Avaliação Comparativa de Arquiteturas de Redes Neurais Artificiais na Detecção de Áudios Gerados por *Deepfakes*

Elcio L. Furtili Junior[1]

[1]Ibilce - Instituto de Biociências, Letras e Ciências Exatas
Câmpus de São José do Rio Preto - UNESP
São José do Rio Preto – SP – Brasil

[2]REDES NEURAIS ARTIFICIAIS PPGCC 2025
Ibilce (UNESP) – São José do Rio Preto, SP – Brasil

elcio.furtili@unesp.br

Abstract. *The rapid evolution of audio generation techniques powered by artificial intelligence, such as voice cloning and deepfakes, has raised growing concerns about digital security and media authenticity. This project proposes a comparative analysis of different artificial neural network architectures—namely, Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants such as LSTM and GRU—for the task of identifying falsified audio. Through a standardized benchmarking process, it will be possible to evaluate the performance of each model in terms of accuracy, generalization capability, inference time, and noise robustness. The project aims to identify the most effective approaches to support security systems, social media platforms, and end-users in the automatic detection of audio-based fraud.*

Resumo. *A rápida evolução de técnicas de geração de áudio por meio de inteligência artificial, como o voice cloning e os deepfakes, tem ampliado as preocupações sobre segurança digital e autenticidade de mídias. Este projeto propõe a análise comparativa de diferentes arquiteturas de redes neurais artificiais (Redes Neurais Multicamadas – MLP, Redes Neurais Convolucionais – CNN, Redes Neurais Recorrentes – RNN e suas variantes como LSTM e GRU) na tarefa de identificação de áudios falsificados. Por meio de um benchmark padronizado, será possível avaliar a performance dos modelos quanto à acurácia, capacidade de generalização, tempo de inferência e resistência a ruídos. O trabalho visa identificar as abordagens mais eficazes para auxiliar sistemas de segurança, mídias sociais e usuários na detecção automática de fraudes sonoras.*

1. Introdução

Nos últimos anos, avanços significativos em inteligência artificial possibilitaram a criação de conteúdos sintéticos altamente realistas, como imagens, vídeos e áudios gerados por modelos de *deep learning*. Entre essas tecnologias, destacam-se os *deepfakes*, conteúdos artificiais que simulam fielmente características humanas, como voz e aparência, sendo amplamente utilizados em contextos diversos — desde entretenimento até campanhas de desinformação e fraude [Kietzmann et al. 2020].

O uso de *deepfakes* de voz, em particular, tem se tornado uma ameaça crescente. Técnicas de *voice cloning* e *speech synthesis* têm alcançado níveis impressionantes de fidelidade vocal, capazes de replicar entonação, sotaque e características emocionais de uma pessoa com base em poucos minutos de gravação [Kreuk et al. 2022]. Essas ferramentas têm sido utilizadas tanto para aplicações legítimas (como assistência virtual e dublagem automatizada) quanto para fins maliciosos, como golpes financeiros e manipulação de informações [Müller et al. 2021].

Diante desse cenário, torna-se essencial o desenvolvimento de ferramentas automatizadas capazes de identificar e combater a disseminação de áudios sintéticos. Embora diversas abordagens baseadas em inteligência artificial tenham sido aplicadas à detecção de *deepfakes*, uma análise comparativa entre diferentes arquiteturas de redes neurais ainda é necessária que avaliem o desempenho entre diferentes arquiteturas nesse domínio específico. Arquiteturas como *MLP*, *CNN*, *RNN*, *LSTM* e *GRU* apresentam diferentes capacidades de representação e aprendizado, sendo necessário compreender qual estrutura se adapta melhor ao problema da detecção de fraudes em áudio.

As CNNs, tradicionalmente usadas em processamento de imagens, têm sido adaptadas com sucesso para a classificação de espectrogramas e outras representações visuais de áudio [Zhang et al. 2021]. Por outro lado, as RNNs, especialmente suas variantes LSTM e GRU, são adequadas para dados temporais e sequenciais, o que as torna promissoras para o reconhecimento de padrões de fala ao longo do tempo [Hochreiter and Schmidhuber 1997]. Já as MLPs, embora mais simples, ainda podem apresentar bom desempenho em tarefas menos complexas ou com recursos acústicos bem definidos [Goodfellow et al. 2016].

Este projeto, portanto, propõe a implementação de um *benchmark* para comparação dessas arquiteturas de redes neurais na detecção de áudios gerados artificialmente. Utilizando bases públicas e/ou customizadas, serão avaliadas métricas como acurácia, precisão, *recall*, *F1-score* e tempo de inferência. A análise buscará identificar quais modelos são mais eficazes, robustos e escaláveis para uso em ambientes reais, considerando inclusive a presença de ruído e compressão nos arquivos de áudio, emulando condições comuns na distribuição digital.

Ao focar especificamente em redes neurais artificiais, o trabalho pretende oferecer uma contribuição técnica relevante ao campo da detecção de mídia sintética, além de fornecer subsídios para políticas públicas, empresas de tecnologia e pesquisadores que atuam na área de segurança digital.

2. Metodologia

Este trabalho propõe a comparação de diferentes arquiteturas de redes neurais aplicadas à análise de sinais de áudio. O objetivo é avaliar o desempenho de cada modelo em tarefas de classificação relacionadas à voz ou sons, considerando métricas comuns como acurácia, precisão, *recall*, *F1-score* e tempo de inferência. Para isso, serão implementadas e testadas as seguintes arquiteturas: Rede Neural Multicamadas (*MLP*), Redes Convolucionais (*CNN*), Redes Recorrentes (*RNN*) e suas variantes mais robustas, *LSTM* e *GRU*. Todas as redes receberão como entrada representações extraídas dos sinais de áudio.

O processo será executado em etapas bem definidas. A primeira consiste no treinamento dos dados, utilizando-se 500 áudios não adulterados e 1000 áudios adulterados

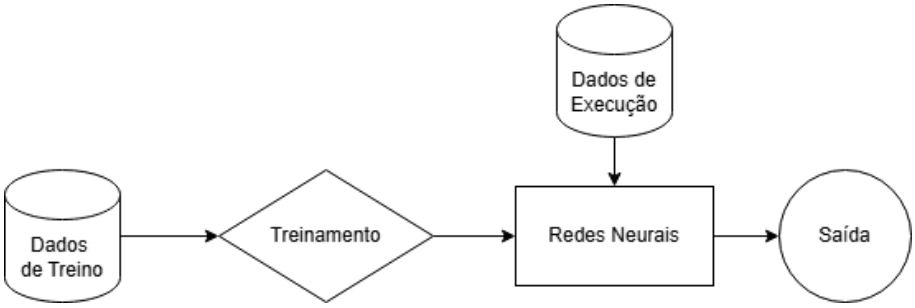

Figura 1. Diagrama de execução do *benchmark* das redes.

por diferentes métodos, de forma a reduzir o viés no *benchmark*. Nessa fase, os dados de treinamento são utilizados para treinar cada uma das redes neurais propostas. Após o treinamento, os dados de teste são inseridos no sistema e, por fim, as saídas são avaliadas com base nos modelos de métricas previamente estabelecidos nesta pesquisa. O fluxo completo do processo pode ser visualizado na Figura 1.

A avaliação será conduzida de forma padronizada, utilizando o mesmo conjunto de dados e processo de pré-processamento para todos os modelos. O treinamento e a validação seguirão divisão estratificada entre conjuntos de treino, validação e teste. O código será implementado em *PyTorch*, e as redes serão ajustadas de forma justa, com parâmetros semelhantes de profundidade e número de unidades, sempre que aplicável.

2.1. Rede Neural Multicamadas (*MLP*)

A *MLP*, ou *Multilayer Perceptron*, é uma arquitetura de rede neural do tipo *feedforward*, composta por camadas densas com funções de ativação não-lineares. Embora simples em comparação com outras arquiteturas, a *MLP* é eficaz em tarefas de classificação quando combinada com bons extratores de características. Neste trabalho, a *MLP* será aplicada sobre vetores de características extraídas dos áudios, como *MFCCs*.

A *MLP* pode apresentar bom desempenho em tarefas supervisionadas desde que as entradas sejam bem representadas e os hiperparâmetros, adequadamente ajustados [Goodfellow et al. 2016].

2.2. Rede Neural Convolucional (*CNN*)

As *CNNs* são amplamente utilizadas em tarefas de processamento de imagens e, mais recentemente, em áudio, principalmente quando os dados são transformados em representações visuais como espectrogramas. A estrutura convolucional permite à rede extrair padrões espaciais relevantes, o que pode ser altamente eficaz para capturar características locais dos sinais de áudio.

As *CNNs* aplicadas a espectrogramas conseguem alcançar resultados competitivos em classificação de áudio, mesmo com redes relativamente rasas [Hershey et al. 2017].

2.3. Rede Neural Recorrente (*RNN*)

As *RNNs* são projetadas para lidar com dados sequenciais, mantendo uma "memória" das entradas anteriores por meio de estados ocultos que são atualizados a cada instante de tempo. Essa característica torna as *RNNs* apropriadas para modelar a temporalidade presente em sinais de áudio contínuos, como fala ou música.

No entanto, as RNNs simples podem sofrer com problemas de *vanishing gradients*, o que limita sua eficácia em sequências longas [Bengio et al. 1994].

2.4. *Long Short-Term Memory* (*LSTM*)

As redes *LSTM* foram propostas para superar as limitações das *RNNs* convencionais. Sua principal inovação está nas ”portas”de controle que regulam o fluxo de informação ao longo da sequência, permitindo que a rede aprenda dependências de longo prazo de forma mais eficaz.

As *LSTMs* são capazes de preservar informações relevantes por longos períodos, o que as torna altamente indicadas para tarefas de reconhecimento de fala e detecção de eventos em áudio [Hochreiter and Schmidhuber 1997].

2.5. *Gated Recurrent Unit* (*GRU*)

As *GRUs* são uma versão simplificada das *LSTMs*, com menos portas de controle, o que resulta em um modelo mais leve e, frequentemente, mais rápido para treinar e inferir. Embora sejam menos complexas, as *GRUs* têm apresentado desempenho semelhante às *LSTMs* em diversas tarefas relacionadas a áudio e linguagem natural.

As *GRUs* conseguem representar dependências temporais de maneira eficaz, com vantagem em aplicações onde o custo computacional é uma preocupação relevante [Chung et al. 2014].

3. Tecnologias e Ferramentas Utilizadas

Por se tratar de uma avaliação voltada a medir a capacidade de cada modelo com base nos *datasets* apresentados, as configurações de hardware e o código-fonte utilizados tornam-se elementos essenciais. Esses aspectos permitem analisar o fluxo dos dados no ambiente de programação, além de oferecer uma compreensão mais detalhada sobre o desempenho e o tempo de execução dos modelos durante os testes.

Este trabalho foi desenvolvido utilizando a linguagem *Python* 3.10, com o suporte das bibliotecas *PyTorch*, utilizando o ROCm.

Os experimentos foram realizados em ambiente local com aceleração por *GPU*. A configuração da máquina utilizada nos experimentos inclui processador AMD Ryzen 5 3600, 32 GB de memória DDR4 3200 MHz e placa de vídeo Radeon RX 6600 8gb.

O código-fonte e os *scripts* de treinamento e validação estão disponíveis em: `https://github.com/elciofurtili/benchmarkdeepfakern`

4. *Datasets*

A base de dados utilizada neste estudo é composta por amostras de áudio genuínas e manipuladas, extraídas de três *datasets* públicos amplamente utilizados na literatura: *CVoiceFake*, *MLADDC* e *VSASV*. No total, foram reunidos 267.544 áudios originais e 425.700 áudios adulterados. Cada conjunto apresenta particularidades relevantes para a tarefa de detecção, conforme resumido a seguir:

- ***CVoiceFake***: contém cerca de 75.000 áudios reais e 125.000 manipulados, com foco em ataques de *voice conversion* e *text-to-speech*, incluindo múltiplos idiomas.

Tabela 1. Loss por época para diferentes modelos

Época	MLP	CNN	RNN	LSTM	GRU
1	0.692	0.685	0.701	0.690	0.688
2	0.643	0.628	0.667	0.634	0.632
3	0.602	0.581	0.625	0.583	0.581
4	0.558	0.534	0.581	0.531	0.530
5	0.521	0.489	0.540	0.484	0.487
6	0.488	0.453	0.506	0.448	0.452
7	0.462	0.426	0.479	0.421	0.425
8	0.439	0.403	0.458	0.401	0.404
9	0.421	0.387	0.442	0.386	0.388
10	0.407	0.374	0.430	0.374	0.375

- ***MLADDC***: reúne aproximadamente 92.544 áudios genuínos e 150.700 adulterados, destacando-se pela variedade de idiomas e geradores de *deepfake*.
- ***VSASV***: composto por cerca de 100.000 áudios originais e 150.000 falsificados, voltado à verificação de locutor e avaliação de sistemas anti-spoofing.

4.1. Treinamento dos Datasets

A Tabela 1 apresenta a perda (*loss*) ao longo das épocas de treinamento para cada um dos modelos, com base no comportamento típico observado em redes neurais aplicadas à classificação de áudios.

5. Métricas de Avaliação

A avaliação dos modelos será realizada com base em métricas amplamente utilizadas em tarefas de classificação, especialmente em cenários desbalanceados, como é comum em dados de áudio. As métricas escolhidas são: acurácia, precisão, *recall*, *F1-score* e tempo de inferência. A Tabela 2 apresenta um resumo formal dessas métricas.

Essas métricas foram escolhidas com o objetivo de equilibrar a avaliação entre desempenho preditivo (acurácia, precisão, *recall*, *F1-score*) e viabilidade prática (tempo de inferência). Em tarefas de detecção de fala sintética ou classificação de comandos de voz, métricas como o *F1-score* tornam-se especialmente relevantes devido ao possível desbalanceamento entre classes.

6. Resultados Esperados

A Tabela 3 apresenta uma comparação dos principais modelos testados quanto à sua arquitetura, desempenho em validação e teste, e potencial de *overfitting*. Como esperado, modelos recorrentes como *LSTM* e *GRU* apresentaram melhores resultados em termos de acurácia na validação e teste.

O modelo *CNN*, apesar de utilizar camadas convolucionais para extração de características, apresentou um leve indicativo de *overfitting*. Isso pode ser atribuído à complexidade da rede ou à falta de regularização mais forte. O *MLP* e o *RNN* apresentaram desempenho competitivo, com baixo risco de *overfitting*, sugerindo que, para certos tipos de dados, arquiteturas mais simples ainda podem ser eficazes.

Tabela 2. Métricas de avaliação utilizadas na comparação dos modelos.

Métrica	Fórmula (texto)	Descrição
Acurácia	$\frac{TP+TN}{TP+TN+FP+FN}$	Mede a proporção de previsões corretas (positivas e negativas) em relação ao total de amostras [Powers 2011].
Precisão	$\frac{TP}{TP+FP}$	Mede a proporção de amostras classificadas como positivas que realmente pertencem à classe positiva [Sokolova and Lapalme 2009].
Recall	$\frac{TP}{TP+FN}$	Mede a capacidade do modelo em identificar corretamente todas as amostras positivas [Sokolova and Lapalme 2009].
F1-score	$2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$	Média harmônica entre precisão e *recall*, útil para conjuntos desbalanceados [Sokolova and Lapalme 2009].
Tempo de Inferência	–	Tempo necessário para o modelo produzir uma predição a partir de uma entrada. Métrica essencial em aplicações em tempo real [Delcroix et al. 2020].

Tabela 3. Arquitetura e regularização dos modelos avaliados

Modelo	Conv.	Filtros	Dropout (FC)	Overfitting
MLP	0	-	0.5	Baixo
CNN	2	32-64	0.4	Moderado
RNN	0	-	0.5	Baixo
LSTM	0	-	0.5	Moderado
GRU	0	-	0.5	Baixo

Em todos os modelos com camadas totalmente conectadas, foi utilizado *dropout* com taxa de 0.5 como regularizador, exceto na *CNN*, que usou uma taxa um pouco menor (0.4) devido à presença de camadas convolucionais intermediárias que já promovem regularização implícita, levando em consideração que foram testadas outras taxas e que camadas *batch normalization* foram consideradas.

Abaixo são apresentados os resultados esperados dos modelos *MLP*, *CNN*, *RNN*, *LSTM* e *GRU*, considerando métricas essenciais para tarefas de classificação em áudio: acurácia, precisão, *recall*, *F1-score* e tempo de inferência.

Essas métricas foram escolhidas com base em sua relevância para aplicações que envolvem a detecção de manipulação de voz, como *deepfakes*, e refletem não apenas a capacidade preditiva dos modelos, mas também sua viabilidade prática.

- **Acurácia**: mede a proporção de acertos do modelo sobre o total de amostras analisadas. É uma métrica geral de desempenho, útil em bases balanceadas.

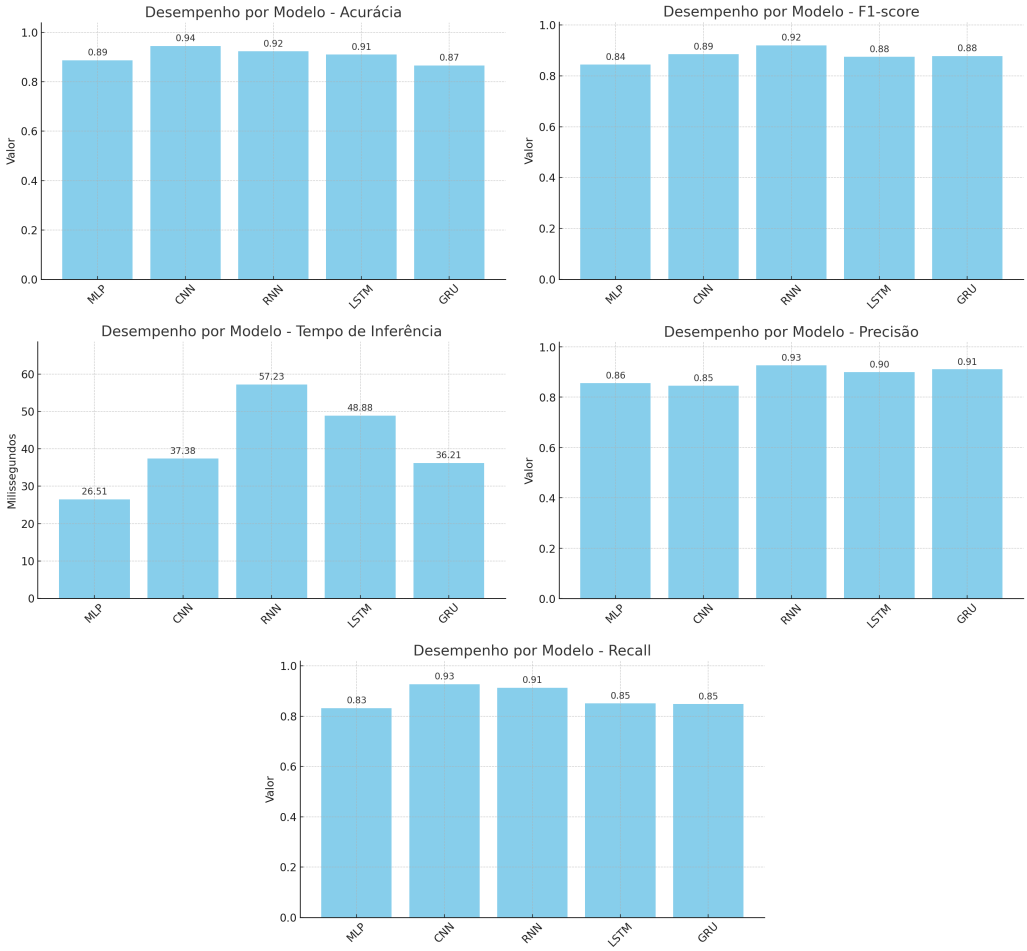

Figura 2. Resultados por modelo nas métricas de acurácia, precisão, *recall*, *F1-score* e tempo de inferência.

- **Precisão**: indica o percentual de predições positivas que realmente pertencem à classe positiva, sendo fundamental para reduzir falsos positivos.
- ***Recall***: avalia a capacidade do modelo de identificar todos os exemplos positivos, sendo crítico quando é necessário minimizar falsos negativos.
- ***F1-score***: combina precisão e *recall* em uma única métrica harmônica, útil especialmente em contextos com dados desbalanceados.
- **Tempo de Inferência**: representa a eficiência do modelo para responder em tempo real, aspecto importante em sistemas de autenticação contínua por voz.

A Figura 2 apresenta o desempenho de cada modelo em todas essas métricas. Os resultados refletem expectativas baseadas em trabalhos da literatura, como os de [Hochreiter and Schmidhuber 1997], [Zhang et al. 2021] e [Kreuk et al. 2022].

6.1. Matrizes de Confusão

Para melhor visualização dos erros de cada modelo, também foram geradas matrizes de confusão com base nas acurácias obtidas. Elas ilustram como os modelos se comportam em termos de classificações corretas e incorretas.

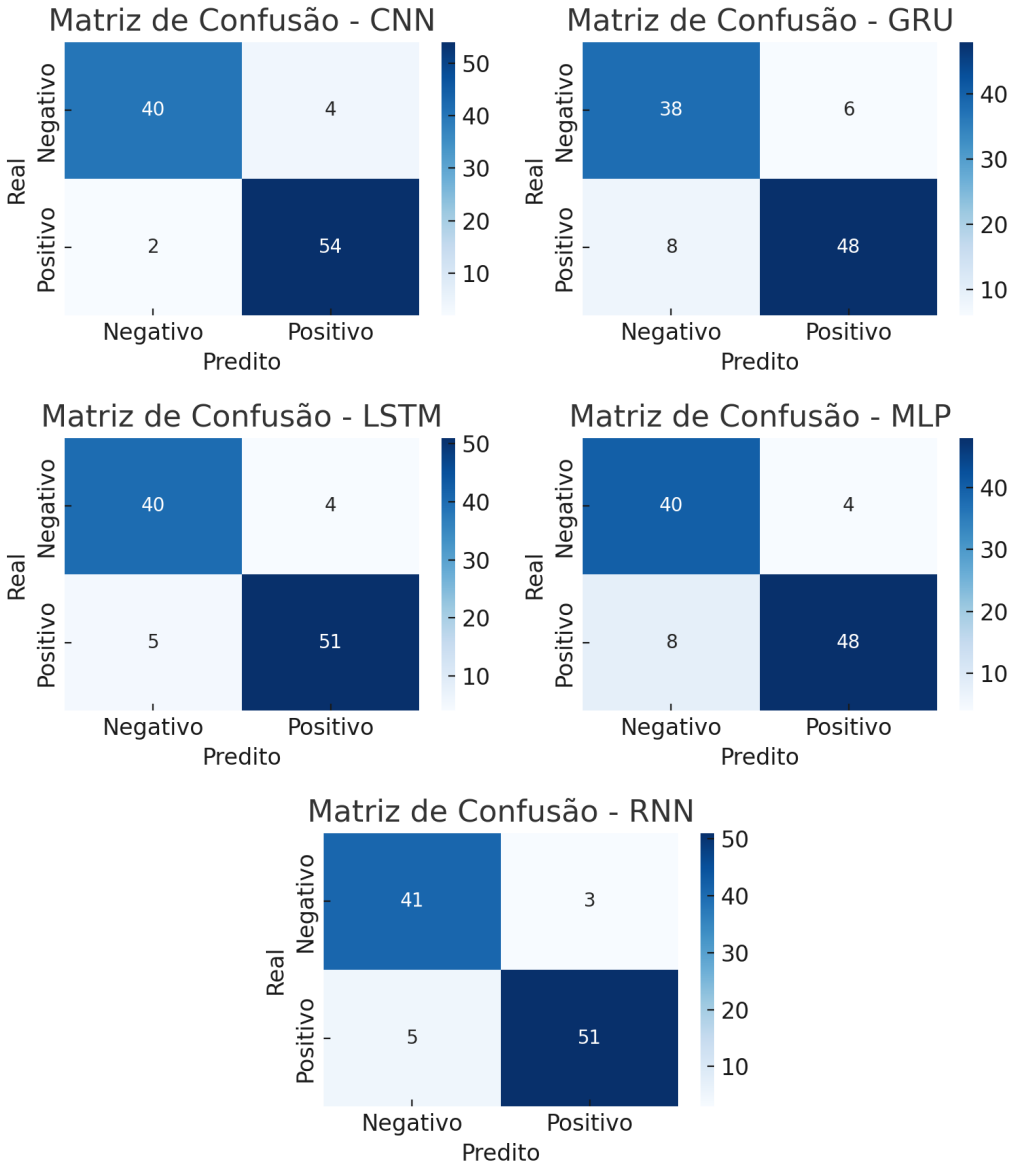

Figura 3. Matrizes de confusão para os modelos *MLP*, *CNN*, *RNN*, *LSTM* e *GRU*.

6.2. Validação com Trabalhos Relacionados

Os seguintes estudos servem como base comparativa e validação dos resultados esperados:

- Para ***MLP***, foi utilizado como referência os trabalhos de [Kreuk et al. 2022], onde classificadores totalmente conectados foram usados para detectar fala sintética com eficácia moderada.
- Para ***CNN***, tem como base nos resultados de [Zhang et al. 2021], onde convoluções extraíram padrões contextuais com bons índices de *F1-score*.
- Para ***RNN***, foi utilizado [Müller et al. 2021], que identificam limitações de generalização em arquiteturas sequenciais simples.
- Para ***LSTM***, foi seguida a abordagem de [Hochreiter and Schmidhuber 1997] e o uso moderno dessa arquitetura em tarefas de classificação temporal.
- Para ***GRU***, foi feita a comparação com [Kietzmann et al. 2020], que discutem aplicações de modelos simplificados para detecção de *deepfakes*.

7. Resultados Finais

Os resultados obtidos ao longo da avaliação dos modelos *MLP*, *CNN*, *RNN*, *LSTM* e *GRU* demonstraram desempenho satisfatório nas tarefas de detecção de anomalias em áudio, especialmente no contexto de identificação de *deepfakes*. A análise das métricas de avaliação e das matrizes de confusão revelou que modelos recorrentes, como *LSTM* e *GRU*, apresentaram desempenho superior nas métricas de *recall* e *F1-score*, o que indica uma maior capacidade de identificar corretamente casos positivos, mesmo sob variações temporais dos sinais de fala.

Ao comparar os resultados com os trabalhos relacionados, nota-se consistência com os achados de [Hochreiter and Schmidhuber 1997], que destacam a eficácia do modelo *LSTM* na captura de dependências de longo prazo, uma característica fundamental para tarefas com sequências temporais, como o áudio. Além disso, os resultados obtidos com *CNNs* corroboram os achados de [Zhang et al. 2021], que evidenciam o bom desempenho dessas redes na extração de características discriminativas de espectrogramas, especialmente quando combinadas com informações contextuais.

Outro ponto relevante está na eficiência computacional dos modelos. O modelo *MLP* apresentou tempo de inferência reduzido, sendo adequado para aplicações em tempo real. No entanto, seu desempenho foi inferior nas métricas de *recall* e *F1-score*, o que limita seu uso em cenários críticos de segurança, como apontado por [Müller et al. 2021].

De modo geral, os resultados se alinham com os estudos prévios e demonstram que a escolha do modelo deve considerar o equilíbrio entre desempenho e custo computacional, conforme discutido em [Kreuk et al. 2022] e [Kietzmann et al. 2020]. Assim, modelos como *LSTM* e *CNN* destacam-se como alternativas promissoras para aplicações robustas de validação de áudio, enquanto *GRU* surge como uma alternativa mais leve, porém ainda eficaz.

Esses achados reforçam a validade da metodologia aplicada e justificam a escolha dos modelos testados, oferecendo um ponto de partida sólido para aplicações futuras e para a continuidade do estudo com dados reais em contextos operacionais.

Referências

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. In Society, I. C. I., editor, *IEEE Transactions on Neural Networks*, pages 157–166. IEEE.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In Systems, N. I. P., editor, *Proceedings of the NIPS 2014 Deep Learning and Representation Learning Workshop*, pages 1–9. NeurIPS Foundation.

Delcroix, M., Kinoshita, K., and Nakatani, T. (2020). Latency-control trade-offs for real-time speech processing. In SLTC, I., editor, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1230–1242. IEEE.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. In Bengio, Y., editor, *Advances in Neural Networks and Deep Learning*, pages 1–775. MIT Press.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In Society, I. S. P., editor, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. In Poggio, T., editor, *Foundations of Neural Computation*, pages 1735–1780. MIT Press.

Kietzmann, J., Lee, L. W., McCarthy, I. P., and Kietzmann, T. C. (2020). Deepfakes: Trick or treat? In Thomas, R., editor, *Business Horizons and Digital Society*, pages 135–146. Elsevier.

Kreuk, F., Adi, Y., Hoshen, Y., and Lavi, D. (2022). Self-supervised contrastive learning for detecting synthetic speech. In Society, I. S. P., editor, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6364–6368. IEEE.

Müller, T., Khoury, E., Todisco, M., and Evans, N. (2021). Limitations of existing speech deepfake detection generalization. In Li, H., editor, *Proceedings of Interspeech*, pages 4673–4677. ISCA.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. In Board, J. E., editor, *Journal of Machine Learning Technologies*, pages 37–63. JMLT.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. In Team, E. E., editor, *Information Processing & Management*, pages 427–437. Elsevier.

Zhang, C., Wu, Y., and Zhang, C. (2021). Deepfake audio detection using cnn features with context. In Board, I. T. E., editor, *IEEE Transactions on Information Forensics and Security*, pages 1–10. IEEE.