An Internet of Things (IoT) Speech Recognition System to improve the performance of Emotion Detection

[1]Dr. Arun Kumar and [2] Dr. Sridhar Manda

Professor, Department of Computer Science and Engineering, MLRIT, Hyderabad, Telangana, India. E-Mail: arunkumar.arigala@gmail.com

Assistant Professor, Department of Computer Science and Engineering, Balaji Institute of Technology and Science, Narsampet, Telangana, India. E-Mail: mandasridhar550@gamil.com

Abstract:

Communicating effectively is the most important step in conveying one's thoughts and ideas to others. Speech is the most effective human communication method. The Internet of Things (IoT) is bringing increasingly intelligent systems to daily life. Wearables, UI, self-driving cars, and automated systems are examples. Most artificial intelligence implementations are voice-based and require minimal user involvement. Because of this, these computer programmes need to be able to fully understand human speech. From a speech percept, it is possible to learn a lot about the speaker's gender, age, language, and emotional state. IoT speech recognition systems frequently include an emotion detection system to better comprehend the speaker's mood. The overall performance of the IoT application can be greatly affected by the performance of the emotion detection system in a variety of ways, and these applications can benefit greatly from this. This study presents a new system for detecting speech emotions based on emotions that improves on the current system in terms of data, extraction of features, and methodology.

Keywords: Affective computing, Emotion detection, AI, Smartkom

1. INTRODUCTION

Speech Emotion Detection is a challenging component. Intelligent systems must mirror human actions to be termed intelligent. Humans may adapt their conversations to their own and their audience's emotions. Machine learning systems can recognise speech emotions. This paper discusses the methodology and tests used to develop a speech emotion detection method. Even though physical indicators like facial expression can help determine an individual's emotional state through the use of physiological signals like electroencephalography, blood volume pulse, and galvanic skin response, the insensitivity of physiological signals to social masking of emotions makes them preferable over physical indicators. Physical signals can be gathered using tethered laboratory sensors or wireless physiological sensors. It's a little more invasive and obtrusive than the first option, but the first one can still get the job done, and it's better than nothing. It is possible, however, to collect physiological signals in a non-invasive and non-obtrusive manner using the second option. Integrated sensor technologies have made it easy for people to quickly start using these sensors that can be worn.

There are a variety of modern-day applications for determining the emotion expressed in a speech perception. Studying how people interact with computers is known as Human-Computer Interaction (HCI) research [1]. The computer system must be able to recognise more than words in order to be effective in an HCI application. The IoT (IoT) field [14], on the other hand, is experiencing rapid growth. Many real-world Internet of Things (IoT) applications. In Internet of Things (IoT) applications[12], voice is critical. According to a recent study, approximately 12% of all Internet of Things (IoT) applications will be able to run entirely on voice commands by 2022.

The selection of a strong emotional speech database is one of the most important aspects of a successful SER system.

Identifying and utilizing the most useful features of

Using machine learning techniques to create reliable classifiers

In both mono-directional and bi-directional voice interactions, understanding the speech signal is critical. IoT and HCI use AI and NLP-based apps to create smart homes and cities[15]. Self-driving automobiles employ voice instructions for numerous purposes. This app's ability to detect the user's mood is a big plus. In cases where the user cannot clearly articulate a spoken command, the user's tone of voice can activate emergency car features. Speech emotion recognition in call centres can divert automated voice calls to customer service reps for further discussion. Lie detectors, criminal investigation analysis, and humanoids all use speech emotion detection.The four main components of our SER system are as follows: To begin with, there is a library of voice samples. In this case, the second feature vector is formed by removing features. Following that, we tried to identify the most important characteristics of each emotion. In a machine learning classification model for recognition, these features are added .Human speech can convey many emotions. Which features to utilise in emotion detection is debatable. Recent research has retrieved energy, pitch, formant, linear prediction coefficients (LPC), mel-frequency melfrequency cepstral coefficients (MFCC), and modulation spectral characteristics. Emotional features were extracted using MFCC and modulation spectral features.

a. Research Questions:

The goal of this project is to make a classification scheme for expressions work better by using portable sensors that people can wear.

When it comes to emotion recognition, how do feature selection methods affect the system's performance?

In what ways can we improve the performance of an emotion classification system that is based on signal processing?

When it comes to capturing emotional shifts over time, how can signals do this?

2. LITERATURE REVIEW

The SER has an issue extracting emotional elements. Researchers believe energy, pitch, formant frequency, LPCC, MFCC, and modulation spectral characteristics convey emotion information (MSFs).Most researchers favour a feature set that combines emotional elements. It's possible to overfit when using a combining feature set because of the high dimensionality and redundancy of speech features, which makes it difficult for most machine learning algorithms. As a result, in order to minimise feature overlap and redundancy, careful feature selection is a must. [2] provides a comprehensive review of feature selection models and techniques. Using feature extraction and feature selection combined improves learning efficiency, reduces computational complexity, builds more generalizable models, and reduces storage needs. Emotion recognition ends with classification. Emotional classification is based on utterances or frames of utterances. Researchers have proposed GMM, HMM, SVM, neural networks, and RNN for speech emotion recognition (RNN). Researchers have developed a modified brain emotional learning model (BEL) for speech emotion identification. Multiple Kernel Gaussian Process Classification combines linear and RBF kernels. VSS uses voice signal segmentation to process textural images. Log-Gabor filters categorise voiced and unvoiced spectrum features. Classifying emotional states with machine learning. In order to classify new observations, these algorithms must first learn from the training images. There is no one-size-fits-all answer when it comes to selecting a learning algorithm; each method has its own advantages and drawbacks. As a result, three different classifiers were used to assess their relative merits.

MLR is a simple and effective machine learning technique for regression and classification. The LRC has been adjusted [4]. The absolute difference between original and forecasted response vectors is an alternative to the Euclidean distance. Machine learning's optimal margin classifier is the support vector machine (SVM). In addition, it has been extensively used in a number of studies on audio emotion recognition [3]. Compared to other classifiers, it is capable of achieving a high level of classification accuracy even with limited training data. [4] provides the theoretical background for SVM. [5] provides a free MATLAB toolbox for implementing SVM. Data scientists and businesses all over the world rely on SVM as the most reliable machine learning algorithm. Separate classes are used to separate them in this supervised learning method, which looks for patterns within them. These features are transferred into a high-dimensional feature space that optimises their hyperplane. SVMs will be simulated using the Lib-SVM toolbox in MATLAB, which uses kernel-based techniques including polynomial kernels and radial basis function kernels.

As Cao et al. [3] wrote, "the emotions expressed by humans are mostly a consequence of mixed feelings." ([3] In order to improve the SVM algorithm, they proposed taking into account mixed signals and selecting the most dominant one. A ranking SVM classifier was used. Ranking SVM applies all personal binary classifier and SVM classifier forecasts to a final multi-class problem. With the ranked SVM algorithm, their system had a 44.40 percent accuracy rate.

Improved preprocessing technology was developed by Chen et al. (Chen et al., 2004). Fisher and Principle Component Analysis (PCA) were used in conjunction with SVM and ANN for preprocessing and classifying data. Each of the preprocessing and classifier algorithms was used in a different combination in each of the four experiments. First, the features of a multi-level SVM classifier were selected using the Fisher method (Fisher + SVM). The second experiment used PCA to lower the SVM classifier's feature dimensions[16]. The third experiment: Fisher + ANN. The data was classified using PCA before ANN [13]. Two significant findings emerged from this set of experiments. In the first place, reducing the system's dimensionality improves its efficiency. When it comes to emotion detection, the SVM classifier algorithm beats out the ANN algorithm. Fisher dimensionality reduction and SVM classification produced an accuracy of 86.50 percent in the winning experiment.

Nwe et al. [5] built a system using MFCC-like properties. A Hidden Markov Model (HMM) and Log Frequency Power Coefficients (LFPC) were used to classify speech emotions. Because they used a dataset that was only accessible to them, their research is not available to the general public. According to them, using LFPC coefficients instead of MFFCC coefficients significantly increases the model's accuracy. The best classification accuracy in their model is 96%.An innovative approach was proposed by

Rong et al. [6] to improve the accuracy of current models. Preprocessing techniques were traditionally used by computer scientists to reduce the number of features. This new system, on the other hand, increased the number of features used to classify. While claiming to have classified audio percepts from a small dataset, they did not reveal the features used in this process. However, neither of those features is language-dependent. Using ERFTrees with a variety of characteristics, they achieved an accuracy of 82.54 percent.

Using a dataset that is more representative of the real world, Narayanan [7] proposes a different approach. There were only two distinct emotions to categorise for his research: happy and angry. He collected data from call centres and performed binary classifications using only those two emotions. The KNN algorithm was used in conjunction with a variety of features, including acoustic, lexical, and other language-based features. In addition, this study focused on the call centre industry and included both male and female participants. Customers of both sexes experienced accuracy increases of 40.70 percent and 36.40 percent, respectively.

3. Methodology

a. Data Collection

A speech signal is a constantly changing sound waveform. When a person pronounces a phoneme, they are able to alter the sound signal by manipulating their vocal tract, tongue, and teeth. Data can be quantified using the features. A better way to represent speech signals is to extract features common to all speech signals in order to get the most information possible out of them. Among the qualities of good features are:

There should be no dependencies between the features. There are a lot of correlations between the features in the feature vector. So, it's important to choose a subset of features that are different and have nothing to do with each other.

- The features should provide context-relevant information. Any further analysis should focus on emotional parts.All data samples should have the same features. It's best to avoid features that only apply to a small subset of data.

- The feature values should be processed. An unmanageable raw feature vector can result from the initial feature selection process. Remove any outliers, missing values, or null values with Feature Engineering.

Two basic categories group emotional speech percepts:1. prosodic characteristics

2. Phonetic features

The energy, pitch, tempo, volume, formant, and intensity are all examples of prosodic features. To a large extent, phonetic characteristics have to do with how words are said in a particular language. The prosodic features, or a combination of them, are used for emotion detection purposes. The emotional content of a song is largely determined by its pitch and volume.

b. Mel Frequency Cepstrum Coefficients (MFCC) Features

MFCC is used to detect speech emotions. Here's some context:For frequency vs. pitch measurement, the word Mel stands for the frequency scale. The formula below can be used to convert a value measured on a frequency scale to a Mel scale.

$$m = 2595 \log10 (1 + (f/700))$$

a) Cepstrum is an acronym for the Fourier Transform of the logarithmic spectrum of a spoken word.

c. Coefficient Computation:

The following equations can be used to figure out the MFCC features of a speech signal:

> In the first step, you need to encapsulate the sound. Using the frame blocking method we talked about earlier, audio signals are split into 20ms to 30ms long frames that overlap by 50%.
>
> The next logical step is to perform a mathematical calculation. The power spectrum for each frame of the signal is computed in this step. A periodogram, another name for the power spectrum, is a tool for determining the frequencies present in a video frame [9]. Each frame's value is multiplied by a Hamming window value in order to narrow the frequency range. The periodogram is the modulus squared of the Discrete Fourier Transform modulus (DFT).

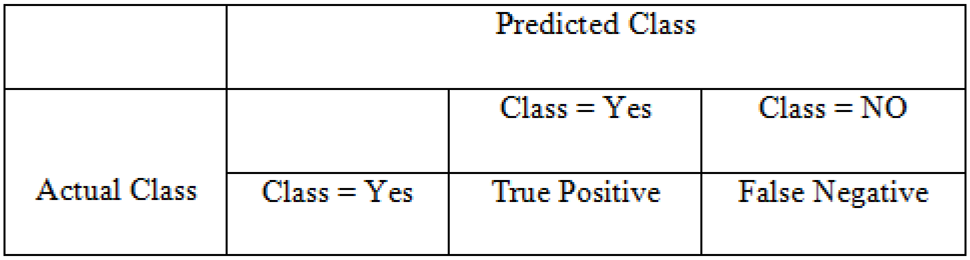
Next, the power spectra can contain a wide range of frequencies that are very close together. The signal's energy values are difficult to calculate because of the signal's frequency variations. Thus, a filter called Mel Filterbank is applied to the power spectrum in order to scale the values. The Mel Filterbank consists of a set of triangular frequency filters. The frequencies become narrower as they approach 0Hz; as they rise higher, they become wider [9]. Each frame's energy is the sum of the power spectrum values and the Mel Filterbank values. However, because the analysis makes use of overlapping frames, the energy values for the various frames will be in close agreement.

Finally, the energising energy must be de-correlated. Use of the DCT function is required for this task. In response to the obtained pitch and energy values, DCT generates a coefficient list. Pitch and energy changes are more consistent in the first 12 to 13 coefficients of each frame. This makes them more useful for analysis. The Mel Frequency cepstral coefficients refer to these lower-level coefficients.

4. Evaluation:

Models that can learn and improve are the most important characteristics of machine learning models. Experts in machine learning evaluate the model's performance even before testing it on real data. Measuring the model's performance with evaluation metrics reveals important model parameters and yields numerical results. The confusion matrix [10] is the most critical metric for evaluating the model.

A confusion matrix has four values used against actual and expected positive and negative classes to compute other metrics. True positives and true negatives are correct positive and negative predictions.There are two types of incorrect predictions: false positives and false negatives. The values in the confusion matrix can be used to derive four important metrics:



	Predicted Class		
Actual Class		Class = Yes	Class = NO
	Class = Yes	True Positive	False Negative

Confusion Matrix

ACCURACY

It is the percentage of predictions that were correct compared to the total number of predictions made. For datasets with equal class distribution, accuracy is the best way to evaluate the model's performance. The following formula can be used to determine accuracy:

(True Positives + True Negatives)/(True Positives + False Positives) = Accuracy

PRECISION

We want the percentage of accurately predicted positive observations. A model's precision determines its accuracy. Even if classes aren't evenly distributed, precision can work. Precision equals true positives plus false positives

precision = true positives or (true positives + false positives)

RECALL OR SENSITIVITY

We're interested in the ratio of correctly predicted positive to all positive observations. A well-performing model has a recall score of at least 50%. Recall can work even if the classes are not evenly distributed. To put it another way, recall = True positives/(False negatives + True positives)

F1 SCORE

It's the average of precision and recall, divided by their respective weights. Uneven class distribution can be measured using the F1 score. F1 can be calculated as F1= 2 * (Recall * Precision) / (Recall + Precision)

5. RESULTS

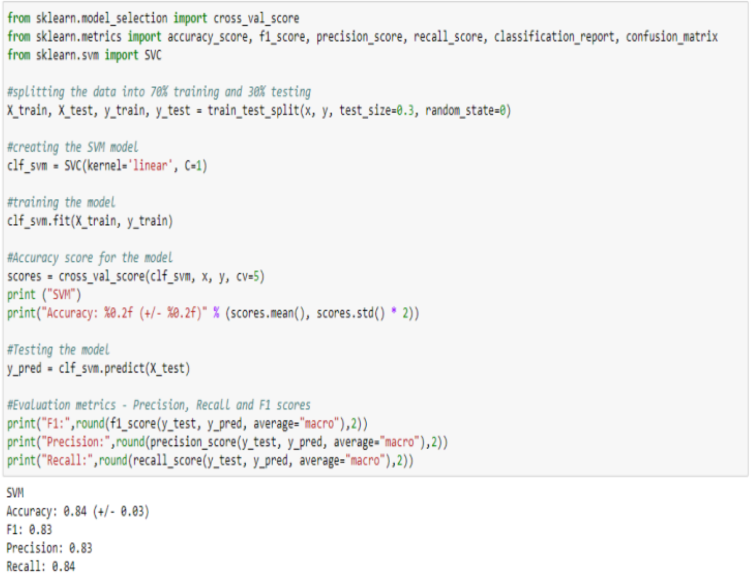
All 34 features were taken into account in this experiment. 2399 audio files were included in the data set. The implementation process is outlined in a concise manner (Fig 2).

To test the model, K-fold cross-validation divides the dataset into training and validation sets.A classifier model is constructed and its parameters are observed using one of the classification algorithms. It is not necessary to fine-tune parameter values at this time.

The data used to train the model is called "training data."

The trained model's accuracy is calculated by comparing it to the validation set.

Finally, the model is put through its paces and evaluated using metrics such as precision, recall, and F1 scores.

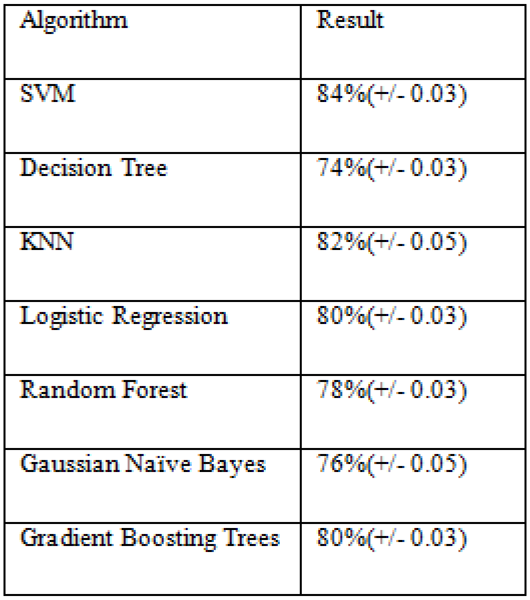


```
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, classification_report, confusion_matrix
from sklearn.svm import SVC

#splitting the data into 70% training and 30% testing
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)

#creating the SVM model
clf_svm = SVC(kernel='linear', C=1)

#training the model
clf_svm.fit(X_train, y_train)

#Accuracy score for the model
scores = cross_val_score(clf_svm, x, y, cv=5)
print ("SVM")
print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

#Testing the model
y_pred = clf_svm.predict(X_test)

#Evaluation metrics - Precision, Recall and F1 scores
print("F1:",round(f1_score(y_test, y_pred, average="macro"),2))
print("Precision:",round(precision_score(y_test, y_pred, average="macro"),2))
print("Recall:",round(recall_score(y_test, y_pred, average="macro"),2))
```

```
SVM
Accuracy: 0.84 (+/- 0.03)
F1: 0.83
Precision: 0.83
Recall: 0.84
```

Fig 2: Code for the implementation

Researchers carried out the same experiment repeatedly with a variety of classification algorithms, and then compared their findings (see Table 2). The accuracy scores were higher than expected. An average of 75% of the students received high marks for their work. The F1 score is the arithmetic mean of precision and recall. SVM's F1 score of 83% makes it a successful algorithm. Tree-based algorithms have been improved. The decision tree classifier improved from 70%+ to 77%+ using Random Forest and Gradient Decent. 80% accuracy using logistic regression. Other metrics average 70%. Nave Bayes and KNN both score 74%+. Algorithm-generated results

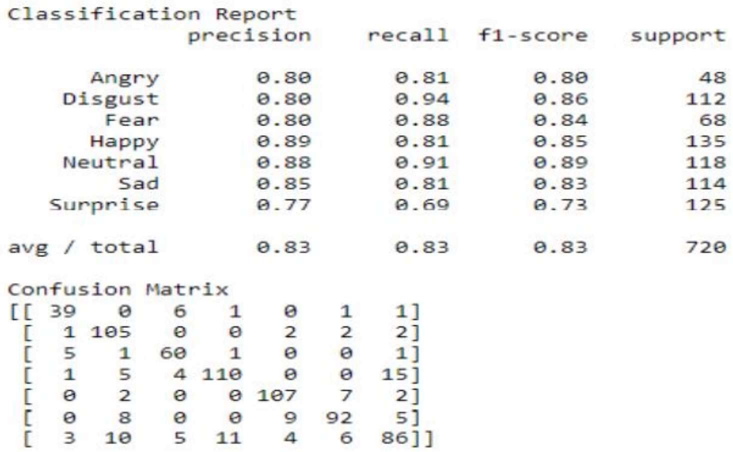


Algorithm	Result
SVM	84%(+/- 0.03)
Decision Tree	74%(+/- 0.03)
KNN	82%(+/- 0.05)
Logistic Regression	80%(+/- 0.03)
Random Forest	78%(+/- 0.03)
Gaussian Naïve Bayes	76%(+/- 0.05)
Gradient Boosting Trees	80%(+/- 0.03)

Fig 3: Performance of Algorithms

Comparison of Results:

Each implementation strategy had a different effect on the various algorithms. The F1 scores can be used to compare accuracy, which is not always a good indicator of a model's quality. Analyzing and contrasting the F1 results from each method yielded some insightful findings. Classifiers such as SVM, KNN, and Logistic Regression perform poorly in the second approach but better in the third approach than in the first. Few tree-based classifiers performed well for the first approach but poorly overall. KNN's performance was constant.



```
Classification Report
              precision    recall  f1-score   support

      Angry       0.80      0.81      0.80        48
    Disgust       0.80      0.94      0.86       112
       Fear       0.80      0.88      0.84        68
      Happy       0.89      0.81      0.85       135
    Neutral       0.88      0.91      0.89       118
        Sad       0.85      0.81      0.83       114
   Surprise       0.77      0.69      0.73       125

avg / total       0.83      0.83      0.83       720

Confusion Matrix
[[ 39   0   6   1   0   1   1]
 [  1 105   0   0   2   2   2]
 [  5   1  60   1   0   0   1]
 [  1   5   4 110   0   0  15]
 [  0   2   0   0 107   7   2]
 [  0   8   0   0   9  92   5]
 [  3  10   5  11   4   6  86]]
```

Fig 4: Classification Report Results

6. CONCLUSION AND FUTURE WORK.

The new era of automation has begun as a result of the increasing growth and development in the fields of AI and machine learning. In most cases, a user's voice commands are used to control these automated devices. If machines could understand the speaker's emotions as well as their words, they would have a significant advantage over current systems (users). Automated call centre conversations, diagnostic tools for therapy, and an automatic translation system are just some of the uses of a speech emotion detection system. Emotion detection systems can be developed using the steps discussed in this paper, and The impact of each step on the final product was tested.. Because of the small number of publicly available speech databases, training a well-trained model proved difficult at first. It was then necessary to conduct a large number of experiments in order to select the best method for extracting features. This was followed by learning about the strengths and weaknesses of every classifying algorithm in terms of recognising emotions. Compared to a single feature, the integrated feature space produced a higher recognition rate at the end of the experiment. The proposed project's efficiency, accuracy, and usability can all be improved in the future. Additionally, the model is capable of detecting mood swings and depressive symptoms. Therapists can use these systems to keep track of their patients' mood swings. Adding a sarcasm detection system to a machine that has emotions is a difficult task. In comparison to other types of emotion detection, sarcasm detection is more difficult because it cannot be determined solely by the speaker's words or tone. Detecting sarcasm in speech can be made easier with the help of sentiment detection that makes use of vocabulary. A speech-based emotion recognition system could be used in a wide variety of ways in the future.

REFERENCES

1. Developer.amazon.com. (2018). Amazon Alexa. [online] Available at: https://developer.amazon.com/alexa
2. Gartner.com. (2018). Gartner Says 8.4 Billion Connected. [online] Available at: https://www.gartner.com/newsroom/id/3598917.

3. H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.
4. L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
5. T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
6. J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009
7. S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.
8. S, Khalid, T, Khalil and S, Nasreen. (2014). 2014 Science and Information Conference, A survey of feature selection and feature extraction techniques in machine learning. PP.372-378.
9. Practicalcryptography.com. (2018). Practical Cryptography. [online] Available at: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstralcoefficients-mfccs/.
10. Exsilio Blog. (2018). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog. [online] Available at: http://blog.exsilio.com/all/accuracy-precision-recall-f1- score-interpretation-of-performance-measures/.
11. Brownlee, J. (2018). A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning - Machine Learning Mastery. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machinelearning/
12. Manda, S., Nalini, ., Kumar, A.A. (2022). Implementation of Bayesian Network Model (BN Trust Model) for IoT Routing. In: Shakya, S., Balas, V.E., Kamolphiwong, S., Du, KL. (eds) Sentimental Analysis and Deep Learning. Advances in Intelligent Systems and Computing, vol 1408. Springer, Singapore. https://doi.org/10.1007/978-981-16-5157-1_31.
13. Kumar, A. Arun, and Radha Krishna Karne. "IIoT-IDS Network using Inception CNN Model." Journal of Trends in Computer Science and Smart Technology 4.3 (2022): 126-138.
14. Sridhar Manda, Dr. Nalini N, “Performance Analysis of Routing Protocols for IoT” International Conference on Electrical, Electronics, Materials and Applied Science - ICEEMAS'17, Swami Vivekananda Institute of Technology, Secunderabad, Telangana, India, 22nd to 23rd Dec 2017, Paper Published in “AIP: Conference Proceedings.
15. Dr. Nookala, Venu and Arun Kumar, A. and Vaigandla, Karthik Kumar, Review of Internet of Things (IoT) for Future Generation Wireless Communications (September 28, 2022). International Journal for Modern Trends in Science and Technology, 8(03): 01-08, 2022, Available at SSRN: https://ssrn.com/abstract=4232170.
16. Dr. Sridhar Manda, Mr. Charanjeet Singh, CVFP: Energy and trust aware data routing protocol based on Competitive Verse Flower Pollination algorithm in IoT, Computers & Security, 2022, 103035, ISSN 0167-4048, https://doi.org/10.1016/j.cose.2022.103035. (https://www.sciencedirect.com/science/article/pii/S0167404822004278).