employing pre-trained language model (PLM-based) for classification and achieving an accuracy score of 98.2%.

The primary goal of a study by Febriansyah *et al.,* [19] was to identify stress using Reddit data. Using natural language processing (NLP) for stress detection, the researchers hoped to enhance mental health evaluations and recommender systems. They gained access to a useful supply of data for their models by taking advantage of social media's extensive usage. They collected data from Dreaddit, a Reddit-based dataset, in order to identify social media users who were under stress. Comprehensive preprocessing was performed on the text data, which included removing stop words, punctuation, URLs, handling reposts, and removing emoji. Posts that had originally been disqualified were added back in to improve the machine learning model. To categorise the text data and detect stress, a number of pre-trained models were used, including Support Vector Machine (SVM), Decision Tree, Naive Bayes, Random Forest, Bag of Words, and TF-IDF. The F1 score, accuracy, recall, precision, and other metrics were used by the study team to assess the stress detection algorithm. Data and labels were used to train each model to categorise new postings as "stressed" or "not stressed." For instance, the SVM categorised messages according to how close they were to a certain hyperplane. Support Vector Machine (SVM) attained 75% accuracy, Naive Bayes 69%, Decision Tree 56%, and Random Forest 60.97%, according to Febriansyah *et al.,* [19].

In a study, Kamite *et al.,* [21] used machine learning algorithms to identify signs of depression based on Twitter data. They presented a system to categorise tweets as depressive or non-depressive by combining multiple tweet characteristics, including language, sentiment, and user interactions. The study's main objective is to treat mental health issues by using data from social media to identify problems early and take appropriate action. Using a dataset of 1,000 tweets that was equally divided into depressive and non-depressive groups, the researchers were able to train and assess their machine learning models. Tokenisation, stop word removal, stemming, feature extraction, and vectorisation were all part of the text data preparation process. To improve the categorisation process, other data including average word length, character count, and word count were extracted. The text data was modified for machine learning techniques using Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation. An equal number of depressed and non-depressive tweets were chosen at random for the purpose of training and evaluating the model in order to guarantee a balanced dataset. Metrics like accuracy, recall, precision, and the F1 Score were used by the researchers to evaluate the performance of the model. Notably, they used the Random Forest method to diagnose depression with an amazing accuracy rate of 99.89%.

The goal of Nijhawan *et al.',* [22] analysis of data collected from Twitter and, on occasion, Facebook was to identify stress markers in people based on their posts on social media. To classify sentiment and anticipate subjects from the text, they used machine learning and natural language processing techniques, such as sentiment and emotion analysis. The 100,042 tweets in their sentiment analysis dataset were classified as either positive (1) or negative (0). A different dataset consisting of 7,934 tweets was also divided into categories based on emotions including fear, rage, sadness, joy, and neutrality. Their algorithms showed promise for mental health analysis by being effective at detecting stress. Cleaning, merging, condensing, and converting were among the pre-processing procedures. Tokenisation was done first, followed by the removal of stop words, and stemming was used to make word forms simpler. The text feature extraction method made advantage of the Bag-of-Words methodology. For the purpose of training and evaluating the models, the dataset was split into subsets for validation and training. One significant drawback was that the training set only contained emoticon-heavy tweets, which could potentially affect the model's performance. The F1 Score and accuracy are included as evaluation metrics. The models with the highest accuracy were the Decision Tree (72%), Random Forest (86.96%), and Logistic Regression (78.99%).

Looking at the trend, this work is dedicated to the development of a stress detection system utilizing machine learning algorithms, specifically on platforms such as Twitter and Reddit. The objectives of this work include:

i. A comprehensive evaluation of various machine learning models for the detection of stress content and
ii. the creation of a user-friendly dashboard for visualizing stress-related data.

Stress is a significant factor with far-reaching health implications [3]. Therefore, a reliable detection system is crucial, particularly for early intervention and personalized care. With the widespread use of social media platforms like Twitter and Reddit, millions of people openly share their feelings and experiences related to stress. This project centres on the challenge of detecting and understanding stress within the context of social media. It underscores the importance of studying user behaviour patterns, emotional aspects, and leveraging machine learning techniques for early stress identification. Given the severe health consequences associated with stress, there is a pressing need for a dependable system to identify it effectively. The ultimate goal of this work is to develop machine learning models capable of distinguishing between normal and stressed behaviour on social media platforms. This system holds the potential to provide timely interventions and enhance our understanding of stress's impact in the digital age.

2. Methodology

This section explains how the project is carried out. A step-by-step process have been followed based on the Cross Industry Standard Process for Data Mining (CRISP-DM) framework [23] as shown in Figure 1, which includes data collection, data preprocessing, feature extraction, model selection, and evaluation. However, it is important to mention that the usage of CRISP-DM framework is up to the evaluation stage. This framework has guided the research design for a structured analysis of stress detection on social media platforms.

2.1 Business Understanding

The objective of this work is to utilize machine learning to identify stress-related content on platforms like Twitter and Reddit. Since many individuals express their stress on these platforms, understanding the prevalence and impact of stress in the digital world is essential. The potential benefits are significant. Healthcare professionals and researchers can gain valuable insights into mental health and stress trends, enabling them to provide better support and make new discoveries. Public health agencies can use this data to implement targeted stress reduction initiatives, while businesses can enhance the well-being of their employees and customers. Ultimately, this work aims to improve people's quality of life in various ways. In summary, our goal is to employ machine learning to detect stress on social media, benefiting healthcare, research, individual well-being, public health, and businesses.

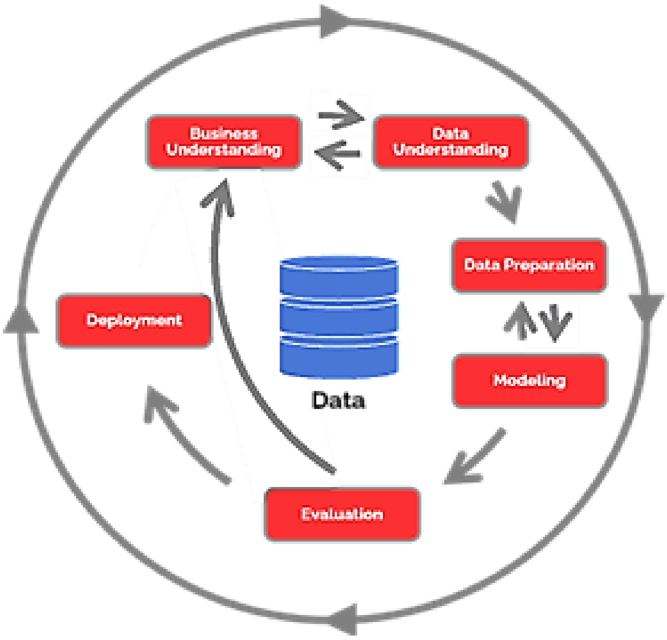


Fig. 1. CRISP-DM framework [23]

2.2 Data Understanding

The dataset used to train and test our social media stress detection model, have been explained in Table 1, originally came from Kaggle [20]. It consisted of over 3,000 rows, containing user ID, Reddit text content, and labels ('0' for no stress and '1' for stress). After cleaning, the dataset contained 3,532 rows, focusing on the 'text' and 'label' columns for analysis. To diversify the dataset, tweets were collected from Apify.com over a week in August 2023, resulting in a diverse dataset of 2,051 tweets, which was cleaned down to 1,228 rows. Each entry included user ID, tweet ID, tweet text, timestamp, language category, likes, and retweets. By combining the Kaggle dataset with the Apify.com dataset, a more comprehensive dataset was created for analyzing stress-related content on social media. This new dataset, as shown in Table 1, consists of 4,760 rows and includes text content, labels, and language information.

Table 1
Dataset used

Dataset	Source	Data type	Number of rows	Number of columns
Reddit	Kaggle	Text	3	3532
Twitter	Apify	Text	7	1228
Integrated	Kaggle and Apify	Text	3	4760

2.3 Data Preprocessing

In this study, there are two separate sets of data to work with, and both were integrated into a new dataset, making a total of three datasets. There are two distinct paths for analyzing this data: one for the Reddit dataset and one for the Twitter dataset. In our methodology involving Reddit data, depicted in Figure 2, importing and meticulously preparing the dataset is done to ensure it is wellorganized and free from inaccuracies during a rigorous cleaning phase. Once the data is prepared, feature extraction is done, where this process uncover meaningful patterns and information within the textual data. Lastly, the process moves to the model evaluation phase, assessing the performance of the models and the accuracy of the predictions. This pivotal step guarantees the reliability and precision of our analysis.

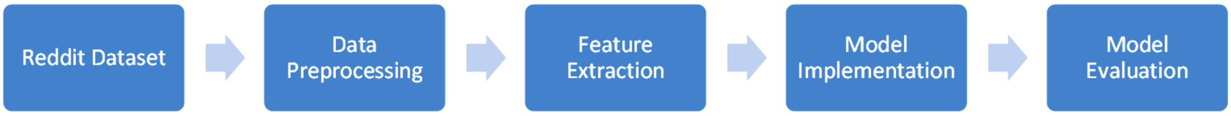


Fig. 2. Reddit data analysis flow

Regarding the Twitter dataset, the process closely mirrors that used for Reddit, with specific adaptations for Twitter, as depicted in Figure 3. Initially, Twitter data is imported and undergoes data pre-processing to ensure cleanliness and readiness. Subsequently, feature extraction is conducted to extract valuable insights from the raw Twitter data. Notably, a distinct step involves sentiment-based labelling, categorizing tweets into stress and non-stress categories. Following labelling, machine learning algorithms are applied in model implementation to predict stress in tweets on social media. Finally, model performance is evaluated. The third dataset, combining Twitter and Reddit data, follows a parallel approach to the Reddit dataset, undergoing similar stages of processing.

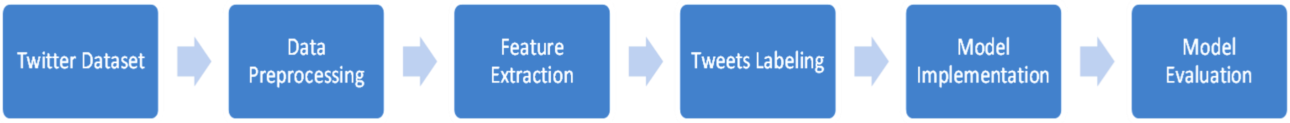


Fig. 3. Twitter data analysis flow

For both Reddit and Twitter data, a thorough data preprocessing phase is implemented to maintain data quality and consistency. This includes addressing missing values, eliminating duplicates, and performing essential transformations on the text data. Natural Language Processing techniques such as normalization, lowercasing [19], expanding contractions, tokenization, stopword removal, punctuation removal, lemmatization, word correction using TextBlob [24], and token concatenation are uniformly applied to ensure effective preprocessing of the text data.

The data pre-processing pipeline consists of several steps, as illustrated in Figure 4. Python libraries are initially imported into Google Colab for analysis. Subsequently, the Reddit and Twitter datasets are imported. Numerical and categorical features within the datasets are identified, and any missing values are addressed. Duplicate data entries are then handled to ensure data integrity. Libraries such as Numpy, Pandas, NLTK, and Matplotlib are utilized to enhance data quality. In the Reddit dataset, no missing values are identified, but there are 21 instances of duplicate entries. While in the Twitter dataset, there are no missing values and during dataset examination, 220 duplicate tweets are discovered.

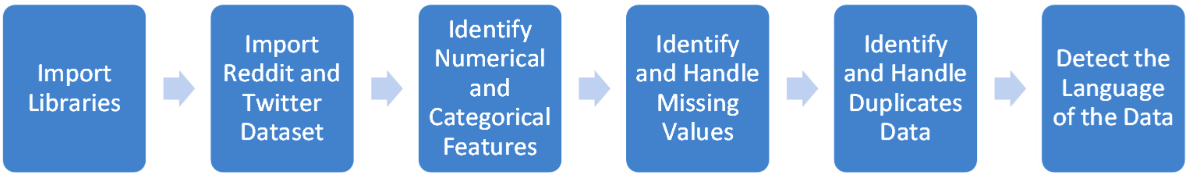

Import Libraries
Import Reddit and Twitter Dataset
Identify Numerical and Categorical Features
Identify and Handle Missing Values
Identify and Handle Duplicates Data
Detect the Language of the Data

Fig. 4. Data Pre-processing workflow

The Reddit dataset includes text in two languages: English and Vietnamese, totaling 3,531 rows, which comprises nearly 99.97% of the original dataset. Conversely, the Twitter dataset contains text in 30 languages, including English, Indonesian, Estonian, Tagalog, and Dutch. For consistency, our analysis focused solely on English tweets. This subset consists of 1,228 rows, representing approximately 59.8% of the original dataset, which initially contained 2,051 tweets.

Data preprocessing stands as a crucial phase in every research endeavor. In this study, the NLTK library is employed for text preprocessing. The comprehensive text preprocessing pipeline encompasses several essential steps: standardizing text to lowercase for uniformity, addressing repeated characters, removing punctuation, tokenizing text into individual words, eliminating common and less informative words (stop words), lemmatizing words to their base forms, correcting spelling errors, and reassembling processed tokens into coherent text. This meticulous preprocessing ensures that the text data is prepared for subsequent natural language processing tasks, such as sentiment analysis or text categorization, facilitating more precise and dependable analysis of social media content.

2.4 Feature Extraction

The study uses TF-IDF (Term Frequency-Inverse Document Frequency) as a technique to turn text data into numbers. It is widely used in Natural Language Processing (NLP) to measure how important words are in documents. This is helpful for stress detection in both Reddit and Twitter data. In another study by Satyendra *et al.,* [9], they also used TF-IDF for Sentiment Analysis of Twitter data. Sentiment analysis needs good ways to measure things in text, and TF-IDF is one of the popular methods. It works by looking at how often words show up in documents and how important they are. So, it helps find the most important words for understanding sentiment in text.

2.5 Twitter Data Labelling Using Sentiment Analysis

This section details the process of labelling tweets in the Twitter dataset for subsequent analysis and machine learning tasks. Sentiment analysis was utilized to categorize tweets into two groups: those expressing positive sentiments (indicating no stress) and those conveying negative sentiments (indicating of stress-related content). TextBlob [24], a widely recognized sentiment analysis tool, was employed to assess the sentiment of each tweet. The labelling system is binary, facilitating machine learning. Tweets with a positive sentiment score (greater than zero) are labelled as 0 (no stress), while tweets with a negative sentiment score (zero or less) are labelled as 1 (stress). This approach simplifies the identification of stress in tweets. Sentiment analysis relies on a sentiment lexicon, i.e. a compilation of words associated with emotions where each word is assigned a sentiment score. By evaluating each word in a tweet against this lexicon, an overall sentiment score is computed. If the majority of words are positive, the tweet receives a positive score, and vice versa. This score aids in determining whether a tweet contains stress-related content, as depicted in Table 2.

Table 2
An example of sentiment score after categorization

Pre-processed Tweets	Sentiment score	Label
imro45 awesome bat thank patting make feel relax whenever stress	1	0
ukwnlvr stressed good	0.7	0
exam stress wanting watch crazy rich asia day exam 😩	-0.112	1

2.6 Data Visualization

In the Reddit dataset, based on Figure 5, the label distribution indicates the presence of 1850 stress-related texts, while 1681 texts are classified as no stress. In the Twitter dataset, the distribution of tweet labels shown in Figure 6 reveals that there are 664 tweets classified as stress-related, while 564 tweets fall into the category of no stress. The integrated dataset, representing the combination of Twitter and Reddit data sources, comprises 2514 stress-related texts and 2245 texts categorized as no stress as shown in Figure 7.

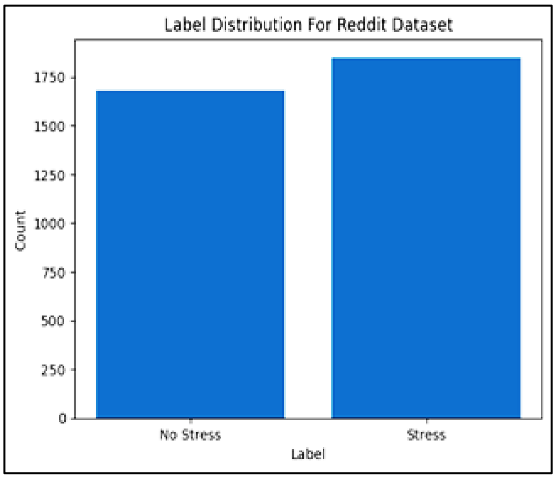


Fig. 5. Label distribution for Reddit dataset

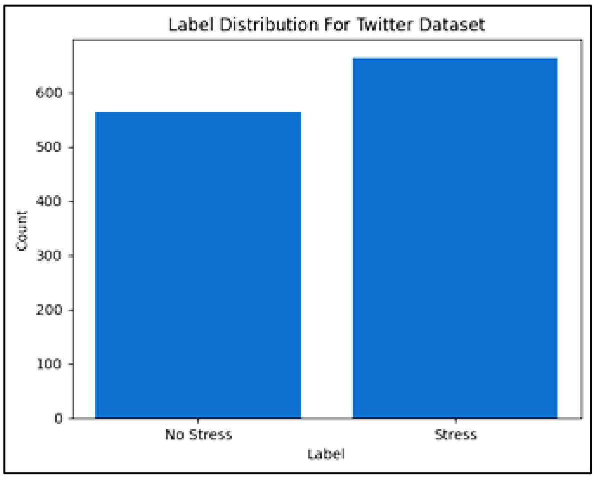


Fig. 6. Label distribution for Twitter dataset

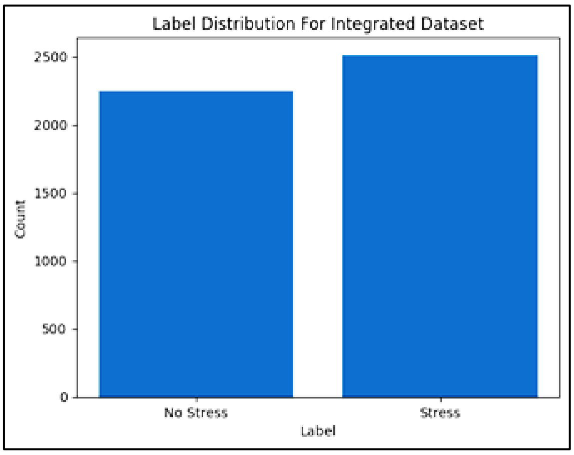


Fig. 7. Label distribution for integrated dataset

Figure 8 displays a word cloud representing the entire Reddit dataset. It highlights words like "know", "feel", "time", "thing", "want", "really", "even", and "going" as highly frequent terms within the dataset. Figure 9 exhibits a word cloud representing the entire Twitter dataset. It highlights words like "stress", " stressed", "time", "people", "need", "go", "life", and "work". Figure 10 presents a word cloud representing the entire Integrated dataset. It highlights words like "know", "feel", "time", "people", "stress", "help", "life", and "want".

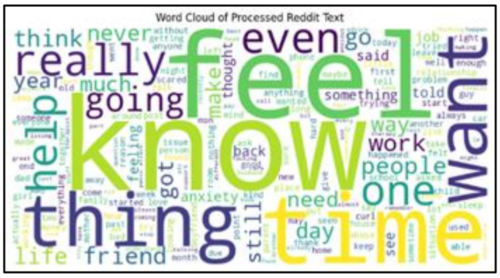


Fig. 8. Word cloud for Reddit dataset

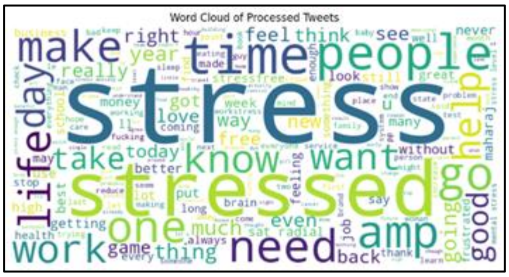


Fig. 9. Word cloud for Twitter dataset

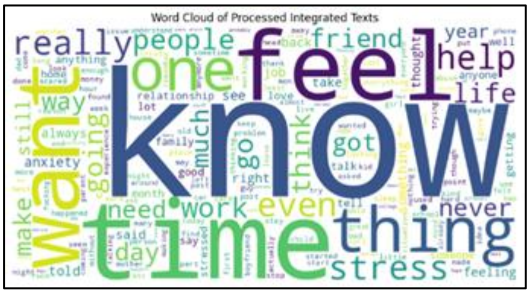


Fig. 10. Word Cloud for Integrated Dataset

2.7 Model Evaluation

This section thoroughly evaluates the performance of stress detection models. Several key metrics are employed to assess the effectiveness of these machine learning algorithms, providing a comprehensive overview of their capabilities. The focus is primarily on five essential performance measures: Accuracy, Precision, Recall, F1 Score, and Receiver Operating Characteristic - Area Under the Curve (ROC-AUC). Each metric serves a crucial role in evaluating various aspects of model performance, facilitating informed assessments of their suitability for stress detection. The evaluation includes the mathematical formulas utilized to compute them [25].

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP\ +FN+\ TN}} \tag{1}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP + FP}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

$$F1\ score = \frac{2 \times precision \times recall}{precision+recall} \tag{4}$$

2.8 Machine Learning Model

Different machine learning methods are being evaluated for stress detection, including Support Vector Machines (SVM) [26], Decision Tree [27], Random Forest [28], and Logistic Regression [29]. These methods were selected based on their effectiveness in text classification and their demonstrated performance in prior research. The techniques will be trained and optimized using the pre-processed data. Initially, the data was divided into 70% for training and 30% for testing to maintain a balanced approach to model training and evaluation. To delve deeper into training, an 80% training and 20% testing split was explored to capture finer data patterns. Additionally, a 90%

training and 10% testing split was tested to assess the model's performance with a larger training set and its ability to generalize well to familiar examples.

3. Results and Conclusions

To thoroughly assess how well the model is performed on the Reddit, Twitter, and integrated datasets for stress detection, a comprehensive approach has been taken to splitting the data. The datasets are divided into different proportions: 70% for training and 30% for testing, 80% for training and 20% for testing, and 90% for training and 10% for testing. To make it clear and easy to understand, tables have been created to summarize the performance of various machine learning models on these datasets, using key metrics like accuracy, precision, recall, F1 score, and ROC-AUC.

3.1 Data Splitting Strategies and Model Performance Evaluation

Table 3 assesses the performance of different machine learning models on the Reddit dataset for stress detection. The dataset is split into three training and testing scenarios, with varying training percentages. Accuracy scores range from 0.60 to 0.71, with Random Forest achieving the highest accuracy of 0.71. Precision and recall are vital, with Random Forest achieving a high recall of 0.80 at 90% training, indicating effective stress text identification. The F1 score, a balance of precision and recall, peaks at 0.76 for Random Forest with a 90% training split. ROC-AUC values above 0.7 show the models are effective in distinguishing stress from non-stress texts. Overall, Random Forest stands out for its high recall, making it useful for capturing most stress-related texts while allowing some precision trade-off.

Table 3
Reddit dataset performance

Dataset			Reddit				
Training data	Testing data	Model	Performance measures				
			Acc	Precision	Recall	F1 Score	ROC-AUC
70%	30%	SVM	0.69	0.7	0.72	0.71	0.69
		Random Forest	0.69	0.68	0.78	0.72	0.68
		Decision Tree	0.60	0.61	0.64	0.63	0.6
		Logistic Regression	0.69	0.7	0.73	0.71	0.69
80%	20%	SVM	0.70	0.71	0.74	0.72	0.7
		Random Forest	0.69	0.69	0.77	0.73	0.69
		Decision Tree	0.60	0.62	0.62	0.62	0.6
		Logistic Regression	0.70	0.71	0.75	0.73	0.7
90%	10%	SVM	0.70	0.74	0.73	0.73	0.7
		Random Forest	0.71	0.72	0.8	0.76	0.69
		Decision Tree	0.61	0.67	0.63	0.65	0.61
		Logistic Regression	0.70	0.73	0.75	0.74	0.69

Table 4 shows the performance of Reddit dataset in our work when compared to Febriansyah *et al.,* [19] work. In our work, we have split the training dataset and testing dataset based on 70:30, 80:20 and 90:10. Then from the accuracy results of splitting, we take an average of accuracy. From Table 4, it can be seen that the accuracy of SVM is differ that is lower from [19] work. While for Decision Tree and Random Forest, the accuracy is higher as compared to [19] work. The result is different due to the splitting that have been done and also due to different preprocessing techniques used in preparing the dataset. From the splitting that have been done, the more data have been used for training, better accuracy can be achieved in testing dataset. From the comparison also, it can be seen that SVM is still the best model to detect stress and it is the same as what have been achieved by [19] work.

The accuracy scores from Table 5 shows slight variations across different models and data splits. They generally fall between 0.68 and 0.81. The SVM model attains the highest accuracy of 0.81, indicating that it correctly identifies stress and non-stress texts with an accuracy of 81%.

Table 4
Comparison with Febriansyah *et al.*, [19] work in Reddit dataset

Model	Reddit dataset	
	Accuracy	
	Splitting (our work)	Without splitting [19]
SVM	0.697	0.75
Decision Tree	0.603	0.56
Random Forest	0.690	0.61

Table 5
Twitter dataset performance

Dataset			Twitter				
Training data	Testing data	Model	Performance measures				
			Accuracy	Precision	Recall	F1 score	ROC-AUC
70%	30%	SVM	0.76	0.76	0.83	0.79	0.76
		Random Forest	0.77	0.74	0.89	0.81	0.76
		Decision Tree	0.68	0.68	0.78	0.73	0.67
		Logistic Regression	0.78	0.77	0.83	0.8	0.77
80%	20%	SVM	0.77	0.77	0.8	0.8	0.76
		Random Forest	0.79	0.76	0.88	0.82	0.78
		Decision Tree	0.68	0.69	0.76	0.73	0.67
		Logistic Regression	0.79	0.79	0.85	0.82	0.79
90%	10%	SVM	0.81	0.83	0.84	0.83	0.81
		Random Forest	0.79	0.8	0.83	0.81	0.78
		Decision Tree	0.76	0.77	0.81	0.79	0.75
		Logistic Regression	0.8	0.82	0.84	0.83	0.8

Table 6 shows, the accuracy scores vary slightly across different models and data splits but generally range between 0.62 and 0.74. The highest accuracy achieved is 0.74, which indicates that the SVM model correctly classifies stress and non-stress texts 74% of the time.

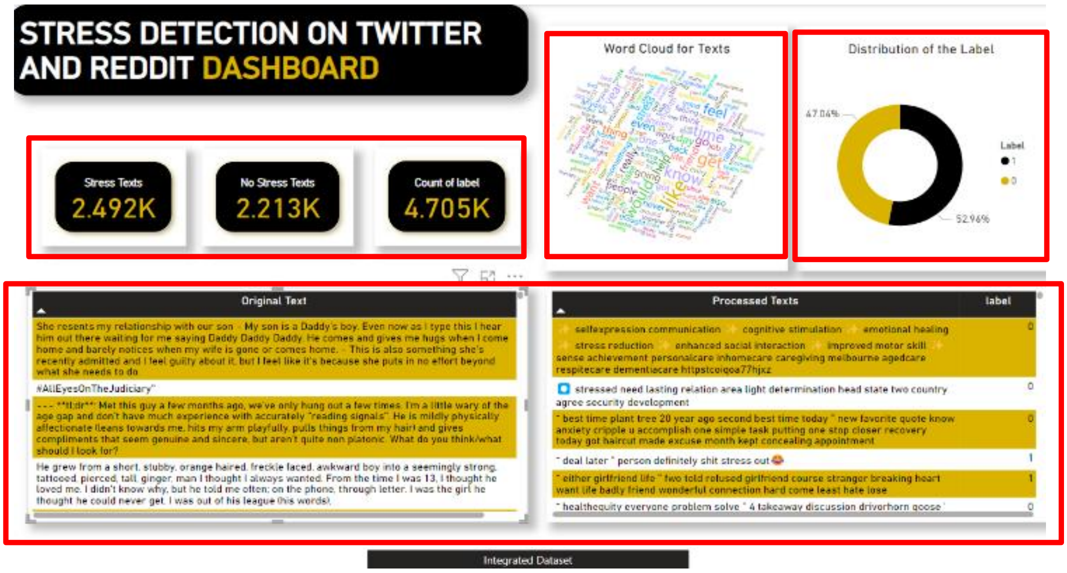
As a conclusion, the model performance evaluation yielded significant insights across the three distinct datasets. It is discovered that SVM, Random Forest, and Logistic Regression models performed well in the Reddit dataset. For the Reddit dataset, the Logistic Regression Random Forest model displayed the most promising performance, achieving an accuracy of 73%, 71%. In comparison, the Decision Tree model performed quite poorly, emphasizing the need of careful model selection. The Twitter dataset exhibited its highest accuracy when employing the SVM model, reaching 81% accuracy with a 90% training and 10% testing data split. Finally, the integrated dataset also saw SVM emerge as the top-performing model, securing a substantial accuracy rate of 74% under the same data split conditions.

Table 6
Integrated dataset performance

Dataset			Integrated				
Training Data	Testing data	Model	Performance measures				
			Accuracy	Precision	Recall	F1 Score	ROC-AU
70%	30%	SVM	0.72	0.73	0.76	0.75	0.72
		Random Forest	0.7	0.7	0.77	0.73	0.7
		Decision Tree	0.62	0.64	0.65	0.65	0.62
		Logistic Regression	0.72	0.72	0.76	0.74	0.71
80%	20%	SVM	0.72	0.74	0.76	0.75	0.72
		Random Forest	0.71	0.72	0.78	0.75	0.7
		Decision Tree	0.62	0.66	0.65	0.65	0.62
		Logistic Regression	0.72	0.73	0.76	0.75	0.72
90%	10%	SVM	0.74	0.75	0.79	0.77	0.74
		Random Forest	0.72	0.72	0.8	0.76	0.72
		Decision Tree	0.64	0.66	0.7	0.67	0.64
		Logistic Regression	0.73	0.74	0.77	0.75	0.73

3.2 Stress Detection Dashboard

Dashboard brings data to life with simple charts, graphs, and word clouds. Users can see the distribution of stress-related tweets and word clouds that highlight the most frequently used stress-related words. These visualizations provide a clear and succinct summary of the data, allowing users to discover trends and patterns at a glance as shown in Figure 11.



Original Text
She resents my relationship with our son - My son is a Daddy's boy. Even now as I type this I hear him out there waiting for me saying Daddy Daddy Daddy. He comes and gives me hugs when I come home and barely notices when my wife is gone or comes home. - This is also something she's recently admitted and I feel guilty about it, but I feel like it's because she puts in no effort beyond what she needs to do
#AllEyesOnTheJudiciary"
- - - **tl;dr**: Met this guy a few months ago, we've only hung out a few times. I'm a little wary of the age gap and don't have much experience with accurately "reading signals". He is mildly physically affectionate (leans towards me, hits my arm playfully, pulls things from my hair) and gives compliments that seem genuine and sincere, but aren't quite non platonic. What do you think/what should I look for?
He grew from a short, stubby, orange haired, freckle faced, awkward boy into a seemingly strong, tattooed, pierced, tall, ginger, man I thought I always wanted. From the time I was 13, I thought he loved me. I didn't know why, but he told me often; on the phone, through letter. I was the girl he thought he could never get. I was out of his league (his words).

Processed Texts	label
✨ selfexpression communication ✨ cognitive stimulation ✨ emotional healing ✨ stress reduction ✨ enhanced social interaction ✨ improved motor skill ✨ sense achievement personalcare inhomecare caregiving melbourne agedcare respitecare dementiacare httpstcoiqoa77hjxz	0
🔲 stressed need lasting relation area light determination head state two country agree security development	0
" best time plant tree 20 year ago second best time today " new favorite quote know anxiety cripple u accomplish one simple task putting one step closer recovery today got haircut made excuse month kept concealing appointment	0
" deal later " person definitely shit stress out 😂	1
" either girlfriend life " two told refused girlfriend course stranger breaking heart want life badly friend wonderful connection hard come least hate lose	1
" healthequity everyone problem solve " 4 takeaway discussion driverhorn goose	0

Fig. 11. Dashboard for Reddit, Twitter and Integrated dataset

Based on Figure 11, the dashboard displays essential data statistics, including the number of stress-related texts, the number of no-stress texts, and the total number of texts in the dataset. This provides users with a quick overview of the dataset's composition. The dashboard features a word cloud representation of the most frequent words in the dataset shown in Figure 11. This visual representation allows users to identify dominant terms and gain insights into the prevalent themes within the text data. A pie chart visualization showcases the label distribution, differentiating between label 0 (indicating no-stress texts) and label 1 (indicating stress texts) in Figure 11. This visualization aids in understanding the balance between stress and no-stress texts in the dataset. Another feature of the dashboard is to illustrate the impact of text preprocessing. The dashboard provides a side-by-side comparison of text before and after processing. This helps users appreciate the transformation that the text data undergoes, from raw unprocessed text to cleaned and structured data.

In this work, focused had been done on creating a binary stress detection model using various machine learning algorithms and assessed their performance. For the Reddit dataset, Logistic Regression Random Forest displayed good results with the accuracy scores 73%, 71%, respectively. On the other hand, the Decision Tree model performed less successfully, highlighting the significance of model choice. Comparison on Febriansyah *et al.'s,* [19] work also had been done for Reddit dataset. It shows a little bit difference as splitting have been done and different pre-processing approach. Additionally, the comparison shows that SVM remains the most effective model for detecting stress, matching the findings of Febriansyah *et al.'s,* [19]. In the Twitter dataset, all models, including SVM, Random Forest, and Logistic Regression, showed commendable results, with SVM achieving an accuracy of 81.30%. For our integrated dataset, SVM and Logistic Regression consistently provided reliable outcomes. SVM excelled across various data splits, with high accuracy (74.37%), precision, F1 score, and ROC AU values. This indicates SVM's strength in stress detection on the integrated dataset.

4. Future Improvement

For future enhancements for this work, improvements in stress-related text detection can be achieved through exploring more advanced feature engineering techniques, such as word embedding (e.g., Word2Vec, GloVe) or deep learning-based approaches (e.g., LSTM, BERT embedding). Another task that can be done is conducting hyper parameter tuning for machine learning models to further optimize their performance. Implementing a real-time monitoring system for social media platforms can also be considered. This is to continuously collect and analyze stress-related content. Investigation on ensemble methods, such as stacking or boosting, can also be considered as this is to combine the strengths of multiple machine learning models.

Acknowledgement

Our deepest gratitude and thanks to Universiti Teknikal Malaysia Melaka (UTeM) and the Ministry of Higher Education Malaysia.

References

[1] Salleh, Mohd Razali. "Life event, stress and illness." *The Malaysian journal of medical sciences: Malaysian Journal of Medical Science* 15, no. 4 (2008): 9-18

[2] Moreno-Smith, Myrthala, Susan K. Lutgendorf, and Anil K. Sood. "Impact of stress on cancer metastasis." *Future Oncology* 6, no. 12 (2010): 1863-1881. https://doi.org/10.2217/fon.10.142

[3] Bannore, Aishwarya, Tejashree Gore, Apoorva Raut, and Kiran Talele. "Mental stress detection using machine learning algorithm." In *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, p. 1-4. IEEE, 2021. https://doi.org/10.1109/ICECCME52200.2021.9590847

[4] Nurika, Okta, Muhamad Hariz Bin Muhamad Adnan, Mohd Hishamuddin Bin Abdul Rahman, and Ahmed Abba Haruna. "Stress Detection and Mitigation through Discovery of Optimal Lecture Delivery Method for STEM-Enrolled University Students." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 40, no. 1 (2024): 87-95. https://doi.org/10.37934/araset.40.1.8795

[5] Guntuku, Sharath Chandra, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. "Detecting depression and mental illness on social media: an integrative review." *Current Opinion in Behavioral Sciences* 18 (2017): 43-49. https://doi.org/10.1016/j.cobeha.2017.07.005

[6] Can, Yekta Said, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study." *Sensors* 19, no. 8 (2019): 1849. https://doi.org/10.3390/s19081849

[7] Iqbal, Talha, Adnan Elahi, Pau Redon, Patricia Vazquez, William Wijns, and Atif Shahzad. "A review of biophysiological and biochemical indicators of stress for connected and preventive healthcare." *Diagnostics* 11, no. 3 (2021): 556. https://doi.org/10.3390/diagnostics11030556

[8] Furia, Leonardo, Matteo Tortora, Paolo Soda, and Rosa Sicilia. "Exploring Early Stress Detection from Multimodal Time Series with Deep Reinforcement Learning." In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1917-1920. IEEE, 2023. https://doi.org/10.1109/BIBM58861.2023.10385549

[9] KVTKN, Prashanth, and Tene Ramakrishnudu. "Semi-supervised approach for tweet-level stress detection." *Natural Language Processing Journal* 4 (2023): 100019. https://doi.org/10.1016/j.nlp.2023.100019

[10] Inamdar, Shaunak, Rishikesh Chapekar, Shilpa Gite, and Biswajeet Pradhan. "Machine learning driven mental stress detection on reddit posts using natural language processing." *Human-Centric Intelligent Systems* 3, no. 2 (2023): 80-91. https://doi.org/10.1007/s44230-023-00020-8

[11] Kumari, Kirti, and Sima Das. "Stress detection system using natural language processing and machine learning techniques." In *WNLPe-Health@ ICON*, p. 45-55. 2022.

[12] Sriramprakash, Senthil, Vadana D. Prasanna, and OV Ramana Murthy. "Stress detection in working people." *Procedia computer science* 115 (2017): 359-366. https://doi.org/10.1016/j.procs.2017.09.090

[13] Chanthiran, Maran, Abu Bakar Ibrahim, Mohd Hishamuddin Abdul Rahman, and Dagmar Ruskova. "Text analytics:" Graphic visualization in education and a scientometric analysis using R tool to explore the impact and trends in classroom learning."." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 40, no. 2 (2024): 1-12. https://doi.org/10.37934/araset.40.2.112

[14] Elbattah, Mahmoud, Émilien Arnaud, Maxime Gignon, and Gilles Dequen. "The role of text analytics in healthcare: A review of recent developments and applications." *Healthinf* (2021): 825-832. https://doi.org/10.5220/0010414508250832

[15] Senave, Elseline, Mieke J. Jans, and Rajendra P. Srivastava. "The application of text mining in accounting." *International Journal of Accounting Information Systems* 50 (2023): 100624. https://doi.org/10.1016/j.accinf.2023.100624

[16] Kovalchuk, Olha, Serhiy Banakh, Mariia Masonkova, Kateryna Berezka, Serhii Mokhun, and Olha Fedchyshyn. "Text mining for the analysis of legal texts." In *2022 12th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 502-505. IEEE, 2022. https://doi.org/10.1109/ACIT54803.2022.9913169

[17] Li, Russell, and Zhandong Liu. "Stress detection using deep neural networks." *BMC Medical Informatics and Decision Making* 20 (2020): 1-10. https://doi.org/10.1186/s12911-020-01299-4

[18] Singh, Prithvipal, Gurvinder Singh, and Sarveshwar Bharti. "The predictive model of mental illness using decision tree and random forest classification in machine learning." In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 01-05. IEEE, 2022. https://doi.org/10.1109/ICACITE53722.2022.9823761

[19] Febriansyah, Mochamad Rizky, Rezki Yunanda, and Derwin Suhartono. "Stress detection system for social media users." *Procedia Computer Science* 216 (2023): 672-681. https://doi.org/10.1016/j.procs.2022.12.183

[20] Singh, Satyendra, Krishan Kumar, and Brajesh Kumar. "Sentiment analysis of Twitter data using TF-IDF and machine learning techniques." In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 1, p. 252-255. IEEE, 2022. https://doi.org/10.1109/COM-IT-CON54601.2022.9850477

[21] Kamite, Sangeeta R., and V. B. Kamble. "Detection of depression in social media via twitter using machine learning approach." In *2020 International conference on smart innovations in design, environment, management, planning and computing (ICSIDEMPC)*, p. 122-125. IEEE, 2020. https://doi.org/10.1109/ICSIDEMPC49020.2020.9299641

[22] Nijhawan, Tanya, Girija Attigeri, and T. Ananthakrishna. "Stress detection using natural language processing and machine learning over social interactions." *Journal of Big Data* 9, no. 1 (2022): 33. https://doi.org/10.1186/s40537-022-00575-6

[23] Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez. "A systematic literature review on applying CRISP-DM process model." *Procedia Computer Science* 181 (2021): 526-534. https://doi.org/10.1016/j.procs.2021.01.199

[24] Aljedaani, Wajdi, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry." *Knowledge-Based Systems* 255 (2022): 109780. https://doi.org/10.1016/j.knosys.2022.109780

[25] Powers, David MW. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." *arXiv preprint arXiv:2010.16061* (2020). https://doi.org/10.48550/arXiv.2010.16061

[26] Evgeniou, Theodoros, and Massimiliano Pontil. "Support vector machines: Theory and applications." In *Advanced course on artificial intelligence*, pp. 249-257. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999. https://doi.org/10.1007/3-540-44673-7_12

[27] Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." *Shanghai Archives of Psychiatry* 27, no. 2 (2015): 130-135. https://doi.org/10.11919%2Fj.issn.1002-0829.215044

[28] Breiman, L. "Random forests." *Machine Learning*." 45, no. 1 (2001): 5-32. https://doi.org/10.1023/A:1010933404324

[29] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. "An introduction to logistic regression analysis and reporting." *The Journal of Educational Research* 96, no. 1 (2002): 3-14. https://doi.org/10.1080/00220670209598786