DG$^{\text{VoiC}}$: Speaker Clustering for Fraud Investigation under Real Call-Centre Conditions

Muhammad Shakeel Akram
Aston University
Birmingham,UK
m.akram5@aston.ac.uk

Amal Htait
Aston University
Birmingham,UK
a.htait@aston.ac.uk

Abdul Hamid Sadka
Aston University
Birmingham,UK
a.sadka@aston.ac.uk

Emma Meisingseth
Domestic & General
Wimbledon, UK
Emma.Meisingseth@domesticandgeneral.com

Karishma Jaitly
Domestic & General
Wimbledon, UK
Karishma.Jaitly@domesticandgeneral.com

***Abstract*—Insurance fraud remains costly and operationally difficult, particularly in call-centre workflows where many customer interactions begin at FNOL. While recent fraud detection methods mainly rely on structured data, text, or images, repeated speaker identity across calls remains underused as an investigative signal. This paper presents DG$^{\text{VoiC}}$, a voice clustering framework for customer verification and cross-profile speaker linking on anonymised real call-centre audio. The approach combines sensitive information-aligned anonymisation, speech-focused preprocessing, sliding-window speaker embedding extraction, and cosine similarity based clustering to identify repeated speakers under real telephony conditions. The method was evaluated on 121 recordings, with a curated reference subset of 56 samples in 22 human-agreed speaker clusters used for validation. The best configuration achieved 96% AMI, 95% ARI, 98% completeness, 100% homogeneity, and 99% V-measure. These results show that speaker clustering can provide a strong additional signal for fraud investigation by helping analysts verify speaker consistency and surface repeated voices across customers.**

***Index Terms*—Fraud, speaker clustering, call-centre**

I. Introduction

Insurance fraud remains a costly and persistent problem for insurers, investigators, and genuine customers [1]–[5]. The total estimated cost of fraud in 2023 to 2024 is £14.4 billion, comprising £9.2 billion affecting individuals and £5.2 billion affecting businesses [6]. In the UK, insurers detected more than £1.16 billion of fraudulent general insurance claims in 2024, with a further 684,800 fraudulent applications prevented at policy inception [7].

This creates strong demand for earlier and more reliable fraud screening, particularly in customer-facing channels. One such channel is the call-centre, especially at FNOL. At this stage, much of the interaction takes place over the phone, which makes identity assessment difficult in real time. Frontline agents must balance customer service with basic verification, while investigators later need to review large volumes of calls to determine whether the same speaker appears across multiple customer profiles. In investigative practice, this can leave cross-case voice reuse underused.

Most existing insurance fraud systems focus on structured data, text, images, or combinations of these modalities. Recent multimodal approaches have shown gains over unimodal baselines, but they do not model repeated speaker identity across calls [8]–[17]. At the same time, research on fraudulent phone calls has largely focused on transcript content rather than speaker reuse, while real operational call-centre audio remains difficult to study because of privacy and biometric data constraints [18]. For example, CallCenterEN releases large-scale anonymised call-centre transcripts, but not the underlying audio, specifically because of biometric privacy concerns [19]. This leaves a gap between recent advances in speaker representation learning and the practical needs of fraud investigation in insurance call workflows.

To address this gap, we introduce DG$^{\text{VoiC}}$, a voice clustering framework for anonymised real call-centre audio. The method is designed to support two related tasks: verifying whether calls linked to a customer are spoken by a consistent speaker, and identifying repeated speakers that appear across different customer profiles. Rather than acting as a standalone fraud decision system, DG$^{\text{VoiC}}$ is intended to support analyst review by surfacing voice-based links under real call-centre conditions.

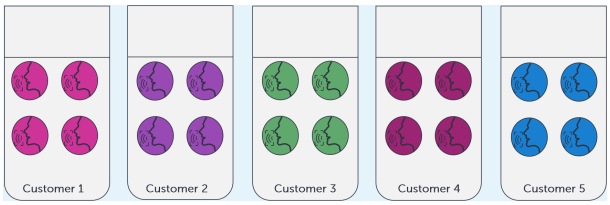
The main contributions of this paper are as follows:

- We present DG$^{\text{VoiC}}$, a practical voice clustering approach for customer verification and cross-profile speaker linking on anonymised real insurance call-centre data.
- We propose a pipeline for long and variable telephony calls, combining word-level aligned anonymisation, speech-focused preprocessing, sliding-window embedding extraction, and similarity-based clustering.
- We evaluate the approach on a real-world dataset and show that ECAPA-based speaker embeddings can achieve strong agreement with human-reviewed speaker clusters.

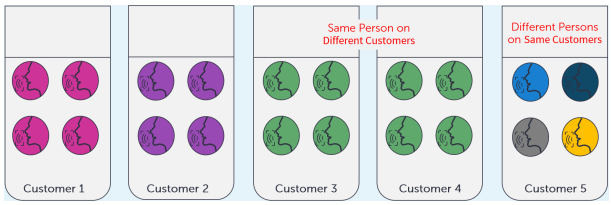
The paper is organised as follows. Section V reviews related work. Section II describes the proposed DG$^{\text{VoiC}}$. Section III outlines the experimental setup, and Section IV presents the results and discussion. Finally, Section VI concludes the paper.

II. Proposed DG$^{\text{VoiC}}$ Voice Clustering Model

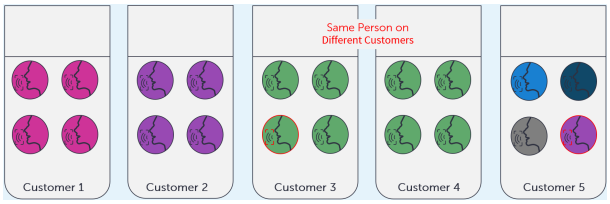
The proposed DG$^{\text{VoiC}}$ framework uses speaker embeddings to detect voice reuse across claims and customer profiles. The aim is to support fraud investigation under real call-centre conditions while keeping false positives low. This is important at the claims stage, where incorrect flagging can affect both



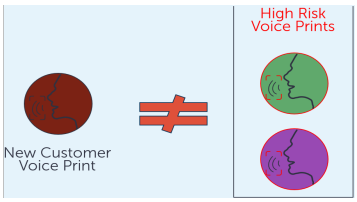
(a) Non-fraudulent scenario



(b) Fraudulent scenario



(c) High risk fraudulent voice prints



(d) Not selling to high-risk voice prints

Fig. 1. Illustrating the scenarios for customer verification and fraud identification.

customer experience and operational decision-making. In practice, the system is designed to help investigators verify whether the claimed customer is the actual speaker and to identify repeated use of the same voice across different profiles.

Fig. 1 illustrates the main investigation scenarios. In Fig. 1a, each customer profile is associated with a distinct speaker, which represents the expected non-fraud case. In Fig. 1b, two types of risk can appear. First, the same speaker may be linked to more than one customer profile, shown by the repeated green voice prints. Second, a customer profile may contain calls from different speakers over time. In this case, some variation is expected across genuine calls, but a clearly separate voice pattern, shown in yellow, may indicate that another person is speaking on behalf of the customer. This creates two investigation needs: verifying whether the customer is the speaker in a given call, and linking suspicious voices that reappear across different profiles.

Fig. 1c extends this idea to recurring high-risk voices. For example, if the voice shown in purple appears under one customer profile and later reappears under another, that voice can be treated as a repeated risk indicator. When such patterns occur with strong similarity across multiple profiles, they may point to organised misuse rather than isolated inconsistency. This supports both claim review and earlier intervention for new or existing customers.

A. Dataset Anonymisation

As the study uses real call-centre data, anonymisation is applied before speaker analysis. PII and other sensitive attributes are detected using NER, Regex, and references from WhisperX word-level time stamps. The corresponding audio regions are then muted using `librosa` and `soundfile`.

B. Speaker Clustering

The customer voice provides a useful signal for fraud investigation. Since operational deployment may require continuous call handling, audio is processed using an overlapping sliding-window strategy. The overlap reduces the risk of missing short but informative speech regions, while very short segments are excluded to avoid unstable embeddings.

Before segmentation, long non-speech regions are removed using Resemblyzer's `preprocess_wav`. This reduces the

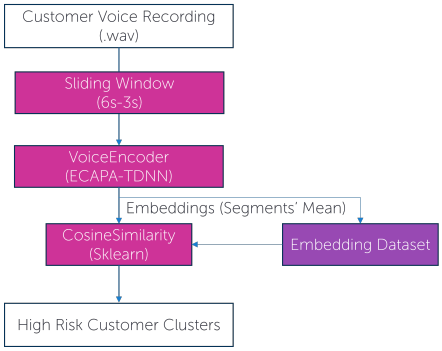


Fig. 2. DG$^{\text{VoiC}}$ model architecture.

effect of silence and low-information regions and keeps the speaker representation focused on active speech. Each valid segment is then encoded using ECAPA-TDNN to produce a fixed-dimensional speaker embedding. Segment-level embeddings for the same customer interaction are aggregated by mean pooling to form a final voice representation.

Cosine similarity is then used to compare embeddings across claims and customer profiles. This makes it possible to detect repeated speakers even when the associated customer details are different. In fraud review, such patterns may indicate linked claims, organised misuse, or a single speaker appearing under multiple identities for human intervention and investigation. This detailed approach is illustrated in Fig. 2

To support investigator triage, speaker embeddings are further assessed using three conditions:

1) cosine similarity greater than 0.718 between candidate voice representations,
2) membership in a cluster containing more than 4 distinct customer profiles, and
3) cluster consistency across customer's recordings.

The cluster-size condition is motivated by the UK Office for National Statistics report (2024), that shows over 93.6% of households have 1-4 members [20]. Therefore, legitimate voice sharing is more likely within small family units than across larger unrelated groups. If these conditions are met, the case ($\geq$9 profiles in a cluster and $>$92% similarity) is assigned a higher voice-risk score to prioritise analyst review.

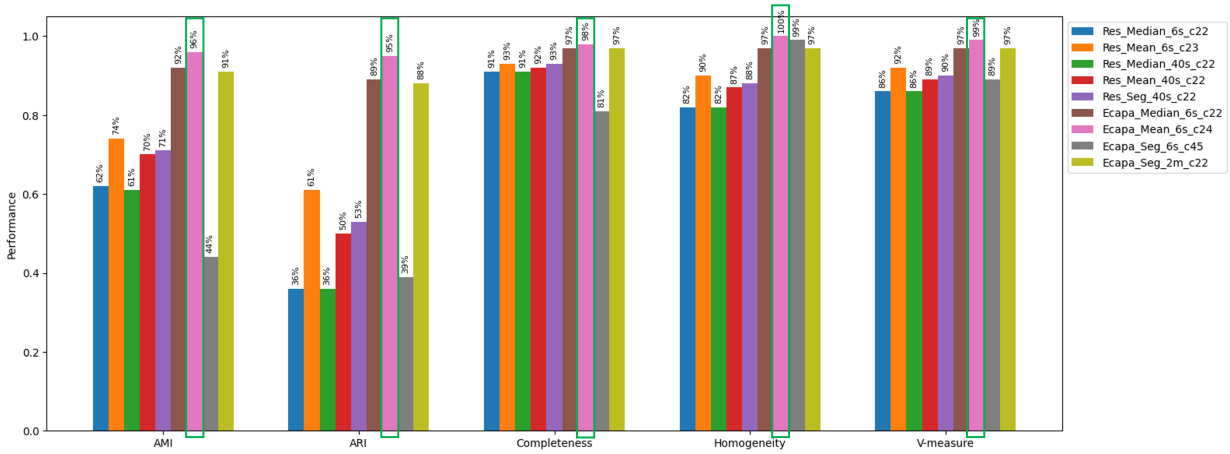


Fig. 3. Experimentation for identifying the best combination for voice clustering model.

Lower similarity, smaller cluster size, or inconsistent clustering leads to a proportionally lower score. The score is intended to prioritise analyst review rather than serve as a standalone fraud decision.

III. Experimental Setup

The study used 121 real call recordings collected from multiple customer profiles. To ensure that speaker consistency could be assessed across repeated interactions, the sample was restricted to customers with at least four available recordings. This provided enough within-customer variation to evaluate where the proposed clustering approach could separate natural voice variability from true cross-profile voice reuse.

To establish a cleaner evaluation subset, the recordings were independently blind reviewed by two fraud-domain practitioners from Domestic and General. Each reviewer listened to the calls and grouped them into speaker clusters based on perceived voice similarity. Only the samples for which both reviewers with high confidence produced the same clustering outcome were retained for validation. This resulted in a subset of 56 samples, organised into 22 agreed clusters, which was used as the clean reference set for modelling and assessment. The selected samples exhibit recording durations ranging from 2min 45s to 34min 38s, with an average duration of 10min 25s.

WhisperX [21] was used to generate word-level timestamps along with a hybrid masking pipeline that combined RoBERTa-based NER [22] with rule-based Regex to detect personally identifiable and sensitive information. The corresponding audio spans were muted using `librosa` and `soundfile`, producing an anonymised version of each call for downstream analysis. For speaker modelling, ECAPA-TDNN [23] was used to extract embeddings. Cosine similarity using scikit-learn [24], was then used to compare embeddings and identify potential voice reuse across customers.

Latency evaluation was performed using `%%timeit` with 20 repeated runs. Experiments were executed on Databricks custom compute (runtime 17.3 LTS ML) with 32 CPU cores, 128 GB memory, and a single 64 GB GPU for both driver and worker nodes.

IV. Results and Discussion

This section reports the clustering performance of the proposed DG$^{\text{VoiC}}$ pipeline on the curated evaluation subset. The aim was to identify the configuration that best separates natural within-speaker variation from true cross-profile voice reuse under real call-centre conditions.

Fig. 3 summarises the best result from each experiment. The evaluation compared two speaker embedding models, Resemblyzer VoiceEncoder (Res) and ECAPA-TDNN (Ecapa), under several segmentation settings. These included overlapping sliding windows of 6 s with 3 s hop size and 40 s with 20 s hop size, as well as single-segment processing using 6 s, 40 s, and 2 min portion. Segment-level embeddings were then aggregated using either mean or median pooling. Clustering was primarily based on cosine similarity with the clustering threshold varied between 0 and 1 in increments of 0.001. DBSCAN and FAISS were also explored, but both performed worse on this real dataset and were therefore excluded from the final model comparison.

Across all tested settings, the results (Fig. 3) were generally strong, which suggests that speaker embeddings remain informative even under noisy and operationally variable call-centre conditions. The best overall configuration used Resemblyzer preprocessing for silence removal, ECAPA-TDNN for embedding extraction, a 6 s sliding window, mean pooling, and a cosine similarity threshold of 0.718 (selected on the experimental sweep as the best development threshold). This setup achieved 96% adjusted mutual information (AMI), 95% adjusted Rand index (ARI), 98% completeness, 100% homogeneity, and 99% V-measure. Leading to 95% accuracy and 0.96 F1 score.

The best-performing configuration produced 24 clusters, very close to 22 clusters in the labels. A closer review showed that the additional clusters were not random errors. In one case, a customer with four recordings had two calls that sounded noticeably different from the other two on re-listening, which led to a split cluster. In another case, one recording contained sustained loud background noise in the later part of the call, which shifted its representation away from the remaining calls from the same customer. This highlights residual acoustic variability as an important target for future work.

For post-hoc operational scoring, cosine similarity was also mapped to a bounded probability score using a sigmoid function to support ranking and triage. Fig. 4 shows example cluster formations from the best configuration. Overall, the

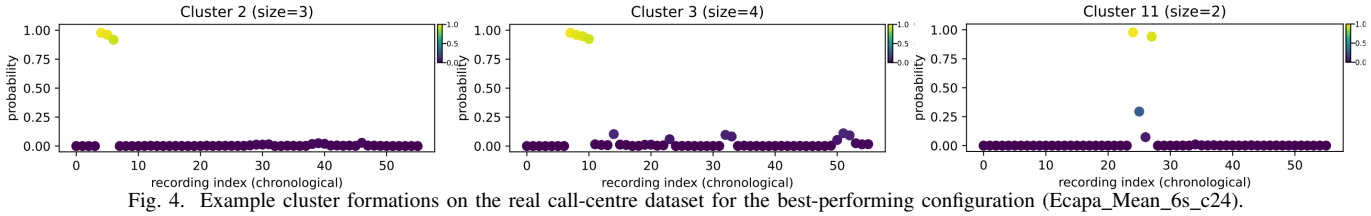


Fig. 4. Example cluster formations on the real call-centre dataset for the best-performing configuration (Ecapa_Mean_6s_c24).

TABLE I
CONTEXTUAL COMPARISON WITH TECHNICALLY CLOSEST APPROACHES.

Ref	Task Focus	Performance	Comments
DG$^{\text{VoiC}}$	Customer-level speaker clustering across calls for fraud investigation	AMI:96%, ARI:95%, C:98%, H:100%, V:99%, A:95%, F1: 0.96, L: 10.08s, EER: 3.85%, FAR: 0.5%, FRR:9.62%	Low–medium complexity; Resemblyzer-ECAPA-Cosine based pipeline designed using real data for real call-centre workflows
[25]	Caller identification and authentication	A: 89%, F1:0.88	MFCC-SVN_linear based supervised models; medium complexity; 6-7s clips
[26]	Multimodal voice phishing and synthetic voice detection	[Voice-only] A: 73%, *L:32.4s	CNN-BiLSTM with MFCC features; medium–high complexity; 20–170s synthetic audios
[27]	Speaker verification	EER:1.7%, FAR:1.7%, FRR:1.7%, DCF:0.1%	ECAPA-based solution on non-english Synthetic data
[28]	Speaker verification and clone detection	Random Forest EER:4.4%, Neural Network A:96%	Dual-model pipeline; higher complexity; ASVspoof data
[29]	Online speaker recognition and clustering	EER: 4.98%, DER: 3.32%	sph-PLDA-based probabilistic back-end; medium complexity;
[30]	Speaker diarization (telephone speech: CALLHOME)	[2-4 Speakers] DER: 7.2%, A: 68.8%	Oracle VAD, ECAPA-TDNN, Mel-filterbank features, EEND-EDA based Neural diarization; higher modelling and deployment complexity

DCF: Detection Cost Function, AHC: Agglomerative Hierarchical Clustering. A: Accuracy, C: Completencess, H:Homogeneity, V:V-measure, L: Latency/recording *Compute: Intel i9-9900K CPU-RTX 3090 GPU-24 GB memory.

results indicate that the proposed configuration can group repeated speakers with high consistency and can support fraud investigation by surfacing cross-profile voice links for analyst review.

In this setup, embedding extraction required an average of 10s/recording (9min 26s for the full dataset), while clustering and customer linking took 84ms/recording (4.69s in total). This results in an average end-to-end processing time of 10.08s/recording.

As most technically related studies report verification- or diarization-oriented metrics such as Equal Error Rate (EER). Although the primary task addressed here is customer-level speaker clustering across calls hence clustering agreement metrics remain the primary evaluation criterion. We additionally derive an auxiliary verification-style analysis for contextual comparison only. This yields 3.85% EER, with a false acceptance rate (FAR) of 0.50% and 9.62% false rejection rate (FRR) at 0.718 clustering threshold.

V. RELATED WORK

Existing work on fraudulent phone calls is closest in operational setting but primarily focuses on transcript content, dialogue semantics, or scam classification rather than persistent speaker identity. In parallel, most insurance fraud research relies on structured or multimodal claim data, while speech-focused studies typically address speaker diarization or verification on benchmark corpora rather than real call-centre fraud workflows [16]–[18], [29], [31]. Recent work on scam automation further shows that modern speech synthesis and recognition technologies can be misused to scale impersonation attacks, reinforcing the need for voice-based controls in telephony environments [32].

Within call-centre contexts, the closest application-level studies address customer identification and caller authentication using voice biometrics. These approaches typically rely on MFCC-based features or supervised speaker identification models and aim to verify known callers rather than to cluster repeated speakers across customer profiles for fraud investigation [25], [26], [33]. As a result, their evaluation protocols, reported metrics, and deployment objectives differ substantially from customer-level speaker clustering.

To the best of our knowledge, no prior work has reported the same end task addressed by DG$^{\text{VoiC}}$, namely customer-level voice clustering across calls under real insurance call-centre conditions to support speaker consistency checking and cross-profile linking. Consequently, direct numerical comparison with the state of the art is not methodologically appropriate, as published systems are evaluated on different tasks, datasets, and metrics, such as DER or EER, rather than clustering agreement measures.

For contextual comparison, Table I summarises the closest technical paradigms in terms of task alignment, reported performance, and relative computational complexity. While end-to-end diarization and verification systems often achieve strong benchmark results, they typically incur higher modelling and deployment complexity. In contrast, DG$^{\text{VoiC}}$ adopts a simpler embedding-based pipeline designed to operate efficiently under real-world call-centre constraints and to support analyst-led fraud investigation rather than automated decision-making.

VI. CONCLUSION

This paper presented DG$^{\text{VoiC}}$ for customer verification and cross-profile speaker linking in insurance fraud investigation. Experiments on anonymised real call-centre audio showed that speaker embeddings can consistently group repeated speakers

under practical telephony conditions, providing a useful signal for analyst-led fraud review. Future work will focus on improving robustness to challenging call conditions, including background noise, within-speaker variability, and customer-side speaker diversity. Additional extensions will explore complementary audio cues, such as prosodic and behavioural features, to enhance voice-based risk scoring alongside speaker clustering, with validation guided by expert assessment against real-world fraud outcomes.

REFERENCES

[1] Home Office and The Rt Hon Lord Hanson of Flint, "Major new crackdown on insurance fraud," https://www.gov.uk/government/news/major-new-crackdown-on-insurance-fraud, October 2024, accessed: 2025-08-01.

[2] Insurance Fraud Bureau, "Ifb's 2024 annual report has been published," https://www.insurancefraudbureau.org/media-centre/ifb-news/2025/ifb-s-2024-annual-report-has-been-published, 2025, accessed: 2025-08-01.

[3] National Association of Insurance Commissioners, "Insurance topics — insurance fraud," https://content.naic.org/insurance-topics/insurance-fraud, 2024, accessed: 2025-08-01.

[4] F. Aslam, A. I. Hunjra, Z. Ftiti, W. Louhichi, and T. Shams, "Insurance fraud detection: Evidence from artificial intelligence and machine learning," *Technological Forecasting and Social Change*, 2022.

[5] A. Ali, S. A. Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: A systematic literature review," *IEEE Access*, 2022.

[6] Home Office, "Economic and social cost of fraud 2023 to 2024," https://www.gov.uk/government/publications/economic-and-social-cost-of-fraud-2023-to-2024, 2026, accessed: 2026-04-01.

[7] Association of British Insurers, "Fraudulent insurance claims continue to top £1 billion," https://www.abi.org.uk/news/news-articles/2025/11/fraudulent-insurance-claims-continue-to-top-1-billion/, 2025, accessed: 2026-04-01.

[8] C. Piehl, "Classification of transcribed voice recordings: Determining the claim type of recordings submitted by swedish insurance clients," Master's thesis, KTH Royal Institute of Technology, 2021.

[9] R. Bäcklund and H. Öhman, "Detection of insurance fraud using nlp and ml: A study on three different nlp-techniques for text classification," Master's thesis, Lund University, 2023.

[10] J.-W. Chang, N. Yen, and J. C. Hung, "Design of a nlp-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance," *Journal of Ambient Intelligence and Humanized Computing*, 2023.

[11] A. Dimri, S. Yerramilli, P. Lee, S. Afra, and A. Jakubowski, "Enhancing claims handling processes with insurance based language models," in *Proceedings of the AAAI Workshop on AI in Insurance*, 2024.

[12] C. M. Gangani, "Ai in insurance: Enhancing fraud detection and risk assessment," *International Journal of Advanced Computer Science and Applications*, 2023.

[13] V. K. Tarra, "Ai in fraud detection: Leveraging machine learning to combat insurance fraud," *International Journal of Innovative Technology and Exploring Engineering*, 2024.

[14] R. A. Perumal, "Innovative applications of ai and machine learning in fraud detection for insurance claims," *Journal of Risk and Financial Management*, 2023.

[15] D. Banulescu-Radu and M. Yankol-Schalck, "Practical guideline to efficiently detect insurance fraud in the era of machine learning: A household insurance case," *Journal of Risk Finance*, 2023.

[16] J. Yang, K. Chen, K. Ding, C. Na, and M. Wang, "Auto insurance fraud detection with multimodal learning," *Data Intelligence*, vol. 5, no. 2, pp. 388–412, 2023.

[17] A. Asgarian, R. Saha, D. Jakubovitz, and J. Peyre, "Autofraudnet: A multimodal network to detect fraud in the auto insurance industry," *arXiv preprint arXiv:2301.07526*, 2023.

[18] Z. Shen, K. Wang, Y. Zhang, G. Ngai, and E. Y. Fu, "Combating phone scams with llm-based detection: Where do we stand?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 28, 2025, pp. 29 487–29 489.

[19] H. Dao, G. Chawla, R. Banda, and C. DeLeeuw, "Real-world en call center transcripts dataset with pii redaction," *arXiv preprint arXiv:2507.02958*, 2025.

[20] Office for National Statistics, "Families and households," 2025, accessed: 2025-08-29. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds

[21] M. Bain, J. Huh, T. Han, and A. Zisserman, "Whisperx: Time-accurate speech transcription of long-form audio," *INTERSPEECH 2023*, 2023.

[22] J.-B. Polle, "Jean-baptiste/roberta-large-ner-english," https://huggingface.co/Jean-Baptiste/roberta-large-ner-english, 2021, fine-tuned RoBERTa model for Named Entity Recognition on CoNLL-2003 dataset.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] R. Sonwane, M. Patil, P. Chauhan, A. Jain, B. Nandwalkar, A. Awate, and M. Shahade, "Trustcaller-voice-based fraud prevention system," in *2024 4th International Conference on Intelligent Technologies (CONIT)*. IEEE, 2024, pp. 1–6.

[26] J. Kim, S. Gu, Y. Kim, S. Lee, and C. Kang, "A multimodal voice phishing detection system integrating text and audio analysis," *Applied Sciences*, vol. 15, no. 20, p. 11170, 2025.

[27] V. Brydinskyi, Y. Khoma, D. Sabodashko, M. Podpora, V. Khoma, A. Konovalov, and M. Kostiak, "Comparison of modern deep learning models for speaker verification," *Applied Sciences*, vol. 14, no. 4, p. 1329, 2024.

[28] A. Nandal and M. Dua, "A hybrid approach to secure automatic speaker verification: integrating clone detection and speaker identification," *International Journal of Speech Technology*, vol. 28, pp. 411–429, 2025.

[29] A. Sholokhov, N. Kuzmin, K. A. Lee, and E. S. Chng, "Probabilistic back-ends for online speaker recognition and clustering," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[30] J. I. Alvarez-Trejos, A. Lozano-Diez, and D. Ramos, "Feature integration strategies for neural speaker diarization in conversational telephone speech," *Applied Sciences*, vol. 15, no. 9, p. 4842, 2025.

[31] L. Serafini, S. Cornell, G. Morrone, E. Zovato, A. Brutti, and S. Squartini, "An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings," *Computer speech & language*, vol. 82, p. 101534, 2023.

[32] G. Gressel, R. Pankajakshan, and Y. Mirsky, "Discussion paper: Exploiting llms for scam automation: A looming threat," in *Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes (WDC '24)*, 2024.

[33] A. Khan and P. Aithal, "Identification of customer through voice biometric system in call centres," *Int. J. Intell. Syst. Appl*, vol. 16, no. 5, pp. 68–78, 2024.