Audio-Based Urgency Detection with SHAP-Enhanced Feature Refinement

Ankit Dighe
Computer Science & Engineering
Sardar Patel Institute of Technology
Mumbai, India
ankit.dighe@spit.ac.in

Sanika Gadkari
Computer Science & Engineering
Sardar Patel Institue of Technology
Mumbai, India
sanika.gadkari@spit.ac.in

***Abstract*—In high-stakes emergency response scenarios, the ability to automatically detect urgency from audio inputs can significantly improve response times and save lives. However, most existing systems focus primarily on emotion recognition or rely on clean, structured datasets, which do not adequately capture the chaotic and stressful nature of urgent real-world communication. This research introduces an innovative method for urgency classification using speech features derived from 911 emergency calls and public datasets such as RAVDESS and LibriSpeech. A carefully balanced and realistic dataset was developed, with noise augmentations to better simulate the acoustic diversity found in real-life scenarios. Baseline performance was established using conventional machine learning algorithms. To improve model interpretability and feature efficiency, SHapley Additive exPlanations (SHAP) has been integrated to guide feature selection via an elbow-based importance cut-off. Performance was assessed using common evaluation metrics, including accuracy, precision, recall, and F1-score. These findings demonstrate that XAI-driven feature selection can simplify urgency detection models without sacrificing predictive performance, offering a promising foundation for real-time emergency response systems.**

***Index Terms*—Audio analysis, Bias detection, Feature selection, SHAP, Urgency detection**

I. Introduction

In critical and life-threatening situations, the ability to automatically detect urgency in spoken language can significantly improve emergency response systems. Unlike traditional systems that rely heavily on human operators, an intelligent urgency detection framework can assist dispatchers in prioritizing cases more efficiently [1]. Such systems can be of value particularly in scenarios involving emergency calls, distress signals, or public safety applications where response time is crucial. Automatic urgency detection in speech could potentially be used in a wide range of scenarios, from healthcare monitoring and critical alarm systems to customer service.

Speech carries a multitude of acoustic cues beyond its linguistic content. These include spectral, prosodic, and temporal features [2]. These features are important in understanding the emotional and physiological states of the speaker, such as stress, anxiety, or calmness. Urgent speech is often characterized by elevated pitch, louder amplitude, faster speaking rate, and increased voice instability [3]. Using these features allows models to understand not just what is being said, but how it is being said, which is critical for urgency classification.

Much of the recent research in speech analysis has focused on emotion detection, sentiment analysis, and speaker recognition. While emotion and urgency are correlated to some extent, urgency encompasses a more task-specific and context-sensitive dimension [4] that emotion classifiers may fail to capture accurately. This study introduces a realistic urgency classification framework using a combination of authentic 911 calls and carefully selected non-urgent speech samples. To bridge the domain gap, targeted synthetic noise augmentations were introduced to simulate diverse acoustic environments. Existing emotion-labeled datasets, such as RAVDESS or CREMA-D, offer a broad categorization of affective states but lack real-world indicators of urgency. As a result, urgency detection remains an underexplored domain, in part due to the unavailability of dedicated, annotated datasets representing urgent and non-urgent speech in realistic environments.

This paper addresses this gap by proposing a robust urgency classification pipeline built on a curated dataset that combines authentic 911 emergency call recordings (urgent class) and selectively sampled utterances from LibriSpeech and RAVDESS (non-urgent class). The audio data was preprocessed and various noise augmentation techniques were applied to simulate real-world acoustic variability. We propose a SHAP-enhanced feature refinement strategy to not only improve performance but also to enhance model transparency. The key contributions of this work are summarized as follows:

1) A balanced and realistic dataset combining real 911 emergency calls (urgent) with carefully selected samples from RAVDESS and LibriSpeech (non-urgent) and enhancing realism through targeted noise augmentation.
2) A SHAP-informed feature selection strategy that identifies and retains the most influential acoustic features through an elbow-based importance analysis, improving model interpretability and enhancing computational efficiency.
3) The paper extends its work to assess the model for unexpected biases with the help of SHAP and extensive testing, thus improving the ethical aspect of this proposed work.
4) A systematic benchmark of multiple machine learning classifiers evaluated before and after SHAP-guided fea-

ture reduction, demonstrating that comparable or improved classification performance can be achieved with a significantly reduced feature set.

The rest of this paper is structured as follows: Section II discusses similar work on urgency and speech-based emotion classification. Section III explains the dataset construction and the rationale behind data selection. Section IV describes the suggested process for feature extraction, augmentation, and model design. Section V presents experimental evaluations, outcomes, and insights derived from applying explainable AI techniques.

II. Related Works

The detection of urgency in speech and text has become an important area of research, particularly in applications that involve emergency response, public safety, and human-computer interaction. Identification of urgency through auditory cues including pitch, energy and speech rate have been the focus of several studies. Landesberger et al. [4] demonstrated that spoken utterances can be classified as urgent or non-urgent using only acoustic features, confirming that urgency is inferable without linguistic content, while Hellier et al. [5] highlighted the interplay between semantics and acoustics in shaping perceived urgency. Gaviria et al. [6] extended this to real-world deployment via a portable deep learning device for audio distress recognition in urban environments. Textual and dialogue-level approaches have also been explored: Akaybicen et al. [7] leveraged LLMs to identify critical conditions from unstructured medical text, Trabelsi et al. [8] applied context-aware and context-agnostic NLP to voicemail urgency detection, and Enzo et al. [9] showed that modelling communicative intentions improves classification by capturing subtle dialogue-level urgency cues. At the multimodal level, Fime et al. [10] integrated signal processing and deep learning across multiple sound categories, while Villegas-Ch et al. [11] deployed a smart city emergency detection system combining acoustic and visual inputs. Regarding explainability, Grinberg et al. [12] applied SHAP to transformer-based audio deepfake detection to identify temporally influential signal segments, and the XAI4EEG framework [13] demonstrated the utility of spectral and temporal SHAP explanations in safety-critical time-series classification, both providing direct precedent for SHAP-guided feature analysis in audio-domain tasks.

III. Dataset

The urgency classification task requires distinguishing between urgent and non-urgent audio samples. Therefore, a thorough dataset was created that was clean and could be easily used from a model point of view.

A. Data Collection

Table I shows the data distribution of the urgent and non-urgent class.

1) Urgent Class: This class is composed of real-world 911 emergency call recordings: audio segments that inherently capture high-stress, emotionally charged, and fast-paced communication. These calls are representative of real emergency situations making them ideal for modeling vocal urgency. Each audio sample was trimmed to a 6-second duration, beginning from the point where the caller starts explaining the nature of the emergency. This ensures that all inputs are aligned and focuses the analysis on the portion of the call most indicative of vocal urgency.

TABLE I
Dataset Statistics

Total Samples	Urgent	Non-Urgent	
		RAVDESS	LibriSpeech
1419	707	432	280

2) Non-Urgent class: In order to construct a well-balanced and representative non-urgent class for the classification task, two publicly available speech datasets were utilized: RAVDESS and LibriSpeech. The RAVDESS dataset is a high-quality, emotional speech and song corpus. The speech subset of RAVDESS dataset [14] provides audio samples of seven emotional expressions including calm, happy, sad, angry, fearful, surprised, and disgusted. For the purpose of creating a non-urgent class, 432 of the 1440 audios were considered. Only neutral, calm, and mildly expressive emotional recordings (e.g., happy-normal, sad-normal) were included. More intense emotions such as anger, fear along with strong disgust and surprise which may mimic urgent vocal cues were excluded to prevent confusion between urgency and emotion intensity. The LibriSpeech dataset is an extensive collection of recorded English speech, designed for automated speech recognition (ASR) research. LibriSpeech contributes structured, emotionally-neutral, and contextually calm speech, representing typical non-urgent conversation tones.

B. Data Preprocessing

The 911 dataset inherently contains background noise, channel distortion, and non-speech artifacts (e.g., static, crosstalk), while RAVDESS and LibriSpeech are studio-quality. This mismatch introduces a domain bias, a model might unintentionally learn to classify urgency based on audio quality or recording conditions instead of genuine speech cues. To enhance the quality of the dataset, a diverse noise injection pipeline was employed on the dataset. All noise injections used in this study were synthesized using audio signal processing techniques. These were chosen to closely mimic real-world distortions commonly found in emergency call environments. The augmentation was carefully stratified as follows:

- 10% (71) of the audio files were modified using GaussianSNR noise with signal-to-noise ratios randomly sampled between 10 dB and 30 dB.
- 30% (214) of the audio samples underwent a combination of bandpass filtering, pink noise, hum noise, and static

bursts. This introduced structured low-frequency and mid-frequency interferences often found in phone lines or analog transmissions.
- Next 30% (214) were processed using a more complex chain that included bandpass and high-pass filters, pink and hum noise, static bursts, dropouts, and μ-law encoding.
- The remaining 30% (213) audios were preprocessed using a mix of compression artifacts and Gaussian noise, simulating audio degradation from packet loss.

Noise injection was applied to the 712 non-urgent recordings to better reflect real-world recording conditions and reduce domain shift. No additional samples were generated, and the dataset size remained unchanged. While preprocessing reduces the acoustic gap, it cannot fully eliminate domain mismatch and models may retain some sensitivity to studio-quality artefacts.

IV. Methodology

A. Overview

In order to classify audio samples as urgent and non-urgent, the proposed approach utilizes acoustic features derived from speech signals. This approach applies a ML model to perform classification. Explainability tools like SHAP have been used to quantify each input feature's value and importance to the predictions of the model. Lastly, the performance of the model has been evaluated using various metrics to examine the impact of feature selection and preprocessing.

B. Feature Extraction

For audio classification, a comprehensive set of 27 acoustic and prosodic features have been extracted from each audio clip. These 27 features have been chosen based on speech and emotion recognition research and are aimed at extracting the spectral and temporal dynamics of human speech that may reflect urgency. Each audio clip was first resampled to a consistent sampling rate and normalized for amplitude. Table II shows the features that were extracted and used for the classification task.

C. Feature Pruning and Dimensionality Reduction

There were 27 audio features in the original feature set. To mitigate the risk of overfitting, feature correlation analysis was performed using Pearson's correlation coefficient. Features exhibiting strong linear correlation ($|r| > 0.9$) were considered redundant. In such cases, one feature from each set of correlated features was kept and the rest were eliminated from the feature set. Non-linear redudancies were addressed by SHAP-guided selection which captures interaction effects.

From the originally extracted 13 MFCCs, only the first three coefficients (MFCC1–MFCC3) were retained, as they capture the dominant spectral characteristics relevant to perceptual features of speech. Higher-order coefficients were excluded due to their sensitivity to noise and limited contribution to coarse-grained tasks such as urgency detection. This selection reduces feature dimensionality while minimizing overfitting and preserving essential information.

TABLE II
Summary of Extracted Audio Features

Feature Category	Features
Spectral Features	• MFCCs • Spectral Centroid • Spectral Bandwidth • Spectral Rolloff
Energy & Intensity	• RMS Energy • Loudness • Zero-Crossing Rate
Pitch & Voicing	• Fundamental Frequency (F0) • Jitter • Shimmer • Harmonics-to-Noise Ratio (HNR)
Formant Frequencies	• Formant F1 • Formant F2
Temporal Dynamics	• Speaking Rate (words per second) • Voice Segment Duration

D. Model Implementation

To establish baselines and evaluate the effectiveness of classification techniques, the following models have been implemented:

- Logistic Regression: A simple linear model used for its interpretability
- Decision Tree: A non-linear model that uses feature splits to produce interpretable decision paths
- Random Forest: An ensemble of decision trees to increase generalization but has less interpretability
- XGBoost: A gradient boosting framework that shows strong performance on tabular data
- TabNet: A DL model designed for data in a tabular form that learns feature importance using attention mechanisms

These models were selected to cover different types of learning paradigms such as linear, tree-based, ensemble and deep learning. Each model was trained using a train-test split of 80:20 and the features were standardized using min-max standardization. To ensure robustness and reduce bias arising from a single split, Stratified 5-Fold Cross Validation was employed during training. Stratification preserved the class distribution across folds, providing a more reliable estimate of model generalization. SHAP-based feature selection was performed independently on each fold for training split only, ensuring no data leakage.

E. Explainability using SHAP

SHAP is an XAI method which assigns an importance value or SHAP value to every feature based on the contribution

TABLE III
SHAP-SELECTED INFLUENTIAL FEATURES PER MODEL

Model	SHAP-Selected Features
Logistic Regression	Spectral Centroid (mean), RMS Energy (mean), Loudness (LUFS), Mean F0, Jitter, Shimmer, Formant F1, Zero Crossing Rate (mean), HNR, Voice Segment Duration, Speaking Rate, MFCC1, MFCC2
Decision Tree	Formant F2, Zero Crossing Rate (mean), Spectral Centroid (mean), MFCC3
Random Forest	Formant F2, MFCC2, MFCC3, Spectral Centroid (mean), Zero Crossing Rate (mean), Voice Segment Duration, Formant F1
XGBoost	Formant F2, MFCC3, Zero Crossing Rate (mean), MFCC2, Voice Segment Duration
TabNet	Formant F2, Spectral Centroid (mean), Voice Segment Duration, MFCC3, Mean F0, Formant F1

of the feature to a particular prediction. SHAP enables understanding the overall behavior of the model and explaining individual predictions. In this work, SHAP has been utilized to identify the most influential audio features in distinguishing between urgent and non-urgent speech.

To identify the optimal feature subset for every model, the Elbow Method was applied to the ranked importance curve. Refer to Table III to see the list of features chosen after SHAP analysis. Features are ordered by descending SHAP contribution and the point where the curve transitions from steep decline to a near-flat region marks the natural cut-off. Features beyond this elbow exhibit diminishing returns and were removed using visual inspection. Fig 1. shows the Elbow plot for XGBoost. The region to the left of the red dotted line contains features which were considered most influential for the model.

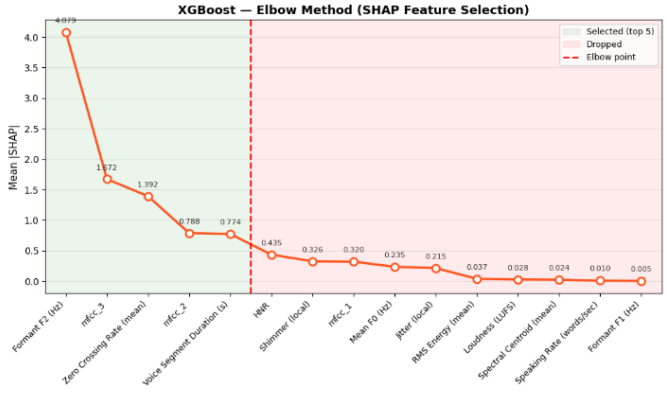

Fig. 1. Elbow plot of SHAP feature importances ranked in descending order of mean absolute SHAP value. The dashed red line denotes the elbow cut-off point, beyond which features exhibit diminishing marginal contribution to model predictions.

TABLE IV
PERFORMANCE OF MODELS WITH ALL FEATURES VS. SHAP-OPTIMIZED FEATURES

Model	Setup	Accuracy	Precision	Recall	F1-score
Logistic Regression	All Features	0.9859	0.9859	0.9859	0.9859
	SHAP-selected	0.9859	0.9859	0.9859	0.9859
Decision Tree	All Features	0.9789	0.9790	0.9789	0.9789
	SHAP-selected	0.9824	0.9826	0.9824	0.9824
Random Forest	All Features	0.9930	0.9930	0.9930	0.9930
	SHAP-selected	0.9965	0.9965	0.9965	0.9965
XGBoost	All Features	0.9930	0.9930	0.9930	0.9930
	SHAP-selected	0.9965	0.9965	0.9965	0.9965
TabNet	All Features	0.9683	0.9685	0.9683	0.9693
	SHAP-selected	0.9894	0.9895	0.9894	0.9894

V. EXPERIMENTATION

A. Evaluation Metrics

Common assessment metrics such as f1-score, accuracy, precision, and recall were used to evaluate the performance of classification models with emphasis on the implications of misclassification in the context of urgency detection from audio. In such scenarios, false negatives can be significantly more costly. Additionally, training and inference times were recorded to evaluate model deployability and the scalability impact of XAI.

B. Results

To establish a reliable baseline, comparisons were made with five classification models. This comparison served two purposes: (1) to identify the most promising model architectures for the urgency classification task, and (2) to quantify the performance improvements gained through subsequent explainability and feature reduction techniques.

1) XAI Results: Table IV presents a comparison between the baseline models and the same models retrained after applying SHAP-based feature selection. The contribution of this analysis is in identifying acoustically interpretable features that characterize urgency. This reduces noise and redundancy while retaining the most discriminative information. The resulting feature subsets are smaller, more meaningful, and yield classifiers that are more efficient in both training and inference.

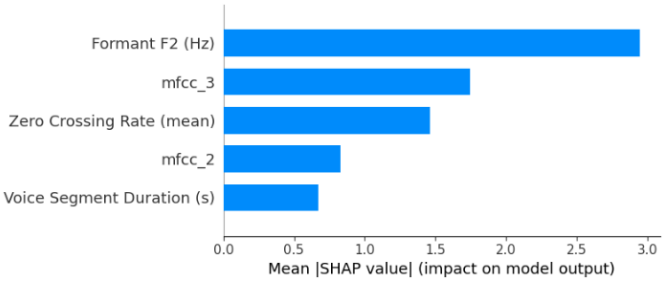

Fig. 2. Summary plot of features selected in XGBoost after using SHAP.

TABLE V
EFFICIENCY OF MODELS BEFORE AND AFTER SHAP-BASED FEATURE SELECTION IN MILLISECONDS

Model	Train (All)	Train (SHAP)	Infer (All)	Infer (SHAP)
Logistic Regression	10.0	5.2	0.6	0.3
Decision Tree	12.50	3.82	0.33	0.32
Random Forest	223.24	197.82	8.23	6.21
XGBoost	45.52	22.12	1.64	1.27
TabNet	2061	2033	19.30	18.90

2) Bias Detection Results: To guarantee that the predictions given by the model are not overly reliant on a single potentially sensitive feature, we conducted bias detection experiments - feature ablation and counterfactual feature sensitivity. For feature ablation, a change of $<1\%$ was observed in metrics like accuracy, F1 score, etc after removing prominent features such as Formant F2. The counterfactual feature sensitivity analysis provides a validation of this conclusion. When Formant F2 values were shifted downward by one standard deviation, only 3.52% of predictions changed, and critically, none of the urgent samples were misclassified as non-urgent as a result of this. When shifted upward by one standard deviation,the prediction change rate fell further to just 0.35%, with only 1 out of 707 samples incorrectly classified. These figures are substantially below the 5% sensitivity threshold commonly used in fairness evaluation literature, indicating that the model's urgency predictions are robust to the kind of speaker-level demographic variation that Formant F2 encodes. Broader evaluation of speaker, gender demographics would require speaker-level metadata which is not present in the current dataset. The convergence of findings from two independent methodologies reinforces the conclusion that the approach does not exhibit significant bias with respect to Formant F2, despite its comparatively higher mean SHAP value (Fig. 2).

C. Computational Efficiency

In addition to the quantitative analysis of the models, we also analyze the computational efficiency and the impact of feature refinement through XAI on these techniques. By assessing factors like training time and inference time the practicality of each model for real-time or resource-constrained deployment scenarios can be evaluated. A smaller model footprint enables deployment on edge devices or low-power environments.

Table V shows the computational costs for all classification models. Among the models tested, logistic regression and decision trees exhibited the lowest training and inference times, making them suitable for lightweight applications. XGBoost and TabNet, while more computationally intensive, offered superior classification performance, particularly after SHAP-based feature refinement.

VI. CONCLUSION AND FUTURE SCOPE

This study offers a practical and interpretable approach to urgency detection using audio recordings, addressing a gap often overlooked in favor of emotion recognition. By combining real 911 calls with structured non-urgent datasets and applying SHAP-based feature refinement, we built a realistic, balanced dataset and improved model performance across multiple classifiers. While the proposed approach demonstrated strong performance on the constructed dataset, further validation on real-world non-urgent recordings would provide a more rigorous assessment of generalization. This lays the groundwork for real-world applications in emergency dispatch, surveillance, and healthcare. Future work may explore multimodal inputs, advanced deep learning models, and context-aware features to boost adaptability in diverse scenarios.

REFERENCES

[1] A. S. M. Mohsin, S. H. Choudhury, and M. A. Muyeed, "Automatic priority analysis of emergency response systems using internet of things (IoT) and machine learning (ML)," Transportation Engineering, p. 100304, Jan. 2025

[2] G. H. Mohmad Dar and R. Delhibabu, "Speech Databases, Speech Features, and Classifiers in Speech Emotion Recognition: A Review," in IEEE Access, vol. 12, pp. 151122-151152, 2024

[3] M. Kobayashi, Y. Hamada, and M. Akagi, "Acoustic features correlated to perceived urgency in evacuation announcements," Speech Communication, vol. 139, pp. 22–34, Mar. 2022.

[4] J. Landesberger, U. Ehrlich, and W. Minker, "Do the urgent things first! - Detecting urgency in spoken utterances based on acoustic features," 28th ACM Conference on User Modeling, Adaptation and Personalization, Jul. 2020.

[5] E. Hellier, J. Edworthy, B. Weedon, K. Walters, and A. Adams, "The Perceived Urgency of Speech Warnings: Semantics versus Acoustics," Human Factors the Journal of the Human Factors and Ergonomics Society, vol. 44, no. 1, pp. 1–17, Mar. 2002.

[6] J. F. Gaviria, A. Escalante-Perez, J. C. Castiblanco, N. Vergara, V. Parra-Garces, J. D. Serrano, A. F. Zambrano, and L. F. Giraldo, "Deep Learning-Based portable device for audio distress signal recognition in urban areas," Applied Sciences, vol. 10, no. 21, p. 7448, Oct. 2020.

[7] F. Akaybicen, A. Cummings, L. Iwuagwu, X. Zhang, and M. Adewuyi. "A Machine Learning Approach for Emergency Detection in Medical Scenarios Using Large Language Models." arXiv preprint arXiv:2412.16341 (2024).

[8] A. Trabelsi, S. Soussilane, and E. Helbert, "Voicemail urgency detection using context dependent and independent NLP techniques," Proceedings of the 14th International Conference on Agents and Artificial Intelligence, pp. 450–456, Jan. 2023.

[9] L. Enzo, B. Nils, F. Benamara, M. Alda, V. Moriceau, and C. Camille, "Speech acts and communicative intentions for urgency detection," Proceedings of the 11th Joint Conference on Lexical and Computational Semantics, pp. 289–298, Jul. 2022.

[10] A. A. Fime, Md. Ashikuzzaman, and A. Aziz, "Audio signal based danger detection using signal processing and deep learning," Expert Systems With Applications, vol. 237, p. 121646, Sep. 2023.

[11] W. Villegas-Ch and J. Govea, "Application of deep learning in the early detection of emergency situations and security monitoring in public spaces," Applied System Innovation, vol. 6, no. 5, p. 90, Oct. 2023.

[12] P. Grinberg, A. Kumar, S. Koppisetti, and G. Bharaj, "What does an audio deepfake detector focus on? A study in the time domain," ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, Mar. 2025.

[13] D. Raab, A. Theissler, and M. Spiliopoulou, "XAI4EEG: spectral and spatio-temporal explanation of deep learning-based seizure detection in EEG time series," Neural Computing and Applications, vol. 35, no. 14, pp. 10051–10068, Sep. 2022.

[14] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE, vol. 13, no. 5, p. e0196391, May 2018.