International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 27, No. 4, August, 2022, pp. 679-684
Journal homepage: http://iieta.org/journals/isi

Recursive LSTM for the Classification of Named Entity Recognition for Hindi Language

Rita Shelke*, Sandeep Vanjale

Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune 411030, India

Corresponding Author Email: ritashelke@gmail.com

https://doi.org/10.18280/isi.270420

Received: 28 May 2022
Accepted: 31 July 2022

Keywords:
NER, Hindi language, BiLSTM, ML

ABSTRACT

NER assumes a key part in Information Extraction from reports (for example email), conversational information, and so forth. Many tongue handling applications, for example, data recovery, question responding to, and machine interpretation, depend on NER. It tends to be challenging to determine the ambiguities of lexical components utilized in a text arrangement. There is too much work has been already done in English language but there is a need to improve accuracy for the NER in Hindi language. In this research researcher are minimize chances of misclassification by using different classifier namely location, name, weather etc. BiLSTM Development of a NER framework for Indian languages is a similarly troublesome task. In this paper, Researcher have done the different research to contrast the aftereffects of NER and typical implanting and quick text implanting layers to examinations the exhibition of word installing with various bunch sizes to prepare the profound learning models. In this paper, Researcher have done the different examinations to contrast the consequences of NER and typical implanting and quick text installing layers to investigations the presentation of word inserting with various group sizes to prepare the profound learning models. The value of the precision of proposed system architecture is 76.13% which is way more than other system architectures. Also, the value of recall and F1-score of proposed system architecture is 71.49 and 74.26 respectively. So, by comparing proposed system architecture with existing SpaCy, CoreNLP and NLTK it is easy to conclude that proposed system architecture is reliable in all the sense.

1. INTRODUCTION

Statistical language demonstrating is important to catch the normality of regular language to work on the exhibition of different normal language applications. Measurable language models give a correctly classified appropriation that is utilized to allocate a correct class (entity) to different semantic units including words, expressions, and whole archives. Numerous NLP applications, for example, machine interpretation, discourse acknowledgment, data recovery, word sense disambiguation, and POS taggers, depend on language modeling (LM). The correct class allotted to the accompanying word in a discourse grouping to be anticipated is gainful to discourse recognizers. A LM is utilized in Machine Translation frameworks to calibrate the likelihood scores of the framework's results to work on the grammaticality and perfection of the interpretation in the objective language.

It has been laid out that brain networks with a solitary secret layer proceed as estimated capacities [1]. In a brain organization, the secret layer addresses a learnt non-direct blend of information data that fills in as a guess. The central brain network engineering is displayed in Figure 1. This model learns both an appropriated portrayal of words and a correct class work for anticipating the following word simultaneously. The conveyed portrayal of words catches syntactic and semantic properties of words, taking into account great speculation to word groupings not saw in the preparation set.

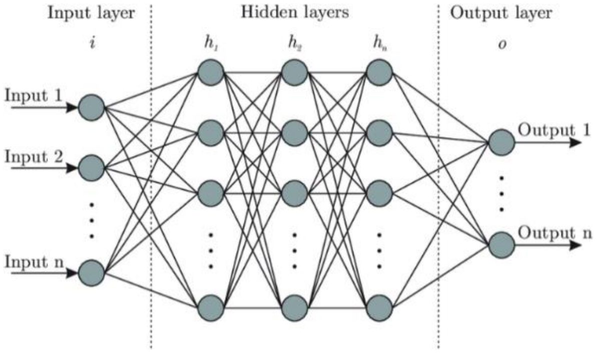

Figure 1. Basic neural network architecture

In assorted spots of the world, various neighbourhood dialects exist, each with its own letter set, signs, and linguistic structure. India [2-5] is the main country with antiquated and morphologically assorted provincial dialects. PCs process information communicated in English by means of standard ASCII codes more effectively than information addressed in other normal dialects. Building the robots' capacity to understand other natural languages, on the other hand, is time-consuming and involves a variety of methods. Many research projects and applications have been developed to handle natural languages for real-time needs, such as (1) Chatbot (2) Text-to-speech conversion (3) Language Identification (4) Hands-free computing (5) Spell-check (6) Summarizing-electronic medical records (7) Sentiment Analysis and so on.

The different methodologies utilised to construct the aforementioned apps are given, with a focus on Indian Regional Languages (IRL).

The internet is no longer monolingual; material in various regional languages is fast expanding. India has roughly 1000 recorded languages and dialects, according to the 2001 census. Much research is being done to make it easier for people to work and communicate with computers in their own native languages. Google supports transliteration in Indian Regional Languages (IRL) such as Kannada, Hindi, Bengali, Tamil, Telugu, Malayalam, Marathi, Punjabi, and Gujarati, along with search in 13 languages. Machine Translation (MT), Sentiment Analysis (SA), Parts-of-Speech (POS) Tagging, and Named Entity Recognition (NER) are the main focused activities on IRL (NER). Machine translation is a type of interlingual communication in which computers translate a source language into a target language while keeping the message intact. The distinguishing proof of offered viewpoints and the direction of thoughts in a piece of message is called as feeling examination. POS Tagging is a technique for naming each word in an expression with a label that demonstrates the grammatical feature it has a place with [6-10]. Substance With a Name Recognition tracks down legitimate names in organized and unstructured materials and afterward bunches them into determined classes of interest. AI strategies and regular language handling methods are fundamentally utilized in the advancement of IRL applications. For English, language handling techniques have been broadly considered. Notwithstanding, given of the variety of morphology and design, there hasn't been a lot of exploration done on IRL. Figure 2 portrays the general model for language handling.

Machine transliteration, pre-processing, lexical and morphological analysis, POS tagging, feature extraction, and assessment are all processes in the general model for language processing. The unstructured natural language is represented by the raw text block in the figure.

2. LITERATURE REVIEW

They took on the challenge of dealing with named entities since the quality of the translation would suffer if they didn't. Only using a rule-based method wasn't enough, so they utilized a hybrid approach, collecting 10,000 phrases from news websites as a corpus and using Stanford's NER tool for name entity identification. The system created 9180 name entities out of a total of 9234, with an accuracy of 83.65 percent precision, 83.16 percent Recall, and 83.40 percent F-Measure value [1].

In their article, they discovered that among Indian languages, Kannada has no capitalization, a lack of bigger gazetteers, a lack of standardization, and a lack of spelling. The number of alphabets in the Kannada language is 50 with vast grammar rules. The diversity of language makes it very difficult to get accuracy of the classifier above 90%. They discovered that there is a lack of annotated data and that the language is strongly agglutinating and inflected. They used Multinomial Nave Bayes classifiers to create a Supervised Statistical Machine Learning system for Kannada Language. They have to employ 22 named entities and a 95170-word corpus. They acquired 83 percent precision, 79 percent Recall, and 81 percent F-Measure value from their constructed model for recognizing named things [2].

Because of the nature of Arabic and the scarcity of linguistic resources, they were able to create achievements that allowed them to overcome the limitations. Because the available corpora aren't annotated with name entities, the relations don't have enough annotated instances to be used for learning techniques. They merged Machine Learning and Genetic Algorithm principles to improve the machine learning method's overall performance. Precision 84.8 percent, Recall 67.6 percent, and F-Measure value 75.22 percent are achieved using a hybrid technique [3].

They discovered a way to overcome the limitations of knowledge corpus availability and resolve ambiguities of named entity identification in Hindi by combining Rule Based Approach and List Look Approach. They discovered various Precision, Recall, and F-Score values for Location, Person, Organization, Date, Money, Direction, and Transportation, among other things. Their method has a 95.77 percent accuracy rate [4].

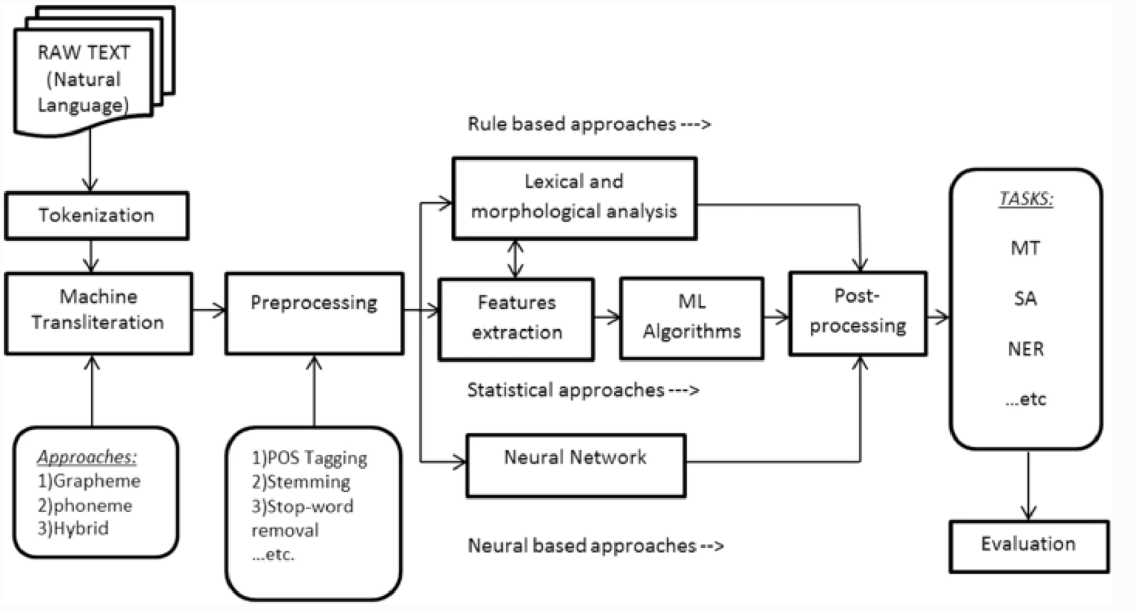

Figure 2. Generic model for language processing

They were able to outperform the constraints of ambiguous names, no capitalization, a lack of resources and tools, a lack of standardization and spelling, the absence of labeled data, and the lack of a large gazetteer in Hindi. They need to try multiple NER methodologies and apply voting methods to improve the performance of Hindi language. This is nothing but the giving same input to the different algorithms and check the result. The result which maximum time occurrence will be the final result among the results of those algorithms. For five testing files, they obtained 71.43 percent Precision, 30.86 percent Recall, and 43.10 percent F-1 Measure using the CRF method. For five testing files, MaxEnt scored 76.92 percent precision, 19.8 percent Recall, and 31.49 percent F-1 measure. For three testing files, they obtained 96.05 percent precision, 86.90 percent recall, and 91.25 percent F-1 measure using a rule-based technique [5].

Because of the complex morphology of the Kashmiri language, they encountered a number of difficulties. In contrast to English, the Kashmiri language lacks capitalization. They needed to run a test for noun recognition and achieved good results utilizing a dictionary gazetteer, lists, and morphological suffix mapping approaches. They needed to identify nouns and got a score of 93.32 percent with 07.75 percent mistakes. In addition to employing NE tags to resolve ambiguity, gazetteers lists and features are used to resolve ambiguity [6].

They have discovered hurdles with the Manipuri language, which is one of the constitutional Indian languages. No capitalization, redundant named things with additional specialized meanings in the dictionary, extremely inflectional language resulting in enormous complicated word forms, free ordering language difficult to compare to others, resource restricted language The evaluation was carried out using the CRF technique, and the results were 81.12 percent recall, 85.67 percent precision, and 83.33 percent Score value. The NER is frequently used to create Manipuri gazetteer lists [7].

For the Missing language, they described named entity recognition. It is a Tibeto-Burman minority in Assam. It's a language with few resources. For feature extraction, they employed a 12 named entity tagset, as well as Roman Script. Support Vector Machine is required for the categorization of identified entities. Because there are few resources for the language, writers must create their own corpus. 16000 data points were evaluated out of 34000, and the results were 90.58 percent recall, 85.14 percent precision, and 87.77 percent F-Score value [8].

No capitalization, scarcity of resources, agglutinative nature feature, free-word order, complexity of spelling variants, borrow words, nested and compound named entities, and many more obstacles for Urdu named entity identification are detailed in their article [9].

They focused on general issues for Punjabi language by employing two machine learning approaches in their paper: Hidden Markov Model and Entropy Markov Model for Punjabi named entity recognition. There is a lack of consistency in spelling. Due to the fact that Punjabi is not as widely used on the internet as other languages, a large gazetteer is not available. There are a large number of popular terms in Indian languages that are also utilised as Named Entities. When compared to English, there is no capitalization feature. The Punjabi language has a scarcity of resources and instruments. The Hidden Markov Model (42k words) and the Max Entropy Model (61k words) of the coaching corpus were obtained from various news items and Punjabi newspapers [10].

They discovered that named entity identification accuracy among the 22 Indian languages isn't comparable to those of foreign languages studied in depth in NER. There is a lot of information on Punjabi Language out there, but it isn't in a usable manner for local people. During this language, there are no web sources for various gazetteer listings. They claimed that employing the 12 Named Entity tagset, a machine learning technique is best suited for NER in Punjabi. They discovered that a context window with word sizes of 3, 5, and 7 leads to the same F-Score score. The experiment was carried out utilizing the Conditional Random Field method [11].

They claimed that Indian languages are inflectional, free-order, and morphologically rich, but that they are deficient in resources. Because Indian languages are unclear, it is impossible to recognize them. One big issue they encountered was that a NER system designed for one domain didn't function well with another. As in Indian languages the same words/ phrases can be used in different domain with different context. So, NER in Indian languages becomes very domain specific. Using multiple NER approaches, a 90 percent F-Measure value was reached in Indian languages. They discovered certain difficulties with the Urdu language during their investigation. They needed to apply the Hidden Markov Model for their NER study. They attained 100 percent accuracy in seven sentences of BBC news in Urdu and 100 percent performance outcomes in tourist corpuses. They discovered that there are several NER tools accessible, but that they are all language dependent. There isn't such a tool, which is a shame language independent [12]. Bangare et al. [13, 14] worked on machine learning. Shelke et al. [15] worked on emotion analysis. Gupta et al. [16] worked on user emotions. Pande et al. [17-19] worked on the spline curve etc. The mathematical model for the classification using deep neural network approach is very important to get higher accuracy [20-22].

3. PROPOSED SYSTEM

Individuals with handicaps were expanding at a bigger scope after the conflicts in the around the world. For giving some assistance to them, word expectation apparatuses were created which can assist them with conveying and furthermore to assist individuals with less speed composing and lacking information on Hindi jargon, this would assist them with growing the exactness of their composing effectiveness.

The suggested system's block diagram in Figure 3 the following is the stages. The initial stage is to gather and categories data on Hindi words, including their synonyms, antonyms, and meanings. Raw text is broken down into little chunks/ words/ sentences called tokens during pre-processing. We eliminated the stop words from the text since we didn't want them to eat up space in our database or take up precious processing time. The text vectorization is performed to convert text into vectors. On the basis of vectors, a neural network is used to forecast related words, synonyms, and use of a particular term.

The different LSTMs are used in recursive fashion. The LSTM is trained for different class of NER. For every class LSTM is trained differently. It helps to reduce the misclassification chances. The worst-case scenario in this case will be recursive LSTM will fail to predict entity in the text but it will not mis-classify it. The sensitivity, selectivity, specificity, and accuracy of the system are evaluated using the performance matrix.

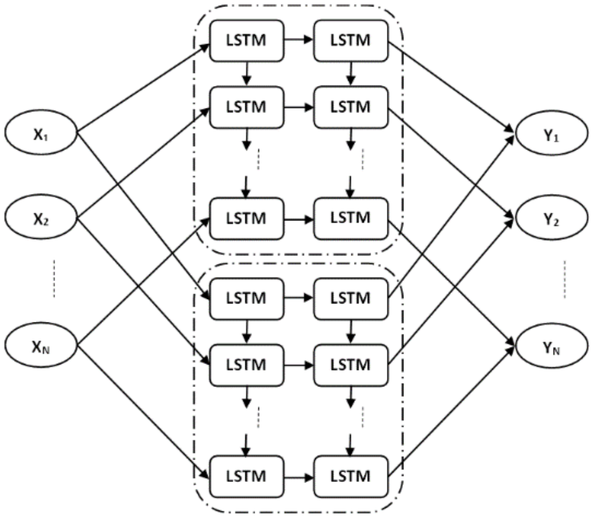

Figure 3. Proposed system

In this paper, Researcher have done the different tests to contrast the aftereffects of NER and ordinary implanting and quick text inserting layers to examinations the presentation of word inserting with various bunch sizes to prepare the profound learning models. The value of the precision of proposed system architecture is 76.13% which is way more than other system architectures. Also, the value of recall and F1-score of proposed system architecture is 71.49 and 74.26 respectively. So, by comparing proposed system architecture with existing SpaCy, CoreNLP and NLTK it is easy to conclude that proposed system architecture is reliable in all the sense.

3.1 Reading the data with dataset reader

The information we have is BILOU labeled and is in CoNLL design. A CoNLL designed information has single word per line with these elements isolated by a space and the following sentence is isolated by a line. In BILOU labeling plan B(Beginning), I(Inside), L(Last), O(Outside), U(Unit). A Dataset Reader peruses a record and converts it to an assortment of Instances. Here we will utilize the conll2003 dataset per user by tweaking it a smidgen according to our prerequisites as we just have NER labeled information and this per user acknowledges a CoNLL2003 dataset which has pos, NER and piece labels as well. We have a climate dataset, with three elements area, weather type, and date. We really want to execute two techniques the accompanying two strategies to peruse and make tokenized examples.

4. RESULT AND DISCUSSION

The results are calculated on the database which is having the sentences in Hindi language about weather enquiry. The length of the dataset is about 2600 sentences. The dataset is tested using proposed system architecture along with three different existing algorithms viz. SpaCy, CoreNLP and NLTK.

The precision, recall and F1-score is calculated graphically illustrated in Figure 4. Performance parameters analysis illustrates that the proposed system is better in every aspect of the performance parameter.

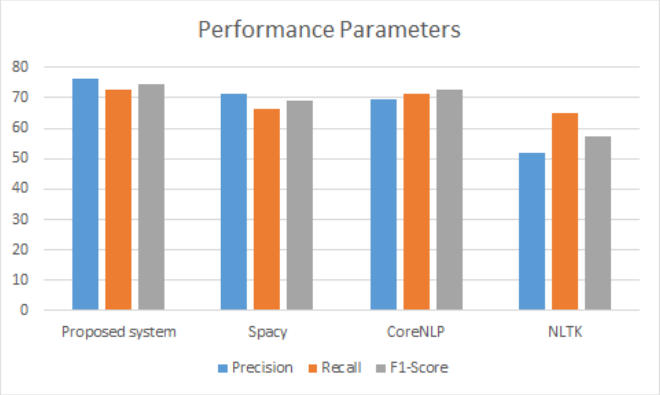

Figure 4. Proposed system

Table 1 gives broad idea about how proposed system performs. The minimum precision value is 51.94% in case of NLTK whereas it is maximum in case of proposed system architecture. Similarly, for Recall and F1-score the proposed system is superior. As recursive methods are utilized with the various LSTMs. The LSTM has been trained for various NER classes. It aids in lowering the likelihood of misclassification. In this example, recursion will be the worst-case scenario. LSTM won't correctly classify an entity in the text but it won't be able to anticipate it.

Table 1. Precision, Recall and F1 score of the prosed system with existing systems taken on the same dataset

	Precision	Recall	F1-score
Proposed system	76.13	72.49	74.26
Spacy	71.36	66.18	69.21
CoreNLP	69.38	71.3	72.82
NLTK	51.94	65.12	57.15

The time complexity of the system is also tested on different hardware platforms [23]. The time it takes to get result on different hardware platforms is tabulated in Table 2.

Table 2. Time complexity of the proposed recursive LSTM of different hardware platforms

Hardware Platform	Time required to get result (in seconds)
CPU, i3 processor, 8GB RAM	3.123
CPU, i5 processor, 8GB RAM	2.123
CPU, I7 processor, 8GB RAM	1.982
GPU, Nvidia K80	0.014

Time complexity of the i7 processor is low as compared to i3 processor which only means that the proposed algorithms time complexity is depends on the hardware platform.

5. CONCLUSIONS

Researcher have done the different examinations to contrast the aftereffects of NER and typical implanting and quick text installing layers to investigations the exhibition of word inserting with various group sizes to prepare the profound learning models. The value of the precision of proposed system architecture is 76.13% which is way more than other system architectures. Also, the value of recall and F1-score of proposed system architecture is 71.49 and 74.26 respectively. Also, the main focus of the system was to reduce the mis-classification rate which is achieved dramatically. Also, by comparing proposed system architecture with existing Spacy, CoreNLP and NLTK it is easy to conclude that proposed system architecture is reliable in all the sense.

REFERENCES

[1] Mathur, S., Saxena, V.P. (2014). Hybrid approach to English-Hindi name entity transliteration. In 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, pp. 1-5. https://doi.org/10.1109/SCEECS.2014.6804467

[2] Amarappa, S., Sathyanarayana, S.V. (2015). Kannada named entity recognition and classification (NERC) based on multinomial naïve bayes (MNB) classifier. International Journal on Natural Language Computing (IJNLC), 4(4): 39-52. https://doi.org/10.5121/ijnlc.2015.4404

[3] Boujelben, I., Jamoussi, S., Hamadou, A.B. (2014). A hybrid method for extracting relations between Arabic named entities. Journal of King Saud University-Computer and Information Sciences, 26(4): 425-440. https://doi.org/10.1016/j.jksuci.2014.06.004

[4] Kaur, Y., Kaur, E.R. (2015). Named Entity Recognition (NER) system for Hindi language using combination of rule-based approach and list look up approach. Int. J. Sci. Res. Manag.(IJSRM), 3(3): 2300-2306.

[5] Srivastava, S., Sanglikar, M., Kothari, D.C. (2011). Named entity recognition system for Hindi language: a hybrid approach. International Journal of Computational Linguistics (IJCL), 2(1): 10-23.

[6] Malik, A.B., Bansal, K. (2015). Named entity recognition for Kashmiri language using noun identification and NER identification algorithm. Journal of Computer Sciences and Engineering, 3(9): 193-197.

[7] Nongmeikapam, K., Shangkhunem, T., Chanu, N.M., Singh, L.N., Salam, B., Bandyopadhyay, S. (2011). CRF based name entity recognition (NER) in Manipuri: A highly agglutinative Indian language. In 2011 2nd National Conference on Emerging Trends and Applications in Computer Science, pp. 1-6. https://doi.org/10.1109/NCETACS.2011.5751390

[8] Hussain, S., Kuli, J.J., Hazarika, G.C. (2016). The first step towards named entity recognition in missing language. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3013-3016. https://doi.org/10.1109/ICEEOT.2016.7755253

[9] Naz, S., Umar, A.I., Shirazi, S.H., Khan, S.A., Ahmed, I., Khan, A.A. (2014). Challenges of Urdu named entity recognition: A scarce resourced language. Research Journal of Applied Sciences, Engineering and Technology, 8(10): 1272-1278. http://dx.doi.org/10.19026/rjaset.8.1095

[10] Singh, J., Lehal, G.S. (2015). Named entity recognition for Punjabi language using Hmm and MEMM. IRF International Conference, Pune, India, pp. 4-8.

[11] Kaur, A., Josan, G.S. (2015). Evaluation of named entity features for Punjabi language. Procedia Computer Science, 46: 159-166. https://doi.org/10.1016/j.procs.2015.02.007

[12] Jahan, N., Siddiqui, M.A. (2014). Urdu named entity recognition using hidden Markov model. IJACKD Journal of Research, 3(1): 1-5.

[13] Bangare, S.L., Dubal, A., Bangare, P.S., Patil, S.T. (2015). Reviewing Otsu's method for image thresholding. International Journal of Applied Engineering Research, 10(9): 21777-21783. https://dx.doi.org/10.37622/IJAER/10.9.2015.21777-21783

[14] Bangare, S.L., Prakash, S., Gulati, K., Veeru, B., Dhiman, G., Jaiswal, S. (2021). The architecture, classification, and unsolved research issues of big data extraction as well as decomposing the internet of vehicles (IoV). In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 566-571. https://doi.org/10.1109/ISPCC53510.2021.9609451

[15] Shelke, N., Chaudhury, S., Chakrabarti, S., Bangare, S. L., Yogapriya, G., Pandey, P. (2022). An efficient way of text-based emotion analysis from social media using LRA-DNN. Neuroscience Informatics, 2(3): 100048. https://doi.org/10.1016/j.neuri.2022.100048

[16] Gupta, S., Kumar, S., Bangare, S.L., Nuhmani, S., Alguno, A.C., Samori, I.A. (2022). Homogeneous decision community extraction based on end-user mental behavior on social media. Computational Intelligence and Neuroscience, Article ID 3490860. https://doi.org/10.1155/2022/3490860

[17] Pande, S.D., Chetty, M.S.R. (2018). Analysis of capsule network (Capsnet) architectures and applications. J Adv Res Dynam Control Syst, 10(10): 2765-2771.

[18] Pande, S.D., Chetty, M.S.R. (2019). Position invariant spline curve-based image retrieval using control points. Int J Intell Eng Syst, 12(4): 177-191. https://doi.org/10.22266/ijies2019.0831.17

[19] Pande, S.D., Patil, U.A., Chinchore, R., Chetty, M.S.R. (2019). Precise approach for modified 2 stage algorithms to find control points of cubic Bezier curve. In 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-8. https://doi.org/10.1109/ICCUBEA47591.2019.9128550

[20] Ladkat, A.S., Patankar, S.S., Kulkarni, J.V. (2016). Modified matched filter kernel for classification of hard exudate. 2016 International Conference on Inventive Computation Technologies (ICICT), pp. 1-6. https://doi.org/10.1109/INVENTIVE.2016.7830123

[21] Ladkat, A.S., Bangare, S.L., Jagota, V., Sanober, S., Beram, S.M., Rane, K., Singh, B.K. (2022). Deep neural network-based novel mathematical model for 3D brain tumor segmentation. Computational Intelligence and Neuroscience, vol. 2022, Article ID 4271711, 8 pages. https://doi.org/10.1155/2022/4271711

[22] Shobana, M., Balasraswathi, V.R., Radhika, R., Oleiwi, A.K., Chaudhury, S., Ladkat, A.S., Naved, M., Rahmani, A.W. (2022). Classification and detection of mesothelioma cancer using feature selection-enabled machine learning technique. BioMed Research International, vol. 2022, Article ID 9900668, 6 pages. https://doi.org/10.1155/2022/9900668

[23] Ladkat, A.S., Date, A.A., Inamdar, S.S. (2016). Development and comparison of serial and parallel image processing algorithms. In 2016 International Conference on Inventive Computation Technologies (ICICT), 2: 1-4. https://doi.org/10.1109/INVENTIVE.2016.7824894