Truth on Trial: Detecting Deception in Judiciary Systems with Graph Attention Networks

Shyam Sathvik
IIT Jodhpur
b22ee036@iitj.ac.in

Neermita Bhattacharya
IIT Jodhpur
b22cs092@iitj.ac.in

Abstract

Lie detection is a complex task that involves analyzing speech patterns and acoustic features to distinguish between truthful and deceptive statements. This project aims to leverage a Graph Attention Network Model (AASIST: Audio Anti-Spoofing System) for binary classification of deceptive and truthful speech across distinct datasets: the English Real-life Deception Detection Dataset (RLDD) and the Romanian Deva Criminal Investigation Audio Recordings Dataset (RODeCAR). The study focuses on evaluating the adaptability and effectiveness of the AASIST architecture for this challenging classification task across different linguistic contexts. Preliminary results on the RLDD and RODeCAR datasets are presented, with ongoing work on the Mandarin dataset. The model achieved an impressive accuracy of 81.81% and an EER of 18.52% on the RODeCAR dataset, while it achieved an accuracy of 63.16% and EER of 52.77% on the RLDD dataset. The codes files, report, and poster are available here.

1. Introduction

Deception detection using speech processing has gained significant attention due to its potential applications in security, forensics, and psychology. Existing research [1], [2], has primarily focused on handcrafted acoustic and prosodic features, such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch contours, energy analysis, and speaking rate, often coupled with traditional machine learning models or deep learning architectures like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs). However, developing systems that generalize robustly across diverse datasets, languages, speakers, and recording conditions remains a significant challenge. Many models trained on specific datasets exhibit poor performance when applied to new, unseen data, limiting their practical utility.

The AASIST framework, initially developed for the related task of audio anti-spoofing (detecting synthetic or replayed speech), has demonstrated a strong capability in capturing intricate and subtle patterns within speech signals. Its architecture, often employing deep learning components and self-attention mechanisms, is designed to distinguish between genuine human speech and sophisticated spoofing attacks. We hypothesize that this capability to model fine-grained acoustic details could also be beneficial for the task of deception detection, where cues indicating deceit might be similarly subtle and complex. Therefore, this project proposes adapting and applying the AASIST framework specifically to the binary classification problem of lie detection, exploring its effectiveness and generalizability using datasets from different languages and contexts (English and Romanian).

2. Project Concept Diagram

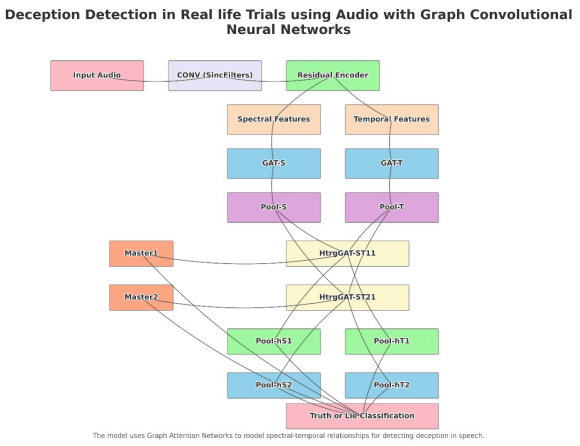


Figure 1. Conceptual Diagram of the AASIST-based Lie Detection System. Speech input is processed through the adapted AASIST model, which outputs a binary classification (Truthful/Deceptive).

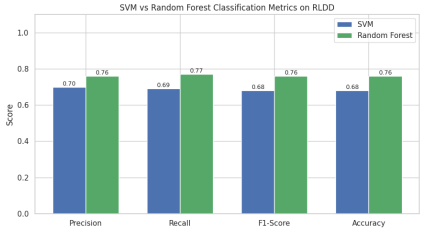
Graph models excel in deception detection, achieving state-of-the-art results on RODeCAR.

3. Databases Utilized

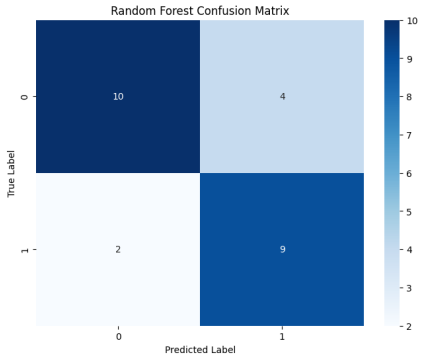
This project utilizes several distinct datasets to rigorously evaluate the robustness and generalizability of the proposed lie detection system across different languages, contexts, and acoustic conditions:

- **Real-life Deception Detection Dataset (RLDD) [4]:** This dataset contains audio recordings sourced from actual courtroom trials in English. It features speakers exhibiting genuine emotional states while providing both truthful and deceptive testimonies. The uncontrolled nature of courtroom recordings, variations in speaker demographics, background noise, and differing emotional intensities make this a particularly challenging and realistic dataset for evaluation.
- **Romanian Deva Criminal Investigation Audio Recordings Dataset (RODeCAR) [3]:** This dataset comprises audio recordings collected during criminal investigations conducted in Romanian. It introduces a different linguistic and acoustic environment compared to RLDD, providing a critical test case for assessing the cross-lingual capabilities and adaptability of our approach.
- **Mandarin Deception Dataset:** (Mentioned in Proposal) Integration and evaluation using a Mandarin Chinese deception dataset are planned for a future phase of this research. This will allow testing the model's performance on a tonal language, offering a third distinct linguistic challenge and further probing the limits of its generalizability.

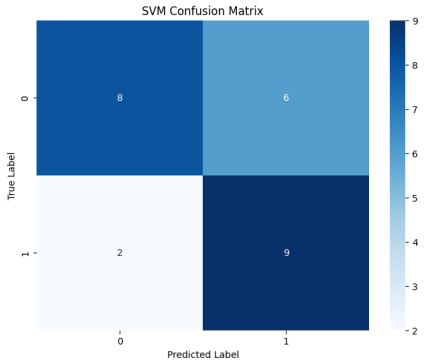
The deliberate use of these diverse datasets is crucial for moving beyond dataset-specific results and towards a more comprehensive understanding of the adapted AASIST framework's potential for reliable deception detection in varied real-world scenarios.



((a)) Performance Metrics of SVM and RF on RLDD dataset



((b)) RF Confusion Matrix



((c)) SVM Confusion Matrix

Figure 2. Performance of traditional machine learning approaches on the RLDD dataset

4. Problem with Existing Work

Detecting deception solely from speech signals is actually an inherently challenging task, and existing approaches face many limitations that hinder their reliability:

- **Dependency on Handcrafted Features:** Many traditional methods rely heavily on pre-defined acoustic or prosodic features (e.g., MFCCs, pitch statistics, jitter, shimmer, energy contours, speaking rate). Usually, such features have been proven to be very useful in case the context of speech is important. This might be the reason why Random Forest and Support Vector Machine performed better on the RLDD dataset (accuracies around 70% since they used MFCC features as well as additional gesture information).
- **Poor Generalization Across Datasets:** Models trained on one specific dataset often fail to generalize effectively when tested on data from different sources or recorded under different conditions. This is because variations in language, recording quality, speaker demographics, the nature of the deception (e.g., high-stakes vs. low-stakes lies), and cultural nuances in expression can drastically change and degrade performance.
- **Data Scarcity and Imbalance:** Compared to other speech processing tasks like speaker recognition or speech-to-text, high-quality, reliably labeled deception datasets are relatively scarce. For example - the RLDD dataset had only Furthermore, existing datasets are of-

ten imbalanced, containing significantly more truthful than deceptive samples (or vice-versa). This imbalance can bias model training and make standard accuracy metrics potentially misleading.

- **Subtlety and Complexity of Deceptive Cues:** The acoustic and prosodic cues associated with deception are often subtle, highly variable between individuals, and potentially non-linear. Capturing these intricate patterns effectively requires sophisticated modeling techniques capable of learning complex feature representations directly from the data, rather than relying solely on pre-engineered features.

The proposal acknowledges that while advanced deep learning techniques like CNNs and LSTMs have been applied, robust generalization remains a primary hurdle. This project seeks to address these limitations by exploring the potential of the AASIST architecture, hypothesizing that its design for detecting sophisticated audio manipulations may be advantageous for capturing the subtle, complex patterns indicative of deception.

5. Proposed Methodology

The core of this project lies in the adaptation and systematic evaluation of the AASIST framework [5] for the binary classification task of speech lie detection. The methodology encompasses the following key stages:

1. **Framework Selection and Adaptation:** The AASIST framework was chosen due to its proven success in capturing subtle acoustic patterns crucial for distinguishing bonafide speech from spoofed audio. Its deep learning architecture, often incorporating self-attention mechanisms, is well-suited for learning complex representations from raw or minimally processed audio. We adapted the original framework by modifying its output layer to produce a single score representing the probability of the input speech segment being truthful. This converts the architecture into a binary classifier suitable for the lie detection task.

2. **Data Preparation and Processing:** Consistent and appropriate data handling is crucial for model training. Each dataset (RLDD, RODeCAR) undergoes a standardized preprocessing pipeline. This typically involves:

    - Loading audio files and potentially resampling them to a common sampling rate.
    - Extracting relevant acoustic features (e.g., spectrograms, or features learned internally by AASIST).
    - Segmenting the audio into manageable chunks or frames suitable for model input.
    - Assigning ground truth labels (Truthful/Deceptive) based on the dataset annotations.
    - Partitioning the data into distinct training, validation, and testing sets to ensure unbiased evaluation.

3. **Model Training Procedure:** The adapted AASIST model is trained independently on the prepared datasets. The training process involves:

    - Defining and loading training configurations, including hyperparameters like learning rate, batch size, number of training epochs, and optimizer choices.
    - Initializing the AASIST model architecture with appropriate weights (either randomly or using pre-trained weights if applicable).
    - Employing an optimization algorithm (e.g., AdamW) to update model parameters based on the calculated loss (typically Binary Cross-Entropy for this task). A learning rate scheduler might be used to adjust the learning rate during training for improved convergence.
    - Incorporating techniques like Stochastic Weight Averaging (SWA) during training to potentially enhance model generalization by averaging weights from multiple points in the training trajectory.
    - Iteratively processing the training data in batches over multiple epochs, computing the loss between model predictions and true labels, and performing backpropagation to update the model's weights.
    - Regularly evaluating the model's performance on the separate validation set during training. Key metrics (e.g., validation loss, accuracy, EER) are monitored to track learning progress, detect potential overfitting, and select the best-performing model checkpoint for final evaluation.

4. **Evaluation Protocol:** Following the training phase, the performance of the selected best model is rigorously assessed on the held-out test set, which the model has never encountered during training or validation. This involves:

    - Generating prediction scores for each sample in the test set using the trained model.
    - Calculating a comprehensive suite of performance metrics by comparing the model's predictions and scores against the ground truth labels. The primary metrics include:

- **Accuracy:** The overall proportion of correctly classified samples.
- **Precision, Recall, F1-Score:** Metrics that provide deeper insight into performance, especially for imbalanced datasets, by quantifying true positives, false positives, true negatives, and false negatives.
- **Equal Error Rate (EER):** The specific error rate where the false acceptance rate (misclassifying deceptive as truthful) equals the false rejection rate (misclassifying truthful as deceptive). Lower EER indicates better overall class separability at this operating point.
- **Area Under the ROC Curve (AUC):** A threshold-independent measure of the model's ability to discriminate between the positive and negative classes. A value closer to 1.0 indicates better discriminative power.

- Analyzing results using confusion matrices and potentially visualizing performance through plots such as ROC curves, Precision-Recall curves, and score distributions to gain further insights into the model's behavior and error patterns.

This systematic methodology ensures a reproducible process for adapting, training, and evaluating the AASIST framework for lie detection across different datasets.

6. Results and Analysis

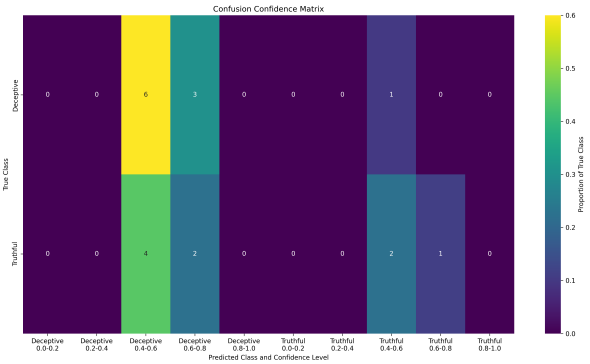
This section presents the quantitative performance of the adapted AASIST model evaluated independently on the RLDD and RODeCAR datasets. Note that the final results presented were obtained using different training configurations (RLDD: 75 epochs, batch size 8; RODeCAR: 50 epochs, batch size 32).

6.1. RLDD Dataset Results

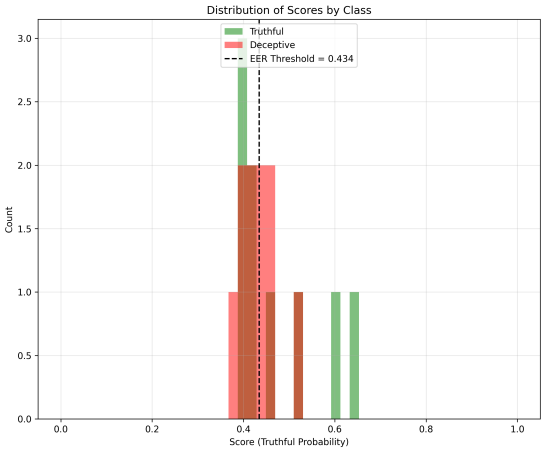
Evaluation on the English Real-life Deception Detection (RLDD) dataset yielded the performance metrics summarized in Table 1.

Table 1. Performance Metrics on RLDD Dataset

Metric	Value
Accuracy	63.158%
Precision (Truthful)	75.000%
Recall (Truthful)	33.333%
F1 Score (Truthful)	46.154%
Equal Error Rate (EER)	52.778%
ROC AUC	0.567



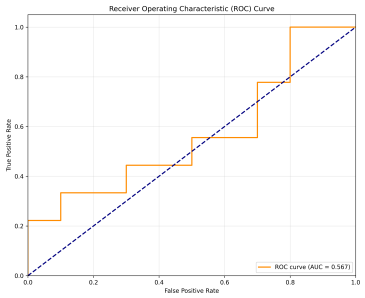
((a)) Confusion Matrix for RLDD



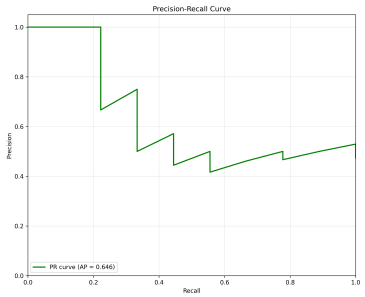
((b)) Score Distribution for RLDD

Figure 3. Confusion matrix and score distribution for the RLDD dataset

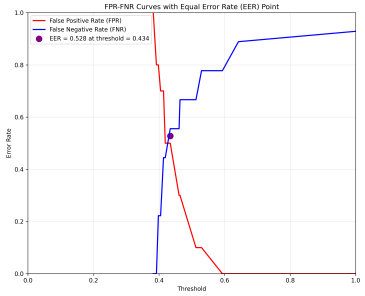
Analysis (RLDD): The results on RLDD show performance moderately better than chance, but highlight significant challenges. Accuracy stands at 63.16%. A notable imbalance exists between precision (75.0%) and recall (33.3%) for the truthful class, indicating the model correctly identifies truthful instances only one-third of the time it encounters them, despite being relatively precise when it does predict 'Truthful'. This suggests a bias towards classifying samples as deceptive. This is in contrast to the start of training, where all samples were being classified as true so the recall was very high. The EER (52.78%) is high, and the ROC AUC (0.567) is only slightly above the baseline of 0.5, indicating poor class separability and limited discriminative power in this specific, uncontrolled courtroom setting. The small test set size (19 samples: 9 truthful, 10 deceptive) should also be considered when interpreting these results.



((a)) ROC Curve



((b)) Precision-Recall Curve



((c)) EER Curve

Figure 4. Performance curves for the RLDD dataset

6.2. RODeCAR Dataset Results

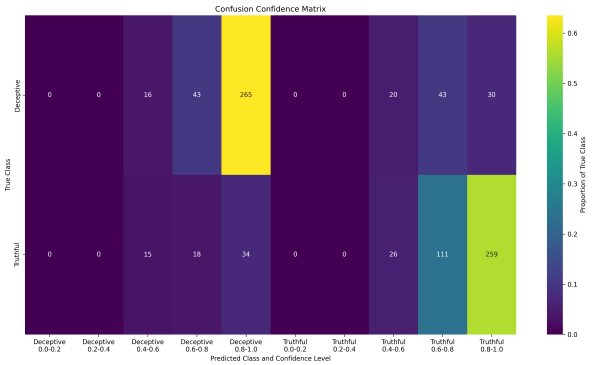
Evaluation on the Romanian Deva Criminal Investigation Audio Recordings (RODeCAR) dataset yielded the performance metrics summarized in Table 2.

Analysis (RODeCAR): In stark contrast to the RLDD results, the model demonstrates significantly stronger performance on the RODeCAR dataset. Accuracy reaches 81.82%. The Precision, Recall, and F1 scores are all substantially higher and more balanced (around 81-85%, assuming these are weighted averages). Most notably, the EER is much lower at 18.52%, and the ROC AUC is high at 0.886. These metrics indicate a considerably better ability of the model to discriminate between truthful and deceptive speech within the context of the RODeCAR dataset. The model achieves good separation between the classes with a much lower error rate compared to RLDD. This dataset also featured a much larger test set size (880 samples: 463 truthful, 417 deceptive).

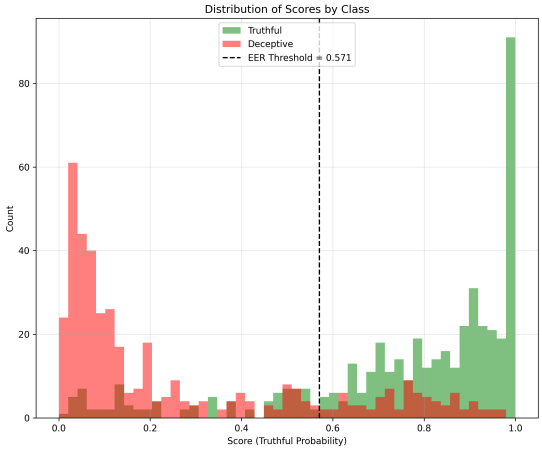
Table 2. Performance Metrics on RODeCAR Dataset

Metric	Value
Accuracy	81.818%
Precision (weighted avg approx.)*	80.982%
Recall (weighted avg approx.)*	85.529%
F1 Score (weighted avg approx.)*	83.193%
Equal Error Rate (EER)	18.520%
ROC AUC	0.886

* The provided results list single values for Precision, Recall, and F1.

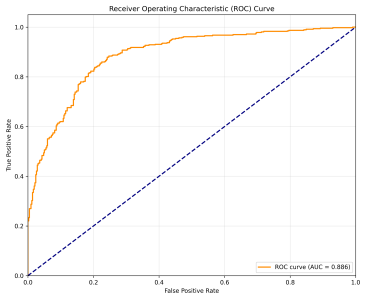


((a)) Confusion Matrix for RODeCAR

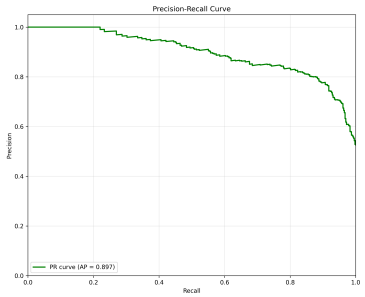


((b)) Score Distribution for RODeCAR

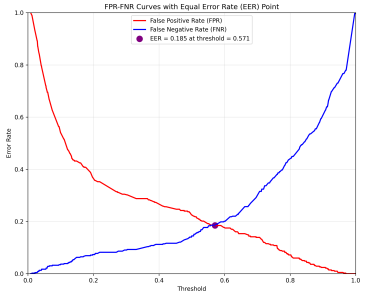
Figure 5. Confusion matrix and score distribution for the RODeCAR dataset



((a)) ROC Curve



((b)) Precision-Recall Curve



((c)) EER Curve

Figure 6. Performance curves for the RODeCAR dataset

6.3. Comparative Analysis

Table 3 directly compares the performance metrics achieved on the two datasets.

The performance difference between RLDD and RODeCAR is substantial across all major metrics, particularly EER and AUC, which measure class separability. Potential contributing factors to this disparity include:

- **Dataset Characteristics:** RLDD involves uncontrolled courtroom audio with high emotional variance and likely more background noise, while RODeCAR might have more controlled recording conditions typical of investigations. The nature, stakes, and linguistic manifestation of deception may also differ significantly.

Table 3. Comparison of Performance Metrics on RLDD and RODeCAR Datasets

Metric	RLDD Result	RODeCAR Result
Accuracy	63.158%	81.818%
Precision*	75.000%	80.982%
Recall*	33.333%	85.529%
F1 Score*	46.154%	83.193%
Equal Error Rate (EER)	52.778%	18.520%
ROC AUC	0.567	0.886
Test Set Size (Total)	19	880
Test Set Size (Truthful)	9	463
Test Set Size (Deceptive)	10	417
Training Epochs / Batch Size	75 / 8	50 / 32

* RLDD Precision/Recall/F1 are for the 'Truthful' class.

- **Language:** Acoustic correlates of deception might differ substantially between English and Romanian.
- **Dataset Size:** RODeCAR is significantly larger (880 test samples vs. 19 for RLDD), providing considerably more data for the model to learn robust patterns and enabling more reliable evaluation statistics.
- **Training Configuration:** The different epoch counts and batch sizes used for the final models might have influenced performance, although the large performance gap suggests dataset differences are likely the primary factor.

Further investigation, potentially involving training on RLDD with configurations similar to RODeCAR (if feasible given dataset size) or analyzing feature representations, would be needed to better isolate the primary causes of this performance disparity.

7. Conclusion

This project successfully adapted a Graph Attention Network (AASIST), and evaluated the framework for speech-based lie detection across two distinct datasets: the English RLDD and the Romanian RODeCAR. The methodology involved modifying the AASIST architecture for binary classification, preparing data from these diverse sources, and employing a systematic training and evaluation process.

The results reveal a striking contrast in performance. On the challenging, real-world RLDD dataset, the model achieved modest accuracy (63.16%) but showed poor class discrimination (EER: 52.78%, AUC: 0.567) and a bias against correctly identifying truthful instances. Conversely, on the larger RODeCAR dataset, the model performed significantly better, achieving high accuracy (81.82%) and demonstrating strong discriminative ability (EER: 18.52%, AUC: 0.886).

This performance disparity underscores the high sensitivity of speech-based lie detection models to dataset characteristics, recording conditions, language, and potentially the nature of the deception itself. While the strong performance on RODeCAR suggests the adapted AASIST framework *can* be effective under certain conditions, the poor results on RLDD highlight the persistent challenges in achieving robust generalization, particularly in uncontrolled, high-variability scenarios. The difference in training configurations and dataset sizes also likely contributed to the observed performance gap.

Future work should focus on investigating the reasons for this performance difference, potentially through cross-dataset experiments and analysis of learned features. Evaluating the system on the planned Mandarin dataset will provide further insights into its cross-lingual capabilities. Despite the mixed results, this project provides valuable insights into applying advanced audio anti-spoofing techniques to deception detection, confirming the potential of the approach while clearly illustrating the significant hurdles that remain in developing truly generalizable lie detection systems.

References

[1] Felipe Marcolla, Rafael de Santiago, and Rudimar Dazzi. Novel lie speech classification by using voice stress. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,*, pages 742–749. INSTICC, SciTePress, 2020. 1

[2] Serban Mihalache and Dragos Burileanu. Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection. *Sensors*, 22(3), 2022. 1

[3] Serban Mihalache, Gheorghe Pop, and Dragos Burileanu. Introducing the rodecar database for deceptive speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6, 2019. 2

[4] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66, 2015. 2

[5] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, 2021. 3