International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.064

Using Machine Learning to Analyse User Psychology in Social Media

MerajFarheen Ansari
Software Developer
Department of Information Technology
Universityof the Cumberlands
Willamsburg, Kentucky,
USA merajfarheenansari25@gmail.com

Vilis Pawar
Assistant Professor, Global Business
School and Research Centre,
Dr. D. Y. Patil Vidyapeeth,
Pune, India, pvilis@gmail.com

Venkata N Inukollu
Assistant Professor
Department of Computer Science
Purdue University Fortwayne,
Indiana USA
dr.v.inukollu@gmail.com

Manisha D Kitukale
Professor
Department of Pharmachemistry
P Wadhwani College of Pharmacy
Yavatmal, Maharashtra,
India kitukalemanisha5@gmail.com

JavangulaVamsinath
Assistant Professor
Computer Science and Engineering
VNRVignana Jyoti Institute of
Engineering And Technology,
Bachupally Hyderabad, Telangana
vamsi.img@gmail.com

Sathiya Priya S
Professor
Departmentof Electronics and
Communication Engineering,
Panimalar Engineering College,
Varadarajapuram, Tamil
Nadu sathiyapriya.anbunathan@gmail.com

***Abstract*—In this study, we attempt to identify the emotion levels, such as positive, negative, & neutral feelings, from postings and comments on social networking sites on depression. Social media sites like Facebook and Twitter are becoming effective for helping those in need who require extra care or attention in terms of mental support. They are also utilized for communication and network development among relationships. There are several depressive support groups on Facebook, and they are quite helpful in giving the sufferers mental assistance. In this study, we attempt to formalize the posts and comments on depression into a succinct lexical database and identify the emotion levels from each occurrence. The complete amount of work has been divided into two sections: sentiment analysis and the use of machine learning techniques to examine the capability of extracting sentiment from such a unique category of texts. To determine the sentiment levels, we used the Python textblob module and typical machine learning techniques on the linguistic characteristics. For each of the classifiers, we have calculated the precision, recall, F-measure, accuracy, and ROC values. Random Forest outperformed the other classifiers, successfully classifying 60.54% of the instances. We think that conducting sentiment analysis on a particular class of texts may inspire additional research into how natural language is understood.**

***Keywords*—Sentiment Levels, User Communication, Machine Learning Algorithms. Accuracy Detection Analysis**

I. INTRODUCTION

Heterogeneous efficiency for various demographic subgroups is a fundamental barrier to the practical application of mental health surveillance models [1-2]. The training data may not be adequately representative of the population, which may produce this behavior, or some groups may be more difficult to forecast with the same data. In-depth data collection and training regimens can be used to solve the first situation, which has been extensively researched in the machine learning literature [3-6]. The latter situation is frequently more nuanced and challenging to handle. The value of the models is reduced if these performance disparities are not acknowledged and addressed. Particularly, if historically underrepresented people do worse, it can exacerbate disparities already in place, such as the underdiagnosis of depression [7].

Social media refers to websites and applications that have been specifically designed to enable people to distribute little amounts of material quickly, effectively, and in real-time. It has altered the mode we use to travel to and the mode we complete an activity and the ability to exchange photographs, reviews, activities, etc. in real time [8–10]. Shops that employ social media as a crucial component of their advertising strategy typically experience a quantifiable cost. But the key to using social media well is to stop treating it like an extra accessory and instead treat it with the same consideration, appreciation, and interest as the rest ofr advertising and marketing activities [11]. Utilizing various channels to engage with customers and develop r brand, increase revenue, and increase website traffic is what social media marketing entails.

The investigators have spoken about huge data gathered from social media that has been extensively evaluated by research academics and employed as significant to crucial insight into human conduct. [12] discussed how big data, machine learning, and analytics algorithms may be used to monitor social media and identify consumers' perspectives on opulent hotels from beginning to end the new visual data analysis and spin into an enhanced managing brand strategy for comfort hospitality managers. The researchers of [13] gathered data from 8434 startup firms on Twitter and created features based on social media using a machine learning model to predict each firm's level of social media participation [14]. The study's findings indicate that deep learning provides the greatest forecast accuracy for engagement levels. It also indicates that the amount of company-generated tweets, retweets, and likes is what matters most in determining the efficacy of social media marketing practices [15]. Social media users are perceived as contributing to marketing material as a result of the increasing interest in social media and user-generated content on websites like Tube, Facebook, and LinkedIn.Using big data analytics, they examine client perceptions and attitudes towards social media in this article.Related Works

14-year-olds are probably at risk for depression as a strategy to deal with social media or interact with others [16]. This is a widespread fallacy even in developing nations; there are depressed and suicidal persons who for a variety of reasons avoid seeing psychologists [17]. Some people believe they will be laughed at and their standing in

DOI: rp-9788770040723.064

society would decline if they visit a psychologist. They ultimately turned to suicide as a way to deal with their sadness [18]. Numerous deaths due to despair have been reported throughout the globe, and numerous of them posted their final Facebook post. They all shared the same trait, which was that they were all profoundly depressed [19]. If they can recognize their depressive stage before they reach the crucial stage, they may prevent suicide.

It is used to evaluate people's emotions in a variety of circumstances [20-21]. People now utilize social media platforms to express their emotions in their native languages, such as English [22]. The text that individuals publish on social media platforms is what we want to analyze for the sentiment. Many individuals in the world experience depression for a variety of reasons, many people who become drug addicts, many people who are unable to eat, sleep, work, or engage in other activities, etc. Because many of them also commit suicide, many families have lost dear family members [23]. Many individuals can live healthier lives like regular people if it doesn't happen. Language is one way that people may communicate their emotions. Individuals often express their feelings by writing and speaking about how they are feeling daily on social media[24].

It is to identify depression using the brain's volumetric characteristics. There is a chance that characteristics from brain SMRI will multiply [25]. Diagnostic values and volumetric data were looked at. The results show how often it is to spot depression [26]. SMRI volumetric features of diagnosing depression are shown in the output findings. The classification accuracy of the function vector is evaluated using a variety of classifiers, including SVM, Ensemble Learning Encoder, and Components are identified. These are comparable in terms of memory, accuracy, & correctness. In contrast to Naive Bayes' accuracy rate of 89.5%, they achieved a 90% overall accuracy. A motion analysis method for disorganized and grammatically incorrect customer remarks made in Arabic slang was put out in this paper's current Arabic Slang Sentiment Words & Idioms Lexicon, [27-28]. The new language was painstakingly put together from websites for microblogging. Furthermore, to categorize opinions as pleased or dissatisfied, the SVM approach was used with SSWIL.

III. Proposed Methodology

Researchers have a suggested technique that, like any other natural language recognition system, incorporates data gathering, information pre-processing, analysis, extraction, and classification, the use of machine learning algorithms, emotion recognition, and evaluation. Anxiety writings are a specific group of textual data that we have selected for this study. We have explored communities on social media sites like Twitter and Facebook to get this kind of information. Nowadays, individuals use social media as a powerful tool for communication, but they are also using platforms to provide psychological or emotional support via constructive posting. Regarding their efforts, they recently established the "Depression Support Group," "Anxiety Awareness," "Anxiety & Depression Support UK & Ireland," "Cure r Depression," and "Essential Thrombocythemia Support Group" Facebook groups for anxiety. The items and opinions that are related to depression have been carefully divided. To confirm that all of the data we obtained are relevant to depression, we used a psychologist to verify the information we had gathered. Table 1 illustrates the statistical characteristics of our dataset.

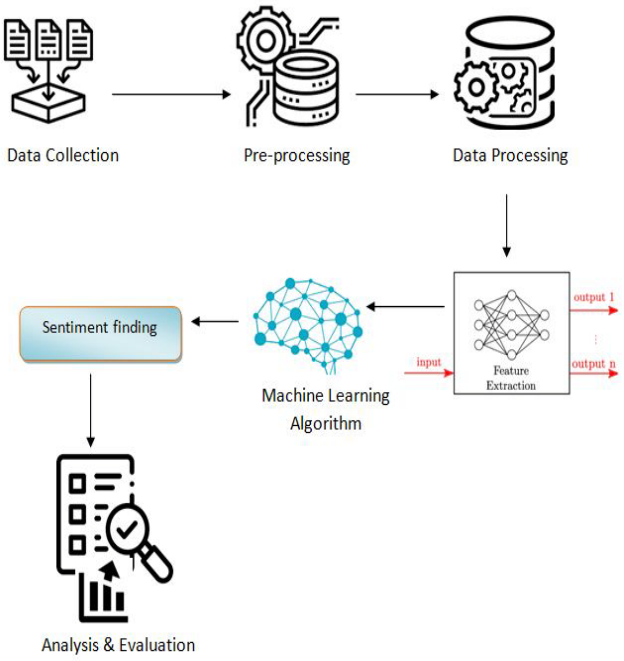

Fig.1. Proposed Methodology

TABLE 1.Overview Of Proposed Approaches

Languages	Wikipedia articles	Training dataset	Evaluation dataset
English	2452	3723	NA
German	2354	NA	3364
Tamil	3562	4785	NA
French	2354	3256	3857
Hindi	2548	NA	3265

TABLE 2. Statistical Properties of the Dataset

Characteristics	Quantity
Data collected size	1021
No. of words	63756
No. of characters	358740
No. of sentences	4407
No. of special characters	2900

Researchers have completed the fundamental pre-processing activities associated with natural reading comprehension, such as data cleansing, as all of the obtained data are in a readable form. Punctuation marks, hashtags, and advertising links have been deleted from the comments and the posts. We have also deleted the HTML entities from promotional postings because they typically contain those elements. As a result, we have left certain marks and special characters in place.

The impartiality and orientation values of each example were extracted using the text blob program after the data had been sanitized. For our dataset, the traditional "V-shape" also appeared. Fig.2 shows that relatively few examples are presented when the polarity values are significantly negative but the subjectivity values are less. It should be noted that

DOI: rp-9788770040723.064

postings with extremely high or extremely low impartiality values do exist. As a result, the scatter plot has a "V shape".

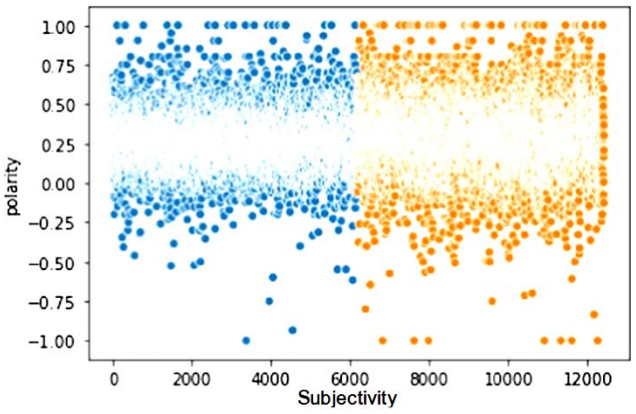

Fig. 2. Subjectivity vs. polarity scatter plot

A. *Feature Extraction*

That process of feature extraction is where this paper makes one of its primary contributions. As opposed to a standard text classification challenge, researchers have retrieved linguistic characteristics at the character, word, and sentence levels. The python NLTK package was used to extract a total of 86 features. They have isolated the special characters and the exclamation point from the 86 linguistic elements. Additionally, we utilized the numbers to determine how these characteristics affected sentiment analysis. These features extracted are immediately input into the training and testing phases of machine learning.10 distinct machine learning models have been used to identify the hierarchical classification issue. The techniques were carefully chosen since they are often utilized in other works of a similar nature. To assess effectiveness, components are identified, and Bayes-based regression and tree-based methods are used.

They have simply relied on the polarization values that were derived throughout the data preprocessing step for sentiment recognition. The inner lexicon collection utilized in the text blob program is quite trustworthy for evaluating the messages from social media or microblogging. Following finishing this phase, the dataset includes cases that are 55.6% positive, 31.2% negative, & 13.2% neutral.

IV. Evaluation and Analysis

Employed the most well-liked measures for performance assessment, including accuracy, ROC, PRC, recall, f-measure, and precision. Efficiency has been regarded as one of the most important measures in the majority of research publications that evaluate classification techniques. We are not presenting the theoretical formula or other characteristics associated with these assessment measures in this study because they are extensively utilized.

A. *Experimental Analysis*

Reliability, ROC, PRC, recall, f-measure, & accuracy were some of the metrics used to gauge success shown in Table 3. In the vast majority of research articles that analyze categorization systems, efficiency has been recognized as one of the most performance metrics. Due to their widespread use, we are not giving the theoretical model or any further features related to these assessment measures in this work.

TABLE 3. Performance Evaluation Metrics

Algorithm Name	Precision	Recall	F-Measure	Accuracy (%)
Naïve Bayes	0.544	0.342	0.333	34.974
Decision Tree	0.509	0.500	0.504	50.035
Random Forest	0.598	0.608	0.548	60.546
Support Vector Machine	0.694	0.577	0.449	56.980
Sequential Minimization Optimization	0.577	0.561	0.718	55.491
Linear Regression	0.522	0.541	0.522	53.83
Proposed System	0.762	0.633	0.713	75.61

Researchers employed a variety of measures to determine the performance, including reliability, ROC, PRC, recall, f-measure, and simplicity shown in Table 4. Productivity has been acknowledged as one of the most important measures in the great majority of research that examines categorization technologies. We are not providing the computational foundation or any more details about these evaluation measures in this study due to their broad use. This sort of outcome demonstrates the data imbalance factors. We employed relatively little data that came from a strong social media environment, the data is unbalanced, which affects the machine learning techniques.

TABLE 4. Statistical Metrics

Models	Kappa	MAE	RMSE	ROC
Naïve Bayes	0.1254	0.4451	0.6598	0.132
Decision Tree	0.1492	0.3470	0.5921	0.143
Random Forest	0.3241	0.3509	0.4192	0.242
Support Vector Machine	0.0677	0.2831	0.5301	0.158
Sequential Minimization Optimization	0.135	0.3446	0.4452	0.112
Linear Regression	0.1242	0.3357	0.4481	0.163
Proposed System	0.0149	0.3856	0.441	0.042

Fig.3 displays the Matthews linear regression data. The Algorithm for Random Forests produces the MCC value with the highest MCC.

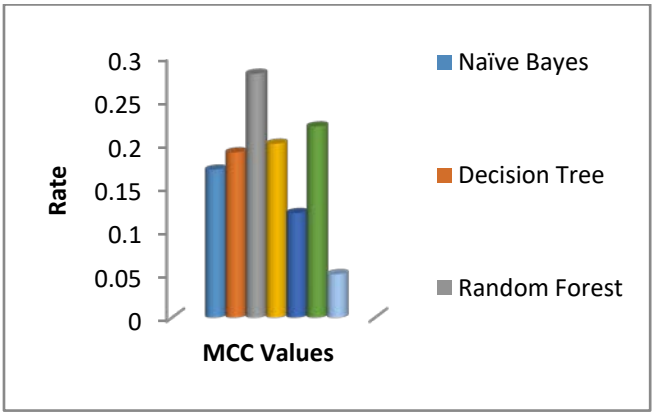

Fig.3.Bar chart of MCC values for each ML classifier

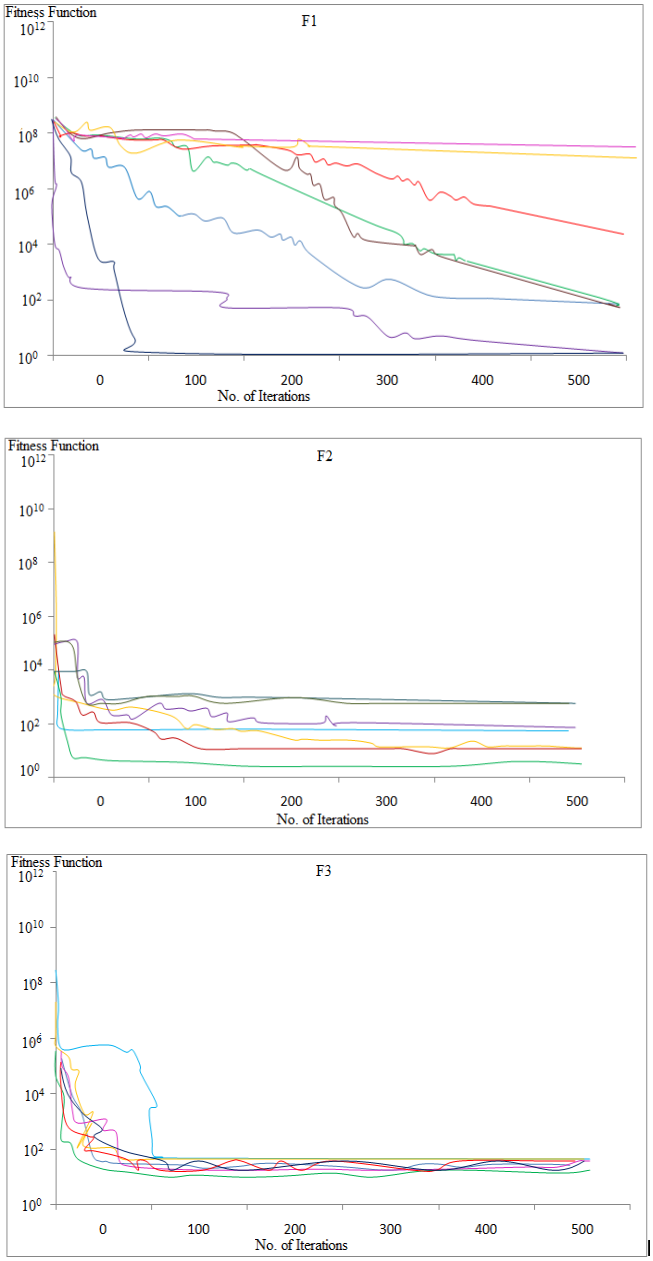

Fig. 4. Comparison of F1-F3 convergence score

The lowest Friedman mean rank across testbed suit programs. As a result, it receives the highest overall score for such functions. Fig.4 depicts the best convergence curves for the implemented methods.

V. Conclusion

Throughout this study, researchers developed an approach for using linguistic elements collected from comments on social media or postings on melancholy. It is exceedingly difficult to develop classifiers that perform better since social networking and microblogging sites use very casual language. We have used 10 classification techniques to investigate the effectiveness of various category classifications. The research methodology part contains the response to the three research problems that were defined. The most effective ML classifier is Random Forest, which has a decent chance of using textual data connected to depression for sentiment analysis.

References

[1] A. Saha, A. Al Marouf, and R. Hossain, “Sentiment analysis from depression-related user-generated contents from social media”, In 2021 8th International Conference on Computer and Communication Engineering (ICCCE), IEEE, pp. 259-264, 2021.

[2] F.F. Nastro, D. Croce, S. Schmidt, R. Basili, and F. Schultze-Lutter, “Insideout project: using big data and machine learning for prevention in psychiatry”, European Psychiatry, vol. 64, no. S1, pp.S343-S343, 2021.

[3] K. Chaudhary, M. Alam, M.S. Al-Rakhami and A. Gumaei, “Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics”, Journal of Big Data, vol. 8, no. 1, pp.1-20, 2021.

[4] L. T. Pugazhendhi, R. Kothandaraman, and B. Karnan, “Implementation of Visual Clustering Strategy in Self-Organizing Map for Wear Studies Samples Printed Using FDM”, Traitement du Signal, vol. 39, no. 2, 2022.

[5] A.S. Uban, B. Chulvi, and P. Rosso, “On the explainability of automatic predictions of mental disorders from social media data”, In International Conference on Applications of Natural Language to Information Systems, Springer, Cham, pp. 301-314, 2021.

[6] P.R. Garikapati, K. Balamurugan,T.L Latchoumi, and G. Shankar, “A Quantitative Study of Small Dataset Machining by Agglomerative Hierarchical Cluster and K-Medoid”, In Emergent Converging Technologies and Biomedical Systems, Springer, Singapore, pp. 717-727, 2022.

[7] P.N. Achyutha, S. Chaudhury, S.C. Bose, R. Kler, J. Surve, and K. Kaliyaperumal, “User Classification and Stock Market-Based Recommendation Engine Based on Machine Learning and Twitter Analysis”, Mathematical Problems in Engineering, 2022.

[8] B. Karnan, A. Kuppusamy, T.P. Latchoumi, A. Banerjee, A. Sinha, A, Biswas, and A.K. Subramanian, “Multi-response Optimization of Turning Parameters for Cryogenically Treated and Tempered WC–Co Inserts”, Journal of The Institution of Engineers (India): Series D, pp.1-12, 2022.

[9] K. Arunkarthikeyan, and K. Balamurugan, “Studies on the impact of soaking time on a cryogenic processed and post tempered WC-Co insert”, Materials Today: Proceedings, vol. 44, pp.1692-1699, 2021.

[10] J. Yarlagaddaa, and M. Ramakrishna, “Fabrication and characterization of S glass hybrid composites for Tie rods of aircraft”, Materials Today: Proceedings, vol.19, pp.2622-2626, 2019.

[11] M. Karami, T.H. Nazer, and H. Liu, “Profiling Fake News Spreaders on Social Media through Psychological and Motivational Factors”, In Proceedings of the 32nd ACM Conference on Hypertext and Social Media, pp. 225-230, 2021.

[12] K.A. Govindasamy, and N. Palanichamy, “Depression detection using machine learning techniques on twitter data”, In 2021 5th international conference on intelligent computing and control systems (ICICCS), IEEE, pp. 960-966, 2021.

[13] J. Kim, S. Hwang, and E. Park, “Can we predict the Oscar winner? A machine learning approach with social network services”, Entertainment Computing, vol. 39, p.100441, 2021.

[14] G.K. Gupta,and D.K. Sharma, “Depression detection on social media with the aid of machine learning platform: A comprehensive survey,” In 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp. 658-662, 2021, March.

[15] S. Smys, and J.S. Raj, “Analysis of deep learning techniques for early detection of depression on social media network-a comparative study”, Journal of trends in Computer Science and Smart technology (TCSST), vol. 3, no. 1, pp.24-39, 2021.

[16] A.P. Venkatesh, T.P. Latchoumi, S. ChezhianBabu, K. Balamurugan, S. Ganesan, M. Ruban, and L. Mulugeta, “Multiparametric Optimization on Influence of Ethanol and Biodiesel Blends on Nanocoated Engine by Full Factorial Design”, Journal of Nanomaterials, 2022.

[17] Q. Zheng, Y. Guo, Z. Wang, F. Andrasik, Z. Kuang, J. Li, S. Xu, and X. Hu, “Exploring Weibo users’ attitudes toward lesbians and gays in Mainland China: A natural language processing and machine learning approach”, Computers in Human Behavior, vol. 127, p.107021, 2022.

[18] E. Sherman, K. Harrigian, C. Aguirre, and M. Dredze, “Towards Understanding the Role of Gender in Deploying Social Media-Based

Mental Health Surveillance Models", In Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access, pp. 217-223, 2021.

[19] T.P. Latchoumi, R. Swathi, P. Vidyasri, and K. Balamurugan, "Develop New Algorithm To Improve Safety On WMSN In Health Disease Monitoring", In 2022 International Mobile and Embedded Technology Conference (MECON), IEEE, pp. 357-362, 2022.

[20] Y.J. Bae, M. Shim, and W.H. Lee, "Schizophrenia Detection Using Machine Learning Approach from Social Media Content", Sensors, vol. 21, no. 17, p.5924, 2021.

[21] R. Kenny, B. Fischhoff, A. Davis, K.M. Carley, and C. Canfield, "Duped by bots: why some are better than others at detecting fake social media personas", Human factors, p.00187208211072642, 2022.

[22] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021). Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[23] F.Janjua, A. Masood, H. Abbas, I. Rashid, and M.M.Z.M. Khan, "Textual analysis of traitor-based dataset through semi supervised machine learning", Future Generation Computer Systems, vol. 125, pp.652-660, 2021.

[24] J. Yarlagaddaa, and R. Malkapuram, "Influence of MWCNTs on the Mechanical Properties of Continuous Carbon Epoxy Composites", Revue des Composites et des MatériauxAvancés, vol. 30, no. 1, 2020.

[25] A. Arora, P. Chakraborty, and M.P.S. Bhatia, "Problematic use of digital technologies and its impact on mental health during COVID-19 pandemic: assessment using machine learning", In Emerging Technologies During the Era of COVID-19 Pandemic, Springer, Cham, pp. 197-221, 2021.

[26] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[27] P Matheswaran, C Navaneethan, S Meenatchi, S Ananthi, K Janaki, and A Manjunathan, "Image Privacy in Social Network Using Invisible Watermarking Techniques", Annals of the Romanian Society for Cell Biology, vol.25, issue.5, pp.319-327, 2021.

[28] A Manjunathan, A Lakshmi, S Ananthi, A Ramachandran, and C Bhuvaneshwari, "Image Processing Based Classification of Energy Sources in Eatables Using Artificial Intelligence", Annals of the Romanian Society for Cell Biology,vol.25, issue.3, pp.7401-7407, 2021.