AN EFFICIENT APPROACH TO DETECT DEPRESSION THROUGH PREDICTIVE ANALYSIS

Sakshi Rastogi[1], Gaurav Kumar Srivastava[2], Sunil Kumar Vishwakarma[3]
Student[1], Assistant Professor [2, 3]
Department of Computer Science & Engineering[1-3]
Babu Banarasi Das University, Lucknow[1-3]
E-Mail ID- sakshirastogi.2607@gmail.com[1], gaurav18hit@bbdu.ac.in[2], sunilvishwakarma83@gmail.com[3]

Abstract.

Nowadays in this world when Machine is learning by algorithms and performing according to the inputs Python has its unique importance. Depression is said to be a feeling in which a person feels having a low mood and a state of strongly not liking someone/something. It has many effects on the body of the person. The symptom which is recognized as the core of depression is not feeling interested in the works or not feeling pleasure in the things that give joy to them earlier. The primary intention of this research is to carry out a comparison amongst the different Algorithms of Machine Learning based on accuracy, precision, sensitivity, F1 score, and Confusion Matrix to find which algorithm gives the best performance on depression data. The final aim of this research paper is to provide a model that will predict depression in the human body.

Keywords. Random Forest Classifier, Extra Trees Classifier, Multi-Layer Perceptron, Support Vector Classifier, F1 Score, Confusion Matrix, Predictive Analytics

1. INTRODUCTION

Depression is said to be a feeling in which a person feels having a low mood and a state of strongly not liking someone/something. The symptom which is recognized as the core of depression is not feeling interested in the works that give joy to them earlier. This can result in the person having a state of sadness, thinking difficulty, problems in paying attention. It can also lead to an increase or decrease in the diet and sleeping time of the person. In this condition person also experience the feeling of dejection, hopelessness and suicidal thoughts.

1.1 Symptoms of Depression

- Problem in Sleeping
- Loss of Interest
- Increment in Fatigue
- Emotions that are Uncontrollable
- Appetite of a Person Changes
- Weight of a Person Changes

2. LITERATURE SURVEY

This section is concerned with all the previous works that are done in this field. It presents a study that gives the comparison between different Algorithms of Machine Learning. They have used algorithms as- Logistic Regression, Random Forest, XG Boost, Support Vector Machine, Ada Boost, K-NN and Decision Tree. They have performed their research on the prediction of liver disease at an early stage. They have compared the above Machine Learning Algorithms on the basis of Accuracy, Precision, Recall, F1 Score, an area under curve and Specificity. They have collected their datasets from UCI Machine Learning Repository. There result in states that Random Forest Algorithm performs the best in terms of accuracy with 83.70%. Random Forest also performs well in the terms of other parameters too. So, they concluded Random Forest as the best algorithm that can be used in predicting Liver Disease. It describes about the machine learning techniques principles and he also described the use of them in the domains of real-world applications. They further describe the challenges and potential they need to perform in their research. On the basis of their goal, they shortly discussed how the methods of machine learning are being used in providing an appropriate way in solving the problems of the real world. The conclusion was that machine learning is built upon the data that is provided to the algorithms for learning purposes and the performance provided by them. It has the algorithms of Machine Learning for doing the predictions on anxiety, depression and stress in their paper. They have gathered their data by making the questionnaire related to their topic. This consists of the data of several cultures and communities which are employed and unemployed. They realize in their research that classes they made were imbalanced at the time when they start making confusion matrix. So they measure f1 score to identify the best accuracy model. They find that Random Forest Classifier as the best model. The conclusion was that the f1 score is the important aspect in finding the best accuracy model. It describes the use of various kinds of machine learning. It also merges the results of the analysis that comes from all the algorithms that were used for performing their research. Their main purpose was to increase the awareness of Machine Learning among the persons. Their conclusion presents that it is necessary for the Machine Learning model to continuously grasp from the past doing that come from countries that are developed, set up algorithms of machine learning mostly for the making enterprises in domestic areas and providing help of the economy in developing industry. It presents about the survey on how machine learning can be used for providing investigation on depression. The methods which they use in their systems are based on the method of detection via posts on social media, syntax and semantic analysis of the person's emotion in order to predict the depression levels of different age groups. Some have performed comparative research on four Algorithms of Machine Learning. For the purpose of reducing attributes, they used CFSSubsetEval. They have collected their datasets from OASIS-Brains.org. They finally concluded as J48 is the best algorithm for the purpose of detecting Dementia.Some have conducted their research on various Algorithms of Machine Learning with the aim of finding the effectiveness. The datasets that were used in this research were from different types of clinics. In these datasets are small, medium and large that can be accessed publicly. The comparison between algorithms was done on the basis of the requirement of accuracy and time in training and testing algorithms. The result implies that K-Nearest Neighbor performed well amongst all the algorithms used.It also presents that social network data gives the opportunity to work on the user's moods and attitudes when they convey messages with the use of social media. The data for the analysis was on the

Facebook data that they collected from an online public source. They gave their analysis on 7146 Comments on Facebook. They got the conclusion as 54.77% depressive person who conveys between mid-night to mid-day & 45.22% depressive person who conveys between mid-days to mid-night. Some have done comparative research amongst some popular Algorithms of Machine Learning. They have used two datasets in order to provide the best efficiency. They have collected all the information including datasets from the UCI Machine Learning storehouse. There first dataset contains 6500 rows & 13 columns & second dataset contains 1055 rows & 13 columns. There result shows that Support Vector Machine performs the best accuracy of 99.38%.

3. PROPOSED METHODOLOGY

The model that we have proposed for this research is described in the Figure 1

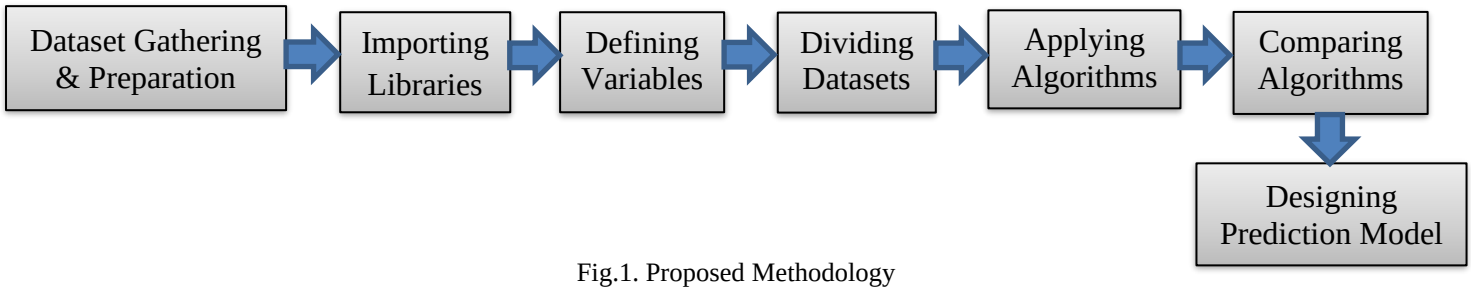

Fig.1. Proposed Methodology

3.1. Dataset Gathering & Preparation

In developing a Machine Learning Model or performing any type of research in Machine Learning the first and most important step is to gather the datasets. We have taken the Depression datasets from an online portal. There were 1290 rows and 27 columns in the original dataset. Table 1 describes the original Depression Dataset.

Table 1. The Original Dataset

1.	Timestamp	Age		obs_ consequence	Comment
2.	8/27/2014 11:29	37	-----	No	NA
3.	8/27/2014 11:29	44	-----	No	NA
⋮	⋮	⋮		⋮	⋮
1259.	11/30/2015 21:25	46	-----	No	NA
1260.	2/1/2016 23:04	25	-----	No	NA

Table 2 describes the final dataset that we have used for performing this research. In the column 'Gen', 0 represents Female and 1 represents Male and in all the other columns 0 stands for No and 1 stands for Yes.

Table 2. The Final Dataset

1.	Age	Gen		phys_health_ consequence	Target
2.	37	0	--------	0	1
3.	44	1	--------	0	1
⋮	⋮	⋮		⋮	⋮
1248.	32	1	--------	0	0
1249.	36	1	--------	0	0
⋮	⋮	⋮		⋮	⋮
1259.	46	0	--------	0	1
1260.	25	1	--------	0	1

3.2. Importing Libraries

Python has plenty of libraries that can be used by the developers for performing different kinds of research like on Machine Learning Models, Robotics etc. All the libraries are predefined and are easily available we just have to import them and use them according to our requirement. We have done the same we have imported the libraries like pandas, matplotlib, sklearn etc. We have used the following libraries-

1. Pandas- We used this library for reading the datasets.
2. Matplotlib, Seaborn- We have used the seaborn library for plotting heat map and matplotlib library for plotting the labels, giving the titles etc. around the heat map.
3. Numpy- We have used this library for designing the prediction model.
4. Sklearn- We have used many modules of this library in this research paper. We have used the modules like model_selection, ensemble, neural_network etc.

3.3. Defining Variables

Defining the variables that are included in the creation of the Model is also the most important step as the Model fits only on variables. For performing this we have to follow this pseudo code-

Step 1. Define the variable X by dropping target column and storing all other columns in it.
Step 2. Put the axis=1 for dropping target column.
Step 3. Define the variable y by storing target variable in it.

3.4. Dividing Datasets

It is necessary to divide the datasets into Training Module and Testing Module so that we can perform the works on Training Module and make predictions on Testing Module. We have divided the whole datasets into Training and Testing Module. We have divided the dataset into 80-20 ratio that means our Training Module contains 80% and Testing Module contains 20% respectively. For performing this we have to certain pseudo code the code is as follows-

Step 1. Define the ratio in which you want to divide the dataset.
Step 2. Define the Test Size according to the desired ratio.
Step 3. Apply the train_test_split module.

3.5. Applying Algorithms

The main step in this research is to apply algorithms. We have applied all Machine Learning Algorithms like Random Forest Classifier, Extra Trees Classifier, Ada Boost Classifier, Decision Tree Classifier and Multi-Layer Perceptron. For performing this we have defined the pseudo-code that is given as follows-

Step 1. Store each algorithm into different variables.
Step 2. Fit each model on the training dataset using the fit method.
Step 3. Use predict method for doing the predictions.

3.6. Comparing Algorithms

The pre-final step in this research is to compare the above-mentioned models. We have compared all the algorithms on the basis of accuracy, precision, sensitivity, F1 score and Confusion Matrix to find the best algorithm amongst all the models. For doing the work the pseudo code is as follows-

Step 1. Use accuracy score for finding accuracy.
Step 2. Use precision score for finding precision.
Step 3. Use recall score for finding sensitivity.
Step 4. Use f1 score for finding F1 Score.
Step 5. Print the results of comparison.

The another parameter that we are giving for comparison is the Confusion Matrix. Confusion Matrix is divided into two rows and two columns. It is plotted using matplotlib and seaborn libraries. We have done this in Three Steps. We have plot the Confusion Matrix of 2 Algorithms at a time and followed by the other Algorithms. Pseudo that is applicable for doing this is as follows-

Step 1. Use confusion matrix for plotting confusion matrix.
Step 2. Use the heatmap for doing the same.
Step 3. Print thr results.

3.7. Designing Prediction Model

The final step in the study is to evaluate a model that tries to predict depression in the human body. This prediction model works on the dataset that is provided to the model. This model has been made with the help of library named as numpy. The input contains the data in the same manner that is defined in the dataset. It implies that the 1st value is the Age of the Person, 2nd is the Gender in the form of 0 and 1, 3rd defines that if the person is having any Family History of Depression?, 4th is asking that if the person is working in any tech company?, 5th is asking that if the person is going through any treatment?, 6th is asking if the person is working remotely? 7th is asking if the person has mental health consequence? 8th is asking if the person has physical health consequence? and finally 9th is the target variable that predicts the depression in human body. The pseudo code for this is shown as follows-

Step 1. Use numpy for designing model.
Step 2. Input the values in the form of array.
Step 3. Reshape the array which was given as input.
Step 4. Use the results of comparing algorithms.
Step 5. Print the results either in the form of 0 or 1 i.e., 0 means Patient doesn't have Depression and 1 implies Patient have Depression.

4. RESULTS & DISCUSSIONS

The application of all above mentioned machine learning methods with all parameters- accuracy, precision, sensitivity, F1 score & Confusion Matrix is described in Table 3. In the case of Confusion Matrix, the first value represents False Negative value, second

represents False Positive value, third represents True Negative value & fourth represents True Positive value respectively. From the below table, it has been clearly shown that Ada Boost Classifier and Multi-Layer Perceptron both have shown the good performance in some parameters but Support Vector Classifier has shown the best performance in all the parameters. In the case of Confusion Matrix, Multi-Layer Perceptron has performed good as it predicts 78 True Positives but Support Vector Classifier has again performed the best as it has given the total 88 True Positives. So, it has been cleared that Support Vector Classifier is the best algorithm amongst all the presented algorithms so we have used the same algorithm for performing the final step of research i.e., designing of the Prediction Model. According to dataset which was given the prediction model performs effectively as it can clearly predict depression in human body.

Table 3. Result Analysis

Sr. No.	Model	Accuracy	Precision	Sensitivity	F1 Score
0	Random Forest	0.480159	0.488722	0.507812	0.498084
1	Multi-Layer Perceptron	0.503968	0.506224	0.953125	0.661247
2	Extra Trees Classifier	0.476190	0.482143	0.421875	0.450000
3	Ada Boost Classifier	0.523810	0.533333	0.500000	0.516129
4	Decision Tree Classifier	0.476190	0.482456	0.429688	0.454545
5	Support Vector Classifier	0.551587	0.549669	0.648438	0.594982

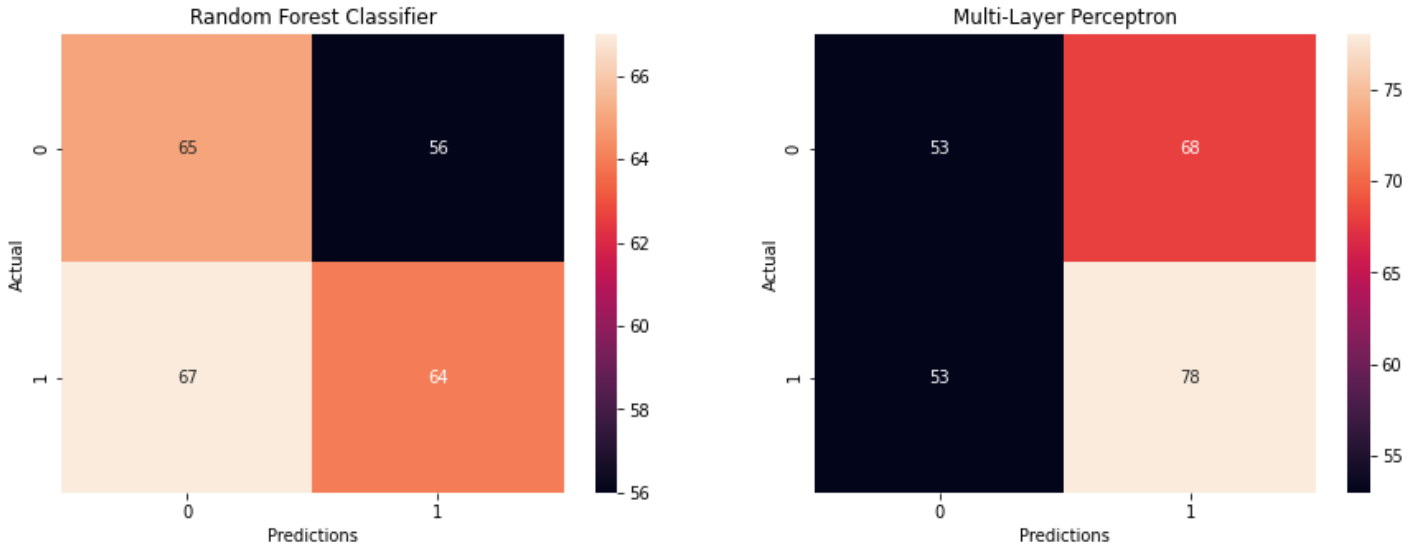

Fig.3. Confusion Matrix of Random Forest Classifier and Multi-Layer Perceptron

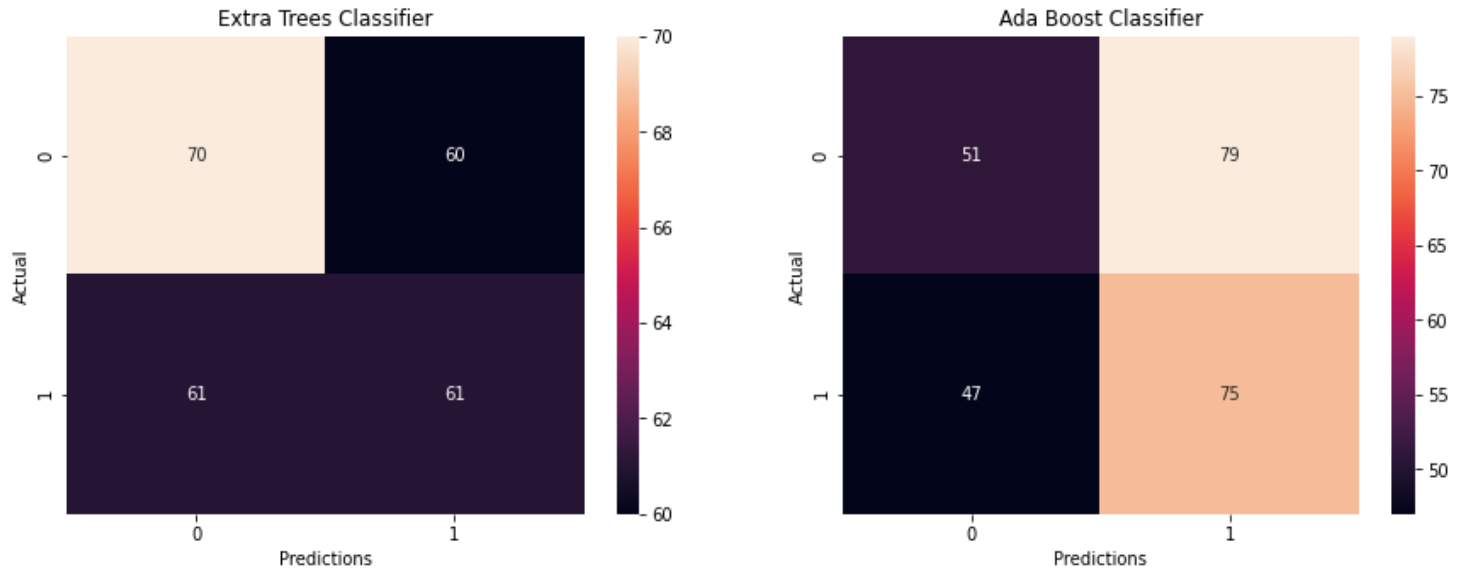

Fig.4. Confusion Matrix of Extra Trees Classifier and Ada Boost Classifier

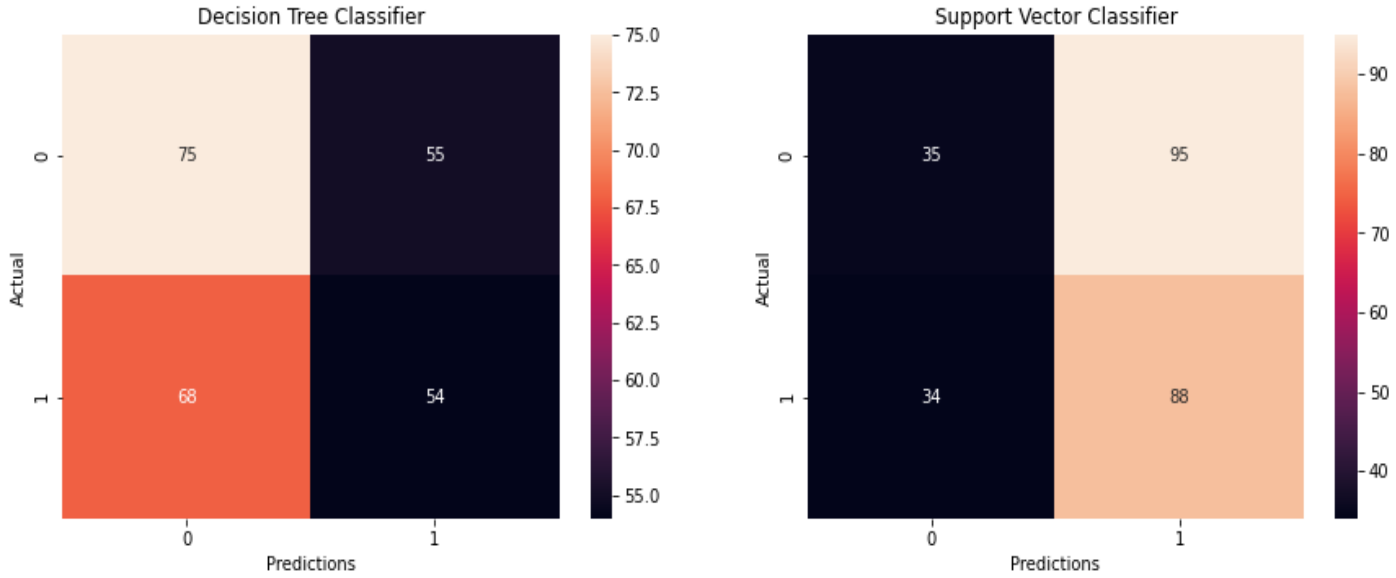

Fig.5. Confusion Matrix of Decision Tree Classifier and Support Vector Classifier

5. CONCLUSION

This paper has shown a comparative study between multiple Machine Learning Algorithms very well. The dataset used in this research is completely authenticated and unique as the proper analysis and editing is done on the dataset. According to the results provided by this research the Support Vector Classifier has performed really well in all the parameters. So, this algorithm can be used for future usage.

This paper has also shown a Model that will predict Depression in the human body according to the given dataset. The model predicts in the form of 0 and 1 i.e., 0 implies that the person is not having Depression and 1 implies that the person is having Depression. In this research we have used the results of the comparative study that is done as the primary objective of this research.

6. FUTURE SCOPE

In future this research will be helpful in the following aspects which are given as follows-

- It will be helpful for researchers to perform study in Predictive Analytics
- It will be helpful in selecting the best Machine Learning Model if the aim is to determine Depression at an initial stage
- It will be helpful in designing the Prediction Model using Machine Learning
- It will be helpful in predicting Depression from the Human Body

7. REFERENCES

[1] M. Ghosh *et al.*, "A comparative analysis of machine learning algorithms to predict liver disease," *Intelligent Automation and Soft Computing*, vol. 30, no. 3, pp. 917–928, 2021, doi: 10.32604/iasc.2021.017989.

[2] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 3, May 2021, doi: 10.1007/s42979-021-00592-x.

[3] A. Priya, S. Garg, and N. P. Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms," in *Procedia Computer Science*, 2020, vol. 167, pp. 1258–1267. doi: 10.1016/j.procs.2020.03.442.

[4] W. Jin, "Research on Machine Learning and Its Algorithms and Development," in *Journal of Physics: Conference Series*, Jun. 2020, vol. 1544, no. 1. doi: 10.1088/1742-6596/1544/1/012003.

[5] G. Srivastava, S. Kumar, H. Pandey, G. Kumar Srivastava, S. Kumar, and H. Pandey, "Modelling of an offline and online software for normalization of microarray data of gene expression by Perl, Bioperl and PerlTk and Perl-CGI," 2019.

[6] Vishwakarma Sunil Kumar, Sharma Birendra Kumar, and Abbas Syed Qamar, "Digital Watermarking for Image Authentication using Spatial-Scale Domain based Techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 4, pp. 2334–2341, Nov. 2019, doi: 10.35940/ijrte.d8215.118419.

[7] D. Bansal, R. Chhikara, K. Khanna, and P. Gupta, "Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia," in *Procedia Computer Science*, 2018, vol. 132, pp. 1497–1502. doi: 10.1016/j.procs.2018.05.102.

[8] M. Diwakar, P. Singh, and A. Shankar, "Multi-modal medical image fusion framework using co-occurrence filter and local extrema in NSST domain," *Biomedical Signal Processing and Control*, vol. 68, Jul. 2021, doi: 10.1016/j.bspc.2021.102788.

[9] Md. 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," *Health Information Science and Systems*, vol. 6, no. 1, Dec. 2018, doi: 10.1007/s13755-018-0046-0.

[10] K. Sethi, A. Gupta, G. Gupta, and V. Jaiswal, “Comparative Analysis of Machine Learning Algorithms on Different Datasets,” 2019. [Online]. Available: www.ccsarchive.org

[11] A. Chakraborty, M. Jindal, M. R. Khosravi, P. Singh, A. Shankar, and M. Diwakar, “A Secure IoT-Based Cloud Platform Selection Using Entropy Distance Approach and Fuzzy Set Theory,” *Wireless Communications and Mobile Computing*, vol. 2021, 2021, doi: 10.1155/2021/6697467.

[12] P. Singh, M. Diwakar, X. Cheng, and A. Shankar, “A new wavelet-based multi-focus image fusion technique using method noise and anisotropic diffusion for real-time surveillance application,” *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1051–1068, Aug. 2021, doi: 10.1007/s11554-021-01125-8.

Biographies

Sakshi Rastogi, currently pursuing Master of Technology from Babu Banarasi Das University, Lucknow. I have done Bachelor of Technology from Invertis University, Bareilly. My research areas are Software Testing and Machine Learning.

Dr. Gaurav Kumar Srivastava, currently working as Assistant Professor in the Department of Computer Science and Engineering at Babu Banarasi Das University, Lucknow, 226028, India. He has completed his Bachelor of Technology in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University, Lucknow, completed his Master of Technology in Computer Science and Engineering from Babu Banarasi Das University, Lucknow and completed Ph.D. in Computer Science and Engineering from Maharishi University of Information Technology, Lucknow. He has published various research articles in International Peer reviewed Journals/ Conferences.

Sunil Kumar Vishwakarma, currently working as Assistant Professor in the Department of Computer Science and Engineering at Babu Banarasi Das University, Lucknow, 226028, India. His research areas are Computer Vision and Image Processing. He has completed his Master of Technology from IET Lucknow and Bachelor of Technology from BBDIET & RC Bulandshahr.