Project Progress Report on

Speech Emotion Recognition (Classification) in real-time using Deep LSTM layers

Submitted in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Ishika Gupta	**2016787**
Rohan Pandey	**2016967**
Yuganshu Joshi	**2017147**

Under the Guidance of

Dr. Ashwini Kumar Singh

Associate Professor

Project Team ID: MP23CSE152

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand

April-2024

CANDIDATE'S DECLARATION

I/We hereby certify that the work which is being presented in the Project Progress Report entitled **"Speech Emotion Recognition (Classification) in real-time using Deep LSTM layers"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering in the Department of Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the undersigned under the supervision of **Dr. Ashwini Kumar Singh, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Ishika Gupta	2016787	signature
Rohan Pandey	2016967	signature
Yuganshu Joshi	2017147	signature

The above mentioned students shall be working under the supervision of the undersigned on the **"Speech Emotion Recognition (Classification) in real-time using Deep LSTM layers"**

Signature
Supervisor

Signature
Head of the Department

Table of Contents

Chapter 1

Introduction and Problem Statement

1.1 Introduction

Human communication is inherently rich with emotional content, and the ability to understand and classify these emotions from spoken language has long been a fascinating challenge in the field of artificial intelligence and machine learning. Recognizing emotions in speech is not only crucial for enhancing human-computer interaction but also holds significant potential across various domains such as mental health assessment, customer service, and entertainment. In today's digital age, there is a growing demand for systems that can accurately identify and respond to human emotions conveyed through speech in real-time.

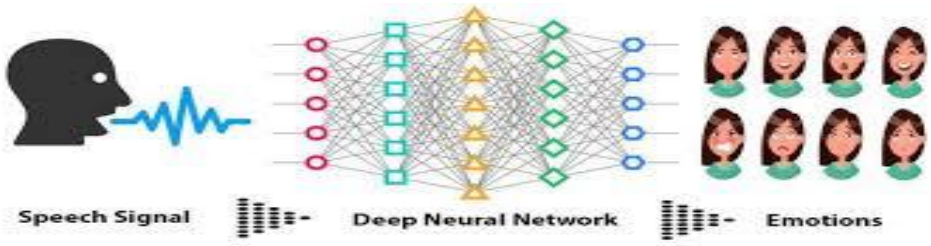

Figure 1.1 speech emotion recognition using deep learning

Speech Emotion Recognition (SER) has received great attention in recent years due to its many applications, from the development of human-computer closely related to customer service in call centers. Emotions expressed through speech play an important role in effective communication, and the ability to recognize emotions in real time has many applications. Deep learning techniques such as short-term temporal (LSTM) networks have been shown to be effective in extracting meaningful patterns from sequential data, making their properties truly important for SER. Based on this background, this work aims to investigate the development of real-time SER using deep LSTM layers[2].

This research focuses on the development of a real-time Speech Emotion Recognition (SER) system, specifically designed to classify emotions expressed in spoken language. The primary objective of this study is to leverage the capabilities of Deep Long Short-Term Memory (LSTM) layers, a subtype of recurrent neural networks (RNNs), to create a robust and efficient model for speech emotion classification.

The utilization of deep LSTM layers is motivated by their proficiency in modeling sequential data, making them well-suited for the complex task of emotion recognition from audio input[1,3].

1.2 Problem Statement

The problem at hand involves building a robust and efficient real-time SER system capable of accurately recognizing emotions expressed in spoken language. This problem can be broken down into several key components and challenges:

- **Data Collection and Labeling:** The first challenge is gathering a diverse and well-labeled dataset of audio recordings that cover a broad spectrum of emotions. Collecting, annotating, and preprocessing such data can be time-consuming and resource-intensive[3,5].
- **Real-Time Input:** Developing a system that can accept and process audio input from a microphone in real-time is non-trivial. Ensuring that the input data matches the model's expectations regarding sample rate, format, and length is crucial for accurate predictions[2].
- **Feature Extraction:** Extracting relevant features from raw audio data is essential. This may involve converting audio signals into spectrograms or extracting features like Mel-Frequency Cepstral Coefficients (MFCCs) to represent the emotional content effectively[1,4,5].
- **Model Architecture:** Designing an effective deep learning model architecture, such as a stacked LSTM network, is a critical aspect. Determining the optimal number of layers, units per layer, and other hyperparameters is essential for achieving high recognition accuracy[3,4].
- **Training and Validation:** Training deep LSTM networks for SER requires careful consideration of training data, validation techniques, and strategies to prevent overfitting. Monitoring and improving validation metrics are essential for model generalization[2,3,4,5].
- **Real-Time Inference:** The system must be capable of efficiently and accurately inferring emotions in real-time. This requires low-latency model predictions while ensuring high prediction accuracy[1,4].

Chapter 2

Objectives

The objectives of the proposed work are as follows:

- **Develop a Real-time Speech Emotion Recognition (SER) System:** Create a system capable of real-time emotion recognition in spoken language, ensuring rapid and instantaneous processing of audio input.
- **Utilize Deep LSTM Layers:** Employ Deep Long Short-Term Memory (LSTM) layers as a fundamental component of the SER system's architecture, harnessing their sequential data modeling capabilities.
- **Achieve High Emotion Classification Accuracy:** Train the SER system to accurately classify a broad spectrum of emotions, including happiness, sadness, anger, surprise, and more, with a focus on minimizing classification errors.
- **Enhance Robustness:** Ensure the SER system's robustness by addressing variations in speech patterns, accents, and background noise, making it adaptable to real-world scenarios.
- **Facilitate Practical Applications:** Enable the integration of the SER system into practical applications, including human-computer interaction, mental health assessment, customer service sentiment analysis, and emotion-driven content recommendations.
- **Optimize for Real-time Constraints:** Develop efficient algorithms and processing techniques to meet the real-time constraints imposed by the SER system, guaranteeing timely and responsive emotion recognition.

Chapter 3

Project Work Carried Out

3.1 Data Preprocessing

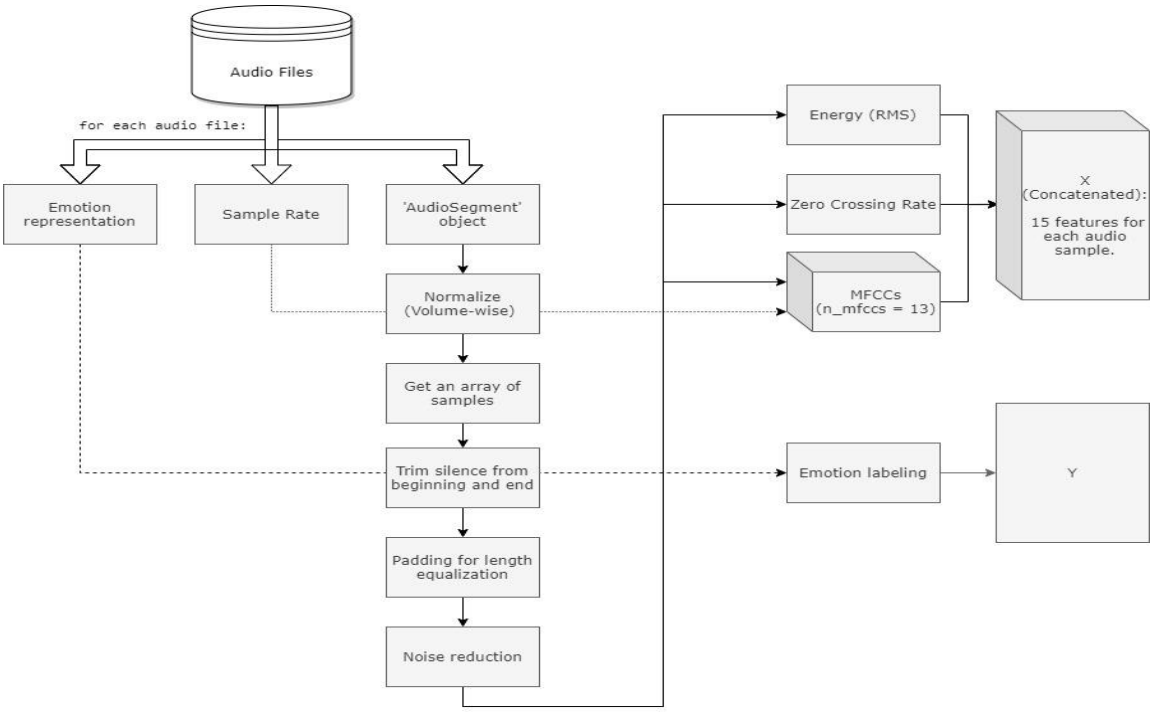

Figure 3.1 Data preprocessing representation

3.1.1 Initial Extraction

The following data is extracted from each audio file:

1. **Emotion representation**
    - RAVDESS: The filename contains a fixed placed int that represents an emotion, e.g. 03 is happy.
    - TESS: The filename contains a string representation of an emotion, e.g 'happy'
2. **Sample Rate**: number of audio samples per second. RAVDESS database was recorded in 48kHz, and the TESS database was recorded in 24.414kHz.

3. The audio is processed in the following order:

- Normalization: The 'AudioSegment' object is normalized to + 5.0 dBFS, by effects module of pydub.
- Transforming the object to an array of samples by numpy & AudioSegment.
- Trimming the silence in the beginning and the end by librosa.
- Padding every audio file to the maximum length by numpy, for length equalization.
- Noise reduction is being performed by noisereduce.

3.1.2 Feature Extraction

The selected features being extracted with librosa for the speech emotion recognition model are:

1. Energy - Root Mean Square (RMS)
2. Zero Crossed Rate (ZCR)
3. Mel-Frequency Cepstral Coefficients (MFCCs)

With frame_length = 2048, hop_lentgh = 512, assuring equally sequential length.

Every 2048 samples (sequence of ~0.058 seconds on average) are being analyzed and translated to 4 sequential feature values (2048 / 512 = 4).

In total, for an audio file lengthed 173056 samples, considering the last sample, 339 sequential values are returned for each feature (173056+1 / 512 = 339).

Emotion representation

There is a different representation of the emotions in each database.

RAVDESS Database

- A RAVDESS filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). The format which carries only emotion expressed by speech is taken as 03-01-X-X-X-X-X.wav, as the 8 emotions are stated in the 3rd part (The 1st 'X' within the file-name format).

TESS Database

- A TESS file name contains the emotion by a direct text, e.g. "YAF_youth_happy.wav".
- To overcome this incompatibility with the RAVDESS representation, "find_emotion" function has been executed.

In addition, classification modeling accepting only values starting from zero, thus "emotionfix" function has been executed for all files, performing an 'n = n-1' process for the emotion representation.

3.1.3 Final Data Setup

In order to input the data into a model, a few adjustments should be made:

- The shapes of the features must be uniform, and in the 3D format:

(batch, timesteps, feature)

- Concatenating all features to a single 'X' variable.
- Adjustment of 'Y' with a 2D shape (keras library requirement)
- Split of X, Y to train, validation, and test sets.
- y_train and y_validation conversion to 'One-hot' vectors for classification purposes (y_test is being converted adjacent to the test)

3.2 Model definition and train

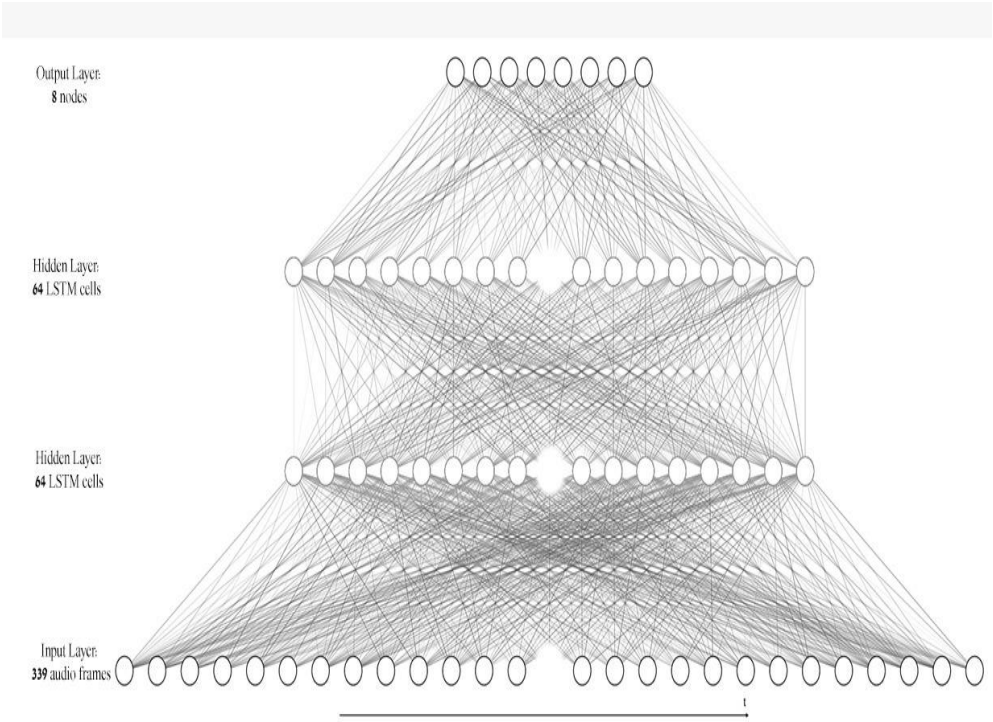

Figure 3.2 LSTM deep learning model used in speech emotion recognition

- The model is executed with keras library, using 2 hidden LSTM layers with 64 nodes, and an output (dense) layer with 8 nodes, each for one emotion using the 'softmax' activation. The optimizer that led to the best results was 'RMSProp' with default parameters.
- The batch size chosen is 23, which is a factor of all samples in the sets; train (3703), validation (368) and test (161).

3.3 Model Evaluation

The model has been evaluated using the following factors:

- A visualization of the loss and categorial accuracy values trend during the train process.
- A confusion matrix for visualizing the number of successful predictions of each emotion: for validation and test sets.
- Model's prediction accuracy rates for each emotion: for validation and test sets.

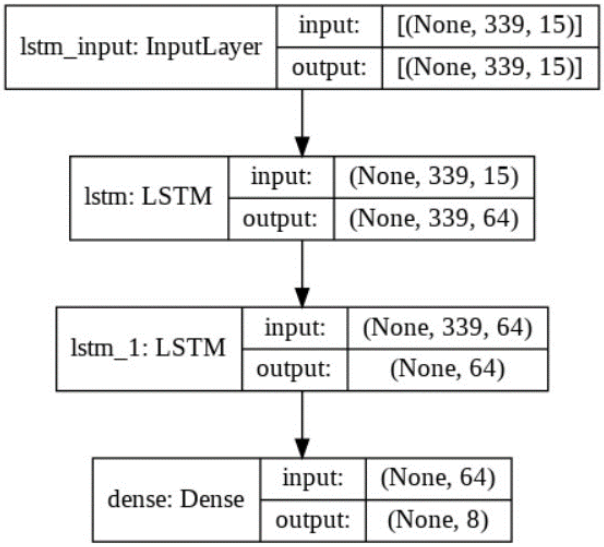

Figure 3.3 Model structure visualization

3.4 Applying Bi-directional LSTM Algorithm

In the context of speech emotion recognition, the use of Bidirectional LSTM (BiLSTM) layers in the provided code enhances the model's ability to capture and understand temporal dependencies in the audio features extracted from speech signals.

- BiLSTM captures dependencies in both temporal directions, enabling the model to understand the sequential dynamics of speech signals effectively.
- The bidirectional nature helps in capturing long-term dependencies in speech signals, which is crucial for recognizing emotional expressions that may span across multiple time steps.
- By considering information from both ends of the sequence, BiLSTM enhances the model's ability to understand and differentiate between various emotional expressions in speech.
- The BiLSTM layers are added to the model architecture. The bidirectional nature of these layers enables the model to capture both short-term and long-term dependencies in the sequential input data.
- The output of the BiLSTM layers is flattened and connected to a Dense layer with a softmax activation function to produce probabilities for each emotion class.

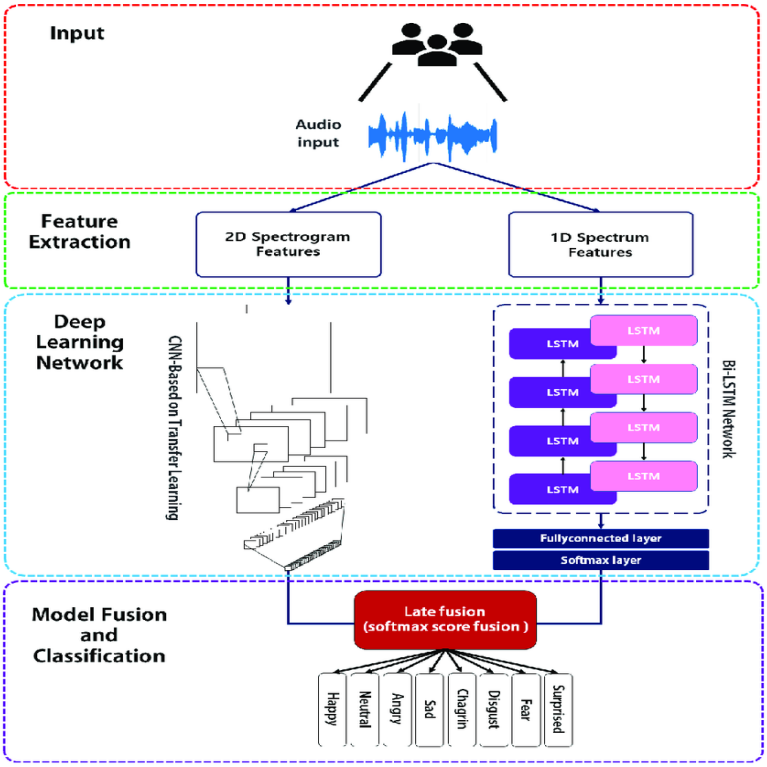

Figure 3.4 Working of Bi-LSTM Algorithm

3.5 Performance Comparison

Comparing the performance of the BiLSTM model (97% accuracy) and the LSTM model (96% accuracy) can provide insights into how these architectures perform on your speech emotion recognition task. Here's a detailed comparison and potential conclusions:

- The Bi-LSTM model shows a slightly higher accuracy compared to the LSTM model.
- The LSTM architecture introduces additional complexity by considering forward contexts. In some cases, this complexity might not lead to a significant improvement, and the Bi-LSTM architecture might perform better on the specific task.
- Consider the nature of the speech emotion data. If the emotions are primarily influenced by short-term patterns, the unidirectional LSTM might not be sufficient to capture these patterns.

In conclusion, the choice between Bi-LSTM and LSTM depends on the specific characteristics of the dataset. While the Bi-LSTM model currently outperforms the LSTM model, further analysis and experimentation can provide more insights into the factors influencing model performance.

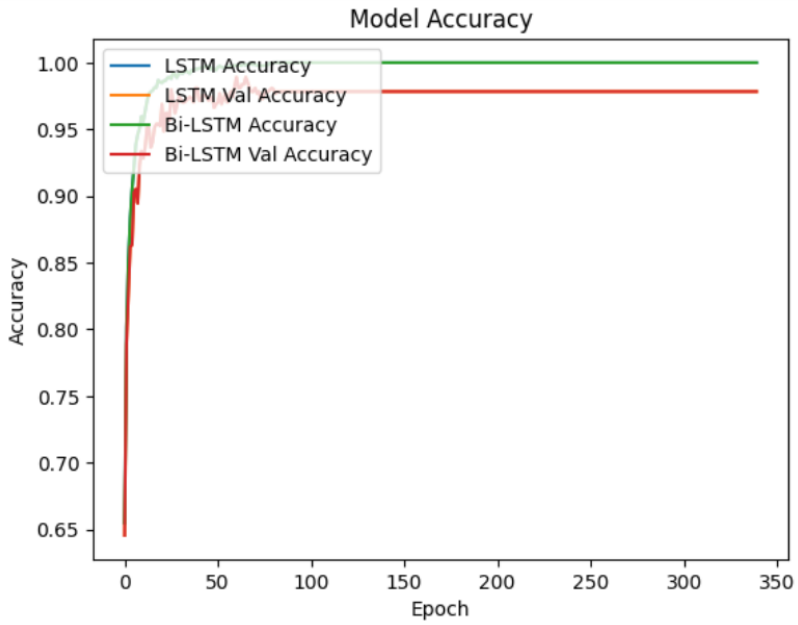

Figure 3.4 Performance of LSTM & Bi-LSTM Model

3.6 Result & Conclusion

As seen, the validation set accuracy of the model had come up to 97.83% and the test set accuracy had reached 96.28% with overfitting in the training process starting around the 100th epoch. Although various regulations have been placed in earlier tryouts, they had restricted the accuracy from reaching its maximum value. A ModelCheckpoint is applied, saving the best weights according to the model's accuracy, thus the overfitting is bypassed.

Also after applying Bi-LSTM algorithm , we achieved the accuracy of 97.83% and the test accuracy had reached 96.28% with the overfitting in the training process starting around 100^{th} epoch.

Within the next part of the study in which the model will be used as a Real-time SER, The inputs will be processed similarly as the data used by the model, to gain similarity and therefore, precision.

Chapter 4

Future Work Plan

The future work plan of our project are as follows:

Sl. No.	Work Description	Duration in Days
1.	Setup Environment	10
2.	Load Pre-Trained Model and Define Emotion List and Functions	18
3.	Real-Time Implementation	25
4.	Testing and Optimization	15
5.	Presentation and Visualization	7
6.	Research paper implementation	25
7.	Final Review	10

Chapter 5

Weekly Task

The report of project work allocated by the supervisor is as follows:

Week No.	Date: From-To	Work Allocated	Work Completed (Yes/No)	Remarks	Guide Signature
1	10-9-2023 to 05-12-2023	Define Project Scope, Objectives and implementing LSTM model.	Yes		
2	06-12-2023 To 10-01-2024	Understanding the working of Bi-LSTM Algorithm.	Yes		
3	11-01-2024 To 08-02-2024	Implementation and model evaluation of Bi-LSTM algorithm.	Yes		
4	09-02-2024 To 07-03-2024	Code Optimization and Final Review	Yes		

References

[1] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, „„EERA-ASR: An energy- efficient reconfigurable architecture for automatic speech recognition with hybrid DNN and approximate computing,IEEE Access, vol. 6, pp. 52227– 52237, 2018.

[2] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, Animage- based deep spectrum feature representation for the recognition of emotional speech, in Proc. 25th ACM Multimedia Conf. (MM), 2017, pp. 478–484.

[3] Mustaqeem and S. Kwon, A CNN-assisted enhanced audio signal processing for speech emotion recognition, Sensors, vol. 20, no. 1, p. 183, 2020.

[4] J. Huang, B. Chen, B. Yao, and W. He, ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network, IEEE Access, vol. 7, pp. 92871–92880,2019.

[5] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv:1409.1556. [Online].

[6] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, Cloud-assisted multiview video summarization using CNN and bidirectional LSTM, IEEE Trans. Ind. Informat., vol. 16, no. 1, pp. 77–86, Jan. 2020.

[7] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, Speech emotion recognition using deep learning techniques.

[8] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, Deep features-based speech emotion recognition for smart affective services, Multimedia Tools Appl., vol. 78, no. 5, pp. 5571 5589, Mar. 2019.

[9] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, Cover the violence: A novel Deep-Learning-Based approach towards violencedetection in movies, Appl. Sci., vol. 9, no. 22.