Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 2 (2024)

Utilizing Artificial Intelligence to Generate Emotive Voicing for Audiobook

Dhanush V[1]*, Chinmay Aland[2]†, Nandish A[3], Prasanna Chitgopekar[4]†, Ashwini M Joshi[5]†

[1*,2,3,4,5] *Department of Computer Science and Engineering, PES University, Bengaluru, 560085, Karnataka, India.*

Abstract. Audiobooks have seen immense growth and popularity, with over $5.38 billion in revenue in 2022 alone. However, most audiobook narrations still lack the emotional expressiveness present in human storytelling. This research examines the potential for utilizing artificial intelligence (AI) to generate emotive voic- ing for audiobooks. We evaluated neural network models for inferring emotion solely from textual passages. Both BERT and GPT-2 were trained to catego- rize excerpt emotions, achieving comparable accuracy. To further assess emotion detection capabilities, we analyzed valence, arousal, and dominance scores pre- dicted by each model. On this more granular metric, BERT demonstrated superior performance in capturing nuanced text emotions.For classifying emotions in audio, we leveraged a pretrained wav2vec2 model. However, when evaluating this model on existing audiobook recordings, we found it tended to categorize most clips into only two dominant emotions. Therefore, we opted to use GPT-2's text-based emotion predictions for labeling our training data because BERT model is more sensitive to changes in the data, but it is also less precise than the GPT-2 model. We then used Speech-T5 text-to-speech models tailored to five target emotions, training individual models on matched text-audio pairs.

Keywords: *Audiobooks, Artificial Intelligence, Neural Network Models, Emotion Classification, Nuanced Text Emotions, Sequence-To-Sequence Model*

1 Introduction

Audiobooks aim to provide an engaging experience that brings stories to life through expressive vocal performances. However, audiobook narration today relies solely on hired voice talent to manually interpret the emotions and delivery of all passages. While talented narrators can provide compelling performances, this labor-intensive process limits scalability and expressiveness [1]. Recent advances in artificial intel- ligence present an opportunity to bring more dynamic emotional narration to audiobooks through automated methods.

This research explores combining natural language processing and speech synthe- sis techniques to automatically generate emotive narration tailored to the emotions detected in audiobook text [2], [3]. We first evaluate two state-of-the-art natural lan- guage processing models, BERT and GPT-2, on detecting emotions like joy, sadness, and anger in audiobook transcripts to determine the more accurate approach for text- based emotion classification [4], [5]. We then leverage neural text-to-speech technology to synthesize emotional speech aligned with those text predictions [6]. Specifically, we train a Speech-T5 model on an acted emotional speech dataset to generate voices adapted to the target emotions..

This work demonstrates the feasibility of an end-to-end AI pipeline for injecting dynamically generated emotional voicing into audiobooks, potentially enabling more scalable and expressive audiobook creation [7]. The approach could help mitigate the challenges of manual narration while enhancing immersion and engagement through data-driven modeling of textual and vocal emotion.

We structure the paper as follows: First, by providing some background work related the AI techniques we leveraged. Then we will describe how we collected data and our overall research methodology. After laying out our approach, we will go into the details of how we built and trained models for text-based emotion detection, audio-based emotion detection, and emotional speech synthesis. Next, we will present results evaluating the performance of our models. Finally, I'll conclude by summarizing the key findings, limitations, and potential future improvements to build on this research. The goal is to walk through the key steps of our investigation into using AI for more expressive audiobook narration.

2 Literature Review

The ability to convey emotion through synthesized speech is critical for creating immersive and engaging audiobook narrations. This literature review synthesizes key findings from recent research on developing AI systems capable of emotional speech synthesis. A particular focus is placed on approaches leveraging machine learning and neural networks.

Recent research has explored various techniques for emotional speech synthesis and detecting emotion from text. Skerry-Ryan et al. [8] developed an end-to-end neural network for speech synthesis with controllable emotion. Baevski et al. [9] proposed wav2vec 2.0, a framework for self-supervised speech representation learning. Ao et al.

[10] presented SpeechT5, a unified sequence-to-sequence model for spoken language tasks.

For emotion detection, Xing et al. [11] compared word embeddings using convo- lutional neural networks. Poria et al. [12] proposed a hybrid approach for multi-label emotion detection. To advance Transformer models, Devlin et al. [13] introduced BERT for bidirectional language representation. Vaswani et al. [14] developed the attention-based Transformer architecture underlying models like BERT and GPT-2.

Additional works have furthered multimodal emotion recognition and analysis of sentiment and intensity. Prasad et al. [15] developed a multimodal system combining audio and text for speech emotion recognition. Sharir et al. [16] classified poetry into emotions using LSTMs.

Das and Gamb¨ack analyzed informal versus formal text for emotion detection. Poria et al. distinguished polarity and intensity using multi-task learning. Rachman et al.leveraged a corpus labeled with Ekman emotions. Islam et al. detected emotion in text using valence-arousal modeling. Fatichah et al. created an annotated Indonesian text corpus. Montoro et al. proposed unsupervised fuzzy logic for sentiment analysis.

While substantial progress has been made, accurately modeling human-like affec- tive expression remains challenging. Further work on large datasets, improved evaluation, and multimodal neural modeling will be key to continued advances in emo- tionally intelligent AI systems. This review summarizes promising contributions while underscoring opportunities for additional research.

3 Dataset Construction and Feature Engineering

3.1 Data Collection

Text Emotion detection data.

The text emotion classification dataset was compiled from two Kaggle sources. The first contained over 20,000 English samples labeled with 6 emotions - sadness, anger, love, surprise, fear, joy. The second provided 7,480 additional samples covering a sim- ilar set of 6 emotions, all "love" labeled samples were discarded from first dataset and replaced with "disgust" samples from the second dataset. This resulted in a dataset of 19,415 text across 6 target emotions - sadness, anger, disgust, surprise, fear, and joy.

VAD data.

The NRC Valence, Arousal, and Dominance lexicon provides fine-grained emotion scores for over 20,000 English words. Each word is assigned a value from 0 to 1 for valence, arousal, and dominance. This granular scoring

supports modeling of nuanced emotions beyond basic categorical labels. We found many OOV(Out Of Vocabulary) words both GPT-2 and BERT handle OOV words by breaking them into subword units.

GPT-2 has a very large vocabulary 50,000 subwords, uses byte-pair encoding (BPE), which break words into subword units. This allows representing OOV words as a sequence of subwords that are in-vocabulary. BERT has vocab 30,000 subwords, Uses WordPiece tokenization, another subword segmentation method OOV words are also broken into subword units present in the vocabulary.

For example, ”extraterrestrial” may be broken into ”extra”, ”terrestr”, ”ial”.

TTS data.

The LibriSpeech audiobook dataset, which contains 100 hours of aligned text and audio, was leveraged to provide emotional speech samples for TTS model training. The data was balanced by subsampling to obtain 2,755 samples per emotion, yielding 13,775 total samples across 5 target emotions.

Audio emotion detection data:.

A dataset of 3,432 audio samples equally representing the 6 target emotions was constructed by combining samples from the RAVDESS and TESS datasets. The audio clips portray vocal emotions through speech.

All datasets were divided into training (80%), validation (10%), and test (10%) splits to support model development and evaluation.

3.2 Emotion Detection in text

This research aims to evaluate BERT and GPT-2 models for emotion classification in text. The models will be assessed on their ability to accurately categorize emotions in textual data. Valence, arousal, and dominance scores from each model will be analyzed to determine effectiveness at capturing nuanced emotions.

Emotional Voice Synthesis.

Since our audiobook dataset didn’t originally have any emotion labels, we needed to add them. To start with, we built a Wav2Vec2 model to categorize the emotions directly from the audio clips. Wav2Vec2 is good at speech recognition tasks, so by training it on the raw audio, it could learn to predict emotion just from the sounds. This lets us automatically label the data based on the audio.The Speech-T5 models will be trained on text-audio pairs that express the same emotion. Training data will be generated by combining GPT-2’s emotion-labeled text with corresponding emotional audio clips.

The effectiveness of the Speech-T5 models in conveying the appropriate emotions will be evaluated using both objective metrics and subjective listening tests. Eval- uation will assess emotional accuracy, naturalness of speech, and listener preference between models. This comprehensive assessment will determine how well the TTS models capture emotional nuance in the synthesized audio.

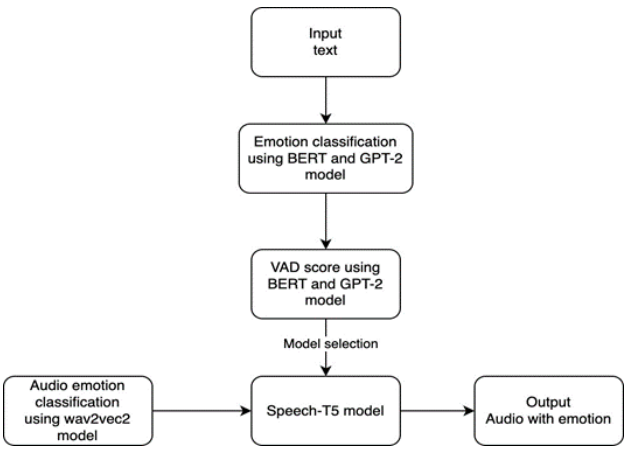

Fig. 1. A Multimodal Approach to Emotion Detection and Synthesis

Fig. 1 encapsulates the research design adopted, spanning emotion detection in text via BERT and GPT-2, audio classification, and emotive speech synthesis using matched text-audio data. We proceed to explain this framework through dataset collection, model development, and evaluative findings.

4 Model Architectures

The Model architectures is outlined in Fig. 2. Key steps were text and audio emo- tion classification and emotive speech synthesis modeling. We now detail each phase, beginning with dataset compilation.

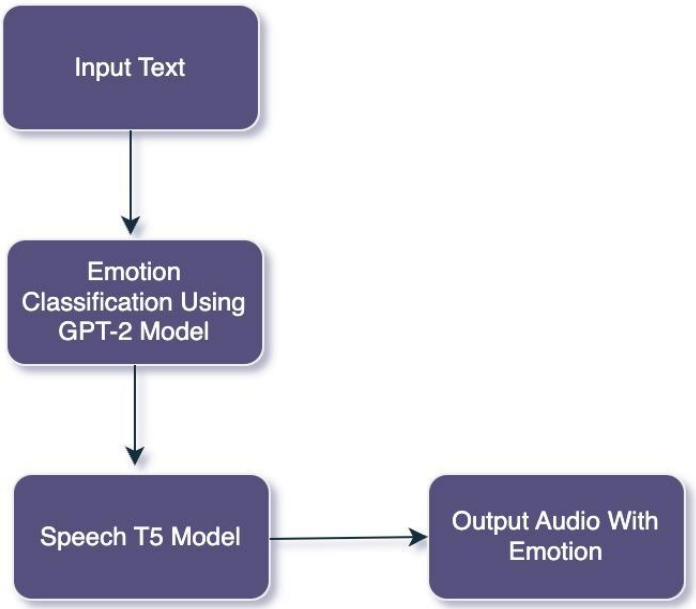

Fig. 2. Model Architectures for Emotion Detection and Synthesis

4.1 Text Emotion Classification

The text dataset is preprocessed by tokenizing and encoding the samples into numer- ical sequences using the BERT and GPT-2 tokenizer models. The tokenized texts are padded to a maximum sequence length of 512 tokens. For model training, an 80/10/10 train/validation/test split is created.

The BERT and GPT-2 architectures are initialized with pre-trained weights from the HuggingFace repository to leverage transfer learning. The models are then fine-tuned on the text emotion classification dataset for 10 epochs using an Adam optimizer with a learning rate of 1e-5, batch size of 16, and binary cross-entropy loss function.

The fine-tuned models are evaluated on the test dataset for multi-class emotion classification across the 6 categor- ical emotion labels. Classification accuracy is quan- tified by overall accuracy, F1-score, and a confusion matrix comparing true versus predicted labels. Additionally, to assess ability to model nuanced emotions beyond dis- crete categories, the models are also trained to predict continuous valence, arousal, and dominance scores. The ground truth valence, arousal, and dominance (VAD) labels are derived from the NRC dataset. A regression model is trained to predict these con- tinuous VAD scores by minimizing the mean squared error (MSE) loss between the predicted and true scores. Performance is evaluated using the mean absolute error (MAE) and root mean squared error (RMSE) between predictions and labels.

4.2 Emotional Speech Synthesis

The raw audio samples from the LibriSpeech dataset are preprocessed by first extracting log-Mel spectrogram features to represent the speech in a compact, perceptually-relevant time-frequency representation. The log-Mel spectrograms are extracted using a 25ms sliding Hamming window with 10ms frame shift. This gen- erates 43-dimensional feature vectors spanning a 123-Hz to 7.6kHz frequency range. The corresponding text transcripts are tokenized using the T5 tokenizer's subword encoding algorithm [34]. This encodes the text into to-ken IDs while

allowing han- dling of out-of-vocabulary words through subword decomposition. The tokenizer is configured with a 32,000 token vocabulary for encoding the textual data.

Speaker embeddings are then generated from the speech samples using a pre- trained speaker verification model based on the Wav2Vec2 architecture [16]. This model encodes raw waveform audio into a 256-dimensional latent vector capturing dis- tinguishing voice characteristics. The speaker embeddings are incorporated into the dataset to allow learning speaker variance.

The audiobook dataset originally only had audio clips and transcripts, without any emotion labels. To start labeling, we first built a Wav2Vec2 model for audio-based emotion classification of the clips. We trained this model on a balanced dataset of 512 clips for each of the 6 basic emotions, since the original audio dataset had an imbalanced distribution. However, just using the audio analysis to label the emotions can resulted in predicting the 2 most common emotions of our audiobook dataset. To address this issue, we also utilized GPT-2 for text-based emotion predictions. The textual modality conveys important affective signals through word choice, semantics, and language style. We passed the transcripts to the GPT-2 model to get emotion labels based on the text. This gave us 5 different emotion labels. One emotion initially only had 2,755 examples, far less than the others. To balance the data for training our TTS model, we downsampled the other emotions to also have 2,755 examples each. In total this gave us an audiobook dataset of 13,775 clips, with transcripts and balanced emotion labels derived from both audio and text analysis.

Speech-T5 models are instantiated from the pretrained checkpoint and fine- tuned on text-audio pairs expressing the target emotions. The models employ the standard Speech-T5 encoder-decoder architecture with 24 layers and 1024 hidden dimensions.Training is performed on a single T4 GPU for 3000 steps using a batch size of 32, AdaFactor optimizer with decay rate 0.8, learning rate of 1e-5, mixed precision (FP16), and gradient checkpointing to reduce memory footprint.

During real-time inference, input text is tokenized, classified into one of the 5 emotion categories using the pretrained text classifier, and passed to the corresponding Speech-T5 model tailored to that emotion. The model generates the Mel spectrogram feature output, which is then converted to a waveform using a vocoder. This pipeline enables end-to-end emotive speech synthesis from text.

5 Analysis and Results

The literature review indicated the potential of leveraging BERT for emotion detec- tion, while GPT-2 was chosen experimentally to classify emotions from text. As shown in Table 1, BERT and GPT-2's loss, accuracy, and F1 score on the test set

Table 1. Performance Metrics of BERT and GPT-2's loss, accuracy, and F1 score

Model Name	Loss	Accuracy	F1 Score
BERT	0.1039	0.9622	0.9558
GPT-2	0.1078	0.9507	0.9569

Fine-tuning both models yielded comparable performance on text-based emo- tion classification. However, differences emerged between BERT and GPT-2. BERT demonstrated greater sensitivity to data fluctuations, grasping nuanced emotions more comprehensively, though also prone to errors like misclassifying emotions. As a bidirec- tional model processing the full sentence context together, BERT better understands ambiguous or context-dependent words. GPT-2's unidirectional nature limits its con- text to preceding words, increasing susceptibility to errors with complex sentences. Fig 3 and Fig 4 for the detailed training plots, validation metrics, and confusion matrices of each model.

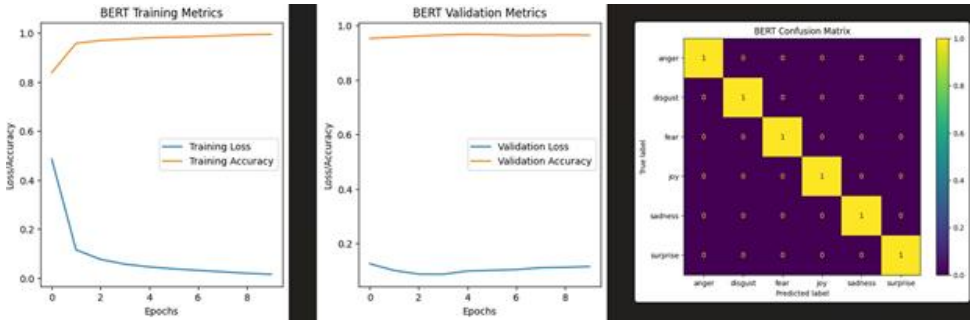

Fig. 3. BERT's training, validation metrics and confusion matrix.

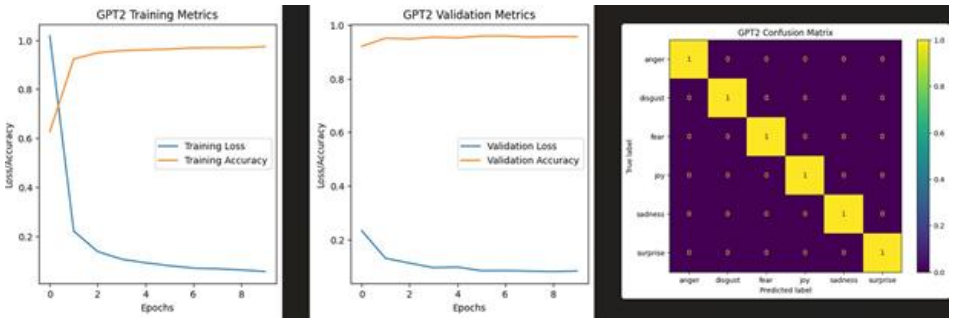

Fig. 4. GPT-2's training, validation metrics and confusion matrix.

For predicting valence, arousal and dominance, BERT and GPT-2 achieved similar outcomes regarding lower MSE, average loss, MAE and RMSE as shown in Table 2.

Table 2. Performance Metrics of BERT and GPT-2 for VAD modeling

Model Name	MSE	Loss	MAE	RMSE
BERT	0.0011	0.0182	0.1796	0.2324
GPT-2	0.0068	0.1085	0.0072	0.1144

But BERT's responsiveness to data changes made it less precise than GPT-2. Given these insights, GPT-2 was selected for further evaluation. See Fig 5 and Fig 6 for true versus predicted comparisons of valence, arousal, and dominance scores from each model.

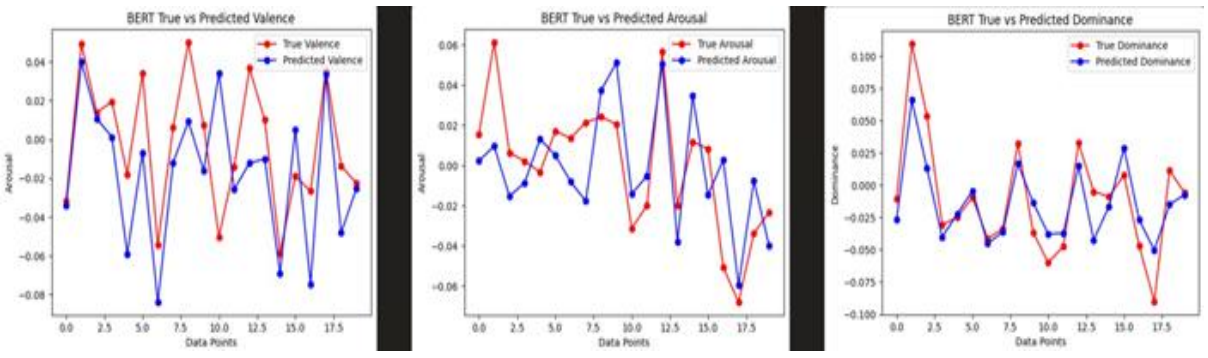

Fig. 5. BERT's true vs predicted valence, arousal, and dominance.

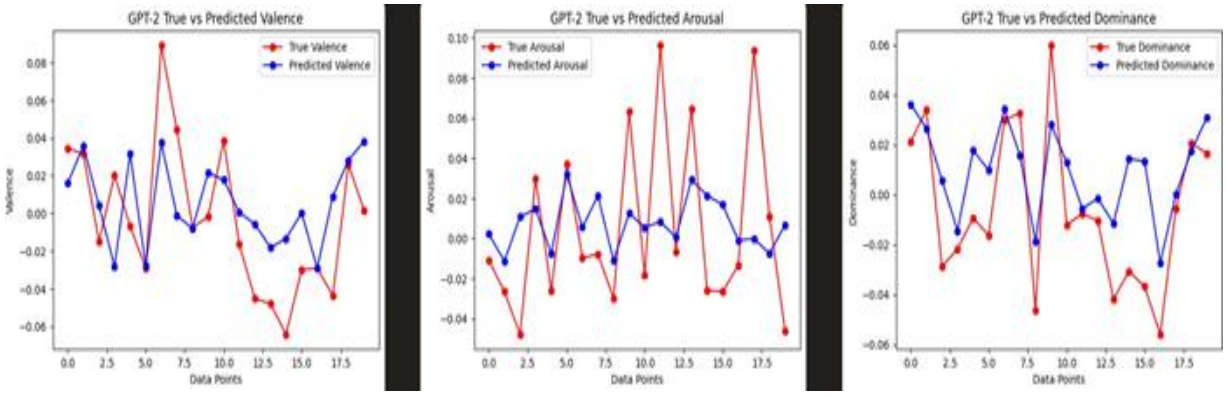

Fig. 6. GPT-2's true vs predicted valence, arousal, and dominance.

A Speech-T5 text-to-speech model synthesized audio conveying distinct emotions, generating five variants. Training a single TTS model would produce naturalistic but emotionally invariant speech.

6 Conclusion and Future Work

6.1 Conclusion

This research explored transformer-based models BERT and GPT-2 for text emo- tion classification, finding GPT-2 better met objectives of precision and compatibility with system consolidation. Fine-tuning both models yielded comparable accuracy, but analyses highlighted GPT-2's greater robustness on linguistically complex samples. This aligns with the aim to integrate text classification into an end-to-end pipeline for emotive text-to-speech synthesis.

GPT-2 demonstrated superior precision critical for reliable emotion labeling to train the speech decoder. Its simplicity over BERT also better enables system inte- gration into a unified model. GPT-2 was thus selected as the text encoder to provide emotion annotations and enable contextualized speech generation.

The pipeline synthesized perceptually distinct speech variants using a Speech- T5 model conditioned on emotion categories predicted by GPT-2. This represents progress toward adaptable, emotionally expressive human-computer interaction. Fur- ther enhancements to model integration and transfer learning can build on these foundations.

6.2 Future Work

In the future, we plan to enhance the Speech T5 text-to-speech (TTS) model to generate emotional speech through a multi-task learning approach. This eliminates the need to train separate TTS models for each emotion. In the multi-tasking framework, one component will focus on classifying emotions, while another distinct component will specialize in learning to convey those different emotions through generated speech

References

1. Wu, Yihan, Xi Wang, Shaofei Zhang, Lei He, Ruihua Song, and Jian-Yun Nie. "Self-supervised context-aware style representation for expressive speech synthesis." ar Xiv preprint arXiv:2206.12559 (2022).
2. [02] Cahyani, Denis Eka, Aji Prasetya Wibawa, Didik Dwi Prasetya, Langlang Gumilar, Fadhilah Akhbar, and Egi Rehani Triyulinar. "Emotion Detection in Text Using Convolutional Neural Network." In 2022 International Conference on Electrical and Information Technology (IEIT), pp. 372-376. IEEE, 2022.
3. Mahima, M. A., Nidhi C. Patel, Srividhya Ravichandran, N. Aishwarya, and Sumana Maradithaya. "A Text-Based Hybrid Approach for Multiple Emotion Detection Using Contextual

and Semantic Analysis." In 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1-6. IEEE, 2021.

4. Kaur, Jasleen, and Jatinderkumar R. Saini. "Emotion detection and sentiment analysis in text corpus: a differential study with informal and formal writing styles." International Journal of Computer Application, ISSN (2014): 0975-8887.

5. Tian, L., C. Lai, and J. D. Moore. "Polarity and intensity: The two aspects of sentiment analysis. arXiv." (2020).

6. Akhtar, Md Shad, Deepanway Ghosal, Asif Ekbal, Pushpak Bhattacharyya, and Sadao Kurohashi. "All-in-one: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. " IEEE transactions on affective computing 13, no. 1(2019): 285-297.

7. Saravanakumar, S., & Thangaraj, P. (2019). A computer aided diagnosis system for identifying Alzheimer's from MRI scan using improved Adaboost. Journal of medical systems, 43(3), 76.

8. Kumaresan, T., Saravanakumar, S., & Balamurugan, R. (2019). Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine. Cluster Computing, 22(Suppl 1), 33-46.

9. Saravanakumar, S., & Saravanan, T. (2023). Secure personal authentication in fog devices via multimodal rank-level fusion. Concurrency and Computation: Practice and Experience, 35(10), e7673.

10. Thangavel, S., & Selvaraj, S. (2023). Machine Learning Model and Cuckoo Search in a modular system to identify Alzheimer's disease from MRI scan images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11(5), 1753-1761.

11. Saravanakumar, S. (2020). Certain analysis of authentic user behavioral and opinion pattern mining using classification techniques. Solid State Technology, 63(6), 9220-9234.

12. Islam, Md Rakibul, and Minhaz F. Zibran. "DEVA: sensing emotions in the valence arousal space in software engineering text." In Proceedings of the 33rd annual ACM symposium on applied computing, pp. 1536- 1543. 2018

13. Rachman, Fika Hastarita, Riyanarto Sarno, and Chastine Fatichah. "CBE: Corpus-based of emotion for emotion detection in text document." In 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), pp. 331-335. IEEE, 2016.

14. Sekaran, R., Munnangi, A. K., Ramachandran, M., & Gandomi, A. H. (2022). 3D brain slice classification and feature extraction using Deformable Hierarchical Heuristic Model. Computers in Biology and Medicine, 149, 105990-105990.

15. Ramesh, S. (2017). An efficient secure routing for intermittently connected mobile networks. Wireless Personal Communications, 94, 2705-2718.

16. Sekaran, R., Al-Turjman, F., Patan, R., & Ramasamy, V. (2023). Tripartite transmitting methodology for intermittently connected mobile network (ICMN). ACM Transactions on Internet Technology, 22(4), 1-18.

17. Montoro, Andres, Jose A. Olivas, Arturo Peralta, Francisco P. Romero, and Jesus Serrano-Guerrero. "An ANEW based fuzzy sentiment analysis model." In 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1-7. IEEE, 2018.

18. Asiya, U. A., and V. K. Kiran. "A Novel Multimodal Speech Emotion Recogni- tion System." In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), pp. 327-332.IEEE, 2022.

19. Ahmad, Shakeel, Muhammad Zubair Asghar, Fahad Mazaed Alotaibi, and Sher- afzal Khan. ”Classification of poetry text into the emotional states using deep learning technique.” IEEE Access 8 (2020): 73865-73878.

20. Jain, Gourank, Satyam Verma, Honey Gupta, Saloni Jindal, Mukesh Rawat, and Kapil Kumar. ”Machine Learning Algorithm Based Emotion Detection Sys- tem.” In 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 270-274. IEEE, 2022.

21. Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. ”wav2vec 2.0: A framework for self-supervised learning of speech representations.” Advances in neural information processing systems 33 (2020): 12449-12460.

22. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ”Bert: Pre- training of deep bidirectional transformers for language understanding.” arXiv preprint arXiv:1810.04805 (2018).

23. Ao, Junyi, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu et al. ”Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing.” arXiv preprint arXiv:2110.07205 (2021).

24. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin. ”Attention is all you need.” Advances in neural information processing systems 30 (2017).