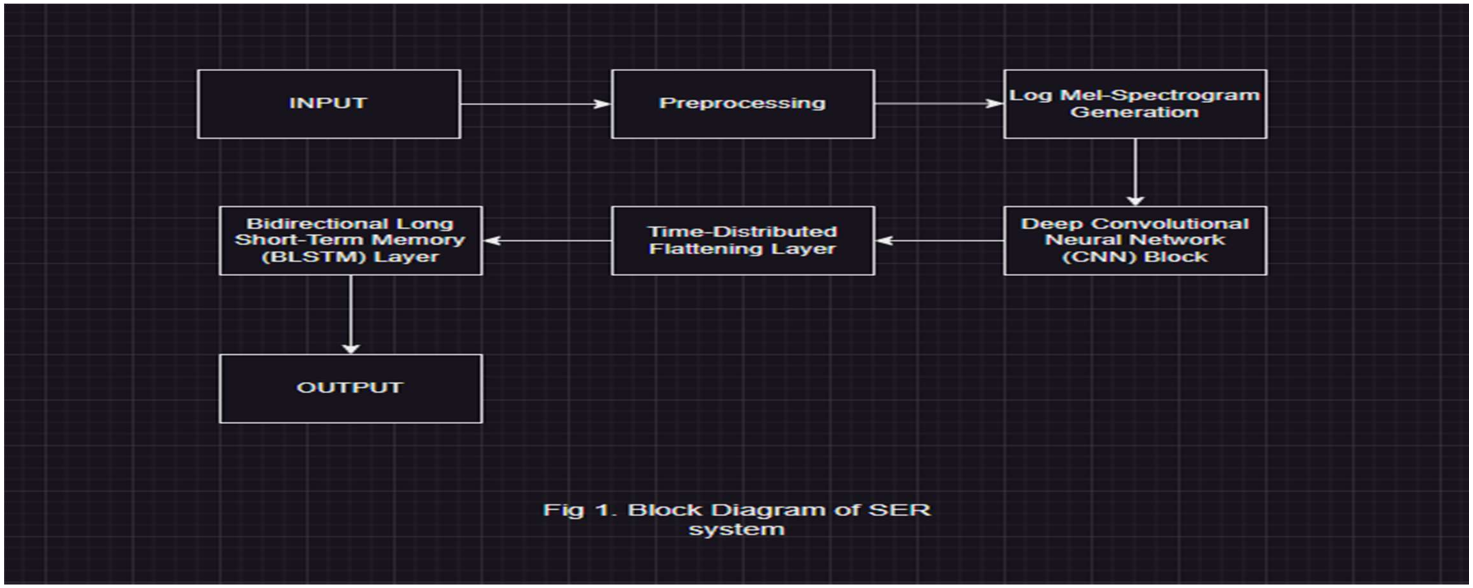
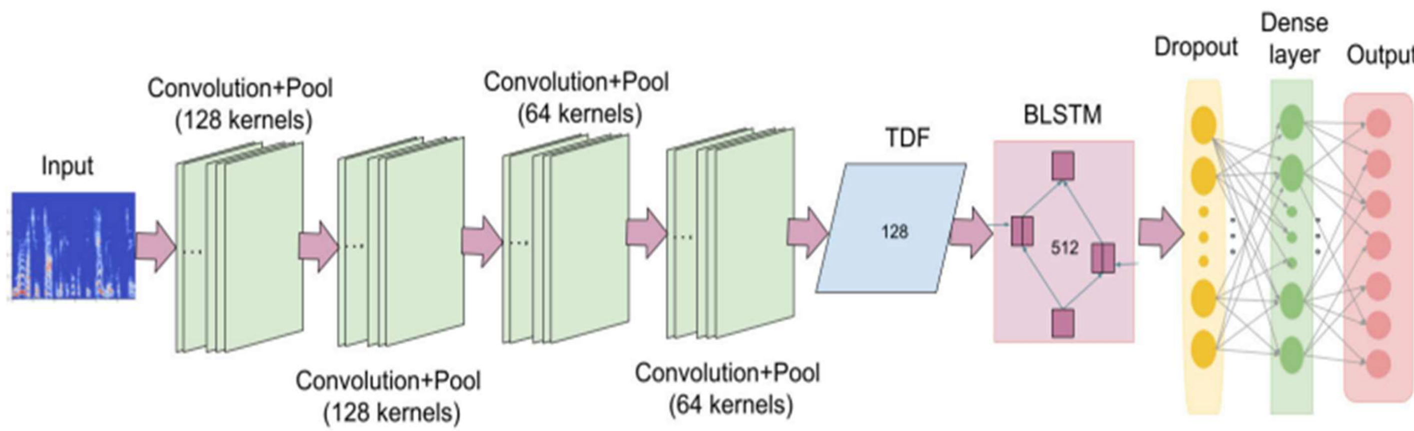
Department of Computer Science and Engineering
Bangladesh University of Business and Technology (BUBT)

CSE 498: Literature Review Records

Student's Id and Name	Md. Shafiul Alam (19202103327)
Capstone Project Title	Speech Emotion Recognition
Supervisor Name & Designation	Jubayer Al Mahmud, Assistant Professor
Course Teacher's Name & Designation	Md. Shahiduzzaman, Assistant Professor

Aspects	**Paper # 1 (Title)**
Title / Question (What is problem statement?)	Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks.
Objectives / Goal (What is looking for?)	**Objective:** The objective of the study is to develop a speech emotion recognition (SER) system using a machine learning approach, specifically focusing on the Bangla language and the American English language. The study aims to investigate and compare the performance of different features and classifiers for accurately identifying emotions from speech signals. Additionally, the study explores the use of deep learning techniques, such as CNNs, LSTMs, and transfer learning, to improve emotion detection in low-resource languages and cross-lingual scenarios. **Goals:** 1. Construct a natural-like human-computer interaction (HCI) system by accurately identifying human emotions from voice signals. 2. Develop a successful SER system by analyzing and classifying speech data to discover embedded emotions. 3. Create appropriate emotional databases, including acted, simulated, audio-only, audio-visual, or facial expression datasets, for the target languages.

	4. Present the state-of-the-art performances achieved by the model for the SUBESCO and RAVDESS datasets.
Methodology/Theory (How to find the solution?)	 Fig 1. Block Diagram of SER system  **FIGURE 2. Architecture of the proposed SER system.**

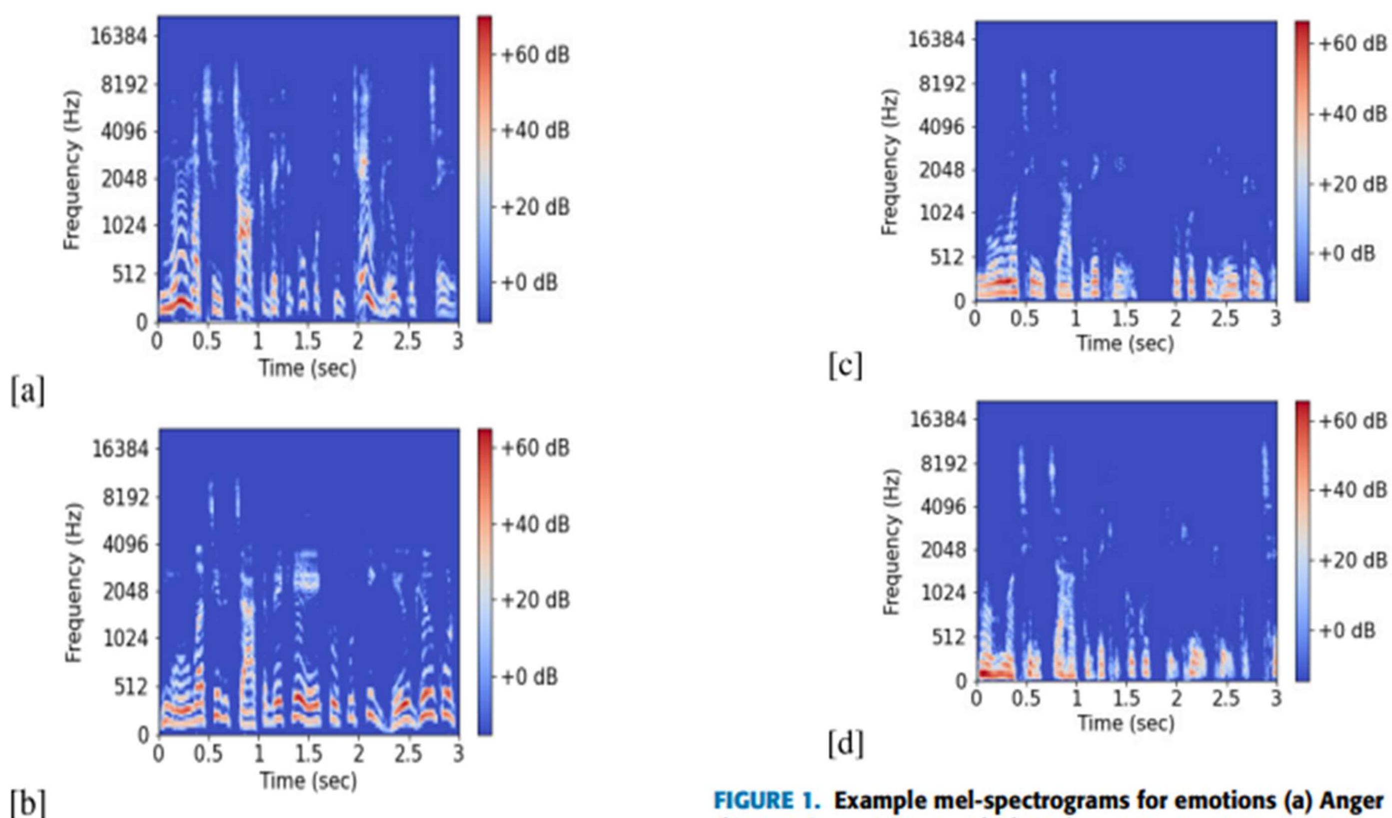
<table>
<tr>
<td>Software Tools
(What program/software is used for design, coding and simulation?)</td>
<td>Software tools that researchers often utilize for developing and evaluating deep learning models include:
<ol type="i">
Python is a popular programming language extensively used for machine learning and deep learning tasks due to its rich ecosystem of libraries and frameworks.
PyTorch is open-source deep learning framework that offers dynamic computational graphs. It provides a flexible and intuitive interface for building and training neural networks.
Librosa is a Python library specifically designed for audio and music analysis. It offers various functionalities for audio preprocessing, feature extraction (e.g., mel-spectrograms, MFCCs), and signal processing tasks.
Pandas is a data manipulation and analysis library in Python. It provides powerful data structures and data analysis tools that researchers often use for handling and processing large datasets.
</td>
</tr>
<tr>
<td>Test / Experiment
How to test and characterize the design/prototype?</td>
<td>

FIGURE 1. Example mel-spectrograms for emotions (a) Anger (b) Happiness (c) Neutral (d) Sadness.</td>
</tr>
</table>

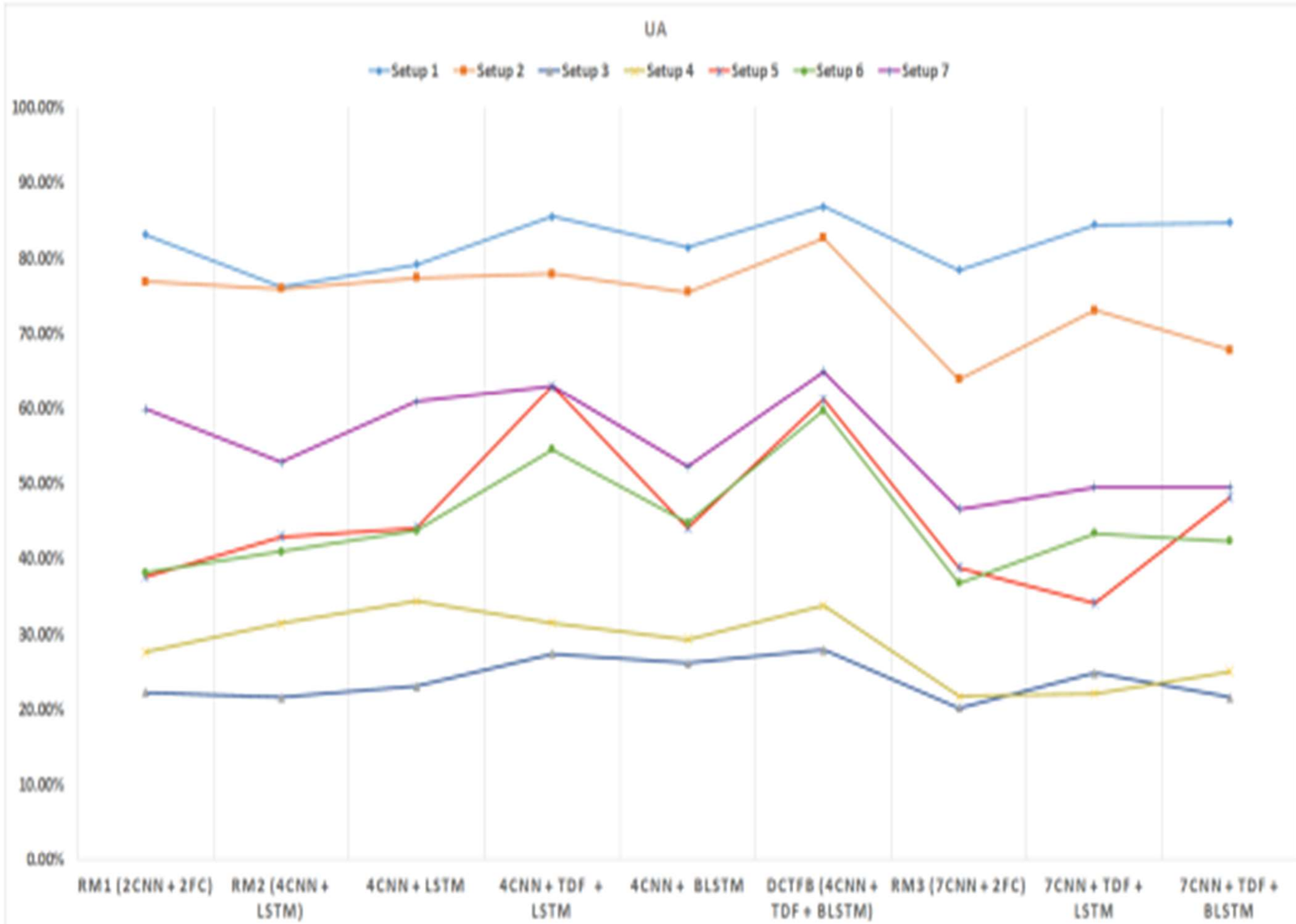
<table>
<tr>
<td>Simulation/Test Data
(What parameters are determined?)</td>
<td>


FIGURE 6. Line graph comparing UAs of all models for all setups.
</td>
</tr>
</table>

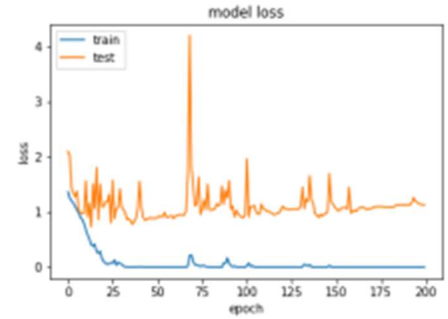


FIGURE 10. **Loss plot of RAVDESS training and testing with the proposed el.**

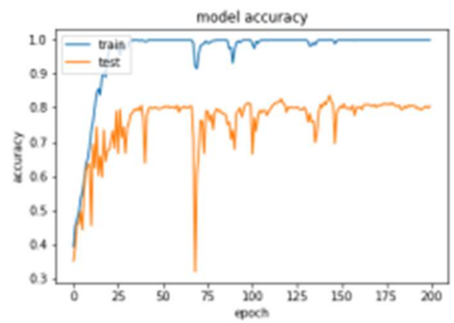


FIGURE 9. **Accuracy plot of RAVDESS training and testing with the proposed model.**

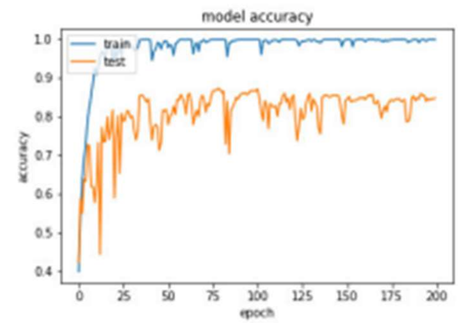


FIGURE 7. **Accuracy plot of SUBESCO training and testing with the proposed model.**

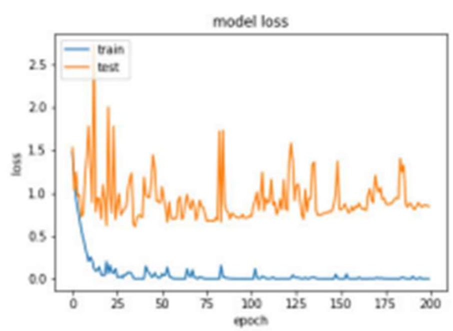


FIGURE 8. **Loss plot of SUBESCO training and testing with the proposed model.**

Result / Conclusion
(What was the final result?)

TABLE 16. **Accuracy matrices (%) for multi-lingual experiment for the proposed model.**

Emotion	Recall (TPR)	Specificity (TNR)	Precision (PPV)	Nvalue (NPV)	F1 Score
Anger	95.88	88.43	64.68	98.98	77.25
Fear	85.88	88.89	63.20	96.59	72.82
Happiness	55.29	92.27	62.25	89.95	58.57
Neutral	57.06	95.51	75.19	90.31	64.88
Sadness	30.59	95.10	59.77	85.21	40.47

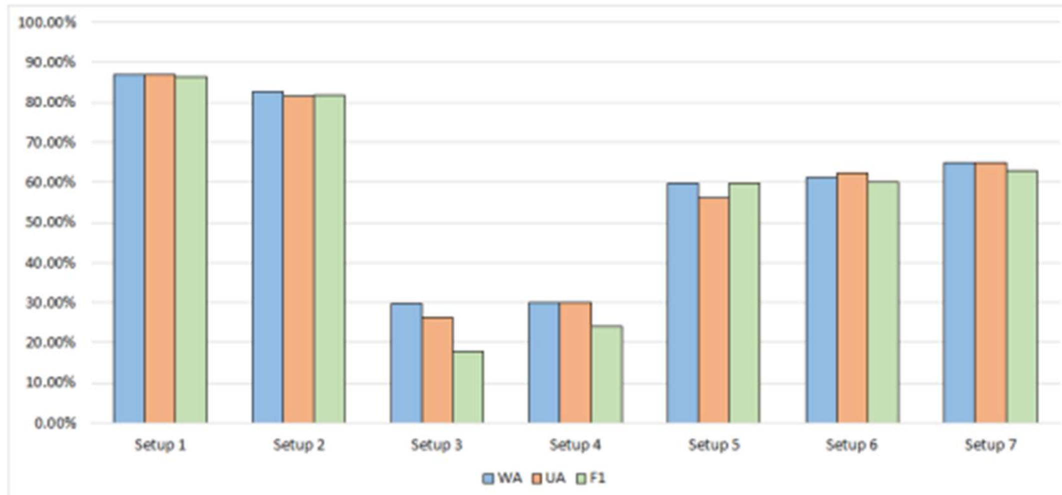


FIGURE 4. **WA, UA and F1 scores for different setups using DCTFB model.**

Obstacles/Challenges
(List the methodological obstacles if authors mentioned in the article)

Conclusion:

In conclusion, this study proposed a novel architecture for speech emotion recognition (SER) and demonstrated its state-of-the-art performance on the SUBESCO dataset for the Bangla language and the RAVDESS dataset for the English language. The proposed model outperformed existing implementations and achieved high accuracy in emotion prediction. Cross-lingual experiments using transfer learning showed satisfactory performance, indicating the potential application of the model to other languages.

Future Works:

1. Explore the use of a multi-dimensional dataset for the Bangla language to improve the performance and generalization capabilities of the SER system.
2. Investigate different deep learning techniques, augmentation methods to further enhance the accuracy of the SER system.

Terminology (List the common basic words frequently used in this research field)	Bangla SER, deep CNN, RAVDESS, SUBESCO, time-distributed flatten.

Review Judgment
(Briefly compare the objectives and results of all the articles you reviewed)

TABLE 5. Model comparisons for SUBESCO (Setup 1).

Model name	WA %	F1%
RM1 (2CNN+2FC)	83.14	82.68
RM2 (4CNN+LSTM)	76.14	76.22
4CNN+LSTM	79.14	79.13
4CNN+TDF+LSTM	85.57	85.56
4CNN+BLSTM	81.43	81.26
DCTFB (4CNN+TDF+BLSTM)	86.86	86.86
RM3 (7CNN+2FC)	78.43	78.15
7CNN+TDF+LSTM	84.43	84.26
7CNN+TDF+BLSTM	84.71	84.71

TABLE 8. Model comparisons for RAVDESS (Setup 2).

Model name	UA %	WA %	F1%
RM1 (2CNN+2FC)	76.92	75.41	74.76
RM2 (4CNN+LSTM)	75.96	74.70	74.43
4CNN+LSTM	77.40	76.62	76.50
4CNN+TDF+LSTM	77.88	76.97	76.84
4CNN+BLSTM	75.48	74.68	74.05
DCTFB (4CNN+TDF+BLSTM)	82.69	81.56	81.99
RM3 (7CNN+2FC)	63.94	62.05	61.88
7CNN+TDF+LSTM	73.08	71.54	71.55
7CNN+TDF+BLSTM	67.79	66.43	66.24

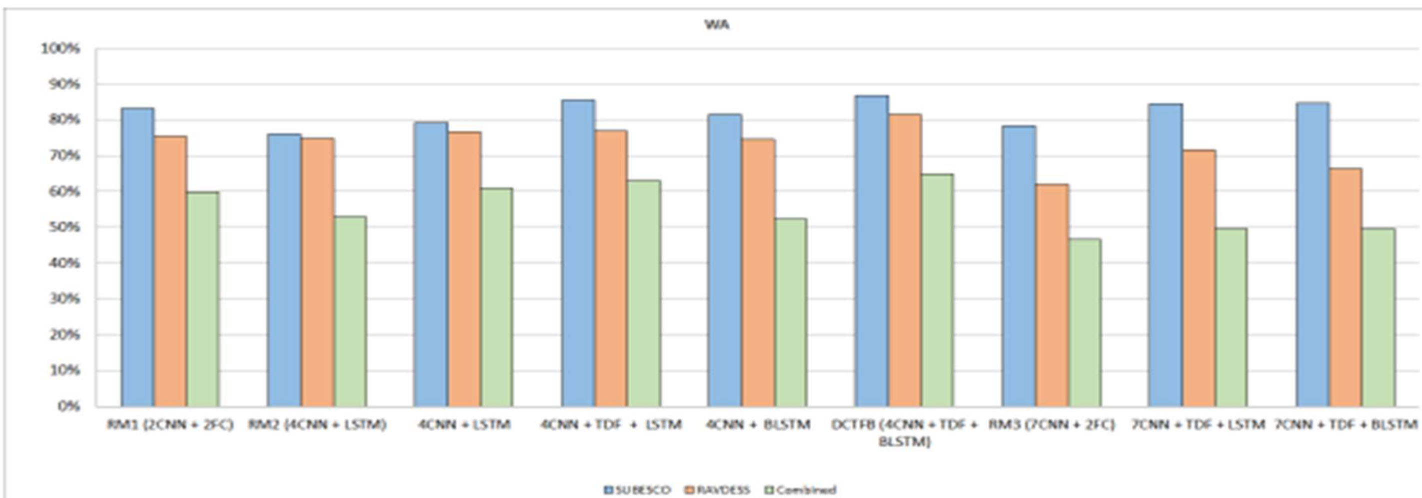


FIGURE 5. WA comparisons for different training datasets.

Review Judgment (Briefly compare the objectives and results of all the articles you reviewed)	

TABLE 7. Accuracy matrices (%) for SUBESCO dataset experiment for the proposed model.

Emotion	Recall (TPR)	Specificity (TNR)	Precision (PPV)	Nvalue (NPV)	F1 Score
Anger	80.00	97.56	84.21	96.77	82.05
Disgust	77.00	96.31	77.00	96.31	77.00
Fear	93.00	98.52	91.18	98.85	92.08
Happiness	83.00	97.56	84.69	97.24	83.84
Neutral	100.00	98.52	91.74	100.00	95.69
Sadness	86.00	99.50	96.63	97.72	91.01
Surprise	89.00	97.09	83.18	98.20	85.99

TABLE 10. Accuracy matrices (%) for RAVDESS dataset experiment for the proposed model.

Emotion	Recall (TPR)	Specificity (TNR)	Precision (PPV)	Nvalue (NPV)	F1 Score
Angry	92.11	97.70	89.74	98.27	90.91
Fearful	65.79	96.59	80.65	92.90	72.46
Happy	76.32	99.42	96.67	94.97	85.29
Neutral	94.64	94.41	85.48	98.06	89.83
Sad	78.95	91.40	65.22	95.51	71.43

Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	In this paper we have to choose CNNs model to use or refer than LSTMs because it improves emotion detection in low-resource languages and cross-lingual scenarios.

Aspects	Paper # 2 (Title)
Title / Question (What is problem statement?)	Acoustic feature analysis and optimization for Bangla speech emotion recognition.
Objectives / Goal (What is looking for?)	**Objective:** The objectives of this study are to investigate the acoustic properties of speech related to emotional expressions, analyze the effects of different emotions on speech characteristics, and identify key features that distinguish various emotions. Additionally, the study aims to optimize feature selection for emotion recognition by exploring different combinations of acoustic features and evaluating their effectiveness. The objectives also include comparing the performance of machine learning models, such as SVM, Random Forest, and XGBoost, in emotion recognition tasks, and determining the feature importance using XGBoost analysis. **Goals:** 1. To contribute to the field of emotional speech recognition and understanding human behavior. 2. To optimize emotion recognition systems by selecting the most relevant acoustic features. 3. To improve the accuracy of emotion recognition models using machine learning techniques. 4. To contribute to feature importance analysis and provide insights into the underlying factors of emotion perception.

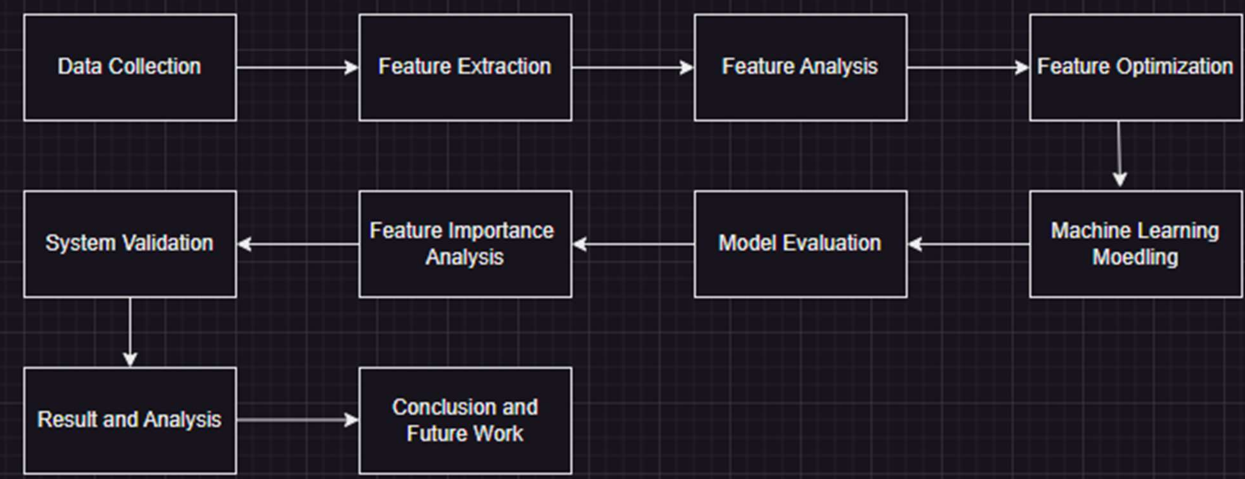
<table>
<tr>
<td>Methodology/Theory
(How to find the solution?)</td>
<td>

Fig 1. Block Diagram of SER System</td>
</tr>
<tr>
<td>Software Tools
(What program/software is used for design, coding and simulation?)</td>
<td>Several software tools can be utilized for acoustic feature analysis and optimization for Bangla speech emotion recognition. Here are some commonly used tools:

i. OpenSMILE (Open-Source Speech and Music Interpretation by Large Space Extraction) is a popular open-source toolkit for feature extraction from audio signals. It provides a collection of pre-defined acoustic features that can be used for speech emotion recognition. OpenSMILE supports various audio formats and offers flexibility in feature customization.

ii. Librosa is a Python library for audio and music analysis. It provides a wide range of functions and tools for feature extraction, including Mel-frequency cepstral coefficients (MFCCs), spectral contrast, and tonal centroid. Librosa simplifies the process of working with audio signals and extracting relevant features.

iii. scikit-learn is a powerful machine learning library in Python. It offers a comprehensive set of tools and algorithms for data preprocessing, model training, and evaluation. scikit-learn includes implementations</td>
</tr>
</table>

of various machine learning models, such as SVM, Random Forest, and XGBoost, which can be used for building emotion recognition models.

iv. MATLAB is a widely used software platform for scientific computing. It provides a range of toolboxes and functions for signal processing, feature extraction, and machine learning. MATLAB's Signal Processing Toolbox and Statistics and Machine Learning Toolbox can be utilized for acoustic feature analysis and optimization.

Test / Experiment
How to test and characterize the design/prototype?

Table 1 Effects of emotions on acoustic attributes for SUBESCO female speakers (Mean/Coefficient of variation).

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Pitch Hz	319/0.17	316/0.15	279/0.17	305/0.13	232/0.09	269/0.13	344/0.13
Intensity dB	81/0.03	78/0.03	77/0.05	78/0.03	73/0.06	77/0.05	79/0.04
F1 Hz	603/0.14	565/0.12	467/0.12	536/0.10	474/0.10	468/0.13	545/0.10
F2 Hz	1,668/0.05	1,625/0.06	1,647/0.09	1,689/0.06	1,666/0.08	1,661/0.08	1,589/0.05
F3 Hz	2,675/0.02	2,634/0.03	2,709/0.05	2,657/0.03	2,719/0.04	2,681/0.04	2,610/0.03
Speech Rate	High/low	Medium	Lower	Higher	Medium/low	Medium/low	Lower
Voice quality	Breathy, chest tone	Grumbled	Irregular	Breathy, sharp	Clear	Resonant	Clear, sharp
Articulation	Tensed	Normal	Normal	Excited	Normal	Slurring	Excited

Table 2 Effects of emotions on acoustic attributes for SUBESCO male speakers (Mean/Coefficient of variation).

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Pitch Hz	242/0.22	220/0.20	209/0.29	193/0.20	138/0.16	167/0.25	240/0.16
Intensity dB	81/0.02	80/0.03	77/0.05	78/0.03	74/0.06	76/0.06	80/0.03
F1 Hz	593/0.14	535/0.14	455/0.16	473/0.15	419/0.14	416/0.14	503/0.13
F2 Hz	1,683/0.06	1,624/0.06	1,617/0.07	1,651/0.06	1,592/0.08	1,573/0.07	1,614/0.06
F3 Hz	2,635/0.03	2,614/0.03	2,604/0.03	2,594/0.03	2,594/0.04	2,558/0.03	2,621/0.03
Speech Rate	Higher	Medium	Lower	High/Normal	Medium/High	Lower	Lower
Voice quality	Breathy, chest tone	Grumbled	Irregular	Breathy, sharp	Clear	Resonant	Clear, sharp
Articulation	Tensed	Normal	Normal	Excited	Normal	Slurring	Excited

Table 3 Effects of emotions on acoustic attributes for RAVDESS speakers (Mean/Coefficient of variation).

	Angry	Fearful	Happy	Neutral	Sad
Pitch (F) Hz	325/0.20	352/0.17	332/0.21	282/0.28	312/0.24
Intensity (F) dB	65/0.10	61/0.12	61/0.10	51/0.15	54/0.13
F1 (F) Hz	648/0.13	598/0.13	630/0.11	575/0.09	559/0.10
F2 (F) Hz	1,597/0.04	1,565/0.06	1,633/0.06	1,568/0.08	1,537/0.07
F3 (F) Hz	2,677/0.04	2,613/0.04	2,629/0.05	2,663/0.06	2,624/0.05
Pitch (M) Hz	202/0.26	191/0.27	185/0.19	149/0.23	164/0.21
Intensity (M) dB	65/0.11	58/0.12	59/0.10	50/0.13	52/0.15
F1 (M) Hz	574/0.11	518/0.10	520/0.08	485/0.08	478/0.09
F2 (M) Hz	1,523/0.06	1,505/0.06	1,498/0.06	1,469/0.06	1,451/0.06
F3 (M) Hz	2,510/0.01	2,468/0.01	2,482/0.01	2,485/0.02	2,438/0.02
Speech Rate	High/low	High/low	High/Normal	High/Normal	Lower
Voice quality	Breathy, chest tone	Breathy	Breathy, sharp	Clear	Resonant, clear
Articulation	Tensed	Slurring	Excited	Normal	Normal/Slurring

F: Female, M: Male

Simulation/Test Data
(What parameters are determined?)

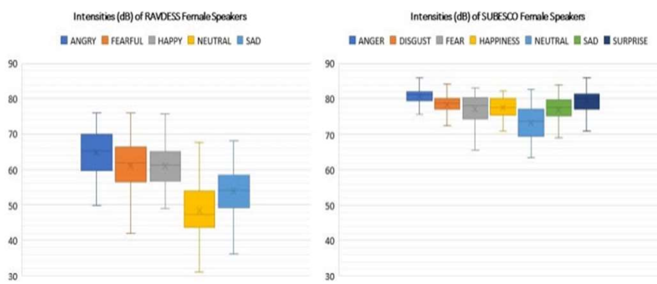


Fig. 1 Emotion-wise intensities for RAVDESS and SUBESCO female speakers.

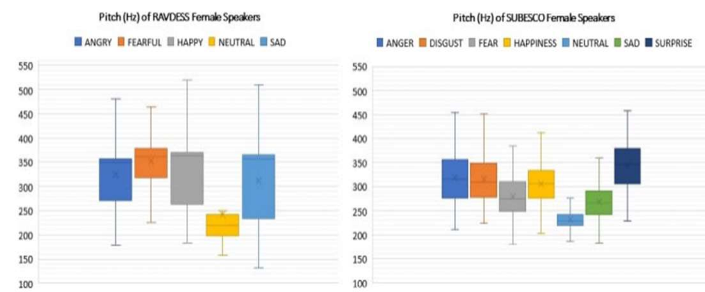


Fig. 3 Emotion-wise pitch for RAVDESS and SUBESCO female speakers.

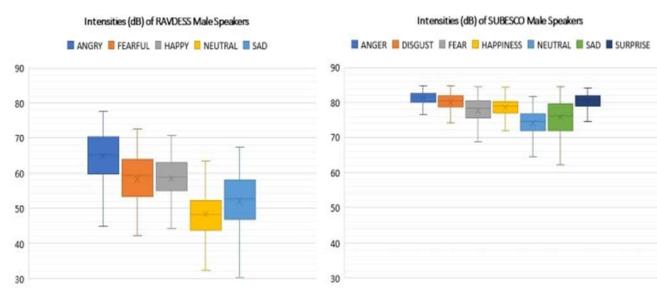


Fig. 2 Emotion-wise intensities for RAVDESS and SUBESCO male speakers.

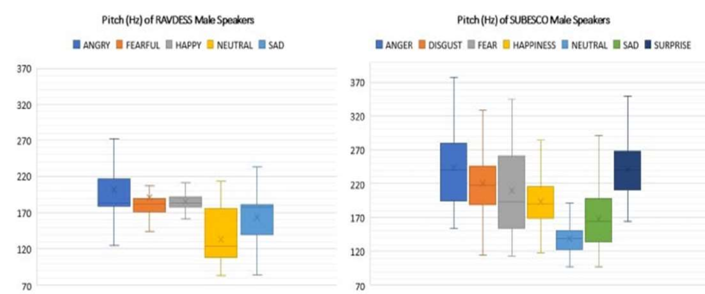


Fig. 4 Emotion-wise pitch for RAVDESS and SUBESCO male speakers.

Result / Conclusion
(What was the final result?)

Table 9 Accuracies for SUBESCO using Setup 1.

	MA%	Pr%	Re%	F1%	MCV%	SCV
SVM	78.48	78.47	78.43	78.17	76.0	0.01
RF	77.43	77.96	77.37	76.62	77.0	0.01
XGB	77.86	77.36	77.78	77.18	76.0	0.01

Table 10 Accuracies for SUBESCO using Setup 2.

	MA%	Pr%	Re%	F1%	MCV%	SCV
SVM	82.90	82.96	82.83	82.49	82.0	0.01
RF	80.90	81.01	80.90	80.33	80.0	0.02
XGB	80.10	78.96	79.11	78.71	78.0	0.02

Table 11 Accuracies for SUBESCO using Setup 3.

	MA%	Pr%	Re%	F1%	MCV%	SCV
SVM	81.62	81.74	81.50	81.23	81.23	0.01
RF	78.10	78.35	78.08	77.39	78.0	0.02
XGB	79.90	79.86	79.88	79.54	78.0	0.01

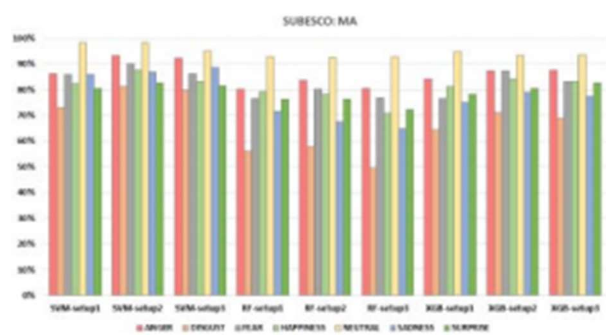


Fig. 7 Emotion-wise recognition rates for SUBESCO dataset.

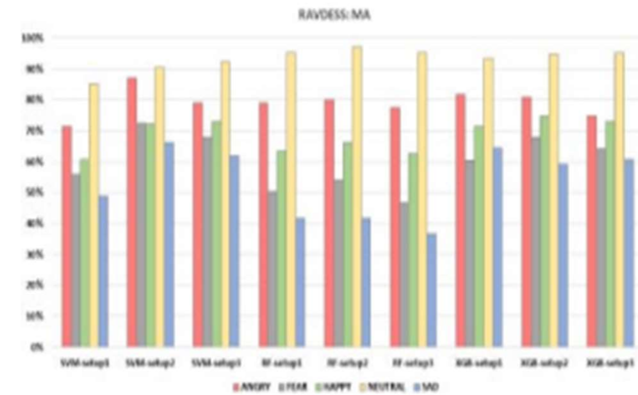


Fig. 8 Emotion-wise recognition rates for RAVDESS dataset.

Table 12 Accuracies for RAVDESS using Setup 1.

	MA%	Pr%	Re%	F1%	MCV%	SCV
SVM	76.17	75.28	74.51	74.70	76.0	0.02
RF	69.73	68.83	67.58	67.60	68.0	0.01
XGB	74.88	73.71	72.98	73.19	71.0	0.02

Table 13 Accuracies for RAVDESS using Setup 2.

	MA%	Pr%	Re%	F1%	MCV%	SCV
SVM	77.62	76.66	76.23	76.33	75.0	0.02
RF	71.50	71.27	69.07	69.20	71.0	0.02
XGB	77.13	75.86	75.36	75.50	72.0	0.02

Obstacles/Challenges
(List the methodological obstacles if authors mentioned in the article)

Conclusion:

In this study, acoustic feature analysis and optimization for Bangla and English speech emotion recognition were conducted. The findings revealed significant emotional cues specific to Bangla speech.

	Future work: Includes exploring more acoustic features, analyzing multiple languages, employing deep learning approaches, developing real-time emotion recognition systems, and investigating multimodal emotion recognition.
Terminology (List the common basic words frequently used in this research field)	Speech emotion recognition, Feature optimization, Machine learning, SVM, Random forest, XGBoost.

Review Judgment
(Briefly compare the objectives and results of all the articles you reviewed)

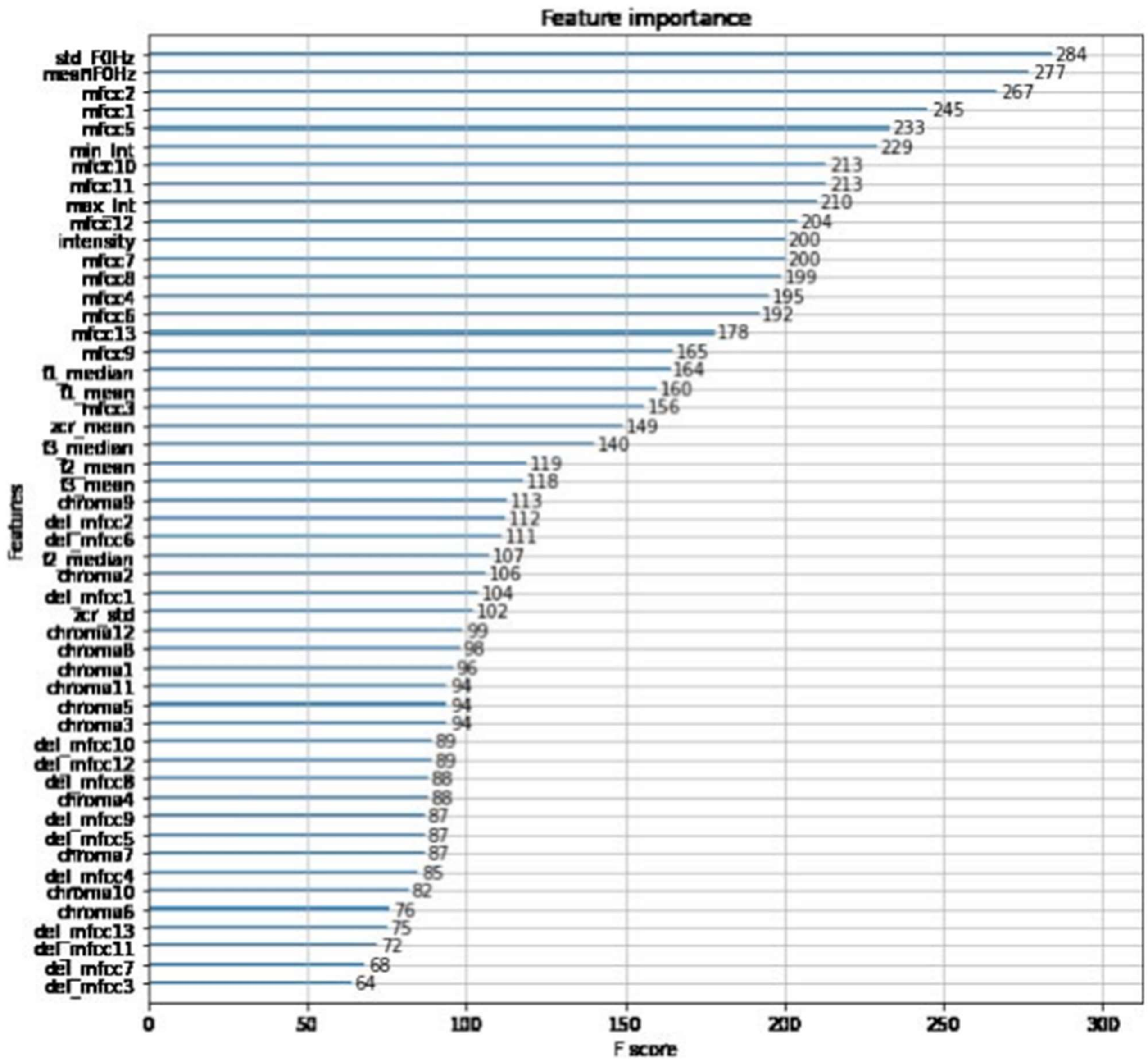


Fig. 5 Feature importance calculated for SUBESCO.

Review Judgment
(Briefly compare the objectives and results of all the articles you reviewed)

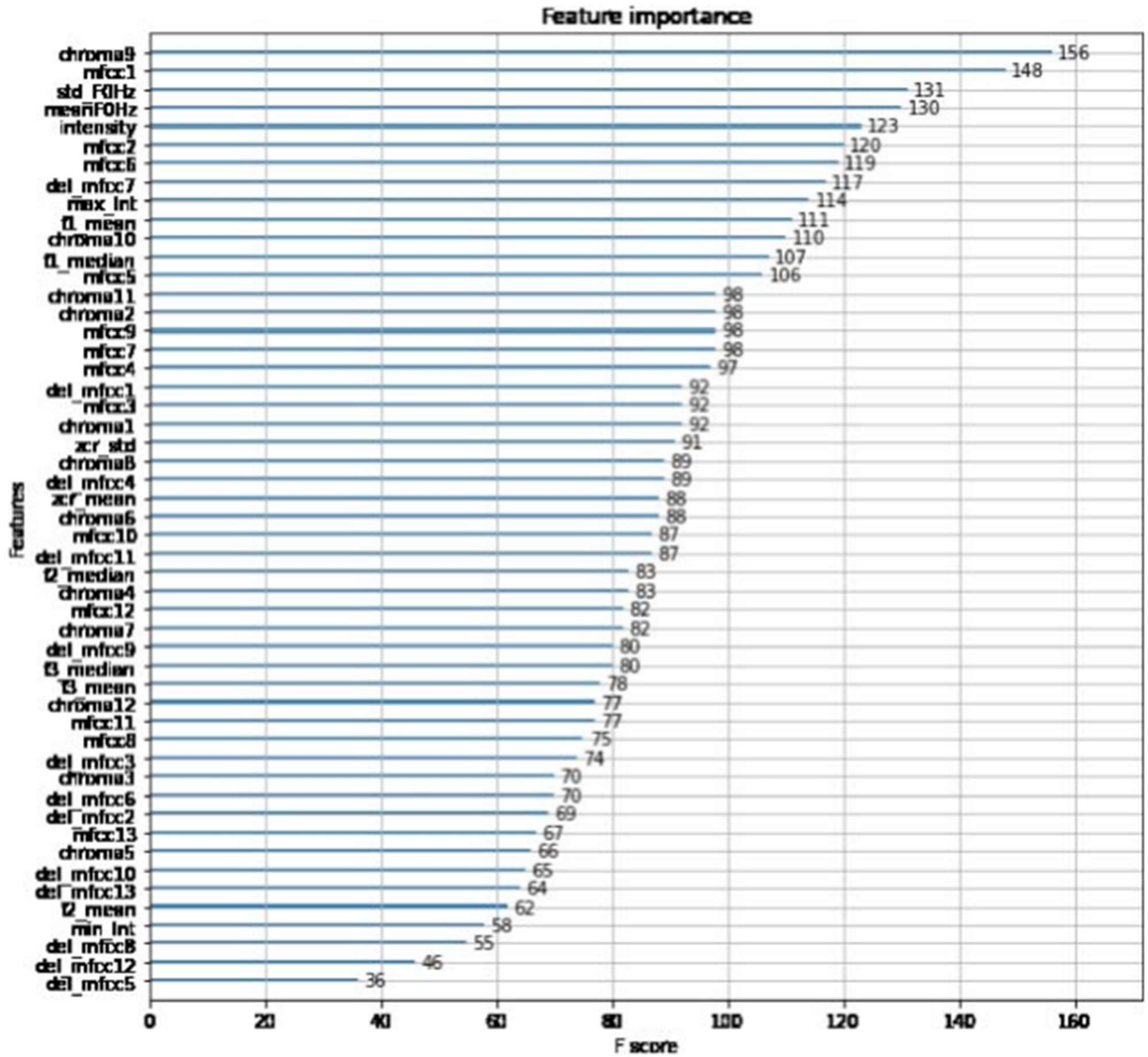


Fig. 6 Feature importance calculated for RAVDESS.

Review Outcome (Make a decision how to use/refer the obtained knowledge to prepare a separate and new methodology for your own research project)	In this paper we have to choose XGBoost model to use or refer instead of SVM, Random Forest because it provides insight into the underlying factors of emotion perception.