iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46473

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII August 2022- Available at www.ijraset.com

Singer Identification by Vocal Parts Detection and Singer Classification Using LSTM Neural Networks

K. Rajashekhar Rao[1], K. Sharvani[2], Ch. Shri Vaishnawi[3], Margaret Marina[4]
[1]*Assistant Professor,* [2, 3, 4]*UG Students, Department of CSE, Sridevi Women's Engineering College, Hyderabad, Telangana, India.*

Abstract: ***Identification of singers is considered an important research area in audio signal processing. It has acquired the scientist's intrigues in two primary branches,1) recognizing vocal parts of polyphonic music, and 2) Classifying Singer. Here, we plan to handle the two issues, simultaneously. Techniques like GMM, SVM, and HMM have previously been utilized in classifying singers. In this work, we proposed a system for singer identification using Deep Learning and Feed Forward Neural Networks, which to the best of our knowledge have not been used for this purpose before. Preprocessing involves examining large sets of audio features to extract the most efficient set for the recognition stage. Our work is divided into several stages. To begin with, the vocal parts of all music files are identified utilizing an LSTM network which can perform well for the time series information, for example, audio signals. Then, at that point, an MLP network is integrated and contrasted with an SVM classifier in order to classify the gender of the singers. At last, one more LSTM network is involved to recognize every singer ID, and contrasted with MLP network in a similar task. In each step, various classifiers are analyzed and the outcomes are looked at, which affirm the effectiveness of our technique contrasted with the best in class.***

Keywords: ***Long Short Term Memory, Mel-Frequency Cepstral Coefficients, Support Vector Machine, Multi Layer Perceptron, Non-Negative Matrix Factorization, Non-Negative Matrix Partial Co- Factorization.***

I. INTRODUCTION

Singer ID is a significant area of examination in signal handling. our Identification system as a subcategory of SPID is a hotly debated issue that has drawn the consideration of scientists. It discovers the ID of a singer in music joined by different instruments. With the development of the music business and the rising comfort of complex recording methods, an enormous number of melodies are delivered and played over the Internet, TV, and radio channels consistently. One of the main highlights related to a melody is the vocalist. Many individuals utilize the voice of a singer to recognize the tunes rapidly.

The characteristics of the vocal line make it more vital, and remarkable for individuals as a suitable way to deal with organizing the music databases.

What makes this issue extremely challenging is because of the backup music instruments, background vocal and recording errors. Hence, the SID framework ought to segregate between various wellsprings of sound. Two primary methodologies are proposed such a long way to handle the previously mentioned issue: 1) MFCC features are to be directly computed from polyphonic music and involve them for classifying singer Id, and 2) utilizing a separation technique for extracting the singer's voice out of the music, trailed by a SPID strategy.

The main methodology depends on the understanding that the processed features are adequately strong and strong to separate the singing voice partitions. One of the important stages in the primary methodology is to identify the vocal segments. The precision of the classification system is improved as it is well trained with spoken information. The creator models the question melody as a GMM built utilizing features extracted from the sung notes of the tune. Then, at that point, the model is contrasted with a vocalist based GMM or a GMM built from one more melody sung by a similar artist. The difference was estimated by the Kullback Leibler uniqueness.

The subsequent methodology (i.e., the partition of the singer's voice) is investigated in and. This work plays out a non-negative network factorization (NMF) to isolate the vocal voice from polyphonic music. A two-stage strategy has been proposed.

The primary stage utilizes non-negative matrix co-factorization (NMPCF) and the subsequent stage separates the pitch and harmonic parts of the artist voices.

Our technique is close to the main methodology, in which we give an evaluation of various characterization techniques for polyphonic music.

We characterize our list of features, and then exploit artificial neural networks and recurrent deep neural networks in these stages and implement them.

©IJRASET: All Rights are Reserved

A. *Machine Learning*

Machine Learning helps us to build a mathematical model for understanding data, we give input data to the model for training the model, and based on this information, the model learns from it, analyzes the pattern, and finds the best way to predict the results, Hence machine learning models are helpful in our system.

B. *Support Vector Machine*

This is one of the machine learning techniques, where This model is trained with input MIR1K dataset which contains gender and singer data, we build two SVM models one is to identify gender and other is to classify singer, Firstly the given data is split into train and test data and then we fit trained input and output data to our model and then given test data input to this model for which output results are predicted using predict function and then compare with actual test output data to calculate accuracy and this accuracy is used to compare with our proposed model called LSTM model accuracy.

C. *Multi Layer Perceptron*

As discussed above, the MLP model was also trained and tested with MIR1K dataset, but the difference between SVM and MLP is, that MLP includes multiple layers to process the given input data which improves the accuracy of results compared to the SVM model, Now the accuracy of MLP is also compared with LSTM model and Finally from this we can conclude that LSTM accuracy is high compared to other two models.

II. RELATED WORK

A. *Mel Frequency Cepstral Coefficients*

MFCC is utilized to separate the novel elements of the human voice. It addresses the transient power range of the human voice. It is utilized to ascertain the coefficients that address the recurrence Cepstral these coefficients depend on the straight cosine change of the log power range on the nonlinear Mel size of recurrence. In the Mel scale, the recurrence groups are similarly divided which approximates the human voice more precisely. Condition (1) is utilized to change the ordinary recurrence over completely to the Mel scale the equation is utilized as $m=2595 \log10 (1+f/ 700)$ (1)

Mel scale and ordinary recurrence scale are referred to by characterizing the pitch of 1000 Mel to 1000 Hz tones, 40db over the audience's limit. Mel recurrence is similarly divided on the Mel scale and is applied to direct space under 1000 Hz to linearize the Mel scale values and separates filter values up to 1000HZ using logs to calculate the power of log of Mel scaled signal. Mel recurrence wrapping is the better portrayal of voice. Voice highlights are addressed in MFCC by partitioning the voice signal into outlines and windowing them then, at that point, taking the Fourier change of a windowing signal. Mel scale frequencies are acquired by applying the Mel channel or three-sided band pass filter to the changed sign. At last change to the discrete structure by applying DCT presents the Mel Cepstral Coefficients as acoustic highlights of the human voice.

B. *Hidden Markov Model*

HMM is characterized as a limited state machine with a fixed number of states. It is a statistical method for characterizing the spectral properties of a speech signal. Probabilities are classified into two types. It tells for a given input model must move from one state to another is called transition in a given time and it is fixed and depends on a given input. In Hidden Markov model the states are not they are not shown directedly and covered up yet the result is apparent which is reliant on the states. Yield is created by likelihood dissemination over the states. It gives the data about the success of states however the boundaries of states are still stowed away.

MFCC is utilized to separate the voice elements from the voice test. Furthermore, HMM is utilized to perceive the speaker based on separated highlights. For this, it first trains the extracted highlights in the configuration of HMM boundaries and to see the log value of the whole voice for recognition. Forward in backward estimation is utilized to prepare the extracted elements and track down their parameters. HMM, model perceived the speaker based on log probability. It recalculates the log probability of voice vector and Relative insensitivity to gap length is an advantage of LSTM over RNNs and hidden Markov models.

C. *Gaussian Mixture Model*

A Gaussian mixture model is a probabilistic model that expects every one of the information focuses is created from a combination of a limited number of Gaussian distributions with unknown parameters.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Gaussian Mixture Model will be utilized as a characterization method. GMM classifier is a sort of classifier which consolidates the upside of parametric and non-parametric methods. GMM doesn't need the capacity of whole preparation vectors to make an order. It is a truly adaptable model that can adjust to include practically any distribution of information. The GMM model is trained with extracted features and when tested with input it produces less accurate results if the voice has background noise and GMM model uses the Expectation Maximization Algorithm to classify results

These are previous work algorithms that have limitations, It can't produce results within in given time and it always does not produce correct results which leads to low accuracy compared to these two models GMM and HMM which are used in the existing system, our proposed model called Long short term memory have high accuracy.

III. PROPOSED ALGORITHM

In our paper we proposed a new model called Long Short Term Memory(LSTM) which is a deep recurrent neural network, it is called deep because many neurons are included to process data by moving from one to another and it includes many hidden layers, LSTM is better compared to standard feedforward neural networks because of its previous connections in the network provides feedback to next layer, LSTM also helps in handwriting recognition, speech recognition, and helps to detect anomaly in network traffic or intrusion detection systems.

An ordinary LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. The forget gate makes a decision whether data have to be remembered and given as feedback to the next layer or not and as we know LSTM has one input layer called the convolution2D layer and hidden layers are MaxPooling2D, Flatten, Bidirectional and these hidden layers are built in the model in repeated fashion and output layer is called Dense Layer. LSTM networks are appropriate for grouping, handling, and making expectations given time series information since there can be slacks of obscure length between significant events in a period series.

IV. TRAINING

LSTM is a supervised model first it is trained in several epochs by shuffling the random data and uses algorithms like optimization, and gradient descent methods are used to calculate gradient which helps us during the optimization process, so it can change the values of the weight of LSTM network with respect to error and corresponding weight, Problem occurs in gradient descent method when error gradients are disappeared exponentially in proportion to the size of the delay between significant events. we have the option to recover these error values from the output layer and this process continues to feed errors to previous layers until the model learns to correctly fit data in the model and remove unwanted data which raises errors.

We build two LSTM models, Identification model is used to identify gender, and the classification model is to identify the singer, First, the model is built by using a convolution 2D input layer and uses above mentioned hidden layers and an output layer called the dense layer which uses activation function like soft max and relu which makes us classify singer ID, and then model is compiled using adam optimizer to calculate loss and metrics, and then fit function is used to fit train input and output data into the model and then weights are saved based on output labels and convert this file to JSON format, the history data file which contains information used to build model is dumped in pickle file, so if once the model is built, it can be used directly instead of building again, and also we can load data from pickle file to calculate accuracy

The MIR1K dataset is given as input to the LSTM model and from this MFCC Features are extracted and trained to model and then the testing process starts where new data or test data is given and compare actual results with predicted results to calculate accuracy.

V. RESULTS AND DISCUSSION

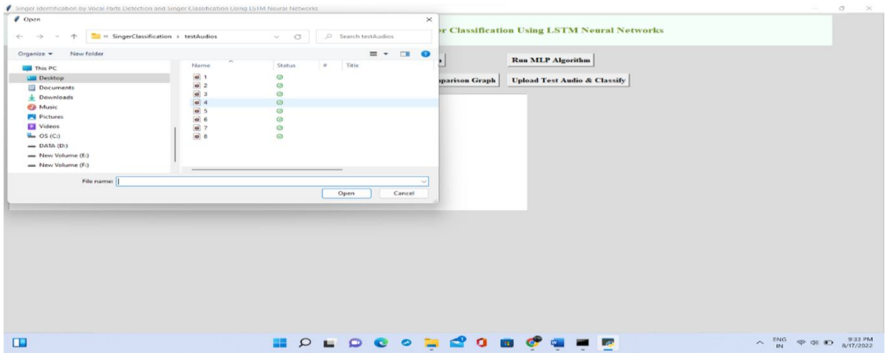


Fig 4.1 Upload audio clip

In the above figure, we can observe after clicking on Upload Test Audio & Classify button, It opens the test audio file dialog box which contains audios of unseen data, and next select any audio clip and click on open, Now using predict function code it extracts mfcc features of a clip using librosa and then ravels it to convert into a sequence of features and these are resized and reshaped into an array and to utilize data processing technique converts it into NumPy array, then give this data into the identification and classification model, argmax function classification depends on maximum probability to which result belongs like if gender voice is of 80% as female voice and 20% as male voice then it produces a result as female, In this similar singer can also be classified and observe the below results after uploading audio clip. It displays results on the window as Uploaded Vocal Parts identified as Female Uploaded Audio file classified as singer name: Annar.

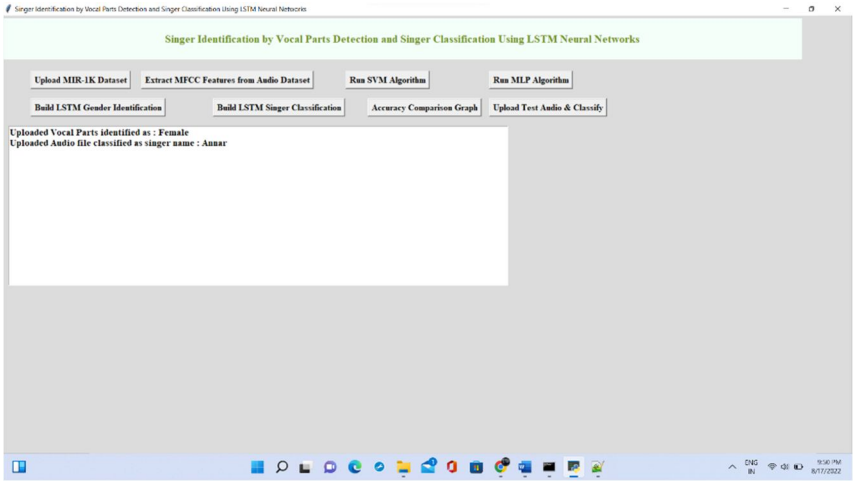


Fig 4.2 Classify Audio

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a supervised system for singer identification that combines several stages of processing blocks involving mainly deep recurrent neural networks. The strategy used here is to first recognize the vocal parts and classify the gender of the singer, and then perform the identification of the singer. Together with a suitable feature vector, the obtained results show the efficiency of our strategy and the proposed system in terms of determining the vocal parts of the singer and incorporating the optimal feature vectors using a deep auto-encoder. Moreover, a similarity measure could be defined to detect similar styles among singers. The use of other processing steps, such as drop-out and batch normalization, will certainly increase the overall accuracy of the system as well.

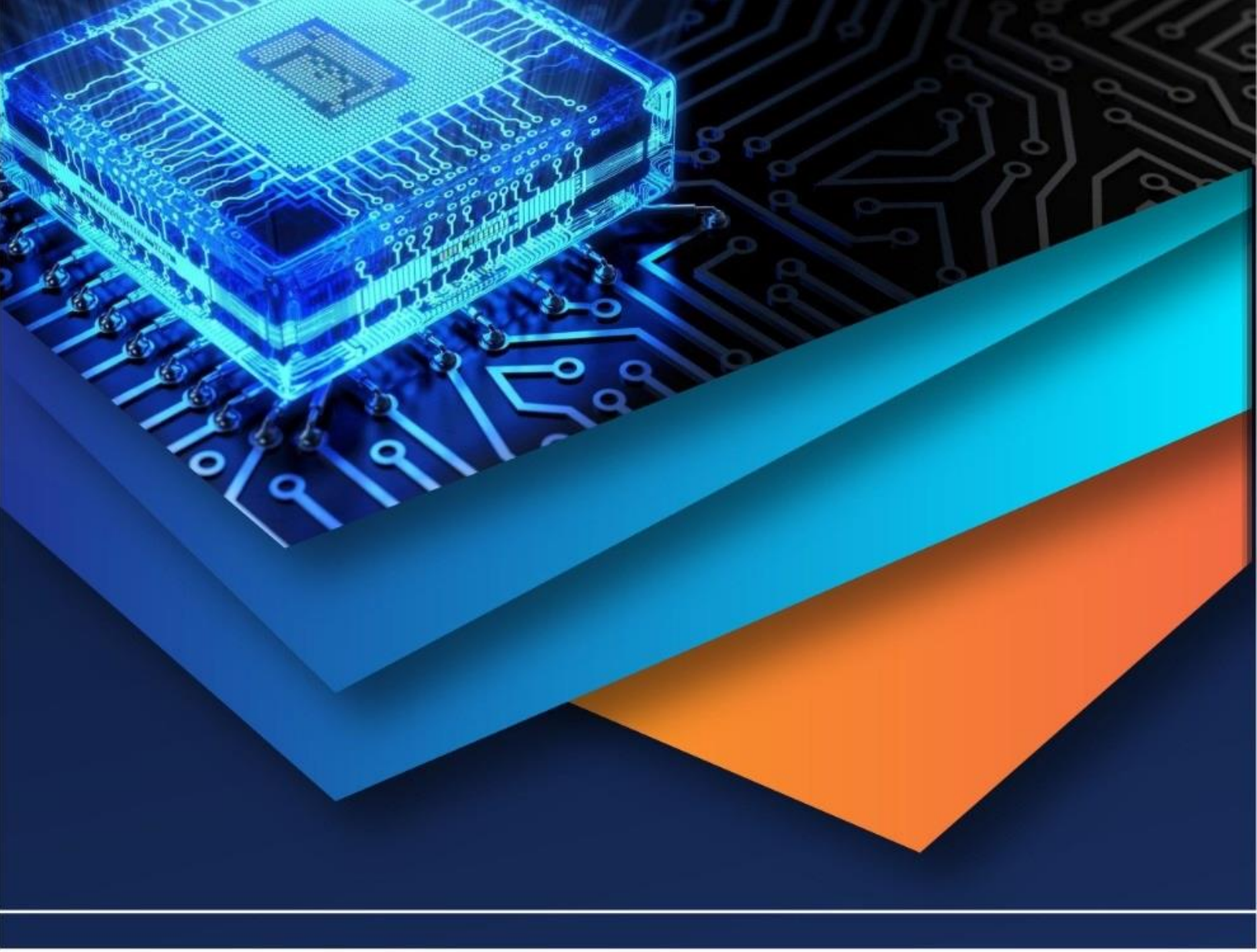
For our future works, it would be intriguing to decide the artist's vocal sorts and integrate the ideal element vectors utilizing a profound Auto-Encoder. Further, a proportion of comparability could be characterized to recognize comparative styles among vocalists. Utilizing further handling tuning steps, for example, drop out and batch normalization will also increase the system overall accuracy.

REFERENCES

[1] survey of audio-based music classification and annotation." IEEE transactions on multimedia 13, no. 2 (2011): 303-319.
[2] Tsai, Wei-Ho, and Hao-Ping Lin. "Background music removal based on cepstrum transformation for popular singer identification." IEEE Transactions on Audio, Speech, and Language Processing 19, no. 5 (2011): 1196-1205.
[3] Pikrakis, Aggelos, Yannis Kopsinis, Nadine Kroher, and José- Miguel Díaz-Báñez. "Unsupervised singing voice detection using dictionary learning." In Signal Processing Conference (EUSIPCO), 2016 24th European, pp. 1212-1216. IEEE, 2016.
[4] Song, Liming, Ming Li, and Yonghong Yan. "Automatic vocal segments detection in popular music." In 2013 Ninth International Conference on Computational Intelligence and Security, pp. 349-352. IEEE, 2013.
[5] Tsai, Wei-Ho, and Hsin-Chieh Lee. "Singer identification based on spoken data in voice characterization." IEEE Transactions on Audio, Speech, and Language Processing 20, no. 8 (2012): 2291-2300.
[6] Regnier, Lise, and Geoffroy Peters. "Singer verification: singer model. vs. song model." In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 437-440. IEEE, 2012.
[7] Zhu, Bilei, Wei Li, Ruijiang Li, and Xiangyang Xue. "Multi-stage non-negative matrix factorization for monaural singing voice separation." IEEE Transactions on audio, speech, and language processing 21, no. 10 (2013): 2096-2107.
[8] Hu, Ying, and Guizhong Liu. "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification." IEEE Transactions on Audio, Speech, and Language Processing 23, no. 4 (2015): 643-653.

[9] Logan, Beth. "Mel Frequency Cepstral Coefficients for Music Modeling." In ISMIR, vol. 270, pp. 1-11. 2000.
[10] Eronen, Antti, and Anssi Klapuri. "Musical instrument recognition using cepstral coefficients and temporal features." In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, vol. 2, pp. II753- II756. IEEE, 2000.
[11] S. Kooshan, H. Fard and R. M. Toroghi, "Singer Identification by Vocal Parts Detection and Singer Classification Using LSTM Neural Networks," 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), 2019, pp. 246-250, doi: 10.1109/PRIA.2019.8786009.
[12] Xulong Zhang, Jiale Qian, Yi Yu, Yifu Sun, Wei Li, "Singer Identification Using Deep Timbre Feature Learning with KNN-NET", ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3380-3384, 2021
[13] Sharmila Biswas, Sandeep Singh Solanki, "Speaker recognition: an enhanced approach to identify singer voice using neural network", International Journal of Speech Technology, vol.24, no.1, pp.9, 2021.
[14] Graves Alex, Navdeep Jaitly and Abdel-Rahman Mohamed, "Hybrid speech recognition with deep bidirectional LSTM", Automatic Speech Recognition and Understanding (ASRU) 2013 IEEE Workshop on, pp. 273-278, 2013.
[15] F. Gers, N. Schraudolph and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks", Journal of Machine Learning Research, vol. 3, pp. 115-143, 2002.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)