XLVI CILAMCE
VITÓRIA - ES - BRAZIL - 2025
XLVI IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGENEERING

Accent-Aware Deepfake Speech Detection in Brazilian Portuguese: Dataset Construction and Model Evaluation

Sofia N. Silva[1], Erick M. B. Santos[1], Katarina Veljovic[1], Karin S. Komati[2]

[1]*Information Technology Coordination*
[2]*Graduate program in Applied Computing (PPComp)*
Instituto Federal do Espírito Santo (IFES), Campus Serra
Av. dos Sabiás, 330 - Morada de Laranjeiras, 29166-630, Serra-ES, Brazil
sofianascimento307@gmail.com, erickmiguelbsantos@gmail.com, katarinaveljovic123@gmail.com, kkomati@ifes.edu.br

Abstract. The rise of digitally manipulated audio content creates new challenges in verifying information authenticity, especially on social media. Advances in artificial intelligence (AI), particularly in text-to-speech and voice synthesis technologies, have greatly enhanced the quality and realism of generated audio. This study addresses the problem of deepfake audio detection and offers two main contributions. First, it introduces the FakeBrAccent dataset, which includes 746 audio samples (373 real and 373 synthetic) in Brazilian Portuguese, featuring regional accents such as Baiano (Bahia), Fluminense (Rio de Janeiro and Espírito Santo), Southern, Northeastern, and Carioca (Rio de Janeiro city). The original BrAccent dataset was used as both the source of real samples and as a reference for simulating accents during the generation of synthetic samples with a text-to-speech tool. Second, the study evaluates the performance of two classification models, Convolutional Neural Networks and XGBoost, on this dataset. The models were tested using standard performance metrics, including accuracy, precision, recall, and F1-score. The findings provide a baseline for future research into synthetic speech detection in Brazilian Portuguese, emphasizing the role of accent variation in model performance.

Keywords: Regional Accents, Text-to-Speech, Synthetic Audio.

1 Introduction

In recent years, advancements in deep learning have enabled the creation of highly realistic synthetic media, known as "deepfakes" [1]. This term describes content, including images, videos, or audio, that has been generated or altered by AI algorithms to convincingly misrepresent something that did not occur. Audio deepfakes are a subset of this phenomenon, referring to any audio in which key attributes, such as speaker identity, have been manipulated or entirely synthesized by AI technologies to retain perceived naturalness [2]. These forgeries are typically produced using advanced deep learning models, such as text-to-speech (TTS) and voice conversion (VC) systems. These technologies present significant challenges for information security, personal privacy, and the spread of disinformation, making the development of robust detection methods a critical area of research [3].

While multiple datasets exist for voice deepfake detection in several major languages, Portuguese resources are scarce [4]. The only known public dataset, H-Voice, contains histograms derived from audio rather than raw waveforms [5]. Brazil is a large and linguistically diverse country, with regional accents that differ markedly across its geographic and cultural regions. This gap highlights the need for a dedicated, accent-aware, waveform-based dataset to properly evaluate and advance deepfake detection models for Brazilian Portuguese. Accordingly, this work proposes the FakeBrAccent dataset. The authentic samples are drawn from the BrAccent corpus [6], which also serves as a vocal reference for the generation of synthetic, accent-aware counterparts via a Text-to-Speech (TTS) system and Voice Cloning.

The study evaluates the performance of two classification models, Convolutional Neural Networks (CNNs) and XGBoost, on the FakeBrAccent dataset. The models were assessed using standard performance metrics, including accuracy, precision, recall, and F1-score. The structure of the article is as follows: Section 2 presents related work. Section 3 describes the FakeBrAccent dataset and the methodological approach. Section 4 reports and discusses the results obtained. Finally, Section 5 concludes the article and outlines directions for future research.

2 Related Work

The study by Yi et al. [7] provides an overview of voice deepfake detection, outlining the main types of artificially manipulated audio and the leading techniques used for their identification. The authors emphasize that text-to-speech and voice conversion tools, driven by deep learning models, have advanced to such an extent that fake audio is becoming increasingly realistic and difficult to detect — posing significant risks to society, particularly in terms of disinformation. Moreover, the study discusses various classification models, including Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), among others. It also highlights models like RawNet, which assist in extracting features from audio samples. These insights provide valuable guidance regarding which approaches and architectures should be considered when designing our own accent detection system.

The study conducted by Young et al. [8] presents a methodology based on matched-guise experiments, a technique related to linguistic variation, particularly accents, with the main objective of creating deepfakes capable of eliminating linguistic distinctions. Four different versions of the same video were produced, featuring variations in intonation and facial appearance, and used in a controlled experimental setting. It was observed that individuals with facial features considered outside the standard of British beauty, combined with a specific accent, were more frequently identified as manipulated videos. This finding demonstrates that deepfakes have revolutionary potential in the field of sociolinguistics and can be applied to other studies that incorporate accent as a relevant factor, as is the case in our work, which aims to contribute to overcoming barriers posed by linguistic diversity.

3 Materials and Methods

The overall workflow of this study, as depicted in the diagram (Figure 1), encompasses dataset construction and model evaluation phases. Initially, real audio samples are sourced from the BrAccent dataset, while synthetic voices are generated by feeding a “Text Input” (e.g., ”Olá, mundo!”) into Speechify. Speechify is a zero-shot voice cloning TTS, that is a system that accepts input in the form of text and a few seconds of a sample of the target speaker’s voice (from BrAccent) to produce speech sound waves similar to the target speaker’s voice [9], to simulate specific accents during the generation of synthetic speech, ensuring accent awareness. Both real and generated fake voices are then integrated into the FakeBrAccent dataset. All audio samples from FakeBrAccent undergo two parallel processing paths: one for converting the audio into an image representation (Mel-spectrogram), which is then fed into a CNN for classification between ”Fake Voices” and ”Real Voices”; and another for MFCC Pre-Processing, generating a feature vector that is fed into an XGBoost for the same classification.

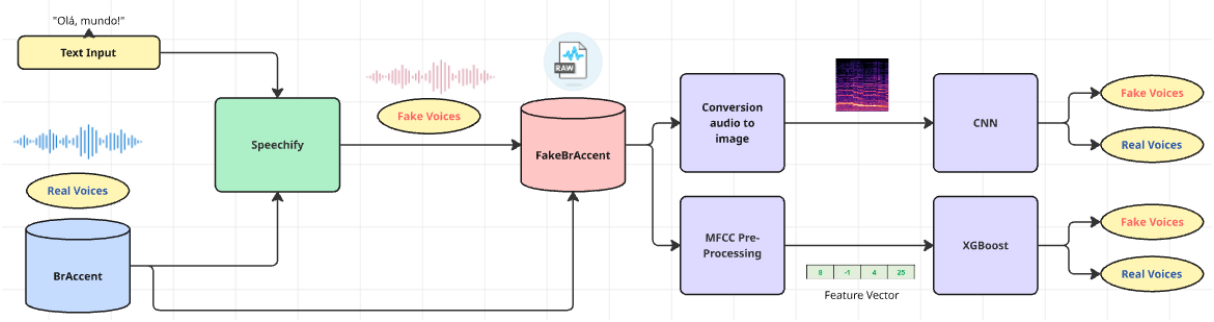

Figure 1. The Process of generating and standardizing fake audios from the Braccent database using Speechify

3.1 FakeBrAccent

The BrAccent dataset comprises 1,403 speech recordings in Brazilian Portuguese captured from native speakers, showcasing a rich regional diversity of accents, including Southern, Nortista, Northeastern, Mineiro, Fluminense, Carioca, and Baiano. To generate the synthetic audio samples, we selected accents with the highest number of samples from BrAccent for use with the Speechify system. The Fake BrAccent-B dataset comprises real and synthetic (fake) audio samples organized by accent and gender. For the Southern, Fluminense, and Baiano accents, the distribution is balanced, with 40 samples each for female and male voices in both real and fake categories. The Northeastern accent includes 40 female and 18 male samples per category, while the Carioca accent contains

40 female and 35 male samples. Overall, the dataset consists of 200 female and 174 male real recordings (totaling 373 real samples) and 200 female and 174 male fake recordings (totaling 373 fake samples). To ensure consistency across the dataset utilized for our experiments, all audio samples—comprising both real and synthetic voices—underwent a series of pre-processing steps. These steps included volume normalization, format unification, and the removal of excessive silence. Finally, the FakeBrAccent dataset is made publicly available on the Kaggle platform[1].

3.2 Model architecture

The first code utilized was developed by Razo [10]. Each audio file is converted into a spectrogram image, which serves as input for the CNNs, designed to process two-dimensional data. The raw audio signal (amplitude versus time) is first transformed into a frequency versus time representation using the Short-Time Fourier Transform (STFT). Subsequently, the Mel scale is applied, simulating human auditory perception, and the resulting spectrogram is resized to 128x128 pixels. The CNN architecture consists of an input layer with 3 channels (RGB) and 128x128 dimensions, followed by two convolutional layers (Conv1 and Conv2) and two pooling layers (Pool1 and Pool2). The convolutional and pooling layers progressively reduce spatial dimensions and increase the number of filters, culminating in a dense layer of 512 units and an output layer of 2 units. The main hyperparameters employed included convolutional layers for the training process, the Adam optimizer for learning, and the ReLU (Rectified Linear Unit) activation function applied to each layer to introduce non-linearity into the model. A batch size of 16 was used, with a learning rate of 0.001, and training was performed for 10 epochs.

The second approach was developed by Chauhan [11] based on the work of [12] and employs the XGBoost decision tree algorithm, which is widely used for classification tasks. XGBoost is an ensemble learning algorithm characterized by high flexibility, strong predictive power, excellent generalization capability, high scalability, efficient model training, and strong robustness. In the experiments, 100 estimators were employed, with a learning rate set to 0.1 and a maximum tree depth fixed at 5. The audio signal was segmented into short frames and then transformed into the frequency domain using the Fast Fourier Transform (FFT). The resulting frequency components were processed through the Mel-Frequency Cepstral Coefficients (MFCCs). These outputs were aggregated over time and combined with additional features to form a complete feature vector, which was subsequently used as input for the XGBoost classifier. The codes used and modified according to the project requirements are available on GitHub for both the CNN[2] and the XGBoost [3]implementations.

Given the limited number of audio samples available in our Portuguese dataset, up to 80 samples per accent, transfer learning techniques were applied. Transfer learning enables the transfer of previously acquired knowledge to a new task, helping to preserve relevant information. Additionally, fine-tuning, a widely adopted method for adapting pre-trained models and improving their performance on new data, was employed. Using these strategies, the five available Portuguese accents in our dataset were adapted for new rounds of model and algorithm training, resulting in a total of 10 models (5 accents $\times$ 2 model types). Training and testing were conducted with 80% of the data allocated for training and 20% for testing.

4 Results and Discussion

The performance of the CNN and XGBoost models in detecting deepfake audio across different Brazilian Portuguese accents is detailed in Table 1. Bold values indicate the best performance for each class/model combination. The most immediate observation is the clear superiority of the XGBoost model over the CNN across all accents and performance metrics. XGBoost consistently achieved higher accuracy, precision, recall, and F1-scores. For example, its lowest accuracy exceeds the highest accuracy recorded by the CNN.

A second key finding is the variation in performance across regional accents, which reinforces the central hypothesis of this study. For XGBoost, performance was particularly strong on the Southern and Carioca accents, while slightly lower, though still robust, results were observed for the Northeastern and Baiano accents. This variation may indicate that certain prosodic or phonetic features of these accents are more difficult for the model to distinguish after synthesis, or that the TTS engine reproduces them with higher fidelity. The CNN model also showed variation across accents, but its overall lower performance makes the trend less evident. It performed best on the Fluminense accent and struggled most with the Northeastern accent. The confusion matrices illustrating the results obtained by the CNN are presented in Figure 2, while those corresponding to the XGBoost model are shown in Figure 3.

[1] `https://www.kaggle.com/datasets/erickmiguelsantos/fake-braccent`
[2] `https://github.com/sofianasilva/DeepFake_Detection_CNN`
[3] `https://github.com/sofianasilva/DeepFake_Detection_XGBoost`

Table 1. Metrics per class/accent for CNN and XGBoost models.

Model	Class/Accent	Accuracy	Precision	Recall	F1-score
CNN	Baiano	66.65%	66.00%	66.00%	65.00%
	Fluminense	62.50%	65.00%	62.00%	61.00%
	Southern	59.38%	60.00%	59.00%	59.00%
	Northeastern	79.17%	81.00%	79.00%	79.00%
	Carioca	39.29%	39.00%	39.00%	39.00%
XGBoost	Baiano	**96.88%**	**97.00%**	**97.00%**	**97.00%**
	Fluminense	**93.75%**	**94.00%**	**94.00%**	**94.00%**
	Southern	**96.88%**	**97.00%**	**97.00%**	**97.00%**
	Northeastern	**87.50%**	**88.00%**	**88.00%**	**87.00%**
	Carioca	**92.86%**	**93.00%**	**93.00%**	**93.00%**

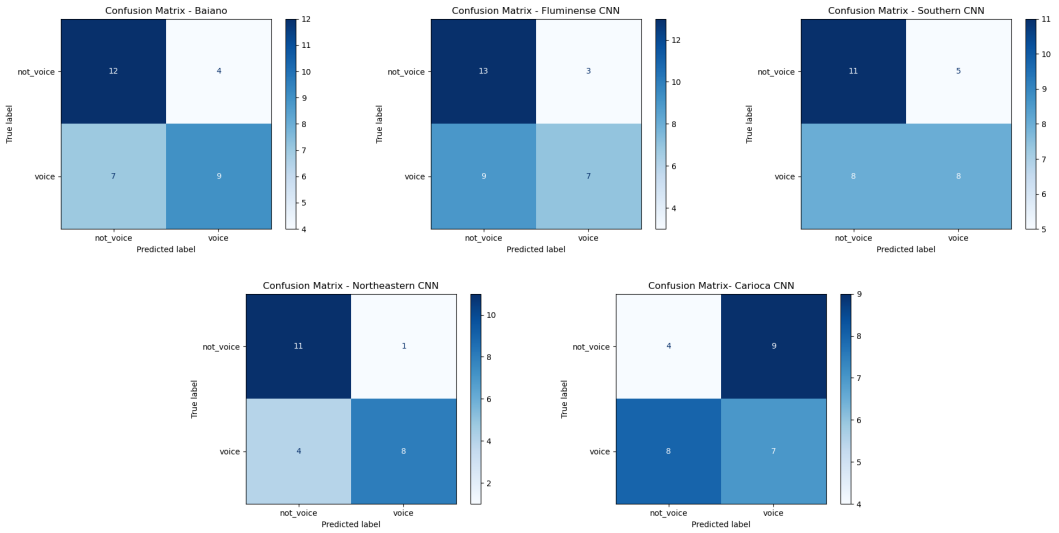

Figure 2. CNN Results.

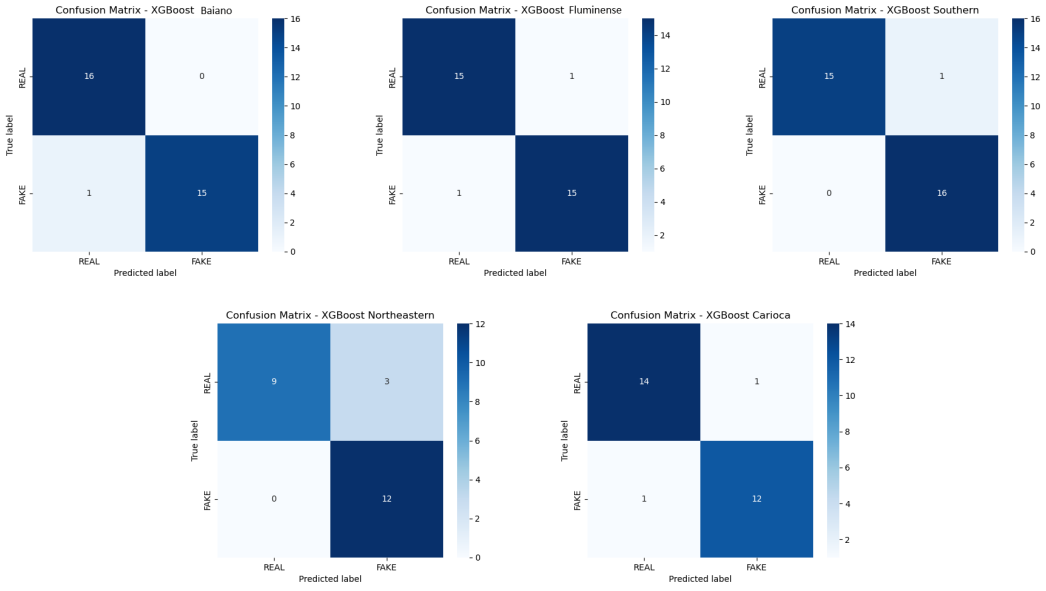

Figure 3. XGBoost results.

5 Conclusions

The results demonstrate that XGBoost achieved the highest overall performance among the two models evaluated. Given the limited dataset size, it is hypothesized that XGBoost outperformed the CNN due to its ability to effectively capture simple and structured patterns. These findings empirically support the importance of accounting for accent variation in the development of deepfake detection systems. A model trained or evaluated on a single accent may fail to generalize across others, potentially resulting in a misleading assessment of its robustness or effectiveness.

Several avenues for future research emerge from this study. First, future work should explore more advanced deep learning architectures, such as transformer-based models or enhanced CNN variants. Second, expanding the FakeBrAccent dataset is a critical next step. This includes incorporating a broader range of Text-to-Speech (TTS) and Voice Cloning (VC) technologies, as different synthesis methods may introduce distinct artifacts that affect model performance. Additionally, increasing the diversity of regional accents and the number of speakers within each accent class would improve the dataset's representativeness and robustness. Finally, performing cross-accent and cross-synthesis evaluations will be essential to assess the generalization capabilities of detection models—advancing toward the development of a truly universal deepfake detector for Brazilian Portuguese. The exploration of additional neural network architectures beyond those used in this work may further strengthen and extend the results achieved.

Acknowledgements. The authors would like to thank IFES (EDITAL PRPPG 08/2025). Professor Komati thanks CNPq for the DT-2 grant (nº 302726/2023-3) and project nº407742/2022-0, also thanks FAPES for project nº 1023/2022 P:2022-8TZV6.

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] E. E. Ferreira, J. O. Andrade, and K. S. Komati. Cross-database in deepfake detection based on a convolutional neural network and vision transformer. In *Workshop de Visão Computacional (WVC)*, pp. 60–65. SBC, 2023.

[2] Z. Khanjani, G. Watson, and V. P. Janeja. Audio deepfakes: A survey. *Frontiers in Big Data*, vol. 5, pp. 1001063, 2023.

[3] J. W. Seow, M. K. Lim, R. C. Phan, and J. K. Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing (Amsterdam)*, vol. 513, pp. 351–371, 2022.

[4] L. Cuccovillo, C. Papastergiopoulos, A. Vafeiadis, A. Yaroshchuk, P. Aichroth, K. Votis, and D. Tzovaras. Open challenges in synthetic speech detection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2022.

[5] D. M. Ballesteros, Y. Rodriguez, and D. Renza. A dataset of histograms of original and fake voice recordings (h-voice). *Data in brief*, vol. 29, pp. 105331, 2020.

[6] N. A. R. Batista and others. Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets. In *Anais do XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, pp. 939–944, Campina Grande, PB. Sociedade Brasileira de Telecomunicações, 2018.

[7] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

[8] N. J. Young, D. Britain, and A. Leemann. A blueprint for using deepfakes in sociolinguistic matched-guise experiments. In *Interspeech*, pp. 5268–5272, 2022.

[9] K. Azizah. Zero-shot voice cloning text-to-speech for dysphonia disorder speakers. *IEEE Access*, vol. 12, pp. 63528–63547, 2024.

[10] H. Razo. human-voice-detection. `https://github.com/hernanrazo/human-voice-detection`. Accessed: 2025-05-27, 2021.

[11] N. Chauhan. Deepfake-audio-detection-mfcc. `https://github.com/noorchauhan/DeepFake-Audio-Detection-MFCC`. Accessed: 2025-05-27, 2023.

[12] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol. Deepfake audio detection via MFCC features using machine learning. *IEEE Access*, vol. 10, pp. 134018–134028, 2022.