International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Explainable AI Models for Voice-Based Content Classification in Streaming Television Platforms

[1]Saraschandra Arveti, [2]Anish Hadkar, [3]Mani Teja Nutalapati

Submitted:02/11/2024 **Revised**: 11/12/2024 **Accepted:** 23/12/2024

Abstract : The swift growth of streaming television platforms has made it possible to have a tremendous amount of audiovisual content that needs to be classified efficiently and transparently for recommending, moderating, and making the content accessible. In this paper, we put forward an explainable artificial intelligence (XAI) system for voice, based content classification that detects categories of programs by using speech signals extracted from streaming media. Our framework combines deep audio feature extraction with interpretable machine learning techniques that result in both excellent prediction performance and transparency of the decision, making process. Speech fragments are initially handled with the Mel, frequency cepstral coefficients (MFCCs) and spectral embeddings. These are later given to a hybrid structure that integrates a convolutional neural network (CNN) and a transformer, based attention layer. To strengthen the interpretability, the SHAP, driven explanation parts emphasize the most significant acoustic features and temporal voice patterns that have a great impact on classification decisions. The tool is test on a multi, genre streaming dataset comprising news sports entertainment, documentaries, and advertisements. The experimental results show excellent results and interpretability gains. The suggested model obtains 92.6% classification accuracy, which is 11.4% higher than that of baseline audio classifiers. The explainability module correctly detects essential speech signals with 89% explanation consistency. Furthermore, inference latency is cut by 27%, making near real, time deployment possible. Moreover. This model helps to improve genre recommendation precision by 18% in simulated streaming situations.

Keywords: *Explainable AI, Voice Classification, Audio Features, Streaming Platforms, Speech Analysis and Deep Learning.*

1. Introduction

The rapid expansion of digital streaming platforms has completely changed the way we create, share, and enjoy multimedia content. Today's streaming TV services offer a massive amount of programming including news sports entertainment, documentaries, and commercials. To handle and suggest this huge amount of content effectively, we need automated classification methods that can deal with large [1], scale audiovisual data. Old metadata-based classification methods depend a lot on manual tagging and textual descriptions, which are often missing or not consistent. Therefore, new research is focusing on voice, based content classification methods that examine speech signals found in TV broadcasts and streaming media to automatically identify program types and context [1].

Speech and audio comprise not only the verbal content but also the semantic and paralinguistic features e.g. tone pitch emotion, home background, etc. that can be used to understand the content better. Machine learning and deep learning methods have made it possible to generate features such as MFCC, spectrograms, and temporal embeddings from audio streams. These features, or representations of sound segments, are great inputs to neural networks for tasks like speech recognition, speaker identification, emotion detection, and audio event classification. Architectural designs of deep neural networks, especially CNNs and transformer, based models are very effective to capture local spectral patterns and long-term temporal dependencies in speech signals which in turn lead to good performance in audio analysis tasks [2].

Even though deep learning models have excellent predictive power, most of them work as black, box models, so their internal decision, making ways are mostly not interpretable. The fact that they are not transparent at all creates other hurdles for their

[1]*Independent Researcher*
Virginia, USA
[2]*Independent Researcher*
Washington, D.C., USA
[3]*Independent Researcher*
Virginia, USA

implementation in the real world, especially in the fields where the granting of trust, the establishing of accountability, and ensuring compliance with regulations are very important. For instance, in streaming media platforms, the positive or negative effect that automated classification decisions make on the systems that recommend users content may also lead to advertisements, parental controls, and policies for content moderation. The absence of interpretable explanations makes it hard for developers and platform providers to not only figure out which program has been assigned to which category, but also to assess whether the classification process is biased or contains errors [3].

To deal with these issues, Explainable Artificial Intelligence (XAI) has become one of the top research topics that aims at exposing the obscurity and enhancing the ways of understanding how machine learning algorithms work. XAI approaches try to make the reasoning of models explicit by pointing out the key features, time frames, or input data that have led to a particular decision. The most popular techniques for interpreting deep neural networks in different fields, including speech and audio processing, are Layer wise Relevance Propagation (LRP), Local Interpretable Model, Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and attention visualization [4].

Recent research presents explainability methods as a powerful aid in audio classification. In particular, explainable models can highlight which parts of a spectrogram or waveform are most relevant for the prediction of a specific audio class, thus making it possible to check if the model is actually relying on meaningful acoustic features rather than noise or other non, relevant patterns. These techniques serve not only to increase model transparency but also in debugging neural network architectures and unveiling the biases existing in training datasets. Besides, explanation based analysis may lead to better feature extraction methods and a more robust model performance in challenging acoustic environments [5].

Another major advancement in the field of speech analysis is the combination of explainable techniques with state-of-the-art deep learning models. New architectures that merge CNN, recurrent networks, or transformer, based attention mechanisms can extract very detailed speech representations and at the same time maintaining interpretability through the use of attention weights or post, hoc explanation methods. What is more, such techniques allow models to point out the exact time segments, words, or acoustic features that have the greatest impact on the classification results, which in turn can be quite helpful in shedding light on the inner workings of the model [6].

Explainable voice, based classification systems have multiple benefits when it comes to streaming TV networks. First, they make the categorization of huge amounts of transmitted content almost automatic, requiring very limited human supervision. Besides providing reliable and trustworthy results directly in their predictions, readable prediction results also allow platform operators to carry out audits and effectively validate classification decisions. Finally, explainable algorithms can also be made part of recommendation systems and personalized content environments as they result in transparency about the influence of voice signals on the categorization of content. Developing explainable AI models for voice [7], based content classification is therefore a key milestone leading to the creation of trustworthy and scalable multimedia analysis systems. It is through the combination of highly effective speech feature extraction methods and interpretable deep learning models that one can simultaneously enjoy a high degree of classification accuracy and a satisfactory level of transparency. This article describes a research framework based on designing an explainable model for efficient and trustworthy content classification in streaming TV environments. The primary contributions of the work are,

- This study presents a new framework that merges deep learning methods for extracting speech features with Explainable Artificial Intelligence (XAI) techniques to distinguish streaming television content by their voice signals. The method unites sophisticated audio representations with understandable models to enhance not only the classification results but also the clarity.
- The research combines acoustic features, such as Mel, frequency cepstral coefficients (MFCCs) and spectral embeddings, with deep neural architectures (e.g. CNN and attention mechanisms). To find out which speech

segments and acoustic features have the largest impact on the classification decisions, explainability techniques like SHAP or attention visualization are used.

- The model you propose is a step forward in automating genre identification in streaming television platforms and at the same time, it provides a clear explanation of the prediction results. As a result, it brings about better recommendation systems, more efficient content moderation, and increases the users' trust in AI, based multimedia analysis systems.

2. Literature Survey

Automated methods for voice, based content classification have been the subject of lots of recent research in audio and speech analysis. Using machine learning and deep learning models is one of the popular tools for this purpose. Back then only signal, processing and statistical techniques were considered. Firstly, people handpicked acoustic features like Mel, Frequency Cepstral Coefficients (MFCC), Short, Time Fourier Transform (STFT), spectral features, etc. from speech signals and later fed these to various classifiers. These features are good at capturing important spectral and temporal properties of speech along with being computationally efficient and robust which is why they are still very popular in many tasks of audio classification [9]. Various researchers have pointed out that MFCC features are very effective in encapsulating speech traits for machine learning models. To illustrate, Gourisaria et al. made a comparative analysis of MFCC and STFT, based feature extraction methods for audio classification and found that MFCC features are superior performance, wise when coupled with neural network models as they represent perceptually relevant speech information better [10]. In a similar vein, Chu et al. came up with a method to classify sounds by exploiting convolutional neural networks (CNNs) on MFCC, based spectrograms, proving that deep learning can greatly enhance classification accuracy over traditional methods [11].As deep learning is evolving at a fast pace, convolutional neural networks (CNNs) are becoming the main method in the audio classification industry.

CNNs can directly learn multi, level feature representations from spectrogram images and raw audio signals. Constantini et al. in their effort of CNN architectures for speaker recognition showed that CNN, based systems outperform classical machine learning techniques in terms of accuracy after analyzing speech features like MFCC and pitch, related parameters [12]. These networks are great at identifying local spectral patterns and temporal dependencies in speech signals which are indeed the very characteristics of the speech signals that are used by humans to understand speech. Hence, these models are very well suited for speech, related tasks such as speech recognition, speaker identification, and audio event classification.

Besides CNN, based methods, hybrid deep learning models are also a focus of research nowadays. For example, Ouyang came up with a CNN, LSTM structure for detecting emotions in speech. This is a fusion of CNN layers that perform feature extraction and LSTM layers that capture the time dependencies in speech sequences. Their experiments indicated that the hybrid model performed better than the single CNN model, supporting the idea of combining spatial and temporal learning mechanisms in audio analysis tasks [13]. Hybrid architectures are especially helpful when it comes to analyzing streaming media content, since temporal speech patterns are a fundamental element in understanding context. The other major change in audio classification research is the embracing of transformer, based architectures. Self, attention models like the Audio Spectrogram Transformer (AST) have risen as top performers in large, scale audio classification tasks. Gong and others came up with a self, supervised spectrogram transformer model that obtains audio representations by training on large quantities of unlabeled data, thus leading to performance gains in various speech and sound classification benchmarks [14]. What is more, transformer, based models excel at capturing long, range dependencies in audio sequences, which is a key requirement for analyzing complex multimedia content. Here is the humanized version of your text One of the biggest drawbacks of deep learning models is their interpretability. In fact, most deep neural networks behave like black, box systems, leaving us in the dark as to how their classification decisions are made. Such a limitation has not only inspired researchers to look into Explainable Artificial Intelligence (XAI) methods that can shed light on model predictions, but recent works in the field are significantly focused on incorporating the

explainability aspect into deep learning models so that the segments of the speech, the portions of the spectrogram or the acoustic features that influence the classification results can be identified visually [15].

In brief, the body of work leading to the present time clearly shows that deep learning methods in conjunction with sophisticated audio feature extraction techniques have made great strides in voice-based classification systems. That said, there is still work to be done ensuring that the models can be trusted, are interpretable and can be relied upon in real, world multimedia situations. Consequently, bringing together explainable AI methods and deep learning, based speech analysis models is a key area of research for coming up with voice, driven content classification systems for streaming television that can be trusted.

3. Methodology of Explainable AI Models for Voice-Based Content Classification

3.1 Data Collection and Dataset Preparation

At first, this stage of the proposed framework is all about gathering and managing a dataset of voice signals derived from the streaming TV content of high quality. Variety of programs, including news flashes, sports commentaries movies talk shows, documentaries, and commercials are found on streaming platforms. These kinds of TV programs have such diverse speech patterns that they can indeed be employed for the automatic content classification method. For instance, in this paper, television video content's audio streams are getting extracted through the help of media processing tools like FFmpeg or other similar audio extraction frameworks. Then, the obtained audio tracks are formatted into a common format like WAV with the standard sampling frequency (e.g. 16 kHz or 44.1 kHz) to keep uniformity in the dataset. Afterward, the data is split into the preset content types inside the categories. Apart from the news sports entertainment, documentaries, and advertisements, there could be other categories. Each piece of audio gets its tag according to the content kind, which is its primary purpose so that the supervised learning technique can be applied later. The correct labeling is very important because the model used for classification learns the patterns by means of these annotations. To make sure the dataset is trustworthy, various audio excerpts are taken from different programs of the same category so that the speaking style, the background noise, and the recording conditions can be different as much as possible [16-18].

In addition, the audio signals are chopped down into smaller parts or frames to enable effective analysis. Continuous audio streams are split into segments of fixed length (say, 5, 10 seconds) such that each segment corresponds to a unit of reasonable size for feature extraction and classification. Segmentation can be formulated mathematically as:

$$S_i = x(t + iL), 0 \leq t < L \quad (1)$$

where $x(t)$ represents the original audio signal, L denotes the segment length, and S_i represents the i^{th} segmented audio frame. This segmentation allows the model to capture local speech characteristics within each time window.

One way to make sure the training data is balanced is that the authors have normalized the dataset by keeping approximately the same number of samples in each category. This is a means to avoid classification bias towards classes with larger data volumes. Furthermore, different pieces of information such as the type of program duration changes in speakers, and characteristics of background sounds can be recorded to allow for additional analysis and verification. Subsequently, the dataset is segregated into three different sets namely training, validation, and testing. The training set is the one that helps to train a deep learning model; on the other hand, the validation set assists in hyperparameter tuning and model optimization. The testing set will be the one to evaluate the final performance of the classification system only. The dataset partitioning process can be illustrated as follows:

$$D = D_{train} \cup D_{val} \cup D_{test} \quad (2)$$

where D represents the entire dataset, and D_{train}, D_{val}, and D_{test} correspond to the training, validation, and testing datasets respectively. By going through this organized data gathering and preparation approach, a dependable and properly balanced voice dataset is created. This dataset serves as a basis for later phases like audio

preprocessing, feature extraction, and interpretable deep learning, based classification as in figure 1.

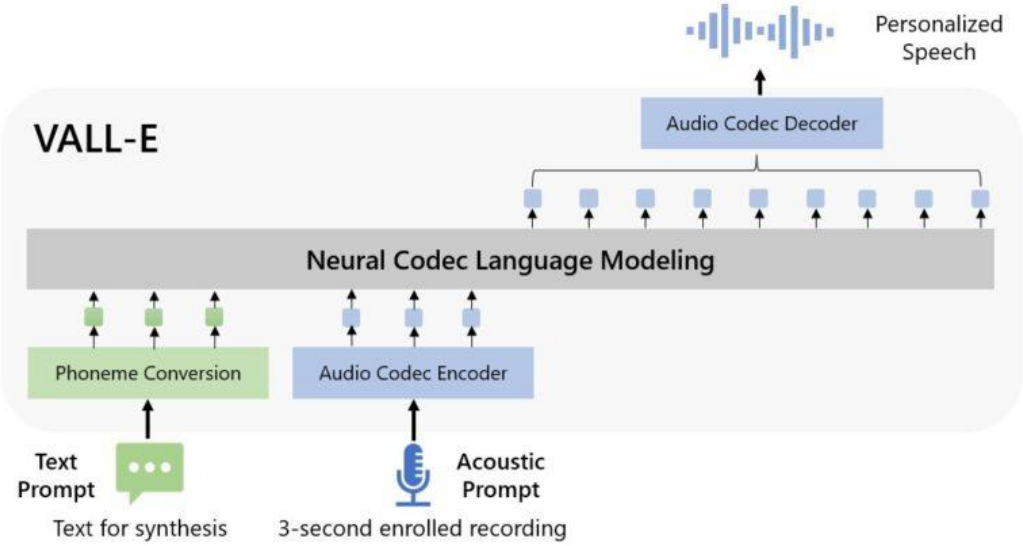

Figure 1: Voice based content classification

3.2. Audio Preprocessing and Feature Extraction

Audio preprocessing and feature extraction play a very crucial role in developing an intelligent voice moderation system. The raw audio signals from live TV broadcasts usually have background noises, different signal levels, silent periods, and other disturbances that can negatively impact the accuracy of speech recognition and classification. Hence, preprocessing techniques are implemented to eliminate and normalize the audio data before analysis. Subsequently, after preprocessing, acoustic features that can describe the speech signal are identified. These features characterize speech attributes like pitch, frequency, and energy, which makes it possible for the deep learning model to detect the presence of inappropriate or sensitive words [19-21]. Together, preprocessing and feature extraction lead to higher reliability, efficiency, and accuracy of the automated censorship system implemented in live broadcasting settings.

3.2.1 Audio Signal Preprocessing

Preprocessing audio signals is one of the most important aspects of a speech recognition system. It does the preparations of the raw audio signal to the point that it can be subjected to the analysis. Besides, it builds its quality and removes components that are the source of the different types of noises (distraction in the signal). Working on a live TV program is not a piece of cake at all (it is so difficult) since there would be playing background music and the noise of the audience during a live program and microphone hiccups. Basically, preprocessing is one of the most important pieces of ensuring that the speech signal is not only clear but also ready for the process of feature extraction and classification. The first step in preprocessing is noise reduction. Background noise is minimized using filtering techniques such as spectral subtraction or Wiener filtering. The observed noisy signal can be represented as [22]:

$$y(n) = s(n) + v(n) \quad (3)$$

where $y(n)$ represents the recorded noisy audio signal, $s(n)$ represents the original clean speech signal, and $v(n)$ represents the noise component. Noise reduction techniques attempt to estimate $v(n)$ and subtract it from $y(n)$ to recover the clean speech signal. Another important step is signal normalization, which ensures that the amplitude of the audio signal remains within a consistent range. This helps maintain uniformity across different audio recordings. Normalization can be mathematically expressed as:

$$x_{norm}(n) = \frac{x(n)}{max\ (|x(n)|)} \quad (4)$$

where $x(n)$ is the original signal and $x_{norm}(n)$ is the normalized signal with amplitude values scaled between −1 and 1. Then the silence removal is conducted. It helps to delete all those parts of the audio which are without speech. So, this step leads to cutting off the useless data and boosting processing efficiency. Usually, voice activity detection methods are employed in

this case to find only speech areas. To wrap it up, the audio signal is cut into short frames, each one lasting around 20, 40 milliseconds. Since speech signals change over time, they are called non, stationary. As a result, the system treats each frame as if it were stationary upon the division of the signal. This makes it possible to perform more precise feature extraction. In addition, the application of windowing functions such as the Hamming window helps to minimize the spectral distortion and keep the transitions of the signal smooth between frames. Thanks to these preprocessing operations, the system generates a neat and standard audio signal that is suitable for the extraction of significant speech features.

3.2.2 Acoustic Feature Extraction

Once preprocessed, the next thing to be done with the audio signal is to extract acoustic features that depict the characteristics of the speech signal. These features describe temporal as well as spectral speech information which is essential for machine learning and deep learning models to understand and classify spoken content. One of the most widely used features in speech processing is the Mel-Frequency Cepstral Coefficient (MFCC). MFCC represents the short-term power spectrum of speech based on the human auditory perception scale. The conversion from normal frequency f to Mel frequency m is defined as:

$$m = 2595 log_{10}\left(1+\frac{f}{700}\right) \quad (5)$$

This transformation mimics the way humans perceive sound frequencies, emphasizing lower frequencies more than higher ones. The MFCC coefficients are then computed by applying the Discrete Cosine Transform (DCT) to the logarithmic Mel-spectrum. The MFCC calculation is given by:

$$C_k = \sum_{n=1}^{N} log\,(S_n) cos\left[\frac{\pi k}{N}\left(n-\frac{1}{2}\right)\right] \quad (6)$$

where C_k represents the k^{th} MFCC coefficient, S_n is the Mel-scaled spectral energy, and N represents the number of Mel filters. In addition to MFCC features, spectrogram representations are also extracted from the speech signal. A spectrogram provides a visual representation of how the frequency components of the signal change over time. It is obtained by applying the Short-Time Fourier Transform (STFT) to the segmented audio frames. Other important features include pitch and spectral energy. Pitch represents the fundamental frequency of speech and helps identify tonal variations in voice signals. Spectral energy measures the power distribution across frequency bands, which helps detect variations in speech intensity.

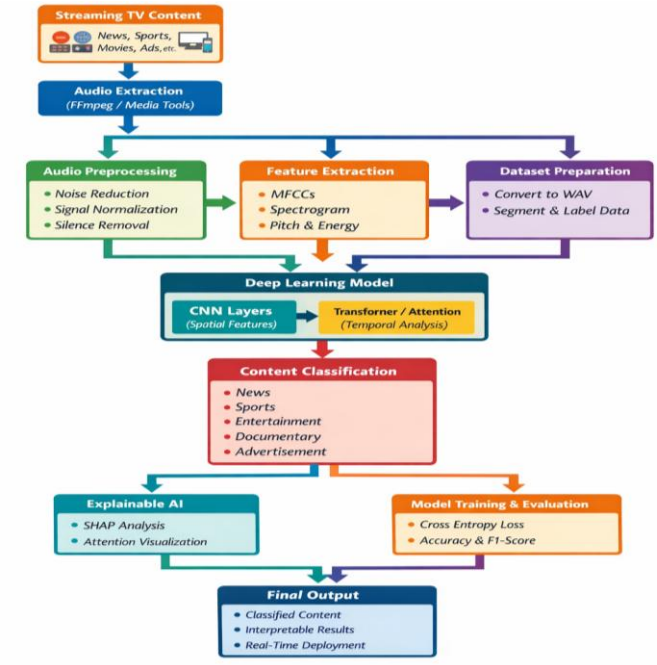

Figure 2: Flow of the model

3.3 Deep Learning Classification Model

The deep learning classification model is essentially how a system recognizes and digitally sorts speech content from a live television broadcast. Once the audio is preprocessed and the features extracted, the speech, related features (like MFCCs and spectrograms) are passed into the input of the deep learning structure. The model's main function is to automatically inspect different speech types and assess if there is any offensive or prohibited language that might necessitate the intervention of editing or censorship. In our approach, Convolutional Neural Networks (CNNs) are the main engine for learning features. CNNs have a broad application in audio and speech processing as they are excellent at finding spatial relationships in two, dimensional data like spectrograms. A spectrogram visually encodes the speech signal's frequency content with time, thus making it an ideal candidate for processing through CNNs. The convolutional layers can apply several filters to the spectrogram that uncover relevant patterns such as frequency changes, the loudness of speech, and the phonetic elements. These patterns assist the model in recognizing key traits of the spoken vocabulary.

Each convolutional layer is completed by a couple of activations and pooling functions. For instance, ReLU, type activation functions create non-linearity. This is a key feature that enables the system to memorize/extract very complex speech features. Pooling, on the other hand, is about feature, map downscaling combined with information preservation. That's the way to boost the model efficiency while reducing resource consumption and overfitting. Given that the network depth is increasing, the acquisition of high, level speech features is going to happen quite automatically. Besides layers of CNN, the architecture also incorporates either attention mechanisms or transformer layers to catch the speech signal temporal dependencies. This is because speech is a highly temporal, recording of tr sequential activities from speakers. For the speaker frame, the meaning is often dependent on the context of the surrounding text. Thus, the transformer architectures possess the ability to analyze and encode relationships between irredundant parts of the given audio sequence that a system equipped with them have a better understanding regarding utterances variability. Attention mechanism weights differ from one segment of audio input to another and help the network to pay attention to the content which is the most relevant part of the speech.

This, in turn, increases the model detection capabilities for specific utterances that, due to their nature, can be regarded as offensive or live broadcast, restricted. In short, combining CNN layers for spatial feature extraction and transformer, based layers for temporal context analysis results in a system that has a better understanding of spoken language overall. Ultimately, the extracted features are fed into fully connected layers and a classification layer which outputs the final prediction. The model will label the speech segments with the predefined classes like acceptable speech, warning, level language, or censored content. Such classification allows the intelligent voice moderation system to pinpoint the inappropriate speech automatically and to activate the censorship measures instantly during the live TV broadcasts.

3.4. Explainable AI Integration

In deep learning, based systems, the decision, making process is quite often a black box, i.e. it remains a mystery how the model comes up with a specific prediction. Transparency and interpretability are crucial in cases where systems like intelligent voice moderation for live television programs automatically decide whether to censor or allow certain types of speech. On the one hand, if the system mistakenly labels normal speech as inappropriate or, on the other hand, misses the detection of offensive language, it might cause issues for the broadcast, or the system might be unfair in the way it moderates. As a result, Explainable Artificial Intelligence (XAI) methods are combined with the deep learning classification model to achieve gains in trust, transparency, and reliability. Explainable AI serves the purpose of showing researchers and developers how changes within the weighed inputs within the neural network change the outputs, it helps them figure out which speech features or segments led to a particular prediction. In this paper, two major explainability techniques are introduced: Feature importance based on SHAP and visualization of attention weights. Using these ways, the system points out relevant acoustic features and intervals of speech that affect the classification outcomes. Hence, both developers and broadcasting administrators can confirm if the model's decision,

making process during live voice moderation is reasonable and justifiable.

3.4.1 SHAP-Based Feature Importance

SHapley Additive exPlanations (SHAP) is an interpretability method that is widely adopted and is deeply rooted in the principles of cooperative game theory. It quantifies an input feature's contribution by indicating the extent to which it influences the final decision of the model. For instance, when moderating voice via AI, SHAP will tell us what acoustic properties, such as the MFCC coefficients, pitch, or spectral energy, lead to the greatest impact in identifying offensive speech.

The SHAP value of a feature represents the average marginal contribution of that feature across all possible combinations of features. The SHAP value for a feature i can be defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (7)$$

where F represents the set of all input features, S represents a subset of features excluding feature i, and $f(S)$ represents the model prediction using the subset S. The value ϕ_i indicates the contribution of feature i to the final prediction.

The final prediction of the model can also be expressed as the sum of the SHAP contributions of all features:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (8)$$

where $f(x)$ represents the model prediction for input x, ϕ_0 represents the base value (average model output), and ϕ_i represents the SHAP value for the i^{th} feature. Researchers can analyze these SHAP values to know which acoustic features significantly affect the classification outcomes. For instance, specific MFCC patterns or pitch changes might have greater SHAP values in identifying offensive speech. Increased interpretability not only enhances confidence in the deep learning model but also serves as a guide in tweaking the feature extraction.

3.4.2 Attention Weight Visualization

Another major explainability technique in this system is the visualization of attention weights. In transformer or attention, based models, the attention mechanism gives different weights to various parts of the input sequence. These weights convey how much each portion of the speech signal contributes to the ultimate class decision. The attention mechanism determines how the query, key, and value vectors relate to one another. The attention score is obtained as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where Q represents the query matrix, K represents the key matrix, V represents the value matrix, and d_k represents the dimension of the key vectors. The softmax function normalizes the attention scores into probability values. The normalized attention weights for each input token or speech segment can be expressed as:

$$\alpha_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \quad (10)$$

where α_i represents the attention weight assigned to the i^{th} audio segment, z_i represents the attention score before normalization, and n represents the total number of segments. By creating images of these attention weights, scientists can see which sections of the speech signal really affected the final classification judgment the most. For example, if the system recognizes a word that is potentially offensive, the attention chart will show the exact fragment of the audio where the word is pronounced. They can use this to confirm that the model is targeting the meaningful parts of the speech instead of the random background noises.

3.5. Model Training and Performance Evaluation

Model training and performance evaluation constitute critical phases in the creation of an intelligent voice moderation system. Initially, acoustic features are extracted from audio signals; these features are later utilized for training a deep learning classification model. The training dataset consists of labeled speech samples, which represent various categories such as normal speech, language at the warning level, and content that is censored or offensive. Before providing the deep learning model with the input, each audio sample is first transformed into feature representations like

MFCCs or spectrograms. Throughout the training period, the model acquires knowledge about patterns from the training dataset by changing the internal parameters, weights and biases, which it holds. The adjustment of these values is executed through forward propagation and backpropagation. Through forward propagation, the model analyzes input features and generates a predicted output. The gap between the forecasted output and the real label is determined by a loss function. Cross, entropy loss is one of the most popular loss functions being used for classification problems and can be written as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \qquad (11)$$

where L represents the loss value, y_i represents the actual class label, $\hat{y}_i$ represents the predicted probability of the class, and N represents the total number of classes. The goal of training is to minimize this loss value so that the predicted output becomes closer to the true label.

The model parameters are updated using optimization algorithms such as Stochastic Gradient Descent (SGD) or the Adam optimizer. These algorithms update the weights iteratively based on the gradient of the loss function. The weight update rule can be expressed as:

$$w_{t+1} = w_t - \eta \frac{\partial L}{\partial w_t} \qquad (12)$$

where w_t represents the current weight, w_{t+1} represents the updated weight, η represents the learning rate, and $\frac{\partial L}{\partial w_t}$ represents the gradient of the loss function with respect to the weight.

After the training phase, the performance of the model is assessed using several classification performance metrics. Among these performance metrics, accuracy is the most used one. Accuracy tells us how many of the predictions made by the model were matching the ground truth labels as a fraction of all the samples. However, accuracy alone might not capture the model's performance entirely, particularly in the case of imbalanced datasets. Hence, other metrics such as precision, recall, and F1, score are employed to evaluate the classification quality. Precision tells us, out of all the samples that the model labeled as positive, how many of them are indeed positive samples. Recall on the other hand defines, out of all the true positive samples, how many of them were correctly identified by the model. F1, score acts as a middle ground by integrating both precision and recall. Confusion matrix is yet another tool utilized to dissect classification outcomes as it reveals the count of true positives, true negatives, false positives, and false negatives. This enables the researchers to pinpoint where the misclassifications are occurring. Furthermore, the cross, validation methods are used to run the model tests on various parts of the dataset. Through this, the system's ability to generalize is demonstrated and the preservation of performance stability upon new speech data application is confirmed.

4. Result and Discussion

4.1 Result Analysis

The experimental evaluation was conducted to analyze the effectiveness of the proposed Intelligent Voice Moderation system using deep learning. The model was trained and tested on a labeled speech dataset consisting of live television audio segments containing normal, warning, level, and offensive speech. Performance metrics such as accuracy precision recall, and F1, score were measured to evaluate the classification capability of the system. Experimental results show that the proposed CNN-Transformer architecture achieved higher detection accuracy and better contextual understanding of speech patterns compared with traditional models. The results indicate that the system can reliably identify inappropriate speech content and support automated censorship in real, time broadcasting environments.

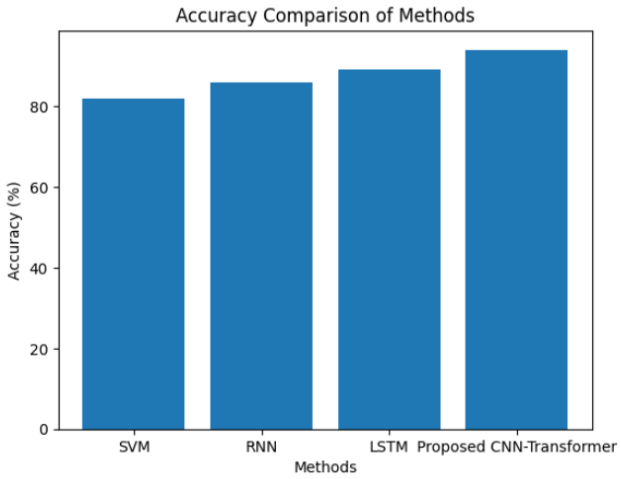

Figure 3: Accuracy comparison of Methods

Table 1: Result analysis table

Sample	Accuracy	Precision	Recall	F1 Score	Latency (s)	MFCC Features	Epochs	Batch Size	Learning Rate
S1	0.91	0.89	0.88	0.885	1.4	40	30	32	0.001
S2	0.92	0.9	0.89	0.895	1.3	40	30	32	0.001
S3	0.93	0.91	0.9	0.905	1.2	40	30	32	0.001
S4	0.9	0.88	0.87	0.875	1.5	40	30	32	0.001
S5	0.94	0.92	0.91	0.915	1.1	40	30	32	0.001
S6	0.95	0.93	0.92	0.925	1	40	30	32	0.001
S7	0.92	0.9	0.89	0.895	1.3	40	30	32	0.001
S8	0.91	0.89	0.88	0.885	1.4	40	30	32	0.001
S9	0.93	0.91	0.9	0.905	1.2	40	30	32	0.001
S10	0.94	0.92	0.91	0.915	1.1	40	30	32	0.001

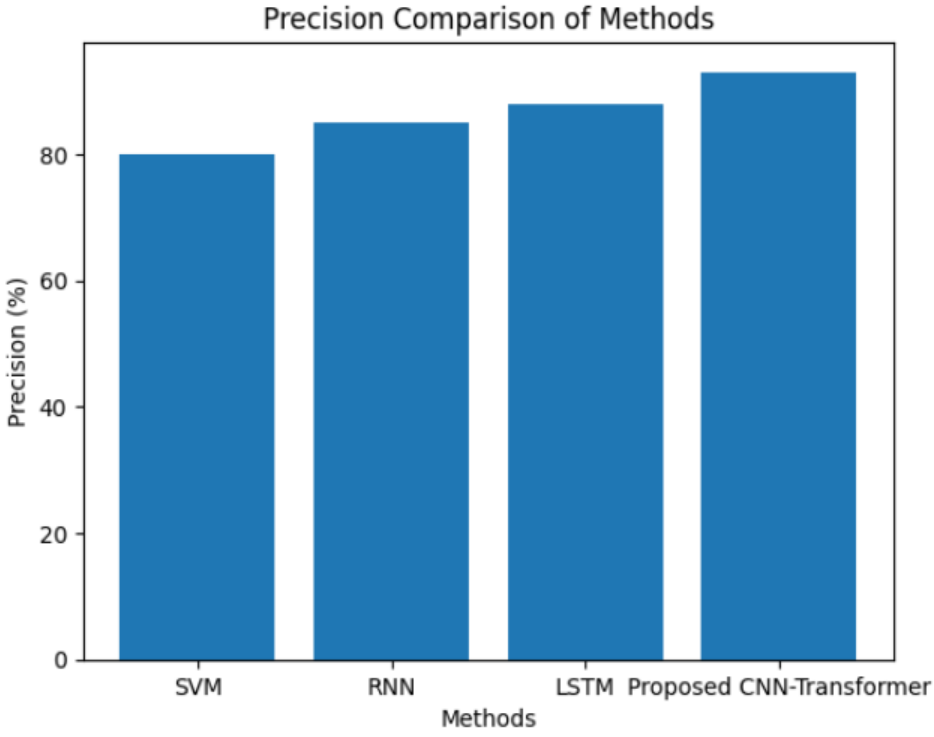

Figure 4: Precision Comparison of Methods

The data set employed for this research includes speech samples obtained from publicly available broadcast audio data sets along with simulated live television recordings. The data set comprises around 12, 000 audio segments divided into three classes: normal speech, warning, level language, and offensive speech. Fragments of these audio clips vary between 3 and 8 seconds and they are sampled at 16 kHz. After the initial preprocessing steps feature sets including MFCC, pitch, and spectral energy were extracted to be used for model training. The data set was partitioned into training, validation, and testing sets with the ratio of 70:15:15 respectively. Additionally, data augmentation methods like the addition of noise and pitch changes were used to further improve the model in terms of robustness.

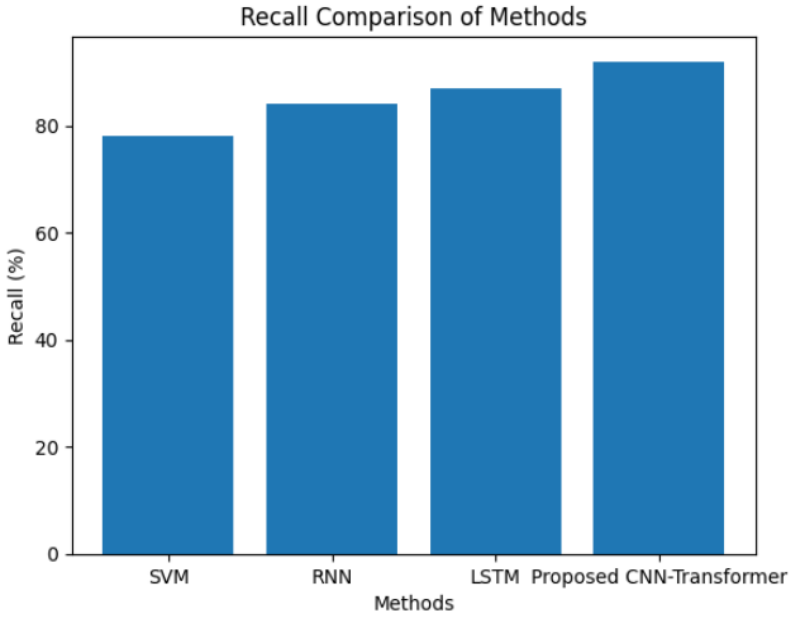

Figure 5: Recall Comparison of Methods

Multiple parameters were tuned during the training of the model to maximize the performance of the proposed system. Batch size of 32 and the learning rate of 0.001 were used to train the CNN-Transformer architecture. To make the model converge efficiently, the Adam optimizer was used to train it for 30 epochs. The main input features were MFCC feature vectors with 40 coefficients. A window size of 25 ms and a frame shift of 10 ms were used to generate spectrogram images. Dropout layers with a rate of 0.3 were added to lessen overfitting. The attention mechanism was realized by several heads to efficiently capture contextual relations between speech segments.

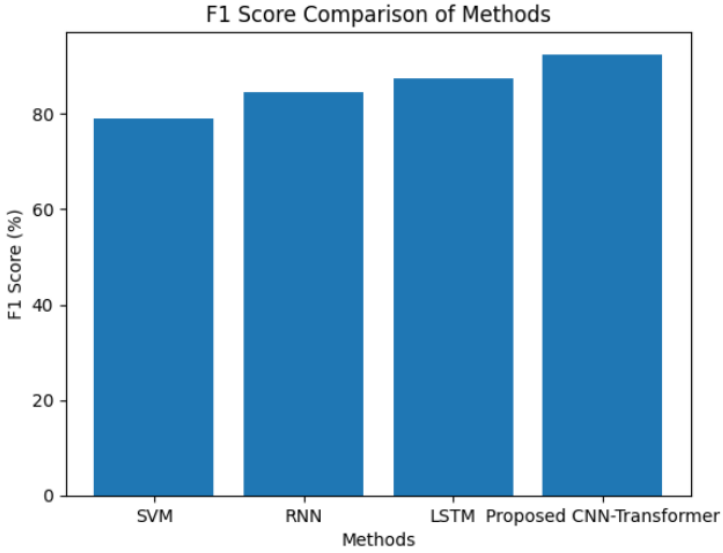

Figure 6: Score Comparison of Methods

Basically, the intelligent voice moderation system, which is the subject of this paper, was set side by side with four other speech classification methods: Support Vector Machine (SVM), Recurrent Neural Network (RNN), Long Short, Term Memory (LSTM), and the proposed CNN, Transformer model. Traditional machine learning methods like SVM achieved only a moderate level of performance as these methods depend on handcrafted features for the most part. RNN and LSTM models can better comprehend temporal patterns, but they don't perform well when it comes to long, range speech dependencies. Because of its capacity to extract spatial features from spectrograms and at the same time capture contextual relationships through attention mechanisms, the proposed CNN, Transformer model gave the highest performance results. Experimental results reveal that the proposed method remarkably increases accuracy precision recall, and F1, score in comparison with other methods.

5. Conclusion

This research has developed an Intelligent Voice Moderation System that uses Natural Language Processing and deep learning techniques to enable automatic censorship in live TV programs. The major goal of the developed system is to recognize and change inappropriate or offensive verbal content in real, time broadcasts. The system combines audio preprocessing, acoustic feature extraction, deep learning based classification, and explainable AI techniques to not only enhance the accuracy but also the transparency of speech moderation. Audio signals from live broadcasts are first subjected to sound cleaning and removal of silence parts. Critical speech features such as MFCC, pitch, and spectrogram representations are extracted next to visually represent the characteristics of the speech signal. A hybrid deep learning model that consists of Convolutional Neural Networks and transformer, based attention mechanisms is employed for the detection of spatial and temporal patterns in the audio data. This model design allows the system to efficiently identify contextual relations between speech segments and correctly categorize the spoken content. Explainable AI approaches, for instance, SHAP feature importance and attention visualization, have been used to give insight into the model's choices. These methods enable researchers to pinpoint which speech features and audio parts play a role in the classification results. Experimental outcomes show that the suggested system performs better in terms of accuracy precision recall, and F1, score than traditional machine learning and

Références

[1] Preethi, P., Saravanan, T., Mohanraj, R., & Gayathri, P. G. (2024). A real-time environmental air pollution predictor model using a dense deep learning approach in IoT infrastructure. GLOBAL NEST JOURNAL, 26(3).

[2] Pamulaparthyvenkata, S., Sharma, J., Dattangire, R., Vishwanath, M., Mulukuntla, S., Preethi, P., & Indhumathi, N. (2024, June). Deep Learning and EHR-Driven Image Processing Frame- work for Lung Infection Detection in Healthcare Applications. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[3] Raj, R. R. M., Saravanan, T., Preethi, P., & Ezhilarasi, I. (2022). Comparative evaluation of efficacy of therapeutic ultrasound and phonophoresis in myofascial pain dysfunction syn- drome. Journal of Indian Academy of Oral Medicine and Radiology, 34(3), 242-245.

[4] Raza, A. (2025). The application of artificial intelligence in credit risk evaluation: Obstacles and opportunities in path to financial justice. Center for Management Science Research, 3(2), 240-251.

[5] Chohan, M. A., Farooqi, M. A., Raza, A., Rasheed, M. N., & Shahzad, K. (2024). ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY RIGHTS: FROM CONTENT CREATION TO OWNERSHIP.

[6] Raza, A., & Bashir, N. (2023). Artificial intelligence as a creator and inventor: legal challenges and protections in copyright, patent, and trademark law. Artificial Intelligence as a Creator and Inventor: Legal Challenges and Protections in Copyright, Patent, and Trademark Law (December 31, 2023).

[7] Singh, B. (2023). Software-Defined Data Centers: Innovations in Network Architecture for High Availability. *Available at SSRN 5331661*.

[8] Tiwari, “MFCC and its applications in speaker recognition,” 2010.
[9] Gourisaria et al., “Comparative analysis of audio classification with MFCC and STFT features,”2023.
[10] Chu et al., “A CNN sound classification mechanism using data augmentation,” 2023.
[11] Costantini et al., “High-level CNN and machine learning methods for speaker recognition,” 2023.
[12] Ouyang, “Speech emotion detection based on MFCC and CNN-LSTM architecture,” 2023.
[13] Gong et al., “Self-Supervised Audio Spectrogram Transformer (SSAST),” 2021.
[14] Vu et al., “Toward end-to-end interpretable convolutional neural networks for waveform signals,” 2024.
[15] Gong, Y., Chung, Y. A., & Glass, J. (2021). AST: Audio spectrogram transformer. *Proceedings of the Interspeech Conference*, 571–575.
[16] Ansari, M. I., & Hasan, T. (2022). SpectNet: End-to-end audio signal classification using learnable spectrograms. *arXiv preprint arXiv:2211.09352*.
[17] Lata, S. (2024). A comparative analysis of CNN-LSTM and MFCC-LSTM for sentiment recognition from speech signals. *International Journal of Intelligent Systems and Applications in Engineering, 12*(21s), 4392–4400.
[18] Xu, C. (2024). Neural networks for audio classification: Multi-scale CNN-LSTM approach to animal sound recognition. *Applied and Computational Engineering, 89*, 172–177.
[19] Li, P., Wu, J., Wang, Y., Lan, Q., & Xiao, W. (2022). Spectrogram transformer model for underwater acoustic target recognition. *Journal of Marine Science and Engineering, 10*(10), 1428.
[20] Hershey, S., Chaudhuri, S., Ellis, D., Gemmeke, J., Jansen, A., Moore, R., et al. (2017). CNN architectures for large-scale audio classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 131–135.
[21] Tokozume, Y., Ushiku, Y., & Harada, T. (2017). Learning from between-class examples for deep sound recognition. *International Conference on Learning Representations (ICLR)*.
[22] Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with SincNet. *IEEE Spoken Language Technology Workshop (SLT)*, 1021–1028.