Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Multimodal Audio Violence Detection: Fusion of Acoustic Signals and Semantics

Shivwani Nadar, Disha Gandhi, Anupama Jawale*, Shweta Pawar, Ruta Prabhu

Department of Information Technology, Narsee Monjee College of Commerce and Economics, 400056 Mumbai, India

* Correspondence: Anupama Jawale (anupama.jawale@nmcce.ac.in)

Received: 10-23-2025 **Revised:** 12-12-2025 **Accepted:** 12-18-2025

Citation: S. Nadar, D. Gandhi, A. Jawale, S. Pawar, and R. Prabhu, "Multimodal audio violence detection: Fusion of acoustic signals and semantics," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 4, pp. 301–311, 2025. https://doi.org/10.56578/ataiml040405.

Abstract: When public safety is considered to be of paramount importance, the capacity to detect violent situations through audio monitoring has become increasingly indispensable. This paper proposed a hybrid audio text violence detection system that combines text-based information with frequency-based features to improve accuracy and reliability. The two core models of the system include a frequency-based model, Random Forest (RF) classifier, and a natural language processing (NLP) model called Bidirectional Encoder Representations from Transformers (BERT). RF classifier was trained on Mel-Frequency Cepstral Coefficients (MFCCs) and other spectrum features, whereas BERT identified violent content in transcribed speech. BERT model was improved through task-specific fine-tuning on a curated violence-related text dataset and balanced with class-weighting strategies to address category imbalance. This adaptation enhanced its ability to capture subtle violent language patterns beyond general purpose embeddings. Furthermore, a meta-learner ensemble model using eXtreme Gradient Boosting (XGBoost) classifier model could combine the probability output of the two base models. The ensemble strategy proposed in this research differed from conventionally multimodal fusion techniques, which depend on a single strategy, either NLP or audio. The XGBoost fusion model possessed the qualities derived from both base models to improve classification accuracy and robustness by creating an ideal decision boundary. The proposed system was supported by a Graphical User Interface (GUI) for multiple purposes, such as smart city applications, emergency response, and security monitoring with real-time analysis. The proposed XGBoost ensemble model attained an overall accuracy of over 97.37%, demonstrating the efficacy of integrating machine learning-based decision.

Keywords: Audio; BERT; Graphical User Interface; MFCC; Stacking ensemble; Violence; XGBoost

1 Introduction

Traditional surveillance systems mostly employ visual data to monitor security concerns. Owing to issues like inadequate lighting, occlusions, or limited camera angles, visual-only methods are frequently less effective [1]. On the other hand, audio-based analysis could offer a different viewpoint or crucial insights into violent events [2] in situations where visual monitoring is impractical or challenging. This study proposed a hybrid audio violence detection system combining acoustic and linguistic features for improved classification accuracy and robustness in order to get around challenges arising from background noise, different speech patterns, and changing acoustic environments [3]. Unlike traditional fusion systems that rely on simple probability averaging, the XGBoost-based fusion model learns how to intelligently weigh and combine predictions from both models, hence capturing complex decision patterns to improve accuracy, recall, and precision [4]. The final model becomes more adaptable and reliable in a range of real-world scenarios by minimizing the errors of the individual base models [5]. Traditional probability-averaging methods often suffer from inflated false positive rates; for instance, prior studies have shown that non-violent but emotionally intense audio (e.g., cheering or loud music) was frequently misclassified as violent due to over-reliance on acoustic probability spikes. These shortcomings motivate the use of a stacked fusion strategy, where the meta-learner could reduce such errors by learning context-specific weighting. To address the need of social safety and other real-world requirements, an user interface was developed and described in the methodology

section. The stacking ensemble learning process that the system employed for violence detection consists of the following (Table 1):

Table 1. Levels of ensemble learning process

Model Level	Description
Base Model at Level 0	A frequency-based model was trained using RF on MFCCs, Chroma Features, Mel-Spectrograms, and Spectral Contrast in order to capture the audio patterns of violent content. A text-based model that analyzes speech transcriptions and employs a revised BERT classifier to detect violent language cues [6, 7].
Meta Learner Model at Level 1	An XGBoost classifier is trained on the probability output from both base models (audio_norob and text_prob) in order to learn an optimal decision boundary, which enhances overall classification performance [8].

2 Literature Review

Advances in artificial intelligence and deep learning have sped up research on audio-based violence detection. Earlier methods combined manually generated audio features like MFCCs and Spectral Contrast with machine learning models like Support Vector Machines (SVMs) and RF classifiers. Conversely, more recent approaches employ deep learning models, which directly learn feature representations from raw audio data, significantly improving accuracy and generalization.

Classical models exhibit distinct trade-offs: SVMs achieve strong margin-based classification but scale poorly with large datasets; Extreme Learning Machines (ELMs) offer faster training times but often underfit complex acoustic patterns; RFs provide robustness to noisy data and strong generalization but require careful hyperparameter tuning. These limitations partly explain the shift toward ensemble and deep learning methods. Some approaches mentioned above are illustrated as follows:

1. Machine Learning-Based Approaches: Several researchers have looked into different machine learning techniques for audio-based violence detection. Durães et al. [3] demonstrated that deep learning outperformed conventional techniques when trained on a range of datasets in the assessment of deep learning models and data augmentation techniques. Mahalle and Rojatkar [9] further investigated the efficacy of ELM for audio-based violence categorization, emphasizing the faster computation time of the models compared to more traditional models.

Further studies have examined lightweight deep neural networks designed for real-world applications and demonstrated that they could increase the accuracy and efficiency of violence detection [10]. Multimodal approaches integrating audio with extra contextual information have been proposed to improve recognition performance in low-resource settings [11]. In the research paper, a cross-lingual model that used few-shot learning techniques was developed to detect violent speech in many languages in order to increase the effectiveness of the method in a range of linguistic settings [12].

2. Deep Learning and Multimodal Fusion: Deep learning has revolutionized the field of violence detection by enabling models to process both textual and auditory data simultaneously. Certain systems have been introduced to detect violent interactions by combining linguistic and speech features. Their results demonstrated that categorization accuracy could be increased by combining data from multiple sources. Similarly, multimodal fusion techniques that used meta-information and deep neural networks have been explored to enhance aggression detection in surveillance applications [13].

As multimodal techniques gain popularity, researchers are looking into how audio may be integrated with written or visual data. Weakly supervised learning methods that used hyperbolic space representations have been developed to improve aggression classification. Other studies have expanded on this concept by introducing a multimodal attention-enhanced feature fusion framework, combining multiple modalities to significantly boost model performance in difficult real-world scenarios [14].

3. NLP for Violence Detection: Beyond acoustic analysis, NLP has evolved into a powerful technique for identifying aggression and violence in speech. Conversational agents are classic examples of advances in NLP [15]. By employing BERT models to detect gender-based violence on social media, research has demonstrated how transformer-based language models may effectively detect harmful information [16, 17]. Building on this, further investigations have proposed a BERT-fasttext model that is more capable of detecting violent language than traditional NLP-based classifiers. Multimodal NLP-based systems combining lightweight neural architecture with specialized audio processing techniques have been introduced, thus facilitating real-time processing for security applications on edge devices. By exploring deeper learning techniques for detecting cases of domestic violence through audio monitoring, additional contributions have highlighted the usefulness of audio surveillance [18]. Other researchers

contributed to this subject by showing how few-shot learning could improve NLP models in cross-lingual contexts, in order to boost their capacity to identify violent speech in multiple languages. Although audio-based violence detection has advanced significantly, certain challenges remain. One major obstacle that results in false positives is the inability to distinguish between actual violent episodes and intense but non-violent situations, such as cheering or loud music. Moreover, it is challenging to apply deep learning models in real-time applications due to computational constraints, thus necessitating optimization techniques to boost efficacy.

The techniques and datasets employed in previous works are summarized in Table 2.

Table 2. Highlights of previous works related to the detection of violent speech

Reference	Methods	Dataset	Results
[19]	BERT-Based model and Large Language Model (LLM)	Data from $\sim$ 420 k Twitter posts spanning a 3-year duration (January 1, 2020 to February 1, 2023)	F1 score of 0.69. In comparison, the detection of antiAsian hateful speech showed a higher effectiveness, with an F1 score of 0.89
[20]	BERT, XGBoost, and RF	Students' Violent Speech (SVS) dataset with 7056 tagged tweets	90% accuracy with BERT
[21]	TensorFlow custom object detection and speech analysis	Custom datasets for 6-image categories. Custom datasets for 2-speech categories.	84% accuracy of the developed system
[22]	MFCC and Stationary Wavelet Transform (SWT)-Based feature extraction. K-Nearest Neighbors (KNN) and Convolutional Neural Networks (CNNs) for classification.	Vera Am Mittag (VAM) corpus German TV talk show recordings	CNNs with feature selection achieved 97.21% classification accuracy
[2]	BERT "Efficiently Learning an Encoder that Classifies Token Replacements Accurately" (ELECTRA) models for NLP	Pre-training dataset collected by crawling 110 Bangla websites (27.5GB, 5.25 million documents). Evaluation dataset for the Violence Inciting Text Detection shared task (texts in the Bangla language).	F1 score of 0.737
[23]	Naive Bayes, SVM, Ensemble classifiers, and Artificial Neural Net	Waseem and Hoyy (2016) English dataset; Fortuna (2017) Portuguese dataset	Improvement of F1 score by 0.1
[10]	Mel-Spectrogram images of speech signals	Image and acoustic data	Improvement in F1 by 8%

As highlighted in the above table, there is an accelerating demand for robust models that could operate in multilingual and low-resource environments. The models discussed so far focus on standard news and tweets datasets or textual datasets. A natural spoken language has seldom been assessed in all these models. By emphasizing the benefits of few-shot learning for cross-lingual adaptability, the current research created new avenues for improving global violence detection systems.

In summary, the subject of audio-based violence detection has seen significant development, progressing from traditional machine learning models to advanced deep learning techniques that make use of multimodal data, NLP, and cross-lingual adaptation. Ongoing research into self-supervised learning, few-shot learning, and multimodal fusion will further enhance the accuracy and practicality of violence detection systems, thus boosting their effectiveness in a range of demanding scenarios.

3 Methodology

This section explains the process of developing a powerful audio violence detection system that integrates a fusion mechanism to enhance the performance of linguistic and acoustic models.

Figure 1 exhibits how an audio input is processed in two parallel branches, one analyzing acoustic features with a RF classifier, and the other converting the audio to text for BERT-based classification. Both branches produce separate Violent/Non-Violent predictions, which are then combined in a decision fusion block. If the fusion score meets the threshold, the final outcome is classified as Violent or it will be Non-Violent. The result, along with intermediate probabilities, transcribed text, and confidence visualization (via a real-time GUI built using Gradio), is displayed to the user.

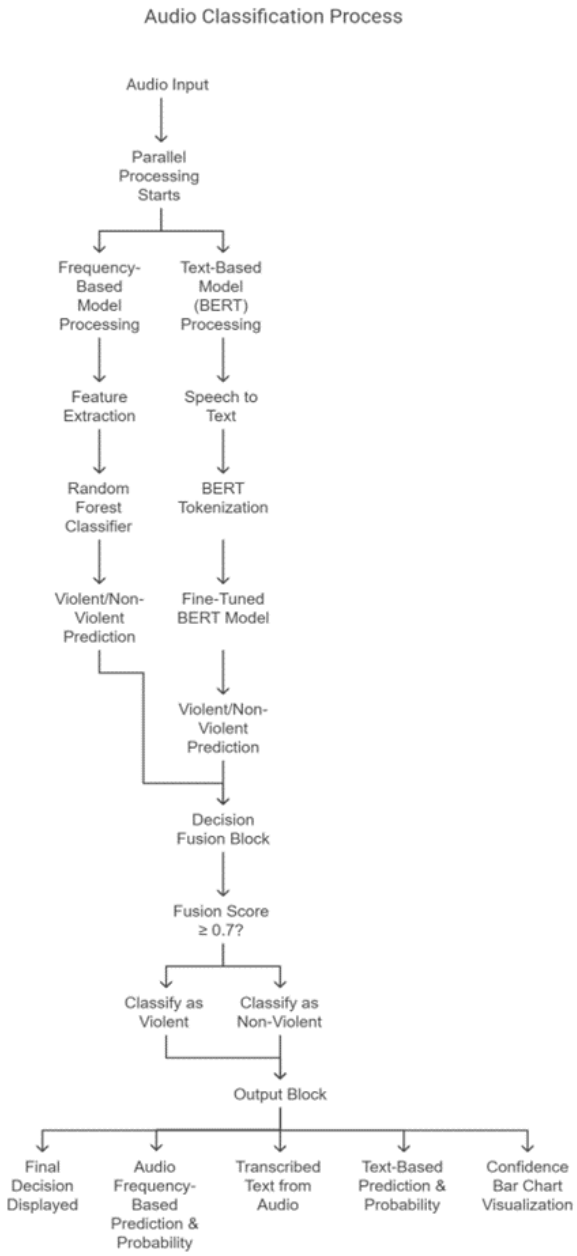

Figure 1. Flowchart of methodology combined the two base models

The dataset used in this study was UBC-NLP/DetoxLLM-7B dataset [24], the latest dataset with a unique feature of a paraphrase detector. Another dataset where audio files were trained were extracted from various YouTube videos of web series and movies. Ground truth for these recorded videos was noted in a form of binary variable, violent or non-violent tone. A list of features extracted from these audio files is set out below..

1. MFCCs: 13 coefficients representing timbre in speech and music;
2. Chroma: 12 values (one for each note in the musical octave);
3. Mel: 128 Mel-Frequency bands, representing energy distribution over perceptual frequency scale; and
4. Spectral Contrast: 7 values describing the difference between peaks and valleys in the spectrum (captures

brightness/timbre).

The three primary components of the system are: (A) Frequency-Based Model with optimized RF; (B) NLP-Based Model improved with BERT; and (C) Fusion Model for final prediction.

A. An Optimized RF Model

The frequency-based methodology focuses on extracting significant auditory features to identify violent patterns in audio. This approach leverages MFCCs, Spectral Contrast, Mel-Spectrogram, and Chroma as primary components for feature extraction. To ensure uniform analysis across all audio samples, a windowing technique was applied, where all input audio files were split into 10-second clips. In cases where the audio duration was less than 10 seconds, padding was added at both the beginning and end to maintain consistency. This preprocessing step ensured that all audio inputs were of a fixed length, which allowed the model to extract meaningful frequency-based features while minimizing bias introduced by varying clip durations. The extracted features were then standardized using scaling techniques before being fed into an optimized RF Model for classification. This model was fine-tuned with hyperparameter optimization, balancing between model complexity and generalization to ensure robust performance in violence detection.

(a) Extraction of Features: Since MFCCs record the short-term power spectrum of audio sources, they are helpful for tasks involving speech. The twelve different pitch classes that help identify the tonal properties of violent noises are displayed in Chroma Features. The Mel-Spectrogram provides a time-frequency representation of sound and displays the energy distribution across frequencies. In the present setup, 40 MFCCs were extracted per frame, thus yielding a feature matrix of $40 \times T$ (where T denotes the number of time frames). The Mel-Spectrogram was computed with 128 frequency bins, while Chroma features used 12 pitch classes and Spectral Contrast captured 7 sub-band contrasts per frame. By calculating the difference between the spectrum's peaks and valleys, spectral contrast aids in the detection of dynamic sound variation. The MFCC calculation formula is presented in Eq. (1):

$$\mathrm{MFCC}(n) = \sum_{k=1}^{K} \log(S_k) \cdot \cos\left(\frac{\pi n(k-0.5)}{K}\right) \tag{1}$$

where, K is the total number of Mel filter banks, n is the cepstral coefficient index, and S_{k} is the spectral power at frequency k. Figure 2 shows the distribution of MFCC frequencies of important features.

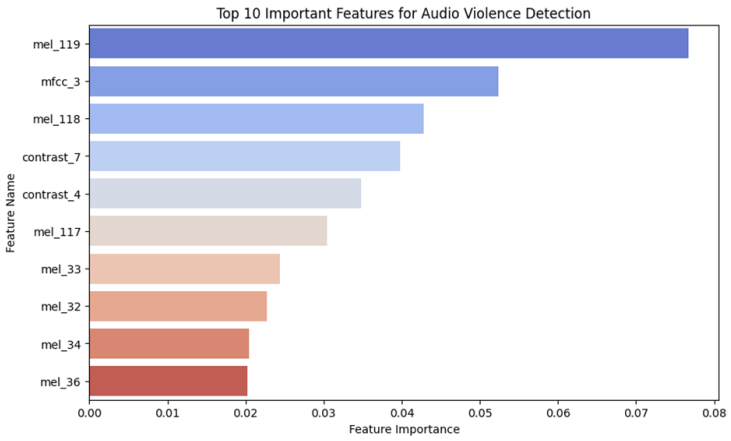

Figure 2. Chart of MFCC frequencies of important features

Figures 2 and 3 illustrates the importance of several audio elements in the XGBoost Fusion Model. Mel_119 (21.1%) and mfcc_3 (13.8%) are the most important traits, followed by contrast_7 (12.0%) and mel_118 (12.4%). The Mel and Contrast features, which are crucial for distinguishing between violent and non-violent sounds, illustrate how the model uses frequency-based analysis. The chart clearly displays a relative ranking of importance, according to which mel_119 is the most influential feature whereas mel_36 is the least influential among the top 10.

(b) Architecture of the Model: The Grid Search CV explored n_estimators $\in \{100, 200, 300\}$, max_depth $\in \{10, 15, 20\}$, and min_samples_split $\in \{2, 5, 10\}$. The optimization objective was to maximize the F1-score on the validation set, in order to ensure a balance between precision and recall rather than raw accuracy [9].

B. Fine-Tuned BERT, an NLP-Based Model

To detect violent content, the NLP-based approach employs speech-to-text transcription and an improved BERT model for text categorization.

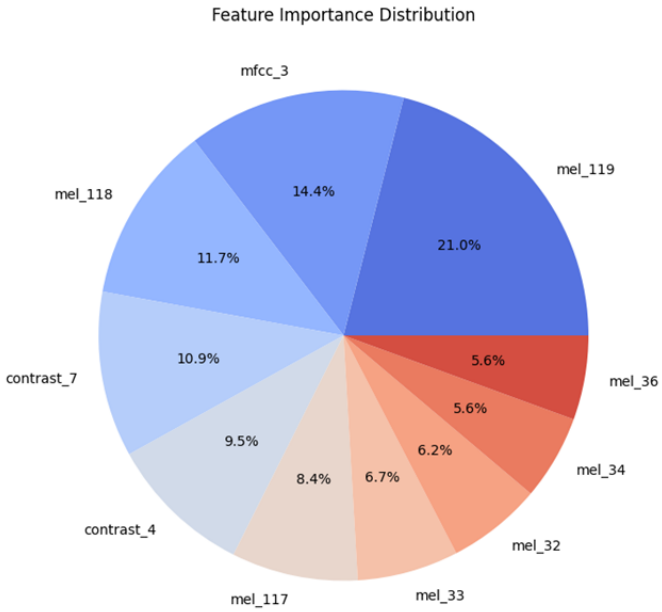

Figure 3. Distribution chart of feature importance

(a) Audio Transcription: To convert spoken words into text, audio recordings are transcribed using Google's Speech Recognition Application Programming Interface (API). This process enables the extraction of linguistic features for further analysis.

(b) Training with a dataset that has been associated with violence: BERT has been improved to distinguish between violent and non-violent words. The architecture of the model is: The input layer consists of tokenized text and attention masks. The contextual associations of the text are recorded by BERT encoder. The Classification Layer has softmax activation and is dense for binary classification. Softmax Activation Formula (used in the output layer of BERT) is shown in Eq. (2).

$$P(y = i \mid x) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_i}} \tag{2}$$

where, z_i represents the output logit for class i, n is the total number of classes (2 in this case: violent or non-violent).

C. Final Prediction Using a Fusion Model

To improve the accuracy of violence detection in an ensemble learning fashion, predictions from the frequency-based RF Model and the NLP-Based Fine-Tuned BERT model are coupled using a meta-classifier [13].

Stacking-Based Fusion Mechanism: Instead of averaging probability scores, the Fusion Mechanism Based on Stacking XGBoost is employed as a meta-classifier to determine the final classification. After being trained on the probability output from the RF and BERT fine-tuned models, this model determines the optimal decision boundary based on both predictions. Eq. (3) shows integration of the strengths of both models, resulting in enhanced accuracy, reduced false positives, and improved adaptability in real-world scenarios.

$$\begin{aligned} P_{audio} &= \text{ Probability of violence from the frequency-based model.} \\ P_{text} &= \text{ Probability of violence from the BERT fine tuned model.} \\ &\text{A feature set is created using these probabilities:} \\ P_{final} &= \text{ XGBoost } (P_{audio} + P_{text}) \end{aligned} \tag{3}$$

Each component of this methodology, ranging from feature extraction with Librosa and the optimized RF classifier to the fine-tuning of BERT for NLP tasks and the XGBoost-based fusion mechanism, is designed to ensure robust and accurate detection of violent audio events, even in challenging environments.

4 Implementation

The entire workflow was implemented and assessed on Google Colab; which provided high performance computing resources for model training and evaluation.

The four primary components for implementing the audio violence detection system are the Frequency-Based Model (Acoustic Analysis), NLP-Based Model (Text-Based Violence Detection), Meta-Learner Model (Ensemble Learning), and a GUI for real-time analysis. The developed system uses Python Libraries, such as Gradio for GUI development, XGBoost for ensemble learning, Transformers (Hugging Face) for NLP-Based Classification, Scikit-Learn for traditional machine learning, and Librosa for audio feature extraction. BERT-Base-Uncased model using the Hugging Face Transformers was used and finetuned for violent text recognition. The Frequency-Based Model distinguishes between violent and non-violent noises by analyzing unprocessed audio signals. Critical acoustic features, such as MFCCs, Chroma Features, Mel-Spectrogram, and Spectral Contrast, which aid in differentiating violent sounds (such as yelling and shouting) from non-violent ones, were extracted using the Librosa package. An optimized RF classifier, chosen for its resilience in processing noisy and high-dimensional audio data, received the standardized extracted features. Using Grid Search CV, hyperparameter tuning was carried out to optimize parameters like minimum samples per leaf (2), maximum depth (20), and number of estimators (100).

With a batch size of eight, ten training epochs, and a learning rate of 2×10^{-5}, the NLP-Based Model fared well in tests of violence classification. WordPiece Tokenization and pre-trained contextual embeddings were utilized to categorize violent and non-violent text segments extracted from transcribed speech using the improved BERT-Based model. The unequal distribution of classes was addressed during training by using class weighting techniques, which ensured that the model learned both violent and non-violent behaviors. Class weights were computed inversely proportional to class frequencies. Eq. (4) shows computation of class weights of BERT model:

$$W_i = \frac{N}{CXn_i} \tag{4}$$

where, N is the total sample size, C is the number of classes, and n_i is the number of samples in class i. This ensured that minority classes (violent or non-violent) contributed equally to the loss function.

The Google Speech Recognition API was applied to transcribe spoken text, which was preprocessed to remove stop words, punctuation, and noise before classification. To further increase prediction reliability, temperature scaling which altered logit distributions before softmax activation was used. Through the calibration of the model's confidence scores, this technique improves categorization stability. Early pausing was implemented using a three-epoch patience window, in order to ensure that training would immediately terminate when validation loss stopped improving. The model successfully mitigated overfitting risks by using dropout regularization inside the hidden layers of BERT and self-attention processes, resulting in strong generalization performance across validation and test datasets.

A meta-learner (XGBoost classifier) was implemented; this was a powerful gradient boosting algorithm known for handling complex feature interactions and improving generalization [13]. The meta-learner training process involved feeding the probability scores from both base models into the XGBoost classifier, which learned the optimal weighting and decision boundaries between the two models. The ensemble model was fine-tuned using Bayesian Optimization, adjusting essential parameters like the learning rate (0.05), max depth (6), and number of boosting rounds (100).

$$P_{final} = \text{XGBoost}(P_{audio} + P_{text}) \tag{5}$$

In Eq. (5), P_{audio} and P_{text} are the probability scores from RF and BERT models are represented by P_{audio} and P_{text}, respectively. Experimental data showed that the XGBoost metaLearner outperformed both standalone models, thus achieving great precision, accuracy, and recall.

The system was evaluated using a diverse dataset that included both violent and non-violent audio samples. The dataset included a range of auditory environments to assess their robustness in real-world situations. An Indian movie was considered to capture the violent tones of characters. All things considered, the hybrid ensemble learning system integrated text-based and acoustic violence detection, as well as employing a revised RF classifier, a BERT-based NLP model, an improved feature extraction pipeline, and an XGBoost meta-learner to increase classification accuracy. The user-friendly GUI ensures that applications for public safety, emergency response, and surveillance are feasible. The scalability, high accuracy, and utility of this implementation contribute to the practicality of automated violence detection in real-time settings.

5 Results and Discussion

With the implementation of GUI, users can upload or record audio, which would subsequently be transcribed and processed in real-time to detect potential violence. It shows the classification results of the frequency-based and text-based models as well as probability scores and a confidence bar plot to facilitate understanding in a few seconds. The implemented GUI provides a user-friendly interface for the real-time audio violence detection system, in which

audios could be uploaded or recorded with a real-time display of results. The processing time of a single audio file is reported to be 20-20 seconds. The GUI, shown in Figure 4, showcases transcribed text, predictions of individual sections along with their confidence scores from the audio-based RF Model and BERT-Based Model.

The GUI was tested on a CPU-only setup (Intel i7, 16GB RAM), achieving average inference latency of ∼1.2 seconds per 10-second clip. GPU acceleration (NVIDIA T4 on Google Colab) reduced latency to ∼0.4 seconds. The lightweight design also allows deployment on edge devices such as Jetson Nano, with real-time processing supported under optimized batch settings.

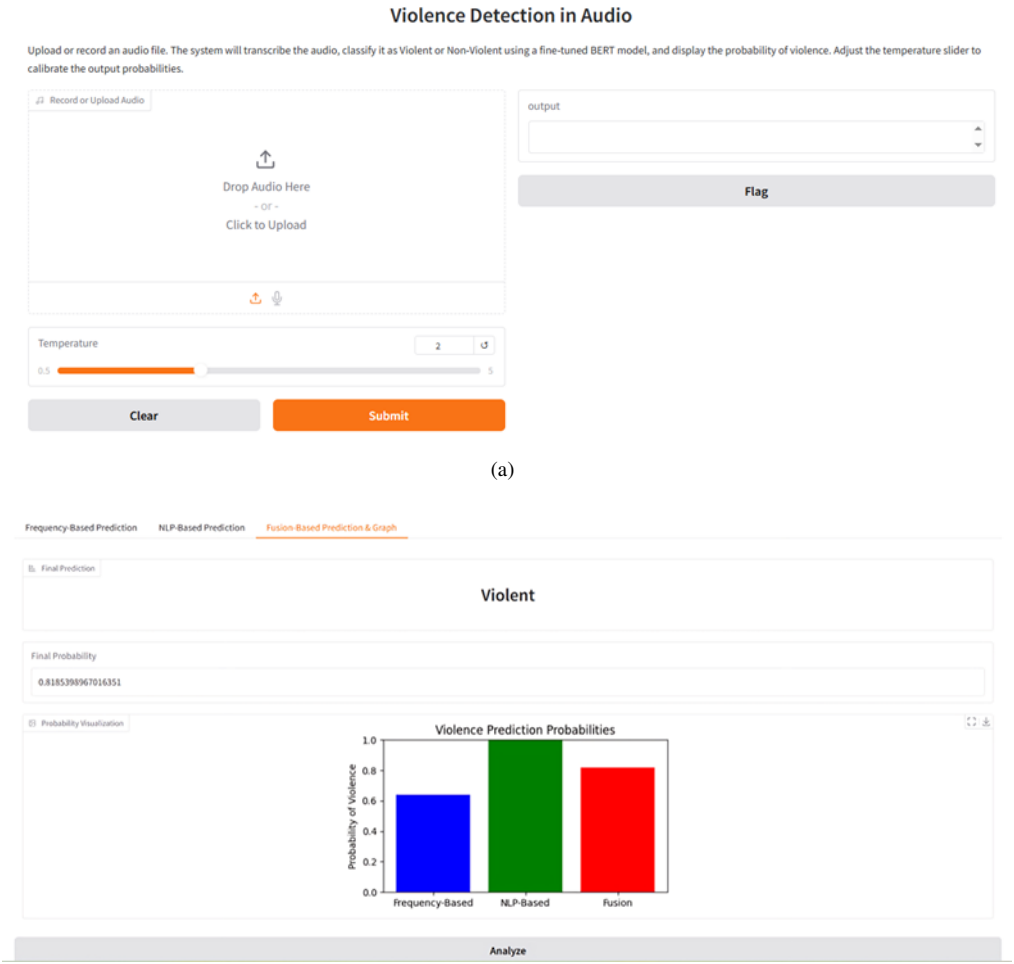

Figure 4. GUI of the multimodel system

The proposed Audio-Text Violence Detection System was evaluated using multiple models, including RF for audio features, BERT for text analysis, and an Ensemble Model (XGBoost) for fusion-based classification. The following subsections display a detailed performance analysis of these models.

5.1 RF Model (Audio-Based Classification)

The RF classifier was trained on MFCCs, Chroma Features, Mel-Spectrograms, and Spectral Contrast to identify violent events based on their acoustic characteristics [3]. Table 3 represents the RF Model which achieved a test accuracy of 9836%, (averaging classwise accuracy from Table 3) demonstrating its ability to classify violent and non-violent audio. However, text-based classification by BERT performed better, thus highlighting the importance of linguistic cues in violence detection. The best performance was achieved by the ensemble model, which combined both textual and acoustic features.

5.2 BERT Model (text-based classification)

The BERT model was trained to analyze the textual transcriptions of audio clips; as represented in Table 4, patterns of violent speech were effectively identified. The BERT model achieved Train Accuracy: 96.86%, Validation

Accuracy: 88.19% and Test Accuracy: 90.37%. The performance of BERT indicates that text-based violence classification is highly effective in identifying harmful language.

Table 3. Classification report—RF model (MFCC alone)

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Non-violent (0.0)	98.0	98.0	98.0	341
Violent (1.0)	99.0	99.0	99.0	515

Table 4. Classification report—BERT model (text-based alone)

Epoch	Training Loss	Validation Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	0.1239	0.3783	88.19	89.63	95.26	92.35
2	0.1818	0.5030	87.80	87.40	97.8	92.30
3	0.0232	0.4752	89.02	91.46	94.10	92.76
4	0.0894	0.0534	88.73	91.65	93.45	92.54

5.3 Fusion Model (Ensemble Approach)

To improve accuracy and robustness, this system integrates both linguistic (BERT) and acoustic (RF) features using XGBoost meta-learner [13]. The dataset shows class imbalance (Violent: 315,695 versus Non-violent: 116,622). To mitigate bias toward the dominant class, oversampling of minority instances and class-weight adjustments were applied during training to stabilize F1-score across both categories.

The model was trained and tested using an optimized dataset of violent and non-violent audio clips. The results below indicate that the ensemble learning approach significantly enhances classification performance as in Table 5.

Table 5. Classification report—fusion model

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Non-violent	97.0	93.0	95.0	116,622
Violent	98.0	99.0	98.0	315,695
Accuracy			97.37	432,317
Macro Average	97.5	96.0	96.5	-
Weighted Average	97.7	97.6	95.9	-

The MFCC-Based model achieves the highest expected accuracy since violent speech often carries distinct pitch and tonal characteristics. However, text-based violence detection remains essential for practical deployment. The fusion model assembling outperformed single classifiers in distinguishing between violent and non-violent audio clips, achieved an overall accuracy of 97.37%, with both high precision and recall. This demonstrated that the integration of textual and frequency-based features improved classification robustness.

6 Discussion

The ensemble approach integrated in the Audio-Text Violence Detection System considerably improves the accuracy rate to 97.37% by combining the content analysis of speech with BERT and feature extraction using RF. This combination is beneficial because it allows greater scrutiny of possible violent incidents through the speech of perpetrators.

The BERT model shows strong performance in violent language classification, delivering a score of 90.37%. The primary shortcoming of BERT is its failure to efficiently recognize violence wrapped in non-verbal sounds or background noises [25]. However, the RF Model demonstrates competence in the analysis of acoustic features and obtains an accuracy rate of 98.36%. Its major problem comes from the inability to tell the difference between violence and non-violence sounds. This problem can be solved by the inclusion of features extracted from texts, thus necessitating an optimal approach to violence detection via both audio and textual analysis.

The findings in this study were in line with earlier research which suggested that multimodal fusion significantly enhanced categorization capacity. Future studies should investigate the merging of few-shot learning with self-supervised learning techniques, as suggested by Sankaran et al. [12], to improve adaptability in a variety of situations. In addition, real-time violence detection systems would be more beneficial in emergency response and security contexts, provided that models for edge deployment could be improved.

7 Conclusions

This work introduced a novel ensemble-based audio violence detection technique to analyze both verbal and auditory data for increasing classification accuracy. By combining a RF classifier for frequency-based features and a refined BERT model for textual analysis, the fusion model could effectively capture the multi-modal nature of violent audio occurrences. An XGBoost meta-learner was added to enhance the decision-making process by balancing the contributions of both models to improve precision and recall. This ensemble learning approach outperformed standalone classifiers by addressing the shortcomings of traditional audio classification techniques and providing a more dependable solution for violence detection. Cross-validation and hyperparameter adjustment were adopted to improve the generalization and robustness of the model, in response to changes in speech patterns and background noise.

Further simplifying practical deployment, real-time GUI facilitates emergency response units, surveillance teams, and law enforcement organizations to engage with the proposed system. These upgrades enable the system to be used in practical safety applications, where it is essential to promptly and accurately identify violent situations.

Overall, this study advanced the field of automated violence detection by creating a scalable, highly accurate, and implementable method. To further improve performance, future research is recommended to investigate multilingual support by integrating variations of BERT models and using multilingual stopword support. Larger datasets and sophisticated deep learning architecture could also be explored in the future. Furthermore, multi-source sensor fusion and video-based violence detection could be added to the proposed system to greatly enhance its suitability for smart surveillance and monitoring public safety.

Author Contributions

Conceptualization, S.N. and R.P.; methodology, S.N.; software, D.G.; validation, A.J., S.P., and R.P.; formal analysis, A.J.; investigation, D.G.; resources, R.P.; data curation, D.G.; writing—original draft preparation, S.N.; writing—review and editing, S.N., A.J., and R.P.; visualization, S.P.; supervision, R.P.; project administration, R.P. All authors have read and agreed to the published version of the manuscript.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] L. Pham, P. Lam, T. Nguyen, H. Tang, and A. Schindler, "A toolchain for comprehensive audio/video analysis using deep learning based multimodal approach (A use case of riot or violent context detection)," *arXiv Preprint*, 2024, Art. no. arXiv:2407.03110. https://doi.org/10.48550/arXiv.2407.03110

[2] S. Page, S. Mangalvedhekar, K. Deshpande, T. Chavan, and S. Sonawane, "Mavericks at BLP-2023 Task 1: Ensemble-based approach using language models for Violence Inciting Text Detection," *arXiv Preprint*, 2023, Art. no. arXiv:2311.18778. https://doi.org/10.48550/arXiv.2311.18778

[3] D. Durães, B. Veloso, and P. Novais, "Violence detection in audio: Evaluating the effectiveness of deep learning models and data augmentation," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 3, p. 72, 2023. https://doi.org/10.9781/ijimai.2023.08.007

[4] Y. Yang and X. Wang, "Research on violent text detection system based on BERT-FastText model," *arXiv Preprint*, 2024, Art. no. arXiv:2412.16455. https://doi.org/10.48550/arXiv.2412.16455

[5] S. Jain, O. C. Phukan, A. B. Buduru, and R. Sharma, "The reasonable effectiveness of speaker embeddings for violence detection," *arXiv Preprint*, 2024, Art. no. arXiv:2406.06798. https://doi.org/10.48550/arXiv.2406.06798

[6] A. Anwar, E. Kanjo, and D. O. Anderez, "DeepSafety: Multi-level audio-text feature extraction and fusion approach for violence detection in conversations," *arXiv Preprint*, 2022, Art. no. arXiv:2206.11822. https://doi.org/10.48550/arXiv.2206.11822

[7] X. Peng, H. Wen, Y. Luo, X. Zhou, P. Y. K. Yu, and Z. Wu, "Learning weakly supervised audio-visual violence detection in hyperbolic space," *arXiv Preprint*, 2024, Art. no. arXiv:2305.18797. https://doi.org/10.48550/arXiv.2305.18797

[8] I. Soldevilla and N. Flores, "Natural language processing through BERT for identifying gender-based violence messages on social media," in *2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)*, 2021, pp. 204–208. https://doi.org/10.1109/ICICSE52190.2021.9404127

[9] M. D. Mahalle and D. V. Rojatkar, "Violence content detection based on audio using extreme learning machine," *Int. J. Recent Technol. Eng. IJRTE*, vol. 9, no. 5, pp. 107–113, 2021. https://doi.org/10.35940/ijrte.E5193.019521

[10] A. Bakhshi, J. García-Gómez, R. Gil-Pita, and S. Chalup, "Violence detection in real-life audio signals using lightweight deep neural networks," *Procedia Comput. Sci.*, vol. 222, pp. 244–251, 2023. https://doi.org/10.1016/j.procs.2023.08.162

[11] A. L. Mohammed, M. D. Swapnil, I. H. N. Peris, R. Nihal, R. Khan, and M. A. Matin, "Multimodal deep learning for violence detection: VGGish and MobileViT integration with knowledge distillation on Jetson Nano," *IEEE Open J. Commun. Soc.*, pp. 2907–2925, 2025. https://doi.org/10.1109/OJCOMS.2024.3520703

[12] A. N. Sankaran, R. Farahbakhsh, and N. Crespi, "Towards cross-lingual audio abuse detection in low-resource settings with few-shot learning," *arXiv Preprint*, 2024, Art. no. arXiv:2412.01408. https://doi.org/10.48550/arXiv.2412.01408

[13] N. Jaafar and Z. Lachiri, "Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance," *Expert Syst. Appl.*, vol. 211, p. 118523, 2023. https://doi.org/10.1016/j.eswa.2022.118523

[14] J. Shin, A. S. M. Miah, Y. Kaneko, N. Hassan, H. S. Lee, and S. W. Jang, "Multimodal attention-enhanced feature fusion-based weakly supervised anomaly violence detection," *IEEE Open J. Comput. Soc.*, vol. 6, pp. 129–140, 2025. https://doi.org/10.1109/OJCS.2024.3517154

[15] F. Zhu-Zhou, D. Tejera-Berengué, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Computationally constrained audio-based violence detection through transfer learning and data augmentation techniques," *Appl. Acoust.*, vol. 213, p. 109638, 2023. https://doi.org/10.1016/j.apacoust.2023.109638

[16] A. Chhabra, A. Sangroya, and C. Anantaram, "Formalizing and verifying natural language system requirements using Petri Nets and context based reasoning," in *MRC@IJCAI*, 2018, pp. 64–71. https://www.academia.edu/download/103247851/paper09.pdf

[17] Z. Wu, S. Jiang, X. Zhou, Y. Wang, Z. Zuo, L. Liang, and Q. Liu, "Application of image retrieva based on convolutional neural networks and Hu invariant moment algorithm in computer telecommunications," *Comput. Commun.*, vol. 150, pp. 729–738, 2020. https://doi.org/10.1016/j.comcom.2019.11.053

[18] A. Bensakhria, "Detecting domestic violence incidents using audio monitoring and deep learning techniques," 2023. https://doi.org/10.13140/RG.2.2.36128.97280/1

[19] G. Verma, R. Grover, J. Zhou, B. Mathew, J. Kraemer, M. De Choudhury, and S. Kumar, "A community-centric perspective for characterizing and detecting anti-Asian violence-provoking speech," *arXiv Preprint*, 2024, Art. no. arXiv:2407.15227. https://doi.org/10.48550/arXiv.2407.15227

[20] M. Chnini, N. Fredj, F. BenSaid, and Y. H. Kacem, "Violent speech detection in educational environments," in *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), Giza, Egypt*, 2023, pp. 1–8. https://doi.org/10.1109/AICCSA59173.2023.10479330

[21] V. Kumari, K. Memon, B. Aslam, and B. S. Chowdhry, "An effective approach for violence detection using deep learning and natural language processing," in *2023 7th International Multi-Topic ICT Conference (IMTIC), Jamshoro, Pakistan*, 2023, pp. 1–8. https://doi.org/10.1109/IMTIC58887.2023.10178618

[22] S. Sekkate, S. Chebbi, A. Adib, and S. B. Jebara, "A deep learning framework for offensive speech detection," in *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC), Marrakech, Morocco*, 2024, pp. 1–6. https://doi.org/10.1109/ISIVC61350.2024.10577928

[23] A. Mehrotra, S. Chaudhary, and G. Sharma, "Violent speech detech in videos using natural language processing," *YMER Digit.*, vol. 21, no. 5, pp. 610–619, 2022. https://doi.org/10.37896/YMER21.05/69

[24] Z. Waseem and D. Hovy, "Dataset: Waseem Dataset," 2024. https://doi.org/10.57702/QAFDBW7H

[25] P. Fortuna, J. Rocha Da Silva, J. Soler-Company, L. Wanner, and S. Nunes, "A hierarchically-labeled Portuguese hate speech dataset," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 94–104. https://doi.org/10.18653/v1/W19-3510

[26] M. T. I. Khondaker, M. Abdul-Mageed, and L. V. S. Lakshmanan, "DetoxLLM: A framework for detoxification with explanations," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 19 112–19 139. https://doi.org/10.18653/v1/2024.emnlp-main.1066

[27] M. K. Gautam, P. K. Rajput, Y. Srivastava, and A. Kansal, "Real time violence detection and alert system," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 1139–1147, 2024. https://doi.org/10.22214/ijraset.2024.59027