



Stress Identification System using NLP and Machine Learning Approaches

¹Abhilash MS, ²Athmaranjan K, ³Akshaya UG, ⁴Meghana KS, ⁵Sanjana BM

¹Final year B.E. Student, ² Associate Professor, ISE, ³Final year B.E. Student, ⁴Final year B.E. Student, ⁵Final year B.E. Student

¹Department of Information Science & Engineering,
¹Srinivas Institute of Technology, Mangaluru, India

Abstract: Abstraction: One of the key psychological conditions that causes us to experience mental or bodily pain is stress. Stress can cause both emotional and physical health problems if we are unable to manage it. Owing to the lengthy history of social isolation, lockdown, fear, mistrust, etc., almost everyone entered the stressed period. In the last three years, online gaming has gained popularity as a substitute for physical activity. Anything that people see in their personal lives can be posted online. On social media, it's customary to analyze, judge, inspire, or identify emotions. Investigating sentiment from social media messages and identifying anxiousness in a group of people based on what they share on their profiles are the goals of this study. This work is divided into two sections: detection using machine learning and data extraction using NLP techniques. The four main stages in this framework are text mining, stress detection, auto summarization, and gathering content from social media. An internet user's mental stress or overload may be lessened by the new paradigm. The new model has employed several types of machine learning algorithms in contrast to earlier approaches. Positive outcomes from the SVM model include a 90% recall rate, 94% precision, and an F1 score that is getting close to 92%. In terms of early stress detection, the new paradigm would benefit society.

Keywords: Stress Recognition, NLP Techniques, ML Methods, ICON Conference 2022, WNLPe-Health Workshop 2022

I.INTRODUCTION

Regardless of age, gender, or race, stress is widely recognized as a common emotion that everyone experiences. What matters most is our response to anxiety, which could be significant or minor [1]. Stress can be either positive or detrimental, particularly during the time of the epidemic [2, 3]. To mention a few, we have created a voting system for the beauty festival, political campaigns, product analysis, and advertising promotion. These media are extensively utilized in all sectors. It is necessary to model and analyze such networks. As the tech industry continues to evolve rapidly, newer innovations are emerging.

Advanced machinery made of textual resources. To efficiently manage aspects of social media like duration, noise, and volatility, text mining techniques become essential. The vast volume of data in the social network makes it necessary to automate data distribution and retrieval within a predetermined timeframe. Interestingly, social media platforms make these enormous records sets accessible to researchers who wish to mine large amounts of data using statistics mining equipment. In the end, these researchers are the best candidates to mine statistics using statistics mining equipment. Statistics mining methods are designed to process large data sets in a variety of statistical styles. We can conclude that statistics mining, or more specifically, web mining, gives social networks the intelligence it needs to develop and operate in a more human-centered and adaptive manner. The growing need for human-computer interaction has made automated stress recognition a rapidly appealing area of study. Researchers have developed multiple methods for evaluating physiological data from sensors affixed to human bodies to identify stress and classify emotions. Only a small number of research, nonetheless, have focused on examining social media postings to assess the stress levels of internet users. We have attempted to identify patterns of online user behavior on major social media platforms, or even employ social media logs for additional research. Our ML models aim to deliver precise and trustworthy findings in the early identification of stress experienced by internet users. In this article, we outline a technique for identifying psychological discomfort from textual data found online. The goal of this project is to create machine learning and NLP methods that can identify anxiety while also examining the subject of conversation in a particular social media message to look into further mental illnesses. This apparatus, along with sentiment analysis, would be helpful for analyzing and classifying the opinions of customers on other unique subjects. This enables us to get important information about the different relationships linking social engagement to fear or dread experienced by internet users. relationships between social interactions and the anxiety or dread experienced by internet users. The following highlight the key contributions of the paper: • Information is extracted from the comment using natural language processing techniques; • Papers on particular themes are analyzed and separated; and • Finally, a suitable strategy for the most accurate detection of stress is chosen. Five distinct sections make up the remainder of the paper. Rerword Similar tasks and methods are covered under Section 2 in the literature review, current research is covered in Section 3, and findings and discussion are covered in Section 4. In the end, we brought the paper to a close by talking about the potential advancements described in Section 5 of the project.

II. LITERATURE SURVEY

Social media has expanded tremendously over the past 10 years, bringing both benefits and drawbacks. Direct communication across cultural and economic barriers is now possible because of the rapid rise of internet and social media networking. While social media offers many benefits [6, 7], it also presents challenges that can negatively influence society.

In the research community, hateful, provocative, and stress-related comments are not uncommon. Even though many studies exist in the field, stress identification is not a newly introduced research topic [8, 9, 5, 10, 7, 7]. Lately, the presence of hate speech and digital hostility has increased significantly [12, 13, 15, 15]. Using offensive and disparaging language through online platforms typically qualifies as hate speech. This might be a reference towards individuals or a group sharing a common goal. In this post, we present our approach to combat harmful language and have drastically curbed it. At present, many individuals use social media to vent their fury and indignation, which is detrimental to other people's feelings. It might have hurt their caste, creed, privacy, faith, and ethnicity, and could potentially be deeply damaging to them. Some comments may be deemed hate speech because of their obscenity, even if they are not intended to offend anyone. In order to determine the most accurate machine learning model to use based on accuracy, the authors [16] first thoroughly examined natural language processing before evaluating a range of machine learning techniques. It is essential that we connect and retrieve facts in everyday scenarios. One of the most popular internet data sources is database systems. The data volume continues to increase, and modern database tools are advancing and having a major influence. Databases are used by almost all online applications for saving and accessing data. The main goal is to provide an easier and more effective way to handle database queries and generate results. Connecting with a diverse audience is made easy by social platforms. These have increased the exposure of the public to atmospheric radiation [17]. Language data obtained through NLP and the results from this media are used by researchers at the Measurement Data Center (ADC). Natural Language Processing (NLP). Depression is the leading cause of impairment and a significant risk factor for suicide. In recent years, social platforms have become primary online information exchange channels. Most people share their ideas, hypotheses, as well as personal thoughts through digital media. The language individuals use on these platforms demonstrates how depression affects communication. As awareness of mental well-being has grown, it has become crucial to understand mental disorders. In the study, the scientists searched Twitter users' tweets to find signs of depression [18]. They used NLP and text analysis methods to accomplish this task. Their CNN and LSTM-based approach allowed them to achieve a 92 percent accuracy rate. Additionally, they examined their model in comparison to logistic regression classifiers. TF-IDF. Some of the latest developments have been compiled here ([19, 20, 21, 22, 23, 24, 9, 9]). Table 1 lists 251 contributors with prior work in stress detection and compares our current model with their assessments of the literature, methods, and tasks. The use of physiological indicators to identify weariness has been the subject of much research. Analyses of physiological data from electrocardiograms, skin conductance responses, and muscle activity signals were conducted in nearly all earlier research. These approaches relied on traditional machine learning algorithms to examine physiological signals in order to identify stress.

Table 1
Comparison table for recently published paper with proposed system based on Stress detection

Source	Measurement	Techniques	Task
Subhani et al. [19]	T test, Distance	Logistic regression, Support vector machine, Naïve bayes	Mental stress detection
Elzeiny and Qaraqe [20]	Electroencephalogram, Hibert-Huanf Transform	K-Nearest Neighbor	Workplace stress detection
Papini et al. [21]	Medical and demographic features	Logistic regression	Posttraumatic stress detection
Jadhav et al. [22]	Textual data, facial expression	Bidirectional Long Short-Term Memory	Text based stress detection from social media.
Dubey et al. [23]	Assisted reproductive technology	Support vector machine	Human spermatozoa detection under oxidativestress
Jebelli et al. [24]	Electroencephalogram	Online Multi-Task Learning (OMTL) algorithms	Stress recognition framework
Das et al. [9]	Electroencephalogram	Backpropagation Neural Network	Cognitive load detection
Zhang et al. [25]	Magnetoencephalography, Electroencephalogram	Support vector machine	Posttraumatic stress detection
Yousefi et al. [26]	Pupildiameter, electrodermal activity	Linear discriminant analysis	Stress detection using eye tracking dataset
Our proposed system	Textual data	Natural language processing with Support vector machine.	Text based stress detection from Twitter.

III. METHODOLOGY

We go into further detail about the dataset collecting and annotation processes in this section. The procedure of gathering the dataset is covered in the first subsection, and the suggested method is explored in the second.

Dataset Collection and Labelling:

The Tweepy API from Twitter, a popular platform for online interaction, was utilized to get our dataset. Specific were identified based on their potential vulnerability to stress, and tweets were first filtered using keywords. For our case study, we then collected tweets from the previous three months. A total of 2978 tweets were chosen for examination after 58 users were identified. The responses of people under stress are illustrated by the instances shown in Figure 1. Similarly, important tweets from our dataset

pertaining to stress are displayed in Figure 2. Physical stress is clearly visible, as shown in Figures 1 and 2. It is evident but underlined that the output result is given in English and retains a standard tone of voice. The work is more difficult because online consumers are typically more implicit. Then, with the help of our staff, we manually classified the tweets as either high-stress or low-stress. of two postgraduate students and five undergraduates. The tweets were independently assigned by each of the seven students. Lastly, we used the majority voting approach to consider the final label. Figure 3 displays the detailed distribution of the various stress class data.



Figure 1: Example of physical images of stress

IV. PROPOSED APPROACH

In this part, we explain how we carried out our approach step by step for stress detection, which combines techniques from both ML and NLP. Text mining plays a critical role in the analysis and processing of non-structured information, which accounts for over 80% of global data. The majority of institutions and organizations currently gather and store vast amounts of data in cloud platforms and data warehouses, and as new data comes in from all directions, this data is growing exponentially every minute. Our suggested study, which uses NLP and ML-based methods to determine stress levels on social media platforms, is illustrated in figure 4. Our framework is divided into four main phases: the first is for gathering social media datasets; the second is for automatically summarizing all posts from a single user, demonstrated in figure 5; the third is for text mining depicted in figure 6; and the fourth is for stress detection, illustrated in figure 7.



Figure 2: Sample Tweets Reflecting Signs of Stress

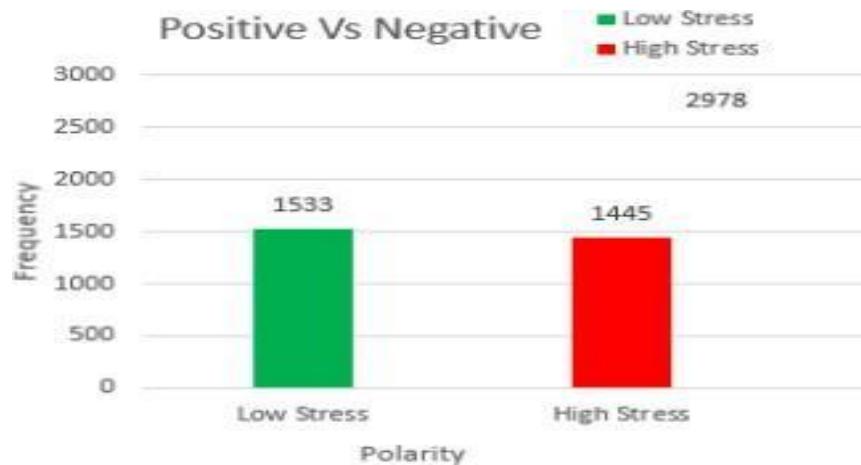


Figure 3: Distribution of stress data Overview of How Stress Levels Are Spread Across the Data

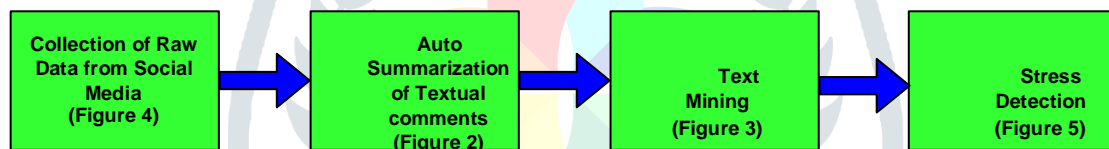


Figure 4: A Glimpse into Our Suggested Approach

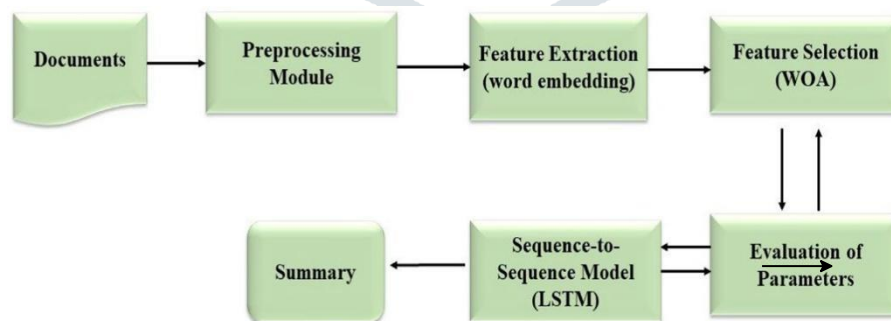


Figure 5: Automatically Generated Summary of User Comments

AUTOSUMMARIZATION:

As illustrated in figure 5, automatic summarization is the process of employing computer techniques to reduce a large dataset into a more manageable portion successfully conveys the key concepts or details included within the original content. All of the comments that the same person has left on the online platform are collected in this section. Not every comment is equally significant for stress detection, according to the core notion of the summary. The technique presented here creates segments from main text using k-means clustering. The length the supplied text determines how many clusters there are. On a larger scale, though, having too few clusters might be troublesome because they can totally change the parent text's intended meaning. A comparable amount of clusters might suggest that the resulting text is too detailed, which runs counter to the summary goal.

TEXT MINING:

We applied text mining (refer to Figure 6) to pull relevant insights from the users comments. The text was initially broken down tokens. Then, using normalization as a strategy, we employed lemmatization and stemming approaches to improve our comprehension of the base form of each token. Furthermore, we have developed a distinct feature set that highlights terms strongly associated with stress by using entity recognition to indentify and extract them. NER is the most commonly used data pre-processing activity. It involves identifying key features within the content and organizing them according to a preset set of criteria. In a book, an entity is a recurring concept or topic that is frequently mentioned or referred to. Comments with were distinguished from highly stressed ones using the data fed into the machine learning classifier. A component of artificial intelligence known NLP converts unstructured text from databases and raw documents into well-structured formats that can be examined or fed into machine learning

algorithms. The proposed text mining structure is illustrated in Figure 6. Among the key approaches to managing nearly 80% of the world's data text mining—an essential technique used to interpret and processing unstructured data. After gathering the required data, we have finally given them the labels of low stress and high stress. Various classifiers have been tested using this information. In this instance, we have set aside 20% of the samples for testing and 80% for training. Count-vectorizer along with term frequency-inverse document frequency (TF-IDF) vectorizers are the two approaches we have experimented with to transform our text input into a format that is appropriate for our classifier. We opted for the count-vector for the latter stage after learning that tf-idf performed remarkably well. For the implementation, our work relied on scikit-learn2 package and the nltk library1.

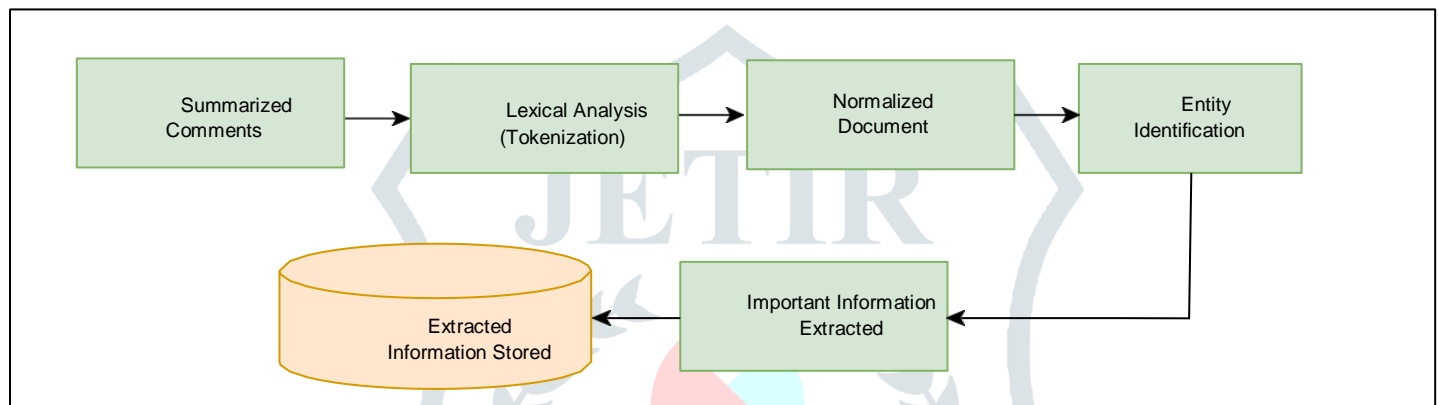


Figure 6: Proposed Text Mining Model

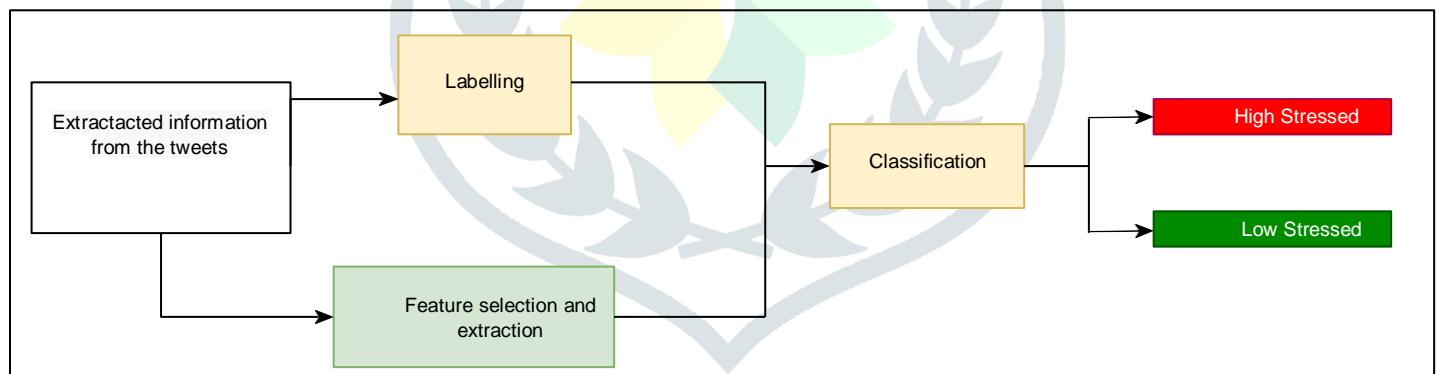


Figure 7: Suggested Approach for Text Detection

V. RESULT AND DISCUSSION

We discussed our observations and experimental outcomeson highlighting experimentation in in this part of the study,while also recognizing its limitations. This research offers automated approach for detecting stress in individuals through analysis of social media content collected through the Twitter API, applying natural language processing, and implementing various machine learning models to help prevent individuals from experiencing stress-related health issues. Additionally, we experimented with basic preprocessing techniques such as deleting stop words, deleting URLs, lowering the capitalization, etc. We have specifically worked with five classification algorithms: logistic regression, naïve Bayes, decision tree, random forest, and support vector machines (SVM). To get the input ready for the classifier, we used a TF-IDF (term frequency–inverse document frequency) vectorizer. Metrics such as accuracy, precision, recall, and f1-score as the performance indicators to evaluate the efficacy of our system. Classifier performance is depicted in Table 2, showing that both random forest and support vector machine classifiers achieved higher accuracy compared to the others.

Table 2
Performance of Proposed Framework

Classifier	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	0.90	0.94	0.90	0.92
Logistic Regression	0.89	0.92	0.88	0.89
Naïve Bayes	0.82	0.86	0.83	0.84
Decision Tree	0.78	0.81	0.76	0.79
Random Forest	0.91	0.93	0.87	0.90

Because there aren't enough examples, we haven't attempted use of neural network-based methods. Additionally, when selecting the tweets for the dataset, we did not take into account those that contained words in languages other than English or had various data modalities.

VI. CONCLUSION AND FUTURE SCOPE

It is not uncommon for individuals to experience stress on a regular basis. However, extreme stress or chronic stress might interfere with our daily routines and endanger our health. Early identification of mental stress can help prevent various stress-related health issues. By social media posts shared during stressful events, this study seeks to improve the accuracy of stress identification. It is possible to use prediction models that use language on social media to identify individual anxiety and depression illnesses. It might improve on traditional screening techniques. Predictive ML techniques could allow diagnose symptoms early, perhaps before they worsen and have more significant psycho-social repercussions. Although we only used a small amount of data for our trials in this study, we believe this method could be extended to handle bigger data sets, which will help us better understand the stress that internet users face.

The ability to experiment with posts in many languages and formats, such as photos, memes, music, and video—all of which are quite prevalent in India—as well as social media elements is another potential area of future development for our work.

REFERENCES

- [1] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, *IEEE Transactions on Affective Computing* 13 (2022) 440–460. doi:10.1109/TAFFC.2019.2927337.
- [2] U.Naseem, I. Razzak, M. Khushi, P. W.Eklund, J. Kim, Covidsentiment: A large-scale benchmark twitter data set for covid-19 sentiment analysis, *IEEE Transactions on Computational Social Systems* 8 (2021) 1003–1015.
- [3] P. Gupta, S. Kumar, R. R. Suman, V. Kumar, Sentiment analysis of lockdown in india during covid-19: A case study on twitter, *IEEE Transactions on Computational Social Systems* 8 (2020) 992–1002.
- [4] S.Greene, H.Thapliyal, A.Caban-Holt, A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health, *IEEE Consumer Electronics Magazine* 5 (2016) 44–56.
- [5] Y. S. Can, B. Arnrich, C. Ersoy, Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, *Journal of biomedical informatics* 92 (2019) 103139.
- [6] S. Dosani, C. Harding, S. Wilson, Online groups and patient forums, *Current Psychiatry Reports* 16 (2014) 1–6.
- [7] K. Kumari, S. Srivastav, R. R. Suman, Bias, threat and aggression identification using machine learning techniques on multilingual comments, in: *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 30–36. URL: <https://aclanthology.org/2022.trac-1.4>.
- [8] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, M. Griss, Activity-aware mental stress detection using physiological sensors, in: *International conference on Mobile computing, applications, and services*, Springer, 2010, pp. 282–301.
- [9] S. Das, L. Ghosh, S. Saha, Analyzing gaming effects on cognitive load using artificial intelligent tools, in: *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, 2020, pp. 1–6.
- [10] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ hasoc 2020: Bert models for hate speech identification in indo-european languages., in: *FIRE (Working Notes)*, 2020, pp. 319–324.
- [11] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ trac-2: deep learning approach for multi-lingual aggression identification, in: *Proceedings of the second workshop on trolling, aggression and cyberbullying*, 2020, pp. 113–119.
- [12] K. Kumari, J. P. Singh, AI_ML_NIT_Patna @ TRAC- 2: Deep learning approach for multi lingual aggression identification, in: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, European Language Resources Association (ELRA), Mar seille, France, 2020, pp. 113–119. URL: <https://aclanthology.org/2020.trac-1.18>.
- [13] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: <https://aclanthology.org/2020.trac-1.1>.
- [14] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, *ACM Computing Surveys (CSUR)* 51 (2018) 1–30.

- [15] K. Kumari, J. P. Singh, Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content., FIRE (working notes) 2517 (2019) 328–335.
- [16] R. Devarakonda, M. Giansiracusa, J. Kumar, H. Shanafield, Social media based npl system to find and retrieve arm data: Concept paper, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 4736–4737.
- [17] S. Chiramel, D. Logofătu, G. Goldenthal, Detection of social media platform insults using natural language processing and comparative study of machine learning algorithms, in: 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), IEEE, 2020, pp. 98–101.
- [18] M. Häberle, M. Werner, X. X. Zhu, Building type classification from social media texts via geo-spatial textmining, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2019, pp. 10047–10050.
- [19] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel, A. S. Malik, Machine learning framework for the detection of mental stress at multiple levels, IEEE Access 5 (2017) 13545–13556.
- [20] S. Elzeiny, M. Qaraq, Blueprint to workplace stress detection approaches, in: 2018 International Conference on Computer and Applications (ICCA), IEEE, 2018, pp. 407–412.
- [21] S. Papini, D. Pisner, J. Shumake, M. B. Powers, C. G. Beevers, E. E. Rainey, J. A. Smits, A. M. Warren, Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization, Journal of anxiety disorders 60 (2018) 35–42.
- [22] S. Jadhav, A. Machale, P. Mharnur, P. Munot, S. Math, Text based stress detection tech niques analysis using social media, in: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2019, pp. 1–5.
- [23] V. Dubey, D. Popova, A. Ahmad, G. Acharya, P. Basnet, D. S. Mehta, B. S. Ahluwalia, Partially spatially coherent digital holographic microscopy and machine learning for quantitative analysis of human spermatozoa under oxidative stress condition, Scientific reports 9 (2019) 1–10.
- [24] H. Jebelli, M. M. Khalili, S. Lee, A continuously updated, computationally efficient stress recognition framework using electroencephalogram (eeg) by applying online multitask learning algorithms (omtl), IEEE journal of biomedical and health informatics 23 (2018) 1928–1939.
- [25] J. Zhang, J. D. Richardson, B. T. Dunkley, Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning, Scientific reports 10 (2020) 1–10.
- [26] M. S. Yousefi, F. Reisi, M. R. Daliri, V. Shalchyan, Stress detection using eye tracking data: An evaluation of full parameters, IEEE Access (2022).