Detection of Emotions Using a Boosted Machine Learning Approach

Y Pragnya
Information Technology
Velammal Institute of Technology
Chennai, India
pragnya1132@gmail.com

N Jagadish Kumar
Department of Information Technology
Velammal Institute of Technology
Chennai, India
jagadishiva@gmail.com

S M Udhaya Sankar
Department of Information Technology
Velammal Institute of Technology
Chennai, India
udhaya3@gmail.com

Allen Manoj
Computer Science
Sri Venkateswara College Of Engineering
Chennai, India
allenmanoj17@gmail.com

Goloti Pavani Achary
Information Technology
Velammal Institute of Technology
Chennai, India
pawanishyam1234@gmail.com

2022 1st International Conference on Computational Science and Technology (ICCST) | 978-1-6654-7655-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCST55948.2022.10040393

***Abstract*—Adolescents' use of social media is unavoidable. Students with greater conscience exhibit more caution when using social media, according to one study. Students said they spend 28 hours per week on social media platform on average. Eyestrain, neck and arm pain, lethargy, poor posture, and a lack in physical activity have all been spurred through the use of social media. It is also being connected to various mental health complications such as suicide, anxiety, loneliness and lack of empathy. Parents, society, and intellectuals are concerned about the explosion in youth social networking use in the modern era because every idea has two opposing ends .This study's objective is to investigate at into the ramification of social media on cognitive health. The next step in uncovering the relationships between these issues in the adolescent generation is to examine and understanding how technology affects the mental wellbeing of the adolescent generation today. Extensive tests on real-world datasets show that our suggested approach on Twitter sentiment analysis shows an accuracy ranging between 88 and 92 percent.**

***Index Terms*—Sentiment Analysis, Social Networks, Cognitive Health, Classification, Boosting.**

I. Introduction

Nowadays social networking has become a salient part of one's life. According to research, users of social media platforms range from approximately 70percent of total of the top half and geriatric to heights of 98.5 percent among individuals below 30. These individuals may have a variety of cognitive ailments or other psychiatric illness. Recent research has looked into social networking sites patterns, the power of social media on cognitive well being, and the prospects for strengthening intervention delivery by leveraging social media's popularity and additional content. Young people commonly report using social media sites like Facebook, Instagram, and Twitter to escape the pressures that endanger their mental health. Depression has become a severe issue that has a negative impact on mental health [1]. The information or points offered by the user are indeed the facts that is helpful for the study to assess the state of mind [2]. Studies have shown that use of beneficial social media like authentic self presentation has led to positive well-being in users [3].

Data science approaches are becoming more common due to the availability of powerful computers and so-called big data from a number of sources. Data-driven approaches have become increasingly relevant for addressing critical scientific concerns. Data science is a concept that combines statistics, data analysis, related methods, and data methods. Data science is already helping several fields of medicine, such as the prevention of heart disease and the treatment of specific cancers. Many researchers have used In order to assess underlying information and expertise on psychological disorders in the huge user-generated content shared on social media, machine learning, a big data engineering approach, is used.

With a concentration on data science for semi healthcare research and ingenuity, task wanting to share for working population capacity building, and early intervention for children with limited resources ecosystems, real studies and initial reports on tearing implementations that haven't been widely tested or put it into practice are being authored. Investigators have become able to discern previously undiscovered patterns or trends in human behaviour and disease trajectories thanks to sophisticated data interpretation, sizable computing power, and access to large amounts of data via latest technological references including internet sites, media platforms, internet search habits, and handheld platforms. Countless investigations and practical methods are showcasing the validity and hope of leveraging big information gathered from the internet to un- cover attitudes and behaviors [4]. To address its limitations, firstly, as the retrieval process depends on data accessibility, it is susceptible to changes in access conditions. Second, it is impractical to manually verify large datasets due to the excess of data as bigger data is not always better data. However, these techniques can also be used to learn more about mental illnesses like post-traumatic stress disorder and depression.

Below is a summary of this paper's contributions. We create a machine learning model to recognise emotions expressed on Twitter and other social media platforms in general. Microblogging and websites for social media altogether. To verify the effectiveness of the given classifier functioned, 1500users were interviewed on digital networking. [5]. We pre- processed big real datasets before using algorithms, and the results provide intriguing insights on the emotions identified in Twitter data analysis that social scientists and psychologists may find interesting. The method for acquiring, analysing, and interpreting social media data presented in this study is novel, thorough, and descriptive. The innovative aspect of this research is the approach suggested for dealing with the opportunities, constraints, and challenges offered by the activities of data retrieval, validation, classification, and filtering.

II. Related Work

According to research, students mostly use social media platforms space is taken away from studying as it is used more for connecting than for academic purposes. Also, it was shown that when using online platforms had a poor link with adolescents' educational excellence; this correlate was considerably reliable than the perks through using social media channels [7]. Another analysis found that the use of online communities caused people to redirect their attention and focus from their academic tasks and toward leisure activities. Social media has indeed been demonstrated to negatively disrupt students' attention, memory, sleep, vision in prior findings.

In accordance with a current data, people did not often breakfast on time or get adequate rest, and that excessive consumption of social media channels had a serious effect on overall physical and emotional well-being. Whereas social media can be a beneficial medium of communication, it also carries a risk to anyone's wellbeing [8]. Social sites does have upsides and downsides, however the way a human utilises these networks determine the impact. Yet, those who had higher levels of self-control were able to manage their social networking use. A recent study found that although students were capable of using social media networks for career purposes, they missed the motivation or willingness to do so.

Depression is a psychological health issue that has become a hot topic for conversation in everyday health problems Melancholy is primarily brought on by a complicated interaction of cultural, environmental, and psychological point of view in its early stages [9]. Depression can occur as a result of major and complex issues . Psychomotor agitation oscillations (fromexcitement to rock bottom) are a symptom of manic depression [10]. Shatte et al. conducted an assessment of 300 papers focusing on machine learning and Data mining applications in mental wellbeing and showed most of these works tackled dementia, Stroke, and sadness as their primary psychiatric illnesses[11]. Also, supervised methodologies were applied in 89 percent of the articles examined to research the specified diseases. In order to forecast mental health status using digital platforms, Chancellor also did a topic textual analysis on research journals[12]. They noticed that heterogeneous academics possess multiple viewpoints on user datasets, which they named "Human-Centered Machine Learning" .

From depression cognitive theory, depressed people have a negative perception of themselves and their surroundings. Online domains like Facebook and Twitter have established a manufacturer platform for trimming analysis by supplying a range of information and contextual factors to analyse user behavior [13]. Linguistic methods and a range of classification techniques were applied to analyze literary data and looked towards how online communities affect individuals' general wellbeing [14]. Twitter was one of the most known social networking services, including around 450 million active users and 150 million tweets broadcasted to a mass audience a day every day. The epidemic of grief, despair, as well as other neurological disorders among People on twitter has already been studied by a few investigators employing Twitter statistics.

Preotiuc-Pietroi et al used Twitter data to investigate the personality of individuals claiming to have post-traumatic stress leveraging wide textual metrics we may diagnose disorders. In compared to depressed persons, his findings demonstrate that people affected were both older and more conscientious [15]. The creators arrive at the conclusion that clients with specific personality or characteristics are more likely to reveal their mental health diagnosis on social media, and that the consequences may not apply to other types of autobiographical text also because linguistic accurate predictions of depression repeated significantly with the linguistic best predictors of character[16].

III. Proposed Work

A. Boosting

Boosting is an ensemble modeling technique for creating a powerful classifier out of many Feeble ones. It's done by putting together a model from an avalanche of basic models. on starting with, the training data is used to develop a model. The second model is then developed, with the goal of correcting the original model's flaws. This process is continuously repeated until the whole data set is correctly approximated or a greater number of models are added as mentioned in fig 1. Extreme Gradient Boosting (XgBoost) is a technique created by researchers at the University of Washington.

B. Adaboost-Classifier

It is constructed by combining a multitude of algorithms to achieve high accuracy. This method is employed to develop an adaptive ensemble. The AdaBoost algorithm utilizes a number of underperforming classifiers to create a powerful, reasonably precise classifier. The core premise behind AdaBoost is to learn the sample data and validate the classifier weights for each iteration in order to generate an accurate predictions for unusual observations. As a core classifier, any algorithm for machine learning that derives weights from the training set is

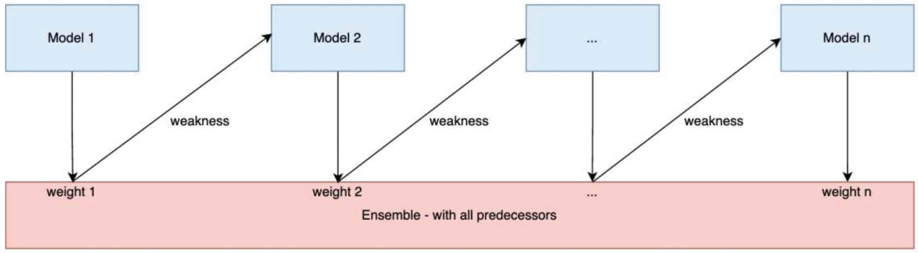

Fig. 1. AdaBoost

suitable.AdaBoost must meet two requirements as mentioned in fig 1.

1) The classifier algorithm must be constantly and interactively trained using various weighted training cases
2) It seeks to exhibit an extraordinary fit for the instances in each and every iteration by reducing training faults.

1)Formulae: B(a) = sign[$\sum_{i=0}^{i} \alpha_i \ B_i \ (a)$]

Algorithm 1 AdaBoost-Classifier

a) Compute the weights w_m =1/n,m=1,2,….,n
b) While $i = 1$ to I:
   i) Fit B_I(a) to the dataset using weights w_j
   ii) Calculate
   $$\text{err}_i = \frac{\sum_{c=1}^{n} w_m \ I(y_m \ \neq B_M \ (x_m \ ))}{\sum_{m=1}^{n} w_m}$$
   iii) Calculate $\alpha_i = \log((1 - \text{err}_p)/\text{err}_p)$.
   iv) Set $w_j \leftarrow w_j \cdot \exp[\alpha_i \cdot I(y_k \neq G_d(x_v))]$, $v = 1, 2, \ldots, n$.
c)Result $B(a)$.= sign[$\sum_{i=1}^{i} \alpha_i \ B_i \ (a)$]

---

7) XGBoost

The XGBoost algorithm was constructed as part of a research project in Washington. In 2016, Tianqi Chen and Carlos Guestrin published their work in the SIGKDD Conference, which kindled the Machine Learning fabrication. The following are a number of the ways during which the algorithm distinguishes itself: It can be used to tackle regression, classification, ranking, and user-defined prediction issues, among other things. Portability: Works on Windows, Linux, and OS X without a problem. All major programming languages are supported, including C++, Python, R, Java, Scala, and Julia and it supports Amazon Web Services various otherenvironment.

8) Naive Bayes - Classifier

It is a categorization technique that is based on the single predictor hypothesis and Bayes' principle as shown in fig2.

Algorithm 2 XGBoost

1) for t=1:K do
2) Initialize $G_0(X_i) = argmin_p = \sum_{i=1}^{N} \ l(y_i, p)$
3) Compute $\nabla G_t \ (X)$
4) Run a new learner function g(X,μ)
5) Predict the best gradient descent stage size $p_k = argmin_p$
6) Output the prediction probability

Based on a Bayesian Classifier, the availability of one character in a class is irrelevant to the occurrence of the second attribute. The Naive Bayes model is quick to develop and is excellent at processing massive amounts of data. Due to its transparency, Naive Inference is believed to surpass even complex classification schemes.

9) Decision Tree

Decision trees are frequently utilized for the purpose of classification also known as regression problems as shown in fig3. It uses a flow-diagram sort of a tree structure to pinpoint the forecast that results from a series of feature-based splits. It begins with the main node called the main parent node and ends with a choice made by leaves. The upside-down orientation of decision trees implies that the root is at the top before being divided into several nodes. In simple terms, decision trees are a collection of fuzzy rule statements. It decides whether the condition is met, and if it is, it goes onto the subsequent node that is tied to the decision.

I. EXPERIMENTAL SETUP

Emotions are subconsciously formed and characterize bodily states. They are basically self-physiologic responses to external or internal stimuli. Feelings, on either hand, are subjective emotional experiences that are driven by conscious thoughts and reflections. There are various kinds of emotions expressed by us every day. The four basic emotions anger, fear, sadness, and happiness are deferentially associated with three core effects: reward, punishment, and stress. This experiment

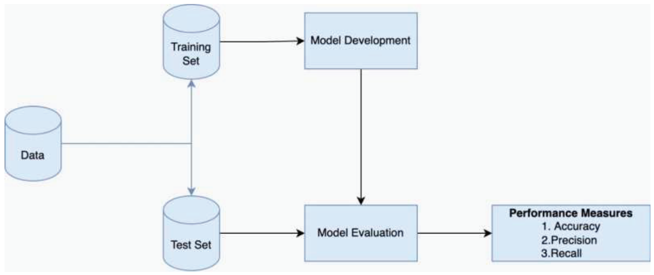

Fig. 2. Na¨ıve Bayes

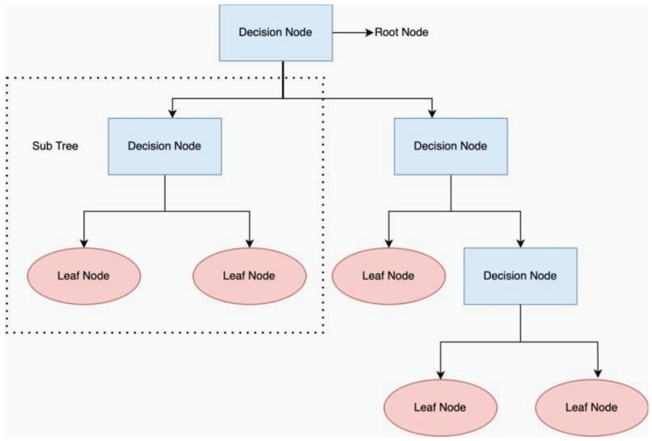

Fig. 3. Decision Tree

detects various conscience displayed in the text that is commented on social media. In these algorithms, XGBoost and AdaBoost are used to analyze the comments and state their emotions. They are categorized as' 0 'for negative emotions and' 1 'for positive emotions. We have taken the primary data set as the sentimental data set from Twitter on which the experiments and graphs are obtained as the result which is explained as follows:

A. Actual Value

TABLE I
TABLE TYPE STYLES

2*Confusion Matrix		Predicted value	
		Positive	Negative
2*True Label	Positive	230(TP)	22(TN)
	Negative	13(FP)	235(FN)

True Positive denoted as TP is the count of patients correctly classified as Depressed. True Negative denoted as TN is the count of patients correctly classified as Not Depressed. False Positive denoted as FP is the count of Not Depressed patients classified as Depressed. False Negative denoted as FN is the count of depressed patients classified as Not Depressed.

1) Preprocessing: According to the psychiatrist's judgment, only 48 of the 1025 indicators recorded during the assessment were used. Some of the needed parameters were not assessed

	Id	Sentiment	Day	SentimentText	Year	Gender
0	1792642663	0	Mon	is so sad for my APL friend............	2018	Male
1	1467810672	0	Tue	I missed the New Moon trailer...	2018	Male
2	1467810917	1	Wed	omg its already 7:30 :O	2018	Male
3	1467811184	0	Thu	.. Omgaga. Im sooo im gunna CRy. I've been at ...	2018	Male
4	1467811193	0	Fri	i think mi bf is cheating on me!!! T_T	2018	Male
5	1467811372	0	Sat	or i just worry too much?	2018	Female
6	1467811592	1	Sun	Juuuuuuuuuuuuuuuussssst Chillin!!	2018	Female
7	1467811594	0	Mon	Sunny Again Work Tomorrow :-| TV Tonight	2018	Female
8	1467811795	1	Tue	handed in my uniform today . i miss you already	2018	Female
9	1467812025	1	Wed	hmmmm.... i wonder how she my number @-)	2018	Female

Fig. 4. Dataset

in the cognition, thus ones with no values or identical values were removed, leaving 1500 records.

2) Sampling: To deal with the class imbalance, the data is categorized into train data and test data with a split ratio of 60:40 using the stratified sampling approach.

3) Selection of Features: Due to the enormous number of features in the data set, feature selection is performed to find the optimal set of features based on significance of features. At each round of boosting, the best feature is detected and added to the ensemble.

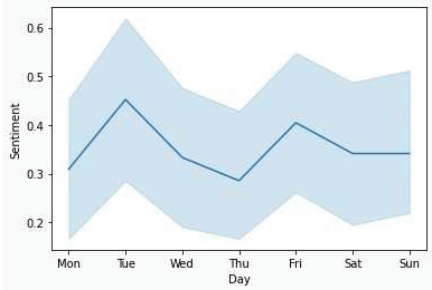

Fig. 5. Graph 1

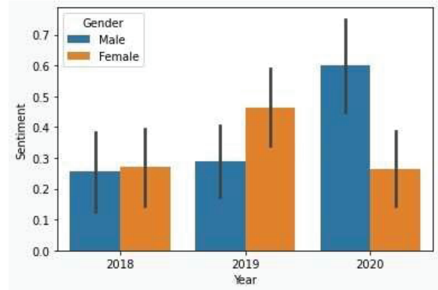

Fig. 6. Graph 2

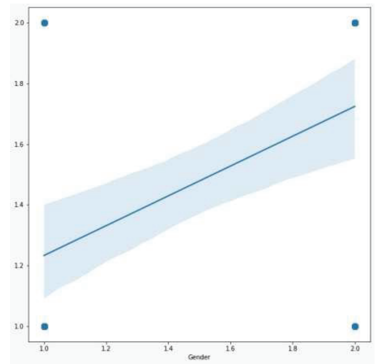

Fig. 7. Graph 3

B. Dataset Description

Our project's goal is to determine people's thoughts and mental health on a daily basis this data is shown in fig4. The below fig7 Dataset, which was taken from Twitter, shows how comments over the previous two years have changed. The tweets of various individuals are included in this data set. Each tweet contains a unique identification number that identifies the author. Additionally, it includes the day on which he or she tweeted. The feeling and the sentimental text are also provided. The emotion is represented by the numerals "1" and "0." Positive emotions are represented by "1," whilst negative emotions are represented by "0"x. This is the foundation for the analysis. The year in which the tweet was posted is also included to show how recent the tweet is. The gender is also given in order to assess the particular gender's mental health. Various algorithms, including Naive Bayes, Decision trees, XgBoost, and AdaBoost, have been employed to determine the implications. Plotted on numerous graphs, the experiment's correctness has been further examined for our purpose.

C. Adaboost

We have taken the primary data set as the sentimental data set from Twitter on which the experiments and conducted. The accuracy of the model data set is 92 percent. Various graphs are obtained as the result which is explained as follows:

1) Graph 1: On the x-abscissas, the days are plotted. On y-abscissas, the sentiment rate is plotted. Based on the rateof tweets and the emotions predicted in the above graph, it was plotted. On analyzing the graph, it can be seen that on Tuesday there are a large number of tweets and the emotion rate is very high. The average between both the positive and negative tweets is taken, and the graph is plotted according to the algorithm. If the graph shows a rise, then it denotes that the positive emotion is more than the negative, and if there is a fall in the graph, it states that there is negative emotion more than positive. Hence, this is concluded from the experimental observation.

2) Graph 2: A bar chart is said to be a diagrammatic representation in which it uses rectangular bars with heights that are proportional to the values which they represent. On the x-abscissas, the years are plotted. On the y-abscissas, the sentiment rate is plotted. Based on the rate of tweets and the emotions predicted by the above bar graph, it is plotted. In this, blue color denotes male emotions and orange color denotes female emotions. As a result, the emotion rate is calculated and plotted every year. In the years 2018 and 2019, female emotions were higher than those of males. In the year 2020, male emotions were higher than those of females.

D. Xgboost

We have taken the primary data set as the sentimental data set from Twitter on which the experiments and conducted. The accuracy of the model data set is 88 percent. Various graphs are obtained as the result which is explained as follows:

1) Graph 3: The above graph depicts gender on the x-axis, where '1.0' denotes male and '2.0' denotes female. On the y-axis, emotions from the data set are plotted by using the XGBoost algorithm. From the above graph, it can clearly be analyzed that in this particular year, female emotions were higher than those of males.

E. Naïve Bayes

We have taken the primary data set as the sentimental data set from Twitter on which the experiments and conducted. The accuracy of the model data set is 84 percent. Various graphs are obtained as the result which is explained as follows:

1) Graph 4: On the x-axis, sentiment is denoted where '0.0' denotes negative comments and '1.0' denotes positive comments. On the y-axis, the years from the data set are plotted by using the Naive Bayes algorithm. From the above

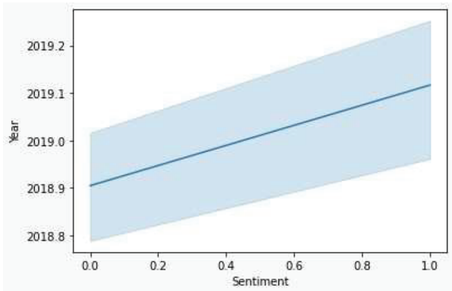

Fig. 8. Graph 4

graph, it is clearly depicted that in the year 2019, positive emotions are higher than negative emotions as compared to the previous year.

F. Decision Tree

We have taken the primary data set as the sentimental data set from Twitter on which the experiments and conducted. The accuracy of the model data set is 86 percent.

IV. Result

Our approach of detecting emotion using boosted machine learning has an average accuracy of 92 percent.

TABLE II
Table Type Styles

Algorithm	Accuracy
AdaBoost	92 %
XgBoost	88 %
Na¨ıve Bayes	84 %
Decision Tree	86 %

We employed a random split ratio of 60:40 of training and testing samples for each trial during the experiment. The experiment was run on an Intel i3 processor with Python. The performance of the classifier is assessed using sensitivity and specificity values.

V. Conclusion

This research presents a novel way to detect mental health states using classification models. In comparison to other models, the Adaboost classifier method of training the data is quick. The method lowers misclassification errors, resulting in higher overall accuracy. We execute the model in parallel threads for the enormous dataset to improve performance. Implementing many other machine learning algorithms, such as deep learning and reinforcement learning, can improve computing accuracy. In the field of biomedical informatics, an efficient predictive model is required to forecast illness risk with high accuracy and minimal computing time. Extensive tests on real-world datasets show that our suggested approach using various classifiers and boosters on Twitter sentiment analysis tasks shows the highest accuracy of 92 percent.

References

[1] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," Heal. Inf. Sci. Syst., vol. 6, no. 1, pp. 1–12, 2018, 10.1007/s13755-018-0046-0.

[2] Grieve R, Watkinson J. The psychological benefits of being authentic on Facebook. Cyberpsychol Behav Soc Netw. 2016;19(7):420–5.

[3] Ilieva RT, McPhearson T. Social-media data for urban sustainability. Nature Sustainability. 2018;1(10):553– 65.

[4] Insel TR. Digital phenotyping: a global tool for psychiatry. World Psychiatry. 2018;17(3):276–7.

[5] D. Dhinakaran, D. A. Kumar, S. Dinesh, D. Selvaraj, and K. Srikanth, Recommendation System for Research Studies Based on GCR, 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, (2022) 61-65, doi: 10.1109/MECON53876.2022.9751920.

[6] Deen, S., Withers, A., Hellerstein, D. (2013). Mental health practitioners' use and attitudes regarding the Internet and social media. Journal of Psychiatric Practice, 19, 454–463.

[7] Livingstone, S., Palmer, T. (2012). Identifying vulnerable children online and what strategies can help them. London: UK Safer Internet Centre.

[8] D. Preotiuc-Pietro, J. C. Eichstaedt, G. J. Park, M. Sap, L. Smith, V. To- bolsky, H. A. Schwartz, and L. H. Ungar, "The role of personality, age, and gender in tweeting about mental illness," in CLPsych@HLT-NAACL, 2015.

[9] J. Zhao and X. Gui, "Comparison research on text pre-processing methods on Twitter sentiment analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.

[10] Udhaya Sankar S.M., Ganesan R., Jeevaa Katiravan, Ramakrishnan M., Ruhin Kouser R, Mobile application based speech and voice analysis for COVID-19 detection using computational audit techniques, International Journal of Pervasive Computing and Communications, Vol. ahead-of-print No. ahead-of-print. (2020), https://doi.org/10.1108/IJPCC-09-2020-0150.

[11] Dhinakaran, D., Prathap, P.M.J. Protection of data privacy from vulnerability using two-fish technique with Apriori algorithm in data mining. The Journal of Supercomputing (2022). https://doi.org/10.1007/s11227-022-04517-0.

[12] Udhaya Sankar, S.M., Christo, M.S., Uma Priyadarsini, P.S. Secure and Energy Concise Route Revamp Technique in Wireless Sensor Networks, Intelligent Automation and Soft Computing, 35(2) (2023) 2337ñ2351.

[13] D. Dhinakaran, P.M. Joe Prathap, D. Selvaraj, D. Arul Kumar, and B. Murugeshwari, Mining Privacy-Preserving Association Rules based on Parallel Processing in Cloud Computing, International Journal of Engineering Trends and Technology, 70(30) (2022) 284-294. doi: 10.14445/22315381/IJETT-V70I3P232.

[14] D. Dhinakaran, and P. M. Joe Prathap, Preserving data confidentiality in association rule mining using data share allocator algorithm, Intelligent Automation Soft Computing, 33(3) (2022) 1877ñ1892. DOI:10.32604/iasc.2022.024509.

[15] Udhaya Sankar, S.M., Abinaya, R., Karunya Jeez, S., Sunju, R. Efficient Data Transmission Technique for Transmitting the Diagnosed Signals and Images in WBSN, 4th International Conference on Recent Trends in Computer Science and Technology, ICRTCST 2021 - Proceedings, (2022) 251ñ256.

[16] Sujithra, T. and Sumathi, Dr. M. and Ramakrishnan, M. and Sankar, S. M. Udhaya, Survey on Data Security in Cloud Environment (May 11, 2020). International Journal of Advanced Research in Engineering and Technology (IJARET), 11(4) (2020) 155- 166.