WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language

M. Sharmila Devi[1], M. Poojitha[2], R. Sucharitha[3], K. Keerthi[4], P. Manideepika[5], C. Vasudha[6]

[1] Assistant Professor in Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.
[2,3,4,5,6] Student, Department of Computer Science and Engineering, Santhiram Engineering College, Nandyal, Kurnool, Andhra Pradesh, India.
DOI: https://doi.org/10.55248/gengpi.2023.4.4.34752

ABSTRACT

Natural Language Processing (NLP) is a field of study that develops software capable of interpreting human speech for mechanical use. Words are the building blocks of advanced grammatical and semantic analysis, and word segmentation is often the first order of business for natural language processing. This paper introduces the feature extraction method of deep learning and applies the ideas of deep learning to multi-modal feature extraction in order to address the practical problem of huge structural differences between different data modalities in a multi-modal environment. In this study, we present a neural network that can process information from several sources at once. Each mode is represented by a separate multilayer sub-neural network structure. It's purpose is to transform features from one mode to another. In order to solve the issues of current word segmentation techniques not being able to ensure long-term reliance on text semantics and lengthy training prediction time, a hybrid network English word segmentation processing approach is presented. This approach uses the BI-GRU (Bidirectional Gated Recurrent Unit) to segment English words and the CRF (Conditional Random Field) model to sequentially annotate sentences, which eliminates the long-distance dependency of text semantics and reduces the time needed to train the network and predict its performance. Compared to the BI-LSTM- CRF (Bidirectional- Long Short Term Memory-Conditional Random Field) model, the experimental results reveal that this technique achieves equivalent processing effects on word segmentation, while also boosting processing efficiency by a factor.

Keywords: BI-LSTM-CRF, BI-GRU-CRF, NLP, Feature Extraction

1. INTRODUCTION

With the rapid development of Internet information technology and the continuous advancement of science and technology, a large amount of data of various types and structures have been accumulated in the real life and scientific research fields. Different information modalities together constitute multi- modal data for the same problem. NLP (Natural Language Processing) is one of the key technologies for realizing human-computer interaction and artificial intelligence. It is listed as the three major elements of artificial intelligence research with voice processing and image processing. In the early days of NLP research, the main focus was on the analysis of language structure, technology-driven machine translation, and language recognition. The current focus is on how NLP can be used in the real world. However, the training of deep frames is a difficult task, and traditional shallow proven methods that have proven effective cannot be transplanted into deep learning to ensure their effectiveness. Another realistic problem is that there is no necessary connection between increasing the layer structure and obtaining better feature representations. For example, in a neural network, the more hidden layers, the less impact the first layer in the back propagation algorithm. When using the gradient descent algorithm, it will also fall into the local optimum and lose the effect of continued transmission. Related scholars have proposed a word segmentation algorithm based on supervised machine learning. This method implements a word-based word segmentation system. The main innovation is to use the maximum entropy model as a tokenizer to automatically label characters. This method has the highest recall rate of 72.9% in the AS2003 closed test experiment. In the method of English word segmentation based on the dictionary and rules, it mainly focuses on the word segmentation algorithm and dictionary structure. The advantages of dictionary-based and rule- based methods are simple, easy to implement, and suitable dictionaries can be formulated according to special scenarios. In addition, in systems that require real- time performance, dictionary-based and rule-based methods are often more suitable because of their high efficiency. The disadvantages are: there is a problem of word segmentation ambiguity; there is no universal standard for word division, so the quality of the dictionary cannot be clearly defined. The dictionary has a great impact on the segmentation result.With the advent of the era of big data, data has become more and more in natural language processing problems. Improving these labeling problems to support parallel computing and being able to perform parallel learning on large-scale training data has also become a research hotspot. Parallel learning is currently supported, including maximum entropy models and conditional random field models. Some researchers have proposed the technical route of``understand first and then segmentation " . The idea of understanding the segmentation first is to solve the lack of global information in the traditional matching segmentation, while the statistical method lacks the structural information of the sentence. Relevant scholars use deep learning to perform sequence labeling in the NLP field. It can also add a sequence labeling model to combine with the output of the previous neural network to extract the best labeling sequence through the Viterbi algorithm. Related scholars have proposed an open domain question answering system based on relationship matching. The problem analysis problem based on relation matching is solved through the associated data in the question answering system. The fragments in the question match the binary relationship in the triples and are automatically collected using the relational text pattern. Existing models do not take into account the importance of different modalities for the

current learning task, but only focus on how to effectively use multiple modalities for feature extraction at the same time. Moreover, the selection of modals and the filtering of harmful modals are not involved, and this issue is also an important issue addressed in this paper. In terms of word segmentation processing, in view of the problems that existing word segmentation methods can hardly guarantee long-term dependency of text semantics and long training prediction time, a hybrid network English word segmentation processing method is proposed. Experimental results show that this method improves the efficiency of natural language processing. In terms of English word segmentation, since traditional machine learning methods cannot solve the long- distance dependencies of texts, it is difficult to analyze the information contained in the problem as a whole and grasp the user's true intention. In order to solve the above problems and save the relevance of the forward and reverse information of the text, this paper uses BI-GRU (Bidirectional Gated Recurrent Unit) neural network and combines the CRF (Conditional Random Field) model to solve the problem of sequence labeling at the sentence level analysis, based on BI-GRU-CRF (Bidirectional-Gated Recurrent Unit- Conditional Random Field) hybrid network English word segmentation processing method. Specifically, the technical contributions of this article are summarized as follows Firstly, a multimodal fusion feature extraction method is proposed. The problem of heterogeneity of multi-modal data is solved through the feature transformation of deep neural networks. Secondly, in view of the problems that traditional neural network models cannot capture the long-distance dependencies of text and the long cost of training and prediction of LSTM (Long Short Term Memory) neural networks, a word segmentation processing method based on BI-GRU-CRF hybrid network is proposed. Thirdly, the proposed method is tested from two aspects, accuracy and timeliness. According to these two sets of experiments, the proposed hybrid network word segmentation processing method has good performance in English word segmentation processing.

2. RELATED WORK

‘‘Deep learning based single image super-resolution: A survey,’’

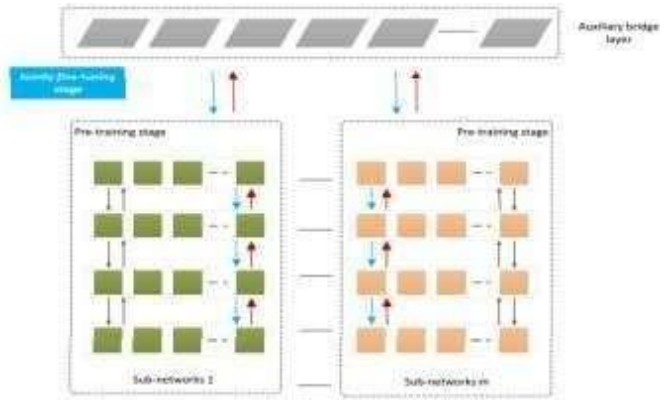
An image may be "super-resolved" by taking a single sample of low-resolution pixels and expanding it into a larger, higher-resolution picture, or by taking many samples of the same image and combining them. Recent developments in this area include interpolation-based, reconstruction-based, and learning- based methods because of its many potential uses. Recently, learning-based approaches have received a lot of interest because of their potential to restore high-frequency information that were lost owing to a lack of resolution. Using Convolutional Neural Networks, this study summarises the vast majority of previous work in the field. Also covered are widespread problems with super-resolution algorithms, including imaging models, the improvement factor, and evaluation criteria.

‘‘Automatic modulation classification: A deep learning enabled approach,’’

This research uses a deep learning technique to examine automatic modulation categorization (AMC), which has important significance in both civilian and military applications. Maximum likelihood based (ML-AMC) and feature based AMC are the two main types of traditional AMCs. However, ML-AMCs are not yet widely used because of their high computational cost and the fact that the manually derived characteristics need specialised expertise. Therefore, we present a fully-automated AMC (CNN-AMC) that uses a convolutional neural network to automatically extract features from a lengthy symbol-rate observation series and a signal-to-noise estimation (SNR). In order to handle the wide range of input dimensions, CNNAMC uses a unit classifier. Due to the difficulty faced by the CNN-complex AMC's model and tasks, an unique two-step training is suggested, and transfer learning is also incorporated to boost retraining efficiency. The simulation findings demonstrate that the CNN-AMC can outperform the feature-based technique, and provide a closer approximation to the ideal ML-AMC when many digital modulation schemes are examined in different circumstances. CNN-AMCs also exhibit a degree of resilience in the face of estimate errors in carrier phase offset and SNR. The deep-learning-based technique outperforms the ML-AMC by a factor of 40-1700 when using parallel computing for inference.

3. METHODOLOGY

Our suggested model's general structure is a hierarchical one. A root network and an upper layer network are the two basic divisions. When training the root network's parameters, the BP (Back Propagation) technique is utilised. In order to make simultaneous adjustments to all of the sub-networks, the function loss established in the auxiliary bridge layer is employed.



There is a lot of variation across the various data types, which is a major challenge for multimodal feature learning tasks. The raw data from each modality exists in its own unique feature space. Most multimodal learning strategies rely on mapping the various modalities onto a common subspace. In particular, deep learning models excel in extracting latent hierarchical feature expressions and transforming raw data into more useful formats for further analysis. The model suggested in this study takes use of deep neural networks' capabilities to filter out the various data modalities. Root network architecture is seen in figure

Various data channels represent various groups of interconnected networks. Here, we refer to the ith hidden layer as h I m and the ith connection weight as w I m, where nm is the total number of hidden layers in the sub-networks m corresponding to the mth mode.

Bi Directional GRU With CRF Model

A Bi-Directional GRU (Gated Recurrent Unit) with CRF (Conditional Random Field) model is a type of sequence labeling model that is commonly used in natural language processing tasks such as named entity recognition and part-of-speech tagging.

The Bi-Directional GRU is a type of recurrent neural network that is capable of processing sequences in both forward and backward directions. This allows the model to take into account both past and future context when making predictions. The GRU cell is a variant of the more commonly used LSTM (Long Short-Term Memory) cell, which is designed to handle the vanishing gradient problem that can occur in recurrent neural networks.

The CRF layer is used to model the dependencies between adjacent labels in the sequence. In a typical sequence labeling problem, the labels assigned to each token in the sequence are not independent of each other, but rather depend on the labels of the neighboring tokens. The CRF layer models these dependencies explicity, allowing the model to make more accurate predictions.

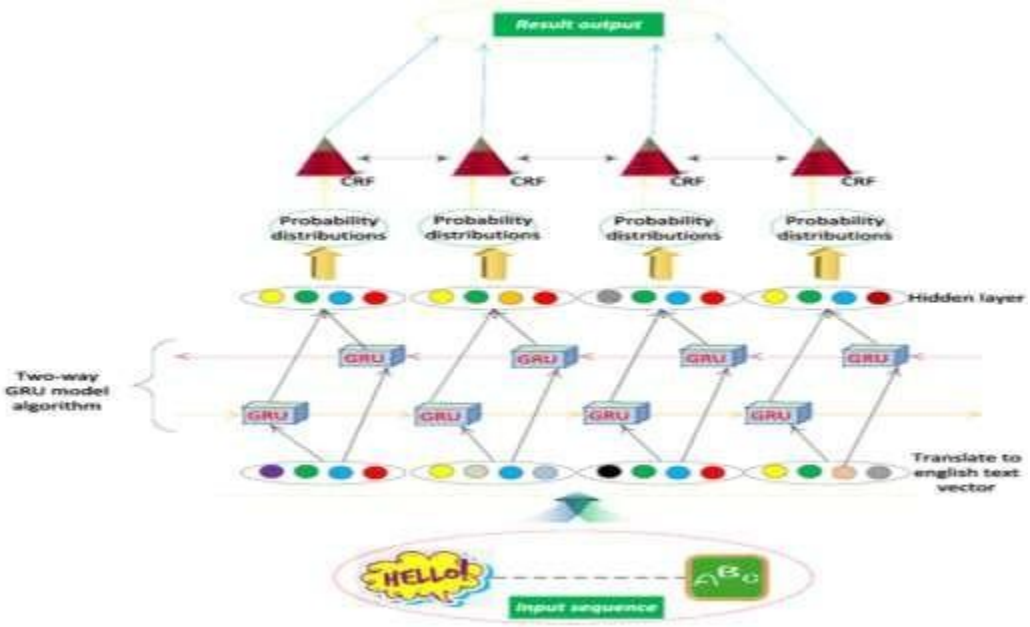


Fig: Schematic diagram of the word segmentation structure of

BI-GRU-CRF hybrid network

The BI-GRU-CRF model is to combine the bidirectional GRU network with the CRF layer. It adds the CRF layer after the output layer of the bidirectional GRU network. The model can effectively utilize the bidirectional GRU network to obtain the past and future information in the input text as the features and predict the whole tag sequence through the CRF layer, so as to achieve the optimal tagging of the text sequence

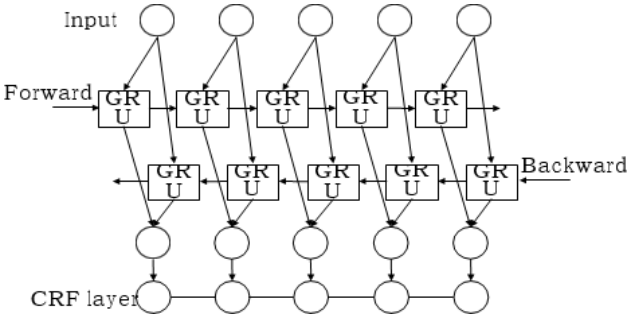


Fig: BI-GRU-CRF network

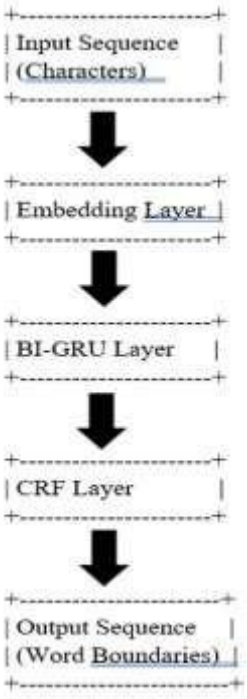


Fig: Architecture of BI-GRU with CRF

In this diagram, the input sequence of characters is first passed through an embedding layer to convert it into a sequence of dense vectors. This is followed by the BI-GRU layer, which processes the sequence in both forward and backward directions to capture contextual information. The output of the BI-GRU layer is then passed through the CRF layer, which models the probability of a sequence of labels (in this case, the boundaries between words) given the input sequence. Finally, the output sequence of word boundaries is generated by the CRF layer.

4. RESULTS AND DISCUSSION

In this study, the author combines NLP with DL to recognise and segment English words, and then compares the efficacy of two different DL neural networks, BI-GRU (Bidirectional Gated Recurrent UNIT) and BI-LSTM (Bidirectional Long Short Term Memory). In the case of neural network models, the best model is the one that produces the least LOSS.

Word segmentation is the process of extracting meaningful information from a given set of data; for example, if the input is "comment sunder questioning," the segmented output will be "comment sunder questioning." In order to achieve this using a neural network, we will first train it with all possible words and their IDs; then, whenever we give such input, the neural network will predict the IDs of the words in the set; finally, it will convert the IDs into words and check them against the.

By uploading some meaning less data to the algarithm, we can get some meaning full data with high accuracy.

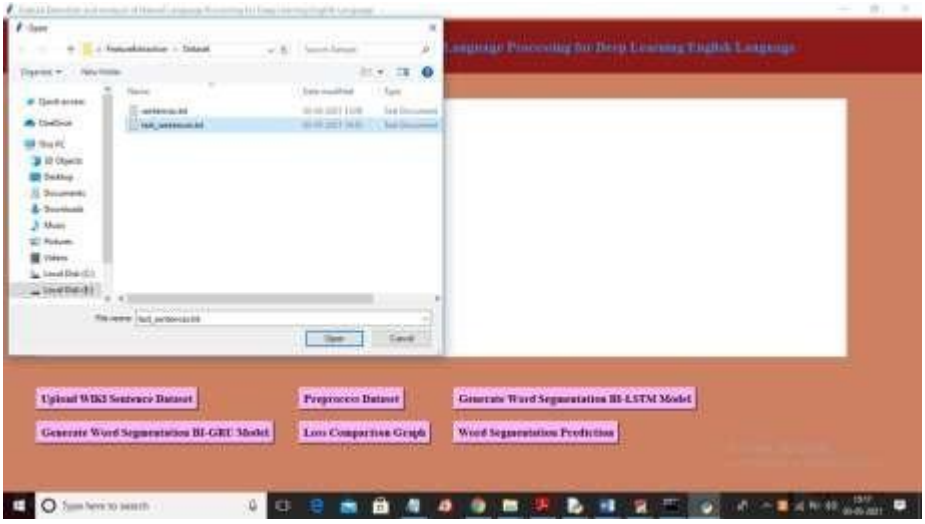


Fig: Word Segmentation Prediction

This is the accuracy graph for the given input text.

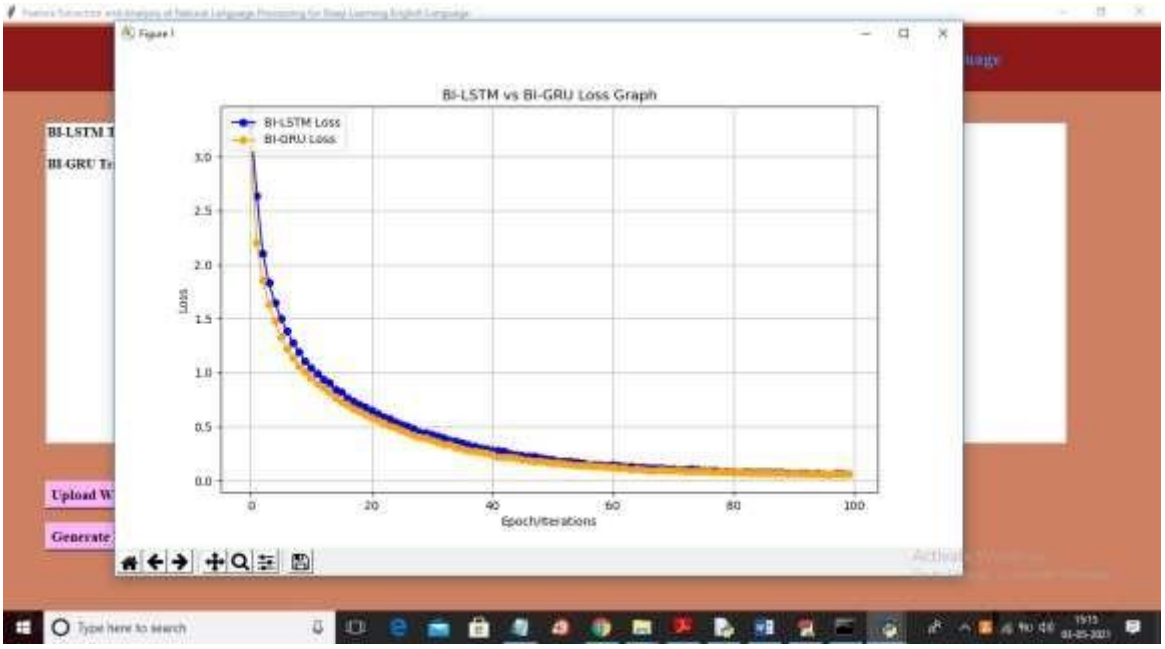


Fig: Loss Comaparison Graph

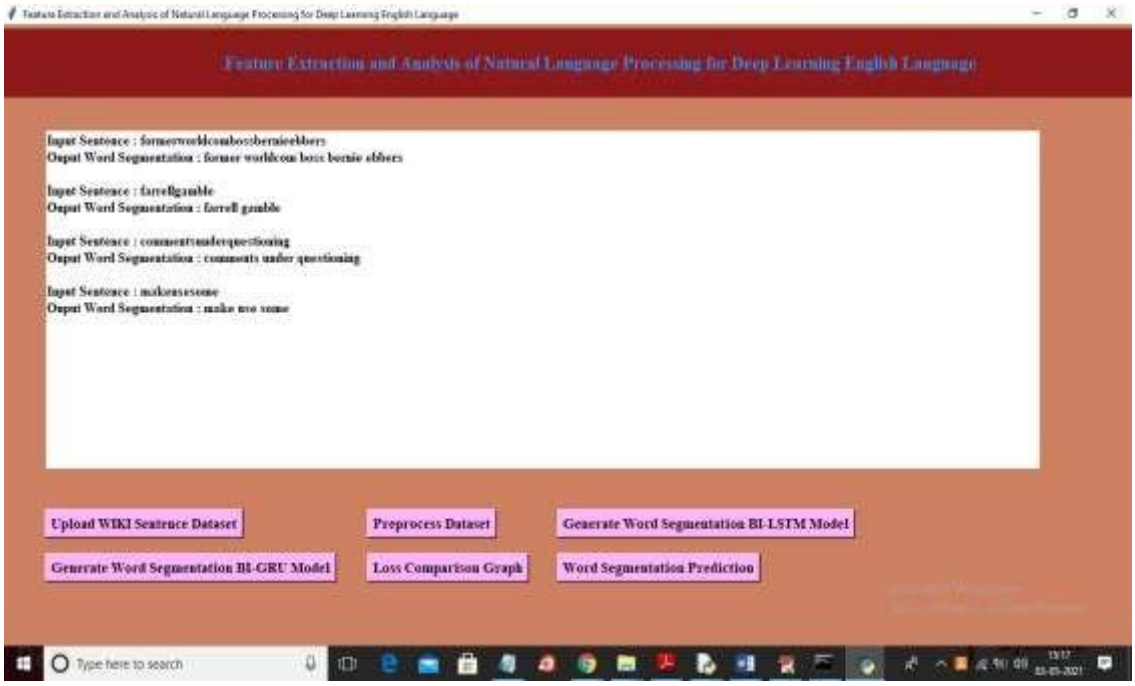


Fig: Word Segmentation

5. CONCLUSION:

In this research, we present a deep neural network–based approach for extracting multimodal shared features, describe the program's full model architecture, and describe how the proposed model was designed and trained. Extensive testing against other models was performed to ensure the viability of the one suggested. The experimental findings validate the effectiveness of the proposed multimodal fusion feature extraction model in efficiently extracting low-dimensional fusion features from the original multiple high-dimensional data. A significant discriminative ability and a reduced feature dimension are both present in the resultant fusion feature expression, demonstrating the robustness of the suggested model. In this paper, LSTM and GRU for English word segmentation are analysed in detail. According to the results of the studies and analyses, both networks are able to address the issue of conventional word segmentation in the text's long-range dependence connection. But LSTM sophisticated structure means that training and forecasting

the data set takes a long period. In comparison to the LSTM, the GRU is a much more straightforward model. There is less time spent on training and prediction, and the structure is straightforward. This research offers a hybrid neural network word segmentation processing approach, combining BI-GRU and CRF models, on the basis that the two-way network is better able to capture the contextual link between meanings. From the experimental data, it is clear that the model suggested in this study outperforms its predecessors in terms of accuracy, and that the proposed technique is 1.62 times quicker than the BILSTM-CRF network word segmentation method in terms of training speed. When compared to the BI-LSTM-CRF network-based word segmentation approach, this one is on average 1.94 times faster. The hybrid network word segmentation approach suggested in this study shows promising results on these two datasets for the purpose of English word segmentation. To further improve the model's capacity for learning, future study may investigate the effect of using a variety of feature extraction and feature selection techniques. Although it does not directly learn from the raw data, the suggested technique instead considers the many characteristics extracted from each kind of data as a separate mode. More investigation is required into how to extract the low-dimensional characteristics used in multi-modal fusion from the raw multi-modal data.

6. REFERENCES

[1] M. Sharmila Devi, Farooq Sunar Mahammad, D. Bhavana, D. Sukanya, TV. Sai Thanusha, M. Chandrakala, P. Venkata Swathi "Machine Learning Based Classification and Clustering Analysis of Efficiency of Exercise Against Covid-19 Infection" , JOURNAL OF ALGEBRAIC STATISTICS, Vol. 13, no. 3, pages. 112-117, June 2022.

[2] M. Sharmila Devi, "A comparative Study of Classification Algorithm for Printed Telugu Character Recognition",International Journal of Electronics Communication and Computer Engineering,Vol.3,no.3,pages. 633-641,2012.

[3] V Lakshmi Chaitanya, "Machine Learning Based Predictive Model for Data Fusion Based Intruder Alert System", journal of algebraic statistics, Vol. 13,no. 2,pages. 2477-2483, June 2022.

[4] V Lakshmi Chaitanya,'' Apriori vs Genetic algorithms for Identifying Frequent Item Sets'', International journal of Innovative Research &Development,Vol.3,no.6,pages. 249-254,June 2014.

[5] B.Swarajya Lakshmi, "Fire detection using Image processing",Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.10 No.2, 2021, pp.14-19, 2021.

[6] B.Swarajya Lakshmi, —"Identity-Based Proxy-Oriented Data Uploading and Remote Data Integrity checking in Public Cloud", International Journal of Research Vol. 5, no.22,pages. 744-757, 2018.

[7] Sunar mohammed Farooq,'' Static Peers for Peer-to-Peer Live Video Streaming'',Inter national journal of Scientific Engineering and Technology Research, Vol.05,No.34, Pages:7055-7064, October-2016.

[8] Farooq Sunar Mahammad, P Bhaskar, A Prudvi, N Yugandhar Reddy, P Jaswanth Reddy, "Prediction Of Covid-19 Infection Based on Lifestyle Habits Employing Random Forest Algorithm", journal of algebraic statistics,Vol.13,No.3,pages.40-45,June 2022.

[9] Sunar Mohammed Farook, K NageswaraReddy,'' Implementation of Intrusion Detection Systems for High Performance Computing Environment Applications'', Inter national journal of Scientific Engineering and Technology Research ,Vol.04, N0.41, Pages:8958-8963, October 2015.

[10] MV Subramanyam, "Automatic feature based image registration using SIFT algorithm", conference of 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12), pages. 1-5, July 2012.

[11] MV Subramanyam, Mahesh, "Feature based image registration using steerable filters and Harris algorithm", pages. 95-99, January 2012.

[12] MV Subramanyam, K Satya Prasad, PV Gopi Krishna Rao,''Robust control of steam turbine system speed using improved IMC tuned PID controller'', Procedia Engineering, Vol.38,Pages. 1450-1456,January 2012.

[13] MV Subramanyam, Giri Prasad,'' A New Approach for SAR Image Denoising'', International Journal of Electrical and Computer Engineering,Vol.5,No.5, Pages. 984-991, October 2015.

[14] O. Bernard et al., ''Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved,'' IEEE Trans. Med. Imag., vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[15] Q. Xia, S. Li, A. M. Hao, and Q. P. Zhao, ''Deep learning for digital geometry processing and analysis: A review,'' J. Comput. Res. Develop., vol. 56, no. 1, pp. 155–182, 2019.