Online at: https://researchtrendsjournal.com **ISSN No:** 2584-282X

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 2; Issue 6; 2024; Page No. 116-121

Received: 01-09-2024
Accepted: 06-10-2024

Deep Ganitrus Algorithm for speech emotion recognition

[1]Siddharth, [2]Avinash Anand and [3]Pooja Upadhyay

[1, 2]M. Tech, Department of Computer Science, Mahakaushal University, Jabalpur, Madhya Pradesh, India
[3]Assistant Professor, Department of Computer Science, Mahakaushal University, Jabalpur, Madhya Pradesh, India

DOI: https://doi.org/10.5281/zenodo.15068822

Corresponding Author: Siddharth

Abstract

Modern automated speech recognition (ASR) challenges are much more difficult than their predecessors' due to the apparent need from practical applications. Over time, the ASR system has improved to handle a wider range of challenges, including a larger vocabulary, more freedom to express oneself, more background noise, more diverse speech, and more languages. There has been a lot of recent activity in the area of speech emotion recognition (SER), which seeks to identify emotional states from signals in spoken language The most recent paper suggests using a deep garnitures algorithm to identify different emotions in a speaker's voice. When the system receives a voice signal, it analyses each individual feature using independent component analysis and the Fisher criteria. Both the computing time and the complexity of the system are drastically reduced by the suggested technique. Consequently, the difficulties of automated speech recognition and speech-emotion recognition are extensively covered in the thesis.

Keywords: Deep, Ganitrus, algorithm, technique and speech emotion recognition

Introduction

Computer games, intelligent voice assistants, and spoken conversation systems are just a few examples of the many possible uses for the future of speech emotion recognition (SER), a growing area of study that aims to build systems that can automatically detect emotions from speech. Nevertheless, present SER systems still have a way to go before they can fully capture the complexities of human emotions, as well as the wide range of human speech patterns and the difficulties in reliably collecting characteristics from these signals. Feature extraction and classification are the two primary parts of a standard SER system. Essential acoustic properties, such as pitch and spectral features, are captured using feature extraction from the spoken stream. After that, the classification part uses the characteristics that were retrieved to categorise the voice signal according to its emotional content.

Computer vision, voice recognition, and natural language processing are just a few of the areas where deep learning has been successful in recent years, solidifying its position as a potent machine learning tool. For several reasons, deep learning works well with SER. To begin with, when it comes to correctly recognising emotions from speech, deep learning models really shine at capturing all the complex interactions between elements. Secondly, deep learning models have special properties that make them good at using big speech datasets. This is especially helpful in SER, where having access to a lot of annotated data makes training the models easier and makes them better at generalising to different speakers and different emotions. The resilience and flexibility of deep learning models in managing variances makes them especially useful in real-world SER applications, even if other models also promise to generalise to new speakers and contexts.

Additional considerations before continuing with the investigation are as follows: Since human emotions are multi-faceted and manifest in a wide variety of ways, creating reliable SER systems is no easy feat. Laughter, smiling, and a high-pitched voice may express happiness, whereas tears, a low-pitched voice, and a frown can represent sadness. Individuals' speech differs according to characteristics such as age, gender, and accent, which adds another layer of complexity to SER systems. A young American woman's speech pattern could vary from that of an elderly British guy, for instance.

Literature Review

Zeng, Taiyao. (2022) [1]. Traditional machine learning methods that were once useful for processing massive amounts of raw, unlabelled audio data are now obsolete in the age of big data. Meanwhile, deep learning models have sparked a lot of interest in automated voice recognition due to their capacity to handle large amounts of data without labels and their ability to process unlabelled input directly. An introduction to deep learning's use in voice recognition is provided in the article. There is an introduction to the recent deep learning research findings in speech recognition, a discussion of the correlation between old and new deep learning models, an analysis of the development trend in deep learning for speech recognition, and an emphasis on how deep learning models should incorporate ideas from old models to create a better deep learning-based speech recognition system.

Tripathy (2023) [6] We show various preprocessing phase modifications that significantly boost the accuracy of the model in this study. Data segregation is improved in the first phase by using Blazeface instead of Haar cascade and verified window selection approaches. Improvements in lip tracking and frame selection according to word length were also shown to significantly affect accuracy. Potentially less space, complexity, and execution time would follow. It is possible to add support for more than one language to this modal. Audio may also be integrated with CNN-automated window selection in the early stages. If a language model is attached prior to softmax activation, the accuracy may be improved. Similarly, the ResNet100 may be used in lieu of CNNs during training to improve accuracy.

Guan (2024) [2] In order to enhance the system's recognition accuracy and capability to deal with complicated situations, this study seeks to investigate the integration approach of deep learning and big language models in voice recognition. An integrated framework for acoustic and language models is constructed using deep neural networks (DNNs), convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and a Transformer-based big language model. Tests conducted on the TIMIT, LibriSpeech, and Common Voice datasets demonstrate that, in comparison to conventional models, the ensemble model achieves noticeably better word error rate (WER) and real-time factor (RTF). The model delivers exceptional results, particularly when it comes to handling different languages and accents. The findings point to a new path for the advancement of speech recognition technology and demonstrate that the system's performance may be enhanced in complicated contexts via the integration of technology.

In the twenty-first century, artificial intelligence (AI) has emerged as a game-changer. The widespread use of AI in healthcare has the potential to revolutionise the medical industry. The effects on anaesthesia are the primary topic of this piece. One of the most important and dangerous parts of anaesthesia, it may cause serious problems that might be fatal, therefore we talk about how it could affect problematic airway management. Patient safety may be greatly enhanced with the precise prediction of problematic airways. We reviewed the research on ultrasonography's predictive power and AI-based models for challenging airway prediction in relation to more conventional methods of airway evaluation. Additionally, we tackle the claim that AI-powered algorithms outperform traditional tests using complicated scales and indices due to their exceptional sensitivity and specificity.

This paper provides a concise overview of the theory and practice behind speech recognition systems, mentions current issues with AI deep learning speech recognition, examines AI deep learning methods for speech recognition, proposes AI deep learning optimisation strategies for speech recognition based on improving the system's targeted feature recognition, conducting AI deep learning simulation training multiple times, and integrating audio and sports features.

Research Methodology

The study presents a new deep ganitrus algorithm (DGA) that uses the Elaeocarpus Ganitrus fruit to identify emotional states and human emotions in speech. The algorithm uses the concept of Rudraksha beads, which are spherical beads found in the fruit of the plant, to categorize emotions such as anger, fear, melancholy, and disgust. The accuracy of identifying these emotions is high, and the study proposes a method that combines these feelings, combining stress, anxiety, and depression.

The DGA technique uses independent component analysis using the fisher criteria and DGN as a classifier for feature extraction. The wake-sleep method is used to make spoken HMIs more efficient and natural by using retrieved variables. The study highlights the importance of understanding the diverse emotions experienced by humans, as they can convey a wide range of emotions.

In traditional medicine, Rudraksha is used to treat various conditions, such as stress, anxiety, depression, palpitations, nerve pain, and psychosomatic diseases. The DGA technique uses various methods to differentiate between different emotions, making it more accurate and natural for spoken HMIs.

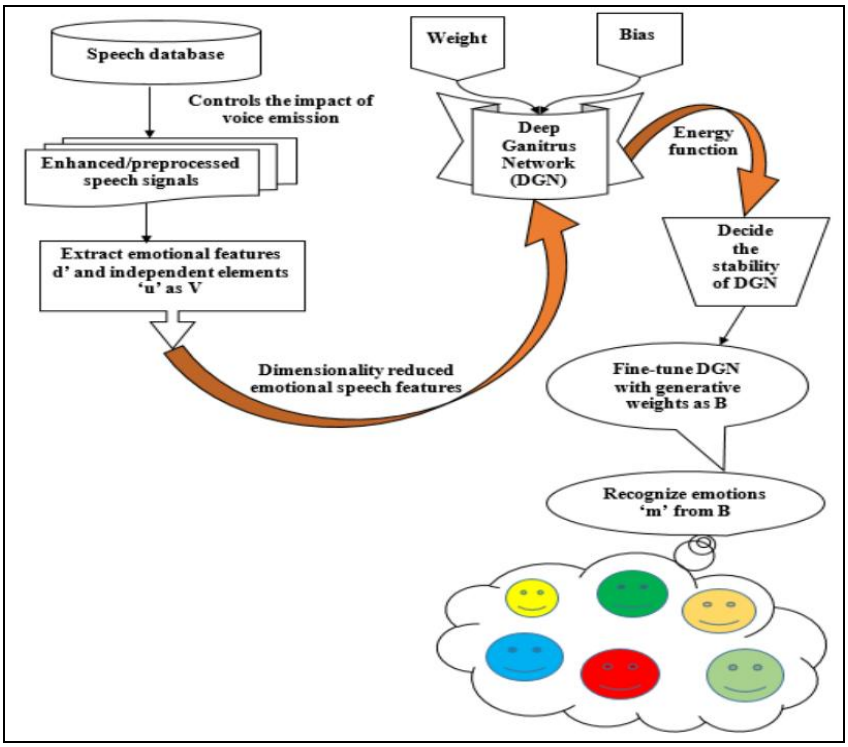

Fig 1: Flow diagram of proposed deep ganitrus algorithm

Just as Elaeocarpus Ganitrus may appear with several faces at once, it can also appear with multiple emotions in the speech signal, all of which can be identified with the use of additional attributes. The suggested DGA analyses the speech signal for its distinct emotional qualities using independent component analysis and the Fisher criteria. Using a greedy wise layer learning of DBN with wake-sleep, the suggested DGA leverages Elaeocarpus Ganitrus categorisation to categorise diverse emotions. Algorithms trained with Deep Learning can handle unlabelled data and understand complicated structures and features without the need for human feature extraction. Algorithms using the DBN technique provide more accurate emotion identification in the quickest time and with the lowest mistake rate by combining some of the most effective, fastest, and optimised findings. A cluster of emotions like anger, joy, sorrow, fear, disdain, boredom, or neutrality could make it difficult to distinguish between individual emotions, even if some of these emotions are plainly visible. The suggested system aims to appropriately recognise the emotions indicated above using greedy layer-wise learning. The following sections provide a concise summary of the steps necessary to accurately recognise different emotions.

Results and Discussion

Here we display the deep ganitrus method that has been suggested for the purpose of emotion extraction. Our suggested method for extracting emotional speech aspects allows for the independent recovery of both the speaker's built-up talents and the person's speaking abilities. It analyses the retrieved passionate data to remove unnecessary data after it observes the final characteristic of the consultant's speech emotional feature at length. In particular, the results are examined after comparing the suggested emotion detection general implementation with several system-based methods.

Experimental Setup

This experiment was carried out using a Windows 10 OS, i5 CPU, and 8 GB RAM platform with Python installed. Theano is a Python package that will greatly improve training performance; therefore, we'll be using it.

Dataset Description

It is crucial for any project to choose the right dataset. We used the Berlin Emotion Database in our SER framework. Joy, sorrow, contempt, wrath, anxiety/fear, neutral, and boredom are the seven emotions represented in its 500 expressions.

Table 1: Emo DB Details

S. No.	No. of Speakers	Gender	Age
1	03	M	31
2	08	F	34
3	09	F	21
4	10	M	32
5	11	M	26
6	12	M	30
7	13	F	32
8	14	F	35
9	15	M	25
10	16	F	31

The main purpose of the speech dataset (Xd) is to get useful information from the dataset, such as the occurrence of different emotions. Think of (Xd) data as a collection of acoustic information representing different emotions, such as anger, joy, sorrow, fear, disgust, boredom, neutral, and so on. Every statement is marked with a Xd and may be spoken in any sequence.

$$V_d = v_1 + v_2 + v_3 + \cdots + v_n$$

A laundry list of additional information is the outcome of having several performers portray a wide range of emotions in every phrase. As an example, a huge set of 2327 features is retrieved from Emo DB, which includes several features. There are 602 distinct characteristics suggested for emotion identification in this work, including SLSC, overall loudness, MPEG-7 descriptors, and the Teager energy operator on autocorrelation.

Enhancement of input speech signal

Damage to the input voice data used for emotion detection might occur during the recording process due to background noise, inconsistency, and irrelevant information. There will be a decrease in detection accuracy and the erroneous category produced by emotion categorization of this raw data. Automatic deletion of unnecessary, blank, or erroneous material improves the input voice signal. A clean signal devoid of contamination or recording fluctuation finally highlights the extraction process.

$m_e = m_1 + m_2 + m_3 + \cdots + m_n$

Equation (4.36) denotes the number of features enhanced/automatically preprocessed voice signals as '*n.*'

Feature Extraction

The speech signal consists of several features that appear in spectacular, flamboyant singularity. A feature based on emotional confidence is one of the components that should be used. By counting the number of lines, one may identify Elaeocarpus Ganitrus, which occurs on several faces inside a single tree. Another analogy is that several emotions may coexist in a single voice transmission. In order to determine the appropriate emotion, we need to isolate its essential features, which include, among other things, pitch intensity, speaking pace, voice quality, tonal force proportion, and supernatural flux. Using factual approaches to extract emotional characteristics from pattern recognition presents a number of issues related to the information dimensions. It is challenging, for instance, to transfer a method that works in a low-dimensional environment to a high-dimensional one. On top of that, the method for handling a low-dimensional issue is often advantageous, efficient, and has minimal processing complexity. Many high-dimensional emotional qualities reflect the extracted datasets in correlation analysis, which likely requires a lot of space and time.

In this case, there are three steps to feature extraction. In the first step, we extract the characteristics of the audio stream and store them separately. In the subsequent phase, the expanded feature vectors are compiled. These feature vectors are composed of separate static and dynamic features. This improved feature vectors are next, if there is one, transformed into small-scale, robust vectors for the recogniser to employ in the last phase.

The improved input signal, ue, which is a blend of speech signals, contains 'n' number of emotional characteristics signals, as will be explained later.

$$m_i = d_{i1}u_1 + d_{i2}u_2 + \cdots + d_{in}u_n \text{ , for all '}i\text{'}$$

Predicted as linear combinations of unidentified emotional components are the voice signals' data variables. The non-Gaussian, latent variables that make up the analysed data are believed to be independent of each other. The seven emotional states-angry, joyful, sad, fearful, disgusted, bored, and neutral-are examined in an autonomous speaker scenario. Consider Equation 4.37 as if all possible combinations mi and independent components bn were random variables instead of flawless voice signals. It is possible to assume, without sacrificing generalisability, that the means of all independent components and aggregate factors are zero.

The above equation is an example of why it is much more reasonable to use vector-matrix description rather than aggregates. Matrix U represents the independent components, while emotional components dia are represented by matrix D.

$$\boldsymbol{m = DU}$$

Occasionally the columns of matrix D given as d_j; the equation can also be written as

$$m = \sum_{i=1}^{p} d_i u_i$$

The process of generating the acquired data via the cooperation of the modules bi is described in Equation 4.39. Since the impartial components are not doing anything, they are not readily apparent. It is also expected that the mixing matrix will not be recognised. Both d and v are estimated using the random vector m. Using this vector in a way that is consistent with widely accepted beliefs is essential. The statistically unbiased hypothesis of the bi component serves as the starting point.

Here, the D matrix is computed using non-Gaussian dispersion. After the estimate technique of a matrix, the opposite framework is achieved, state W, and we will acquire the autonomous component vector by

$$\boldsymbol{V = UD}$$

In this study, we apply the Fisher Criterion to minimise emotional speech characteristics, and the yield of the feature vector is then employed in the event determination optimisation. Here, 'U' is the extricated independent component.

Training for speech emotion recognition

The DG network is built using Theano in this project. If you need to do quick numerical calculations on your CPU or GPU, you may use the Theano Python module. Whether you choose to build your own deep learning models or utilise one of the many wrapper libraries available, this core package for Python deep learning has you covered. Because of this, we can evaluate mathematical processes, such as multi-dimensional arrays, more effectively. In the beginning, the DG network was taught to recognise different emotions by using the greedy layer-wise deep learning technique in conjunction with the wake-sleep algorithm. Half of the speech data in the Berlin dataset is used for

testing purposes, while the other half is used for training purposes. The model with the lowest error rate is selected after over a hundred different combinations of hyperparameters are tested. The settings for the generative pre-processing and processing may be seen in Table 2. The layers were trained for 475 epochs at a learning rate of 0.01.

Table 2: Hyperparameters and training statistics of the DGN

Number of layers	5
Units per layer	50
Learning rate	0.01
Number of epochs	475
Recognition accuracy	0.985

Performance analysis

Because the suggested Deep Ganitrus technique is a recognition algorithm, it is critical to think about the model's presentation in conjunction with the estimates. Case in point: accuracy, false rejection rate (FRR), and false acceptance rate (FAR). The following provides context for the SER's evaluation metrics:

FAR

It has conversed as follows, and it quantifies the percentage of the model's all-out efforts to sense the emotions to the proportion of incorrectly experienced sentiments.

$$FAR = \frac{\text{Number of false recognitions}}{\text{Total number of attempts}}$$

FRR

It is defined as follows and indicates the percentage of false rejection (FR) feelings relative to the total tries made by the method.,

$$FRR = \frac{\text{Number of false rejections}}{\text{Total number of attempts}}$$

Accuracy: Measuring the accuracy involves comparing the framework's positivity and negativity, which show how far off the mark it is when applied to characteristic data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Table 3: Recognition performance of DGA over Berlin database

Emotions	Recall/sensitivity	Precision	F1-Score	Specificity	Accuracy	FAR	FRR
Anger	0.975	0.992	0.987	0.989	0.989	0	0
Sadness	0.977	0.963	0.981	0.971	0.982	0.02	0.01
Joy	0.974	0.988	0.971	0.984	0.988	0	0
Neutral	0.981	0.975	0.972	0.988	0.984	0.01	0.02
Boredom	0.992	0.986	0.987	0.983	0.982	0.03	0.01
Fear	0.978	0.993	0.998	0.988	0.989	0	0
Disgust	0.991	0.982	0.979	0.991	0.981	0.03	0.03
Average	0.981	0.982	0.982	0.984	0.985	0.01	0.01

Analyses of the recognition performance of our suggested DGA framework are shown in Table 3. Using recall/sensitivity, F1-score, FAR, FRR, specificity, and accuracy, it provides a performance analysis of six distinct emotions: sorrow, joy, neutral, boredom, fear, and disgust. In order to test DGA's emotion identification capacity, performance metrics were computed for each identified emotion. In Figure 2, we can see that the proposed DGA has a high accuracy % when it comes to recognising these seven emotions. The average values for recall, precision, F1-score, FAR, FRR, specificity, and accuracy are 0.981, 0.982, 0.01, 0.01, and 0.985, respectively.

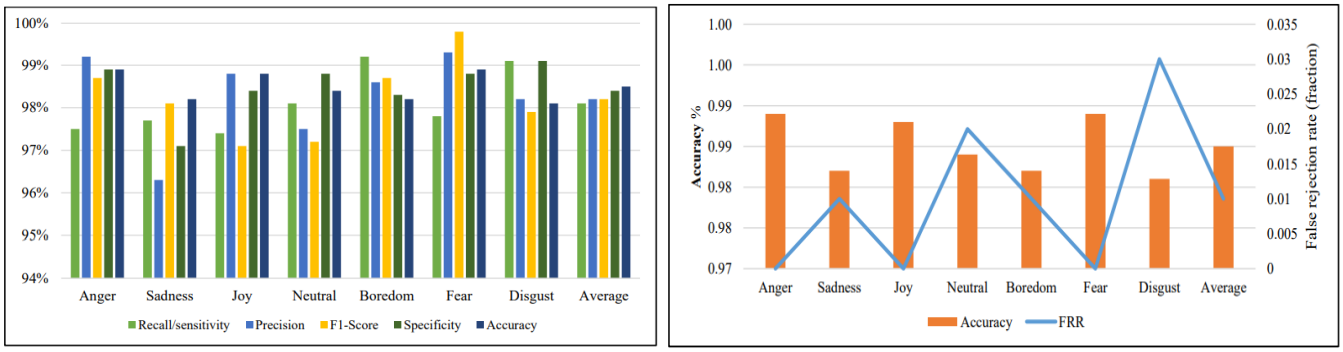

Fig 2: Emotion recognition performance of DGA

For the Deep Ganitrus Algorithm's accuracy percentage, see Table 4 for the confusion matrix. For optimal recognition accuracy, use the bolded values in the matrix.

Table 4: Confusion matrix for accuracy % of DGA

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sad
Anger	98.9	0	0	0.75	0.11	0	0
Boredom	0	98.2	0	0.23	0.47	0.24	0.13
Disgust	0.71	0	98.1	0.52	0.57	0.56	0.34
Fear	0.63	0.34	0	98.9	0	0.13	0
Joy	0.32	0.32	0.56	0	98.8	0.37	0
Neutral	0	0.62	0	0.42	0	98.4	0.32
Sad	0	0.89	0.15	0.53	0	0.14	98.2

The high recognition accuracies of DGA are shown in Table 3 Based on our findings, the recognition rate for neutral, sad, disgusted, bored, and other emotions is lower than for other emotions. This is because adding all of the emotions to the network eliminates the possibility of neutral being mistaken for another emotion. Because these neutral speech segments are so similar when employed with the fixed-length model, they might be mistakenly assigned to distinct emotions, leading to confusion. Sadness and contempt are also less accurately represented. Potentially, memories of melancholy and boredom are diminished while other non-neutral feelings are recalled more vividly. On the other hand, rage and terror are properly identified. For grief, fear, contempt, and rage, the error rate for a single emotion is 0.01. Anger, fear, disgust, and other negative emotions work together to produce this mistake rate. The findings show that the DGA framework improves the accuracy of emotion identification for some.

Conclusion

In the last few decades, the field of human emotion identification via speech signal assessment has grown in prominence. Given the wide variety of human speech patterns, voice qualities, cultural backgrounds, environmental factors, etc., emotion identification via speech signals remains a challenging problem. However, it is more challenging to identify human emotions using speech cues as there are no formal definitions for emotions. Unfortunately, the state-of-the-art in signal processing still falls short of the ideal SER accuracy and success rate when it comes to detecting and characterising the emotions conveyed by speech signals. Alternatively, the deep ganitrus algorithm (DGA) is suggested as a framework for speech-emotion detection that can extract different emotions from voice signals. The EmoDB database is used for the experiments. The features are extracted using independent component analysis with fisher criteria, and the DBN is fine-tuned using the wake-sleep method. The suggested DG algorithm-based study attains 98.5% recognition accuracy with very low FAR and FRR values of 0.01.

References

1. Zeng T. Deep learning in automatic speech recognition (ASR): A review. In: Proceedings of the 10th International Conference on Information and Communication Technologies; c2022. p. 23. DOI: 10.2991/978-2-494069-51-0_23.
2. Yadava G, Thimmaraja. Development of noise robust real-time automatic speech recognition system for Kannada language/dialects. Engineering Applications of Artificial Intelligence. 2024;135:108693.
3. Kiran B. Automatic speech recognition through artificial intelligence. 2023;5(6):Nov-Dec.
4. Mukhamadiyev A, Khujayarov I, Djuraev O, Cho J. Automatic speech recognition method based on deep learning approaches for Uzbek language. Sensors. 2022;22(10):3683. DOI: 10.3390/s22103683.
5. Mehrish A. A review of deep learning techniques for speech processing. Information Fusion. 2023;99:101869.
6. Mehrish A, Majumder N, Bhardwaj R, Poria S. A review of deep learning techniques for speech processing. arXiv. 2023. DOI: 10.48550/arXiv.2305.00359.
7. Al-Fraihat D, Sharrab Y, Alzyoud F, Qahmash A, Maaita A. Speech recognition utilizing deep learning: A systematic review of the latest developments. Human-Centric Computing and Information Sciences. 2024;15:15. DOI: 10.22967/HCIS.2024.14.015.
8. De Lima TA, Da Costa-Abreu M. A survey on automatic speech recognition systems for Portuguese language and its variations. Computer Speech and Language. 2020;62:101055.
9. Chen Y, Zhang J, Yuan X, Zhang S, Chen K, Wang X, Guo S. SoK: A modularized approach to study the security of automatic speech recognition systems. arXiv. 2021. DOI: 10.48550/arXiv:2103.10651.
10. Xia K, Xie X, Fan H, Liu H. An intelligent hybrid-integrated system using speech recognition and a 3D display for early childhood education. Electronics. 2021;10(11):1862.
11. Ahmad A, Mozelius P, Ahlin K. Speech and language relearning for stroke patients-understanding user needs for technology enhancement. In: Proceedings of the 13th International Conference on eHealth, Telemedicine, and Social Medicine; Nice, France; c2021. p. 20-23.
12. Sodhro A, Sennersten C, Ahmad A. Towards cognitive authentication for smart healthcare applications. Sensors. 2022;22(10):2101.
13. Avazov K, Mukhriddin M, Fazliddin M, Young I. Fire detection method in smart city environments using a deep-learning-based approach. Electronics. 2021;11(1):73.
14. Khamdamov U, Abdullayev A, Mukhiddinov M, Xalilov S. Algorithms of multidimensional signals processing based on cubic basis splines for information systems and processes. Journal of Applied Science and Engineering. 2021;24(2):141-150.
15. Musaev M, Khujayorov I, Ochilov M. Automatic recognition of Uzbek speech based on integrated neural networks. In: World Conference on Intelligent Systems for Industrial Automation; 2021; Cham, Switzerland. p. 215-223. Springer.
16. Tripathy T. Homoeopathy in NIPAH Virus. South Asian Res J App Med Sci. 2023;5(5):96-99.