Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multimodal AI-Based Frustration and Escalation Risk Prediction in Banking Call Centres Using Voice Prosody, Text Sentiment, and Interaction Metadata

Bhavya Sri Sanku

Independent Researcher, USA

ARTICLE INFO

Received: 18 June 2026

Revised: 01 Aug 2026

Accepted: 15 Aug 2026

ABSTRACT

This study proposes a multimodal approach to the prediction of customer frustration and escalation risk in call centres of financial institutions based on voice (prosody) and text (sentiment) as well as interaction (metadata). Low, medium and high frustration class es were generated using the IEMOCAP dataset, and various machine-learning models were tested on it. These results indicated that the XGBoost model with Audio, Text and Metadata performed best with 85.7% accuracy, 85.7% F1-score and 0.961 ROC-AUC. There was generally a large amount of predictive information compared with audio and metadata features. Results indicate that multimodal fusion could significantly contribute to the earlier and more accurate identification of the customers' dissatisfaction and thus to the proactive customer service intervention.

Keywords: Multimodal Learning; Customer Frustration; Voice Prosody; Text Sentiment; Interaction Metadata; XGBoost; Escalation Risk.

I. INTRODUCTION

Background

Retail banking call centres are highly valuable places that customer experience depends on the performance of the service, and operational efficiency and responsiveness rely on Artificial Intelligence (AI). The conventional sentiment analysis techniques are limited to analysing positive or negative text sentiment content that omits the voice and emotional behaviour elements in the process of expression [1]. Phrases of frustration can be represented in voice and language, and may arise prior to being directly expressed in language. The interaction information, such as number of repeats, transfer frequency, and hold time, is also an indicator of an overall failing CEX [2]. Multimodal deep learning can present this opportunity via audio data, text and structured call details to enhance customer service and improve the ability. In order to predict customer escalation and frustrations at a early stage of the call at the present banking service delivery arrangements.

Problem Statement

A lack of sentiment in the text may not be detected in a text-based sentiment model, but can be conveyed through tone, repeated transfers and a longer hold time or speech pauses. This invariably creates less escalation activity in the bank's call centres to address earlier as customers' voices can escalate even before there is clear negative language [3]. This is necessary to take advantage of the characteristics of speech prosody, text sentiment, and interaction metadata in order to require a multimodal prediction model.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Research Aim and Objectives

Aim

The aim of the research is to help AI models accurately predict customers' frustrations and escalation in multi-modal applications.

Objectives

- To analyse voice prosody features related to customer frustration in the context of Banking Phone Calls.
- To evaluate whether text sentiment metrics can detect emerging customer issues that the customer is experiencing during the conversation.
- To assess and perform analysis of interaction metadata on patterns for potential escalation risk of customers.
- To integrate multimodal features into a predictive model for automated frustration detection.

Novel Contribution

The multimodal deep learning approach to predicting customer frustration, as both voice prosody features and textual sentiment analysis of the interaction as well as interaction metadata is execute in this paper. The approach is based only on textual information, while the proposed model includes two dimensions of data, such as linguistic, behavioral, and emotional [4]. This also evaluates the features at an audio level, as well as at a transcript/call-process level, both separately and together, through comparative modelling. The framework is specifically designed for banking call centres where customers can have to wait, move to different departments and have to repeat and take multiple calls, before finally getting a satisfactory response [5]. Innovativeness is attributed to its ability to integrate multimodal fusion with early risk prediction, which allows for timely interventions, better service recovery, and proactiveness in banking customer experience management strategies.

II. LITERATURE REVIEW

Multimodal Deep Learning Approaches for Customer Frustration Detection in Calls

The meaning of customer emotions is complex, with the help of multiple signals analyzing. The process is more effective than using a single signal, so multimodal deep learning is a successful method to handle the emotion of customers [6]. In call centre situations, customer needs can be voiced verbally, vocally, and as interaction style. High-level representations can be derived from audio and text through deep learning models, and structured features provide further contextual information on operations [7]. The following are the most popular techniques for combining heterogeneous features like convolutional neural networks, recurrent neural networks, transformers, and multimodal fusion architectures.

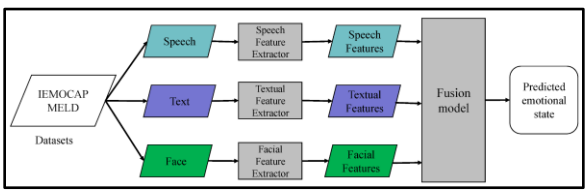

Fig 1. A Survey of Deep Learning-Based Multimodal

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A combination of modalities can actually be more complete to capture the vehicle's emotional state than a unimodal model, since the weakness of one modality could be compensated for by the other [8]. Words can be neutral but delivered at a higher speed or with an irritated tone of voice. However, there are challenges related to data synchronization, lack of modalities, computational power and feature imbalance in multimodal. Therefore, to solve customer frustration detection applications, interpretation and robustness, especially predictive performance, are still challenging tasks that can be accomplished with the help of fusion strategies.

Voice Prosody Analysis for Emotion Recognition and Escalation Risk Prediction

The pitch, speaking rate, loudness, pauses, and vocal intensity may contain information about stressful or frustrated customer attitudes [9]. This is important to note that Voice Prosody can be a key source of customer emotional information. The acoustic features that are usually derived from speech-based emotion recognition systems include Mel-frequency cepstral coefficients, fundamental frequency, energy, spectral characteristics and temporal patterns [10]. Auto-Learning relevant representations from audio signals is possible by deep learning methods such as convolutional neural networks, recurrent architectures and speech transformers [11]. As an example, there may be a ton of dissatisfaction from banking call centres when customers don't outright display negative emotions, but rather through prosodic changes.

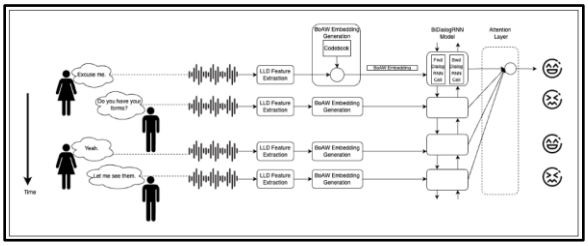

Fig 2. Optimizing Speech Emotion Recognition with Machine Learning

However, there are factors such as speaker age, sex, accent, language, background noise, microphone quality and an individual's communication style that can affect the operation of voice-based prediction. These differences introduce restrictions or challenges for generalisation between customers and across the different environments [12]. Prosody needs to be analysed in a multimodal way to differentiate authentic frustration from typical variations in conversation as a speech mode in order to better make reliable predictions of escalation risk.

Text Sentiment Analysis for Understanding Customer Emotions in Banking

Sentiment scores are popular for determining positive, negative, and neutral sentiment from customer conversations, customer complaints, customer reviews and customer support communications [13]. Auto-speech recognition can be relied upon to convert the conversation on the call center to a text file and analyzed using natural language processing techniques in the case of a banking call center [14]. The traditional systems are mainly based on lexical characteristics and machine learning classifiers, modern systems are more and more utilizing transformer-based models that can quantify contextual connections between words and sentences.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

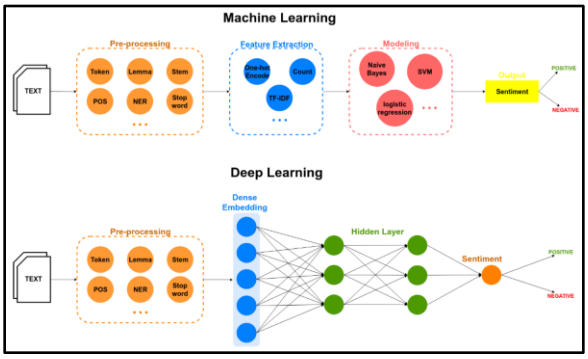

Fig 3. Sentiment Analysis Based on Deep Learning

A text's words indicating dissatisfaction, urgency, repeated complaints, and desire to escalate can be detected by analyzing the text [15]. Neutral statements or sentences can be a sign of great discontent, especially during financial transactions with customers when they try to be polite and neutral. A situation where sarcasm is used, mistakes made by the speaker, speeches that aren't fully complete, domain specificity, or transcription errors can also exacerbate readings of a text model [16]. This is possible that only relying on transcript information can have missed out on information regarding important feelings that can have been picked up by voice tones, call interactions, etc., aspects of vocal behavior.

Interaction Metadata and Multimodal Fusion for Early Escalation Prediction

Interaction Metadata is discerned without classified material and can demonstrate the dissatisfaction of customers [17]. Call centre features that can be relevant to banking include hold time, transfers, length of call, return calls, agent delay, complications received before, how many times the agent is interrupted [18]. Satisfaction can go down as a result of a long wait or multiple transfers, and may escalate to higher levels of conflict, even before there is overt talk in speech and text. This can then improve emotional learning that's gained from audio and transcripts.

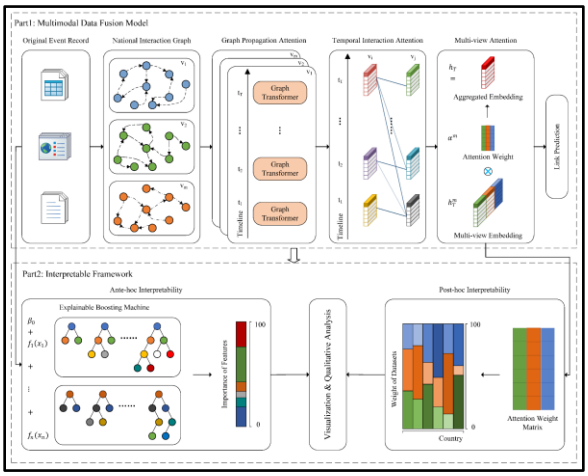

Fig 4. A Novel Multimodal Data Fusion Framework

In multimodal fusion, early fusion, late fusion, and intermediate fusion are used to combine these various signals [19]. Early fusion is performed prior to model training, and late fusion is performed by fusing the prediction performed by each of these models from different modalities. A pretrained deep neural architecture performs intermediate fusion by learning joint representations [20]. The choice of a fusion

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

strategy is crucial as different modalities can be useful to the prediction in different ways [21]. The concept of effective fusion gives rise to a better approach to escalation detection in early warning that can take into account the language meaning, voice, and context of interaction. The issues are with regard to synchronisation, missing information, interpretability, and complexity.

Research Gap

Past studies are mainly on customer attention by text, sound, and operational features, thus minimizing early detection of customer frustrations [22]. The multimodal combination of voice prosody, textual sentiment and interaction metadata in the context of banking call centre applications is limited [23]. The phenomenon of differences in contributions in relation to the task and the expectation of escalation, as such, a major lack of methodology and/or application-based research.

III. METHODOLOGY

Research Design

The experiment follows the quantitative research approach using an experimental research design to design and test a multimodal deep-learning model [24]. This research is aimed at predicting the customer's frustration for further escalation in the bank call centre. The IEMOCAP dataset used in this study was obtained from the Kaggle data repository. The project will test out unimodal and multimodal approach designs and explore differences between the two.

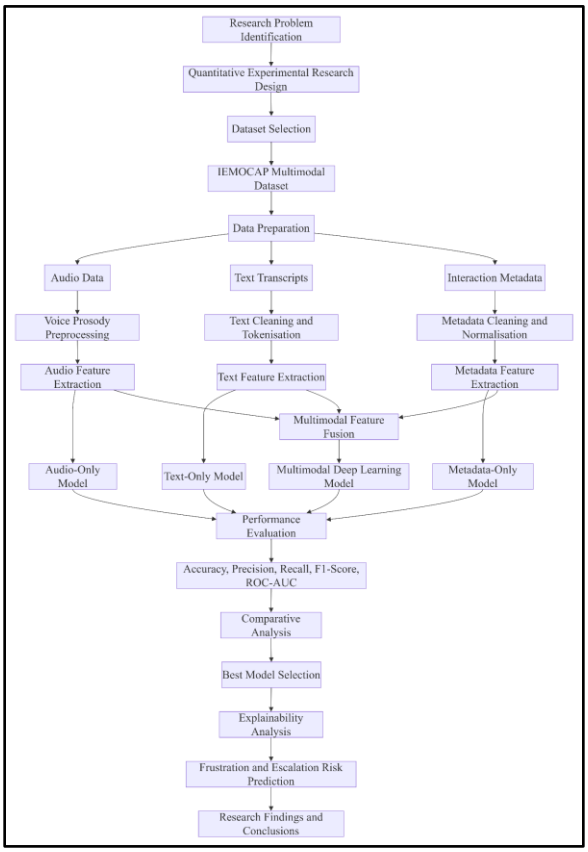

Fig 5. Research Flow Diagram

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dataset Selection and Description

This research enjoys the benefit of data from the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset from Kaggle. This is research in new directions in the development of a database capturing Emotional Embodied Human Interaction. This data consists of approximately 12 hours of audiovisual material collected in two types of dyadic interactions such as scripted and improvised. This has speech, video, facial motion captured, text transcription, and extracted multimodal features. Emotional expressions are coded for valence, activation and dominance scores as well as specific emotion like high frustration, medium frustration, and low frustration [25]. The audio data and the transcript can be employed in voice prosody and sentiment analysis.

Data Preparation and Pre-processing

The IEMOCAP dataset can be formed through the combination of audio files, textual transcription, emotion labels, and interactions. The prosodic feature extraction of pitch, energy, MFCCs, speaking rate, and pauses that the acoustic data will be processed, where necessary to remove noise, silence, segment and normalize the signals. The emotion-keeping symbols that don't matter will be filtered, the letters can be lowercased, and the texts will be tokenised [26]. All missing or ambiguous information will be identified and addressed. Manipulative metadata, like hold time, number of transfers, and call duration, will be normalized. Finally, each modality will be synchronized and will be divided into training, validation, and testing sets.

Feature Extraction

Feature extraction will transform raw multimodal data into numerical representations suitable for model training. Audio recordings will provide prosodic features such as pitch, energy, MFCCs, speaking rate, pause duration, and spectral characteristics. Text transcripts will be transformed into contextual embeddings and sentiment representations using a transformer-based NLP model [27]. Structured interaction metadata will include hold duration, transfer count, call duration, repeat contacts, and response delays. Each modality will therefore generate a separate feature vector before multimodal fusion.

TABLE I: Emotion-to-Frustration Mapping Strategy

IEMOCAP Emotion	Frustration Level	Justification
Angry	High frustration	Anger is intense emotional activation which is related to dissatisfaction.
Sad	Medium frustration	Sadness is a less active negative affect.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Neutral	Low frustration	Speech that does not contain explicit negative emotion is called neutral.
Happy	Low frustration	Emotional positive condition reflects contentment.

1. Audio Energy

$$E = \frac{1}{N}\sum_{n=1}^{N} x_n^2$$

E = average signal energy

x_n = amplitude of the n^{th} audio sample

N = total number of audio samples

2. Speaking Rate

$$SR = \frac{W}{T}$$

Where:

SR = speaking rate

W = total number of spoken words

T = speech duration in seconds

3. Pitch Representation

$$F_0 = \frac{1}{T_0}$$

F_0 = fundamental frequency or pitch

T_0 = duration of one vocal-fold vibration cycle

4. Sentiment Score

$$S = P_{pos} - P_{neg}$$

Where:

S = overall sentiment score

P_{pos} = probability of positive sentiment

P_{neg} = probability of negative sentiment

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Metadata Feature Vector

$$X_M = [H, TR, CD, RC, RD]$$

Where:

H = hold duration

TR = transfer count

CD = call duration

RC = repeat-contact frequency

RD = response delay

6. Final Multimodal Feature Vector

$$X_{Fusion} = X_A \oplus X_T \oplus X_M$$

Where:

X_A = extracted audio/prosodic features

X_T = textual sentiment and language features

X_M = interaction metadata features

$\oplus$ = concatenation operation

The **main formula to highlight in your methodology** should be:

Target Variable Development

The variables that will be used to step represent customer frustration and escalation risk (Low, Medium, High) and will be the target variable. These risk levels will be derived from the intensity of the emotion and properties of the emotion using IEMOCAP emotion labels. Target = (Y = 0,1,2); 0 = low frustration, 1 = medium frustration, 2 = high frustration or escalation at risk.

Model Development Architecture

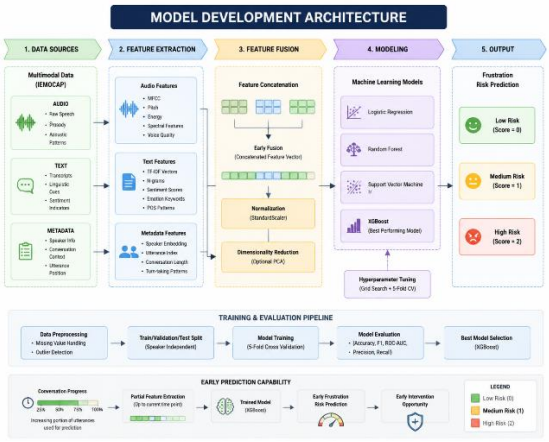

Fig 6. Model Development Architecture

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The architecture integrates audio, text, and metadata features through fusion, model training, and evaluation to predict frustration-risk levels.

IV. RESULTS AND FINDINGS

Frustration Class Distribution:

	Frustration Level	Count	Percentage
0	High	2144	28.84
1	Low	3581	48.18
2	Medium	1708	22.98

Fig 7. Frustration Class Distribution Table

7,433 observations were found in three categories of frustration. Low frustration (3,581) accounted for the biggest percentage of Marine casualties at 48.18%, with High frustration casualties at 28.84%. 1,708 or the 22.98% of cases is medium frustration. A moderate class imbalance can be seen in the distribution, justifying a stratified sampling design and a class-sensitive modelling evaluation measure to guarantee a reliable comparative modelling process for each of the three classes of predicted failures in a banking call.

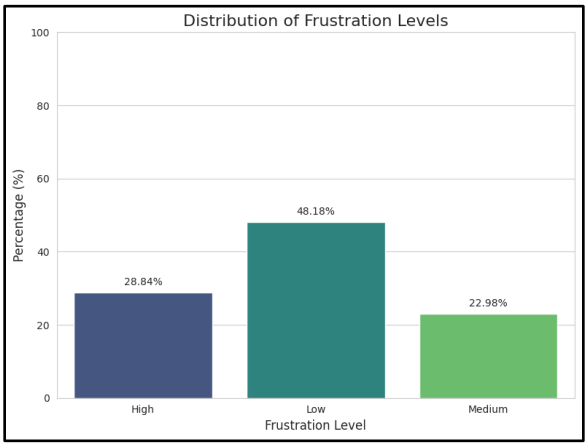

Fig 8. Distribution of Frustration Levels

As can be seen in the histogram, there are similar proportions of High (28.18%) and Medium (22.98%) annoyance, while Low (48.18%) is the more common option. Therefore, it is safe to say that almost the entirety of observations are rated as low level of risk while the other half of the observations is made up of medium or high frustration level. This distribution shows the suitability of using multimodal prediction in this distribution, but in practice other than accuracy measures should be used to check which were based on taking sensitivity of each class into account.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

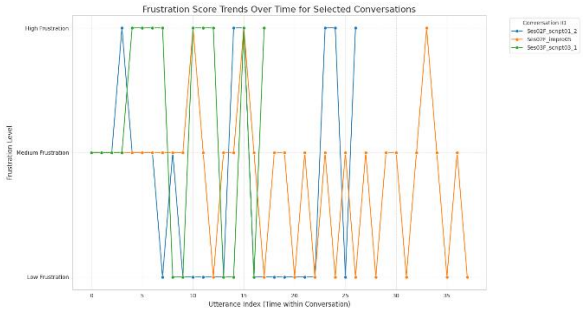

Fig 9: Frustration Score Trends Over Time for Selected Conversations

The figure illustrates how frustration levels change across utterances for selected conversations. Different trajectories show variations between low, medium, and high frustration states, highlighting emotional fluctuations throughout interactions and supporting temporal analysis of frustration-risk development.

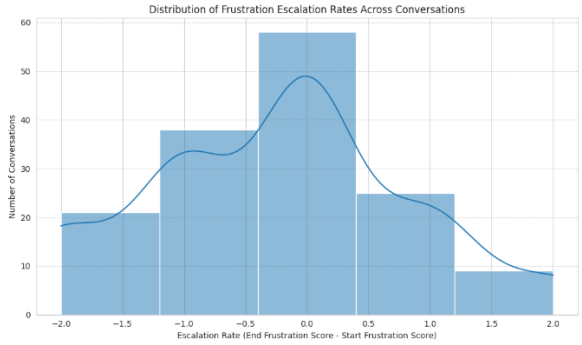

Fig 10: Distribution of Frustration Escalation Rates Across Conversations

The figure presents the distribution of escalation rates calculated from changes between initial and final frustration scores. The spread indicates that conversations experience different emotional progression patterns, with some showing increased frustration while others demonstrate reduced escalation risk.

*** Performance metrics calculated and stored in 'metrics_df'.

Model	Dataset	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	Audio Only	0.547	0.537	0.547	0.537	0.705
	Text Only	0.839	0.840	0.839	0.839	0.944
Random Forest	Audio Only	0.581	0.575	0.581	0.558	0.737
	Text Only	0.824	0.825	0.824	0.823	0.928
	Audio + Text	0.846	0.847	0.846	0.845	0.953
	Metadata Only	0.420	0.367	0.420	0.374	0.504
	Audio + Text + Metadata	0.841	0.843	0.841	0.840	0.952
Support Vector Machine	Audio Only	0.579	0.583	0.579	0.549	0.732
	Text Only	0.837	0.839	0.837	0.836	0.943
	Audio + Text	0.852	0.854	0.852	0.851	0.953
	Metadata Only	0.481	0.544	0.481	0.316	0.506
	Audio + Text + Metadata	0.852	0.854	0.852	0.851	0.953
XGBoost	Audio Only	0.585	0.582	0.585	0.571	0.737
	Text Only	0.824	0.825	0.824	0.823	0.944
	Audio + Text	0.853	0.853	0.853	0.852	0.960
	Metadata Only	0.424	0.374	0.424	0.381	0.508
	Audio + Text + Metadata	0.857	0.858	0.857	0.857	0.961

Fig 11. Comparative Model Performance Metrics

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The figure represents the comparative performance appraisal of various machine learning models with the aid of single and mixed modalities. The best performance of the Text Only features were obtained in Logistic Regression Accuracy of 83.9, ROC-AUC of 0.944, whereas the Audio + Text features worked better with the Random Forests algorithm (Accuracy: 84.6, ROC-AUC: 0.953). Audio + Text + Metadata Support Vector Machine gave a good result and the model had a 85.2% accuracy and a ROC-AUC of 0.953. XGBoost showed optimal results especially when multimodal fusion of Audio, Text, and Metadata was used, and the identification accuracy and the precision, recall and F1-score, and ROC-AUC were 85.7, 0.858, 0.857 and 0.961 respectively. All these findings affirm that the usefulness of combination of complementary modalities in enhancing frustration prediction is better than inputs from single sources.

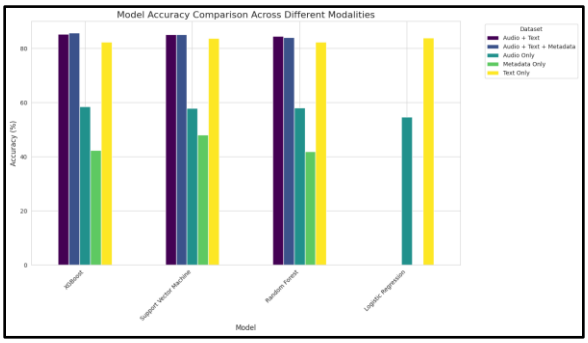

Fig 12. Model Accuracy Comparison Across Different Modalities

The results in terms of accuracy indicate the superiority of multimodal feature combination. With Audio, Text and Metadata data, XGBoost is able to achieve 85.7 percents and Support Vector Machine 85.2 percents. Text-only models are very successful overall at 82-84% and audio-only models are about average at 55-59%. If the performance of textual information in the banking calls is substantial, and the performance of metadata is highly unlikely to be significant, it is logical that not only are there lots of differences in performance, but that the gains in predictive accuracy that can be realised by fusing the two will also be seen to be substantial and consistent across the various data sets.

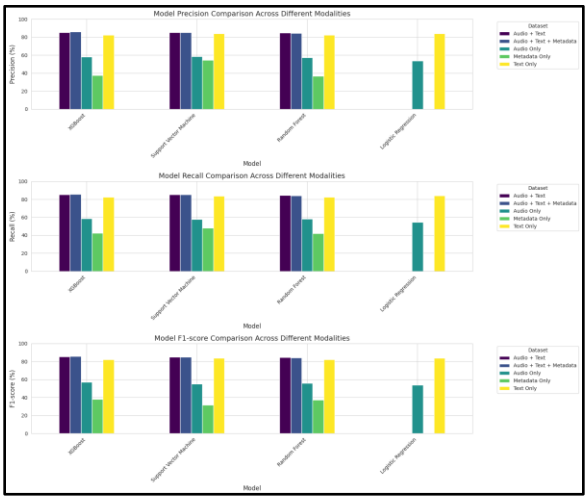

Fig 13. Precision, Recall, F1-Score and ROC-AUC Comparison

The same scores are found when looking at the metrics of precision, recall and F1-Score and comparing it to the audiovisual and text inputs, both which come in at over 84%. There is a mix of from about 40%

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

performance in audio-only models to quoting at a loss in metadata-only models. Although the panel has been distracted by the efforts required to scale up the art work, the overall differences between the classes are indeed comparable, as evidenced by the large area under the ROC curve (AUC) value for the time being (0.96). This illustrates that the classes are more sensitive, balanced and reliable for prediction in all cases.

```
Best Performing Model: XGBoost
Trained on Dataset: Audio + Text + Metadata
F1-score: 0.8571
Accuracy: 0.8574
ROC-AUC: 0.9610

Retrieved best model object: XGBoost (from Audio + Text + Metadata)
```

Fig 14. Best Performing Multimodal Model Summary

The accuracy of the best trained model is found with the model that is trained on the dataset Audio, Text, and Metadata with an accuracy of 85.74% and an f1 score of 0.8571 and ROC AUC score of 0.9610. The results demonstrate good classification and good discrimination of the different levels of frustration. The results confirm the benefits of multi-modal fusion, the combination of linguistic and acoustic features and user interaction improves the prediction accuracy; and the overall best model will always be easily the XGBoost.

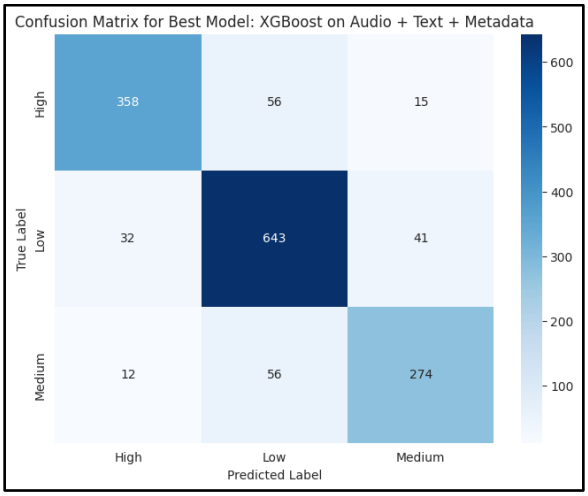

Fig 15. Confusion Matrix for XGBoost Multimodal Model

The confusion matrix indicates that 358 cases were identified as High, 643 as Low and 274 as Medium, which were correctly identified. Misclassification is not serious because in each case there are 56 High cases misclassified with a prediction of Low, and 56 Medium cases misclassified with a prediction of Low. The model appears fairly well suited to understand Low frustration, but isn't so much suited for Medium and High. If the diagonal dominance is strong however, the multiclass prediction performance is reasonably reliable.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

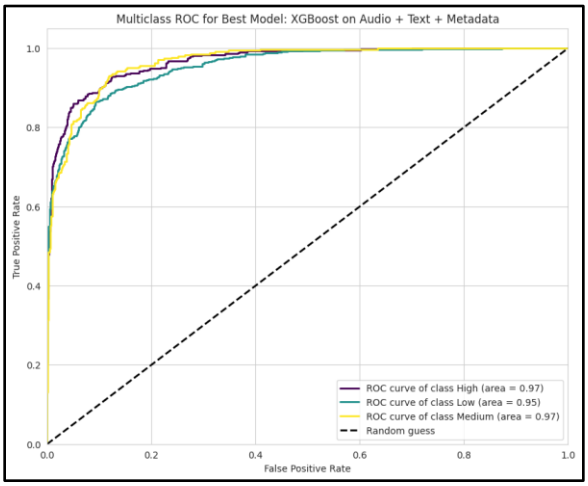

Fig 17. Multiclass ROC Curves for XGBoost Multimodal Model

The ROC curves have shown great class discrimination of XGBoost with multimodal data. With a score of about 0.97 AUC, High and Medium frustration achieve nearly the same; while Low frustration is nearly 0.95. In all three curves, the curve separation, which occurs before customer dissatisfaction escalates, is reliable (shown as above the dotted line which represents the random-separator), and power to detect risk of curve escalation is excellent.

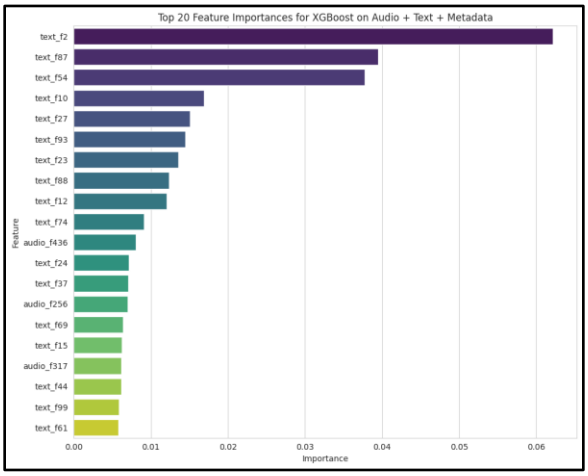

Fig 16. Top Twenty Feature Importances for XGBoost Multimodal Model

The results from the feature importance graph suggest that textual features are more important as these are dominant chapters in the multimodal classification process, which consists of text features (text_f2, text_f87, text_f54, etc.), as well as audio features (audio_f436, audio_f256, and audio_f317, etc.). The ranking highlights that textual feature contributes significantly to the prediction of their frustration, while acoustic feature contributes with discriminative information significantly increasing the performance of the multimodal systems in real applications.

Discussion

The findings showed that integration increase is better than the individual modalities when predicting the outcomes of frustration. The best performing model was the XGBoost based on the union of audi, text and

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

metadata features which performed at 85.7% accuracy on the test data, 85.7% F1-score on test data and had 0.961 ROC-AUC on test data. This is well suited for the objective of the research of Multimodal Takeoff and Escalation detection. The text features did the best job of giving the emotion, adding audio information'd completext the emotion [28]. The outcome of metadata was not as good revealing that metadata on their own are not sufficient but useful in the fusion. The results show that the proposed architecture is well suited for the task and emphasize the value of using linguistic, acoustic and interaction data together for enhanced banking call centre risk prediction.

Table 1: Utterance-Level Frustration Risk Mapping

Utterance Pattern	Emotion Indicator	Risk Level	Score	Meaning
Normal interaction	Neutral/positive emotion	Low	0	No strong frustration signal
Mild negative change	Sadness or negative tone	Medium	1	Possible dissatisfaction
Strong negative emotion	Anger or high emotion0061l intensity	High	2	High frustration-risk signal
Low → Medium	Increasing negative signals	Rising Risk	+1	Frustration is increasing
Medium → High	Stronger negative response	Escalation Risk	+1	Higher risk of escalation
High → Low	Reduced negative signals	Recovery	-1	Frustration is decreasing
Stable Low	Consistent normal emotion	Low Risk	0	Stable conversation
Stable Medium	Continuous negative emotion	Medium Risk	1	Ongoing dissatisfaction
Stable High	Continuous strong emotion	High Risk	2	Persistent frustration risk

V. CONCLUSION

Conclusion

The study demonstrates that using the combination of multimodal learning yields a higher accuracy in predicting the frustration of a customer and their conversion into an escalation issue vis-à-vis the use of one or the other type of data source. The model with high accuracy (85.7%), F1 score (85.7%) and ROC-AUC (0.961) was the computed model of XGBoost with audio and text metadata associated to the interactions performed in the evaluated systems. The most important component was the text features, and the audio and the metadata were complementary components of information to aid in the overall prediction. The findings of this study validate the proposed framework and indicate that if one combines the three mentioned elements, the ability to predict customer dissatisfaction at a much earlier stage can be achieved, which creates an opportunity to take appropriate steps in a banking call centre.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Future Direction

The framework should be tested in the future using operational data and actual banking calls made at the call centre rather than interaction variables. Emerging models such as multimodal transformer models, wav2vec-based speech encoder, attention-based fusion network represent other potential models in deep learning architectures to better improve prediction accuracy [29]. Should there be any way that frustration can be detected 'on the go' while talking, real-time deployment should also be taken into account. Ensure Explanation, including SHAP values, is integrated and utilised to improve transparency, identify critical escalation flags [30]. Need to be more focused on multilingual calls, accents/voice, background noise, privacy, fairness, data security, model generalisation and evaluation need to be done on different banks and customers.

REFERENCES

[1] Anand, S., Miglani, S. and Anand, R., 2025. Multimodal AI for enhanced emotional intelligence in the cloud computing era: A comprehensive approach. In *Establishing AI-Specific Cloud Computing Infrastructure* (pp. 193-218). IGI Global Scientific Publishing.

[2] Bromuri, S., Henkel, A.P., Iren, D. and Urovi, V., 2021. Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, *32*(4), pp.581-611.

[3] Shah, S., Ghomeshi, H., Vakaj, E., Cooper, E. and Fouad, S., 2023. A review of natural language processing in contact centre automation. *Pattern Analysis and Applications*, *26*(3), pp.823-846.

[4] Serrano, S., Serghini, O., Esposito, G., Carbone, S., Mento, C., Floris, A., Porcu, S. and Atzori, L., 2025. Review and comparative analysis of databases for speech emotion recognition. *Data*, *10*(10), p.164.

[5] Płaza, M., Kazała, R., Koruba, Z., Kozłowski, M., Lucińska, M., Sitek, K. and Spyrka, J., 2022. Emotion recognition method for call/contact centre systems. *Applied Sciences*, *12*(21), p.10951.

[6] Lackovic, N., Montacié, C., Lalande, G. and Caraty, M.J., 2022. Prediction of user request and complaint in spoken customer-agent conversations. *arXiv preprint arXiv:2208.10249*.

[7] Bahad, P., Chauhan, D. and Bharawa, D., 2025, May. Sentiment analysis of helpdesk calls: enhancing customer support through natural language processing. In *International conference on recent advancements and modernisations in sustainable intelligent technologies and applications (RAMSITA 2025)* (pp. 385-398). Atlantis Press.

[8] Lackovic, N., Montacié, C., Lequilliec, C. and Caraty, M.J., 2023, June. Healthcall corpus and transformer embeddings from healthcare customer-agent conversations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[9] Ahmed, A., Sivarajah, U., Irani, Z., Mahroof, K. and Charles, V., 2024. Data-driven subjective performance evaluation: An attentive deep neural networks model based on a call centre case. *Annals of Operations Research*, *333*(2), pp.939-970.

[10] Malk, N.A. and Diwan, S.A., 2025. Artificial intelligence in speech emotion detection: trends, challenges, and future directions. *International Journal of Ethical AI Application*, *1*(2), pp.19-29.

[11] Rajkumar, S.C., Yuvasini, D., Selvarajan, S. and Sivakumar, N.R., 2025. A zero-shot LLM framework for multimodal grievance classification, urgency scoring, and abuse detection in civic feedback systems. *Scientific Reports*, *16*(1), p.2299.

[12] FEROLDI, F., 2018. Voice analysis and business applications. A state-of-the-art.

[13] Budžys, A., Medvedev, V. and Kurasova, O., 2021. User behaviour analysis based on similarity measures to detect anomalies. In *DAMSS: 12th conference on data analysis methods for software systems, Druskininkai, Lithuania, December 2–4, 2021.*. Vilnius University Press.

[14] Anand, T., Panwar, S., Sharma, S.K., Rastogi, R. and Gupta, M., 2024. Voice and Speech Recognition Application in Emotion Detection: A Utility for Future Trends. In *Federated Learning and AI for Healthcare 5.0* (pp. 242-268). IGI Global Scientific Publishing.

[15] Flanagan, O., Chan, A., Roop, P. and Sundram, F., 2021. Using acoustic speech patterns from smartphones to investigate mood disorders: scoping review. *JMIR mHealth and uHealth*, *9*(9), p.e24352.

[16] Burgess, J., Carlon, D., Doyuran, E.B., Huang, H., Leckie, C., Matich, P., McCosker, A., Richardson, M., Angus, D., Goldenfein, J. and Kelly, M., 2025. Voice AI and authenticity: current issues and emerging challenges.

[17] Hwang, S., Liu, X. and Srinivasan, K., 2021. Voice analytics of online influencers. *Available at SSRN 3773825*.

[18] Kabir, M.M., Mridha, M.F., Shin, J., Jahan, I. and Ohi, A.Q., 2021. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *Ieee Access*, *9*, pp.79236-79263.

[19] Bhardwaj, V., Thakur, D., Gera, T. and Sharma, V., 2023. Enhanced dialectal speech recognition in Punjabi using pitch-based acoustic modeling. *Ingenierie des Systemes d'Information*, *28*(6), p.1557.

[20] Smyth, H.K., Nyhan, J. and Flinn, A., 2023. Exploring the possibilities of Thomson's fourth paradigm transformation—The case for a multimodal approach to digital oral history?. *Digital Scholarship in the Humanities*, *38*(2), pp.720-736.

[21] Leix Palumbo, D. and Prey, R., 2024. Sounding out voice biometrics: Comparing and contrasting how the state and the private sector determine identity through voice. *Big Data & Society*, *11*(4), p.20539517241297889.

[22] Leal, S.S., Ntalampiras, S. and Sassi, R., 2024. Speech-based depression assessment: A comprehensive survey. *IEEE Transactions on Affective Computing*, *16*(3), pp.1318-1333.

[23] Azeemi, A.H., Qazi, I.A. and Raza, A.A., 2025. A Survey on Data Selection for Efficient Speech Processing. *IEEE Access*, *13*, pp.109234-109253.

[24] Perumandla, R., Bae, Y.H., Izaguirre, D., Hwang, E., Murphy, A., Hsu, L.J., Sabanovic, S. and Bennett, C.C., 2025. Developing conversational speech systems for robots to detect speech biomarkers of cognition in people living with dementia. *arXiv preprint arXiv:2502.10896*.

[25] El-Masri, A., Riedl, M.J. and Woolley, S., 2022. Audio misinformation on WhatsApp: A case study from Lebanon. *Harvard Kennedy School Misinformation Review*, *3*(4), pp.1-13.

[26] Mohd, T.K., Nguyen, N. and Javaid, A.Y., 2022. Multi-modal data fusion in enhancing human-machine interaction for robotic applications: a survey. *arXiv preprint arXiv:2202.07732*.

[27] Mbelekani, N.Y. and Bengler, K., 2025. Irisk: towards responsible AI-powered automated driving by assessing crash risk and prevention. *Electronics*, *14*(12), p.2433.

[28] Noori, S., 2025. Predictive Modeling of Emotional and Behavioral Patterns in Cluster B Users Using AI Web Viewers on Social Media. *Diqiu Kexue*, *49*(2).

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[29] Dritsas, E., Trigka, M., Troussas, C. and Mylonas, P., 2025. Multimodal interaction, interfaces, and communication: a survey. *Multimodal Technologies and Interaction*, *9*(1), p.6.

[30] Chen, J., Seng, K.P., Smith, J. and Ang, L.M., 2024. Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications. *IEEE Access*, *12*, pp.88779-88818.