risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 10/06/2025
Aceitação/Acceptance: 30/08/2025

Análise de metodologias para classificação de deepfakes

Guilherme Veiga Santos Pinto[1], Rudimar Luiz Scaranto Dazzi[2], Anita Maria da Rocha Fernandes[3]

guilhermeveigarj@edu.univali.br; rudimar@univali.br; anita.fernandes@univali.br

[1,2,3] Mestrado em Computação Aplicada, Universidade do Vale do Itajaí, R. Uruguai, 458 - Centro, Itajaí, CEP 88302-901, Itajaí, SC, Brasil.

DOI: 10.17013/risti.59.21–35

Resumo: Este trabalho investiga a crescente sofisticação dos deepfakes e a necessidade de aprimorar os métodos de detecção. O objetivo é avaliar o desempenho de diferentes modelos de Deep Learning na classificação de vídeos manipulados, buscando uma acurácia mínima de 80%. Foram analisadas as arquiteturas LSTM, CNN, GRU e CvT, além da aplicação da Lei de Benford como método estatístico auxiliar. A base Celeb-DF V2 foi utilizada por sua fidelidade e complexidade. O modelo LSTM apresentou a melhor acurácia no treinamento, enquanto o GRU obteve melhor desempenho na avaliação prática. Já o CvT mostrou limitações tanto em desempenho quanto em custo computacional. A Lei de Benford mostrou-se promissora, mas inconclusiva sem um parâmetro de correlação. Os resultados reforçam a importância de atualizar continuamente os mecanismos de detecção para acompanhar a evolução dos vídeos sintéticos e mitigar riscos associados à desinformação.

Palavras-chave: Detecção de DeepFake; Manipulação de Vídeo; Aprendizado Profundo; Redes Neurais; Lei de Benford.

Analysis of Deepfake Classification Methodologies

Abstract: This study investigates the increasing sophistication of deepfakes and the need to improve detection techniques. The objective is to evaluate the performance of different Deep Learning models in classifying manipulated videos, aiming for a minimum accuracy of 80%. The architectures LSTM, CNN, GRU, and CvT were analyzed, along with the application of Benford's Law as a statistical aid. The Celeb-DF V2 dataset was selected for its realism and complexity. The LSTM model achieved the highest training accuracy, while GRU showed superior performance during evaluation. The CvT model underperformed in both accuracy and computational cost. Although promising, Benford's Law yielded inconclusive results without a correlation metric. The findings highlight the importance of continually updating detection methods to keep pace with synthetic video advancements and mitigate risks linked to misinformation.

Keywords: DeepFake Detection; Video Manipulation; Deep Learning; Neural Networks; Benford's Law.

1. Introdução

O avanço das redes neurais artificiais tem impulsionado significativamente a manipulação de conteúdos multimídia, como imagens e vídeos. Ferramentas de inteligência artificial amplamente acessíveis, como FaceApp, Stable Diffusion WebUI e Roop-Unleashed, permitem a criação de deepfakes cada vez mais realistas, facilitando a produção de vídeos sintéticos de alta fidelidade com poucos recursos técnicos (Rana et al., 2022). Embora frequentemente utilizadas de forma recreativa, essas tecnologias vêm sendo empregadas em golpes e fraudes sofisticadas, como revelado por reportagens da BBC e por autoridades de Hong Kong, que relataram prejuízos causados por chamadas de vídeo falsas utilizando técnicas de FaceSwap em tempo real (Machado, 2024; Chen & Magramo, 2025). Tais avanços têm como base técnica o uso de Redes Adversárias Generativas (GANs), que combinam redes generativas e discriminativas para criar conteúdos altamente realistas (Goodfellow et al., 2014; Oberoi, 2018; Hui, 2018).

Nesse contexto, torna-se essencial questionar a eficácia dos métodos de detecção e classificação presentes na literatura científica. Apesar da rápida evolução das pesquisas em resposta às ameaças à privacidade e à segurança da informação, os modelos atuais precisam ser constantemente reavaliados diante do aumento do realismo e da acessibilidade das ferramentas de criação de deepfakes. Técnicas como Machine Learning, Deep Learning e Blockchain vêm sendo amplamente estudadas (Rana et al., 2022), mas mesmo com bons resultados estatísticos, há dúvidas quanto à sua robustez frente aos vídeos sintéticos mais recentes e sofisticados.

Este estudo propõe avaliar diferentes abordagens de classificação de deepfakes, utilizando o Deep Learning (DL) como base principal. Foram testados modelos como Long Short-Term Memory (LSTM) com arquitetura ResNext50, Convolutional Neural Network (CNN) baseada na arquitetura Xception, arquitetura Gated Recurrent Unit (GRU), Convolutional Vision Transformer (CvT), além da aplicação da Regra de Benford para análise de padrões numéricos. A base de dados utilizada foi a Celeb-DF V2, composta por vídeos reais e manipulados de celebridades (Li et al., 2019). O objetivo é investigar se essas técnicas ainda oferecem precisão na distinção entre vídeos autênticos e manipulados, considerando os desafios impostos pelas falsificações modernas e hiper-realistas.

O restante do artigo encontra-se estruturado nas seguintes seções: a seção 2 apresenta a fundamentação teórica; a seção 3 trata dos trabalhos relacionados; a seção 4 descreve o desenvolvimento do trabalho; a seção 5 apresenta os resultados do trabalho; e por fim a seção 6 traz as conclusões da pesquisa e os trabalhos futuros.

2. Fundamentação Teórica

Esta seção apresenta uma síntese dos principais temas que fundamentaram a elaboração deste estudo.

2.1. Deepfakes

O termo “deepfake” refere-se a mídias sintéticas geradas por técnicas avançadas de aprendizado profundo, especialmente GANs. Desde sua introdução, a tecnologia

evoluiu significativamente, tornando-se cada vez mais difícil distinguir conteúdos reais de sintéticos. De acordo com Pei et al. (2024), essas redes consistem em dois modelos neurais: o gerador, que cria conteúdo falso, e o discriminador, que avalia a autenticidade deste conteúdo. O treinamento contínuo entre esses modelos resulta em mídias altamente realistas, dificultando a distinção entre o real e o falso.

Recentemente, modelos de difusão têm emergido como uma alternativa promissora na geração de deepfakes. Esses modelos aprendem a transformar dados de uma distribuição simples para uma complexa, permitindo a criação de conteúdos ainda mais convincentes. Ferramentas como o Stable Diffusion exemplificam essa abordagem, oferecendo resultados impressionantes na síntese de imagens e vídeos. Além das GANs e dos modelos de difusão, técnicas de transferência de estilo e mapeamento de características faciais são empregadas para aprimorar a qualidade dos deepfakes. Essas técnicas permitem a substituição de rostos em vídeos, mantendo expressões faciais e movimentos labiais coerentes com o conteúdo original. O Roop-Unleashed é um exemplo de ferramenta que utiliza essas metodologias para criar vídeos sintéticos de alta fidelidade (Pei et al., 2024).

Com o avanço das técnicas e o aumento da capacidade computacional, a geração de deepfakes tornou-se mais acessível, permitindo que indivíduos sem expertise técnica produzam conteúdos falsificados com relativa facilidade. A evolução contínua dessas tecnologias levanta preocupações sobre a capacidade dos modelos atuais de detecção em identificar deepfakes. À medida que os métodos de geração se tornam mais sofisticados, é de extrema importância desenvolver técnicas de detecção igualmente avançadas para mitigar os riscos associados à disseminação de mídias sintéticas maliciosas (Lee et al., 2024).

2.2. Fundamentos e Arquiteturas de Deep Learning

O DL é um subcampo do Machine Learning baseado em redes neurais artificiais, com estrutura inspirada no cérebro humano, e tornou-se um pilar da inteligência artificial moderna graças ao avanço computacional e à abundância de dados (Goodfellow, Bengio, & Courville, 2016). Essas redes aprendem representações hierárquicas dos dados por meio de múltiplas camadas, como destacam LeCun, Bengio e Hinton (2015), utilizando métodos como o Stochastic Gradient Descent para otimizar funções de perda. As principais arquiteturas incluem as CNNs, as Redes Recorrentes (RNNs) e os Transformers, todas amplamente aplicadas em tarefas como classificação de imagens, reconhecimento de fala e processamento de linguagem natural.

Dentre as RNNs, destaca-se a arquitetura LSTM, proposta por Hochreiter e Schmidhuber (1997), que foi desenhada para resolver o problema do vanishing gradient e modelar dependências sequenciais de longo prazo. A LSTM realiza este controle através de uma célula de memória e três mecanismos de portas (entrada, esquecimento e saída), permitindo seu uso em tarefas que exigem a retenção seletiva de informações em sequências extensas (Gers, Schmidhuber, & Cummins, 2000; Graves, 2012). Para simplificar essa estrutura, surgiram variantes como as GRUs, que alcançam desempenho comparável com um número reduzido de parâmetros (Cho et al., 2014). As GRUs utilizam apenas duas portas (atualização e reinicialização) para controlar o fluxo de

informações, oferecendo uma alternativa mais leve e eficiente para tarefas sequenciais (Chung, Gulcehre, Cho, & Bengio, 2014; Bahdanau, Cho, & Bengio, 2014).

As CNNs, por sua vez, são fundamentais no processamento de imagens, utilizando camadas convolucionais que extraem características como bordas e texturas (LeCun, Bottou, Bengio, & Haffner, 2002). Sua estrutura é composta por camadas convolucionais, pooling e totalmente conectadas, com funções de ativação como ReLU (Simonyan & Zisserman, 2014). Para tarefas de detecção de deepfakes, que exigem alta precisão e eficiência, destacam-se arquiteturas avançadas como a Xception (Extreme Inception). Este modelo substitui as convoluções padrão por convoluções separáveis em profundidade, que desacoplam a filtragem espacial e a combinação de canais, reduzindo a complexidade computacional e otimizando o fluxo de informação por meio de conexões residuais (Chollet, 2017).

Combinando o melhor das CNNs e dos Transformers, a arquitetura CvT propõe uma abordagem híbrida para processar imagens, integrando convoluções no processo de tokenização e nos mecanismos de atenção (Wu et al., 2021). A inovação central do CvT reside na utilização do Convolutional Token Embedding e em blocos hierárquicos. Esta estrutura permite a extração eficiente de padrões locais antes da autoatenção e a modelagem global de dependências, proporcionando um modelo mais robusto e eficiente em comparação com modelos puramente baseados em Vision Transformer (Dosovitskiy et al., 2020; Tolstikhin et al., 2021).

2.3. Regra de Benford

A Regra de Benford, também conhecida como Regra do Primeiro Dígito, descreve a distribuição não uniforme dos primeiros dígitos em conjuntos de dados numéricos naturais. Segundo Benford (1938), essa distribuição indica que o dígito 1 tende a aparecer como primeiro dígito com maior frequência do que o dígito 9. Essa característica é expressa por (1) que determina a probabilidade de cada dígito entre 1 e 9 aparecer na primeira posição. Hill (1995) complementa afirmando que cerca de 30,1% dos números legítimos começam com o dígito 1, enquanto apenas 4,6% iniciam com o dígito 9. Essa distribuição é comum em dados que abrangem múltiplas ordens de magnitude, com crescimento exponencial, sendo aplicável a áreas como contabilidade, finanças e ciências naturais.

$$P(d) = log_{10}\,(1 + \frac{1}{d}) \tag{1}$$

A aplicação da Regra de Benford na detecção de fraudes se baseia na ideia de que conjuntos de dados manipulados tendem a se desviar dessa distribuição esperada. Nigrini (1996) destaca que padrões suspeitos podem indicar manipulações, sendo essa abordagem utilizada em auditorias de relatórios financeiros, registros contábeis e dados eleitorais (Durtschi et al., 2004). Nigrini e Mittermaier (1997) apontam que empresas fraudulentas frequentemente apresentam relatórios com distribuições anômalas, tornando a Regra de Benford uma ferramenta valiosa para auditores. No entanto, Hill (1998) adverte que a regra não se aplica a todos os conjuntos de dados, especialmente os

compostos por valores fixos, limitados ou artificialmente definidos, exigindo assim o uso complementar de outras técnicas estatísticas.

Na literatura recente, a Regra de Benford também vem sendo aplicada à detecção de manipulações em imagens digitais. Vishnu (2021) descreve uma metodologia que envolve a conversão dos pixels da imagem para o domínio da frequência por meio da Transformada Discreta de Cosseno (DCT). Em seguida, extrai-se o primeiro dígito dos coeficientes da DCT e verifica-se, por meio de uma visualização gráfica, se a distribuição desses dígitos segue o padrão esperado da Regra de Benford.

3. Trabalhos Relacionados

Esta seção apresenta a síntese dos sete trabalhos analisados que exploram metodologias de DL e técnicas estatísticas aplicadas à detecção de deepfakes. A seleção destes estudos foi definida por um processo sistemático de busca em bases de dados bibliográficas. As strings de pesquisa foram construídas para identificar publicações recentes (2018-2025) focadas nas arquiteturas de redes neurais (RNNs, CNNs, CvT) e no método da Regra de Benford para classificação de deepfakes. A aplicação de critérios de inclusão e exclusão resultou no conjunto de sete artigos, que servem de base para a comparação de desempenho e custo computacional.

Os estudos de Kularkar et al. (2023) e Mogulkoc e Yuce (2024) enfatizam modelos que combinam aprendizado profundo. Kularkar propõe uma abordagem híbrida com ResNext e LSTM, alcançando 94,09% de precisão ao analisar quadros de vídeo para capturar inconsistências sutis. Já Mogulkoc e Yuce comparam CNN, LSTM e GRU, observando melhor desempenho com GRU (85% de acurácia), o que destaca a importância do uso de modelos capazes de explorar as dependências temporais em vídeos manipulados.

Complementando essa linha de pesquisa, os trabalhos de Wodajo e Atnafu (2021, 2023) investigam o uso de arquiteturas baseadas em Transformers. Em 2021, os autores introduzem o CvT, que combina CNN e Vision Transformer (ViT), atingindo 91,5% de acurácia. Em 2023, evoluem para o GenConViT, uma arquitetura mais complexa que incorpora Autoencoders e Swin Transformers, alcançando impressionantes 95,8% de acurácia e excelente generalização em diversos datasets. Ambos os trabalhos reforçam a tendência atual de utilização de arquiteturas híbridas com mecanismos de atenção para melhor desempenho.

Por outro lado, três estudos analisam a aplicação da Lei de Benford como ferramenta estatística para detecção de manipulações em imagens digitais e deepfakes. Fernandes e Antunes (2023) utilizam a Transformada Rápida de Fourier (FFT) e a métrica Cramer-Von Mises (CVM), para extrair padrões numéricos de imagens e obtêm um F1-score médio de 90,47%, com destaque para a correlação de Pearson como melhor métrica. Já Vishnu (2021) emprega DCT e DFT para transformar as imagens e compara várias métricas de desvio estatístico, apontando a DCT como mais eficaz. Bonettini et al. (2020) ampliam essa abordagem ao utilizar a DCT em imagens divididas em blocos e extrair vetores de características alimentados em classificadores Random Forest, alcançando acurácia superior a 99% em grandes conjuntos de imagens de GANs.

Em resumo, os trabalhos revisados se agrupam em duas grandes vertentes: modelos de aprendizado profundo com redes neurais (CNN, LSTM, GRU, Transformers) e abordagens estatísticas com a Lei de Benford. Os primeiros apresentam alta acurácia, mas com maior custo computacional e necessidade de treinamento robusto. Já os métodos baseados em Benford destacam-se pela eficiência com recursos limitados. A tabela 1 a seguir detalha essas diferenças em termos de abordagem, datasets, desempenho e limitações.

Trabalho	Abordagem	Desempenho	Dataset
Kularkar et al. (2023)	LSTM	94,09%	*FaceForensics++, CelebDF* e DFDC
Mogulkoc e Yuce (2024)	CNN, LSTM e GRU	85% (GRU) 83% (LSTM) 60% (CNN)	DFDC
Wodajo e Atnafu (2021)	CvT	91,5%	DFDC
Wodajo e Atnafu (2023)	*GenConViT*	98,5% (DFDC) 98,28% (TIMIT) 97% (FF++) 90,94% (Celeb-DF V2)	DFDC, *DeepfakeTIMIT, FaceForensics++ e Celeb-DF v2*
Fernandes e Antunes (2023)	Regra de Benford, FFT e *Cramer-Von Mises* (CVM)	90,47%	Dataset do próprio autor.
Vishnu (2021)	Regra de Benford, DCT e FFT	Imagens manipuladas apresentaram 0,0173 de desvio (distância euclidiana), contra 0,0175 para imagens originais.	Dataset do próprio autor.
Bonettini et al. (2020)	Regra de Benford e DCT	87.96%	Datasets gerados via *Cycle-Gan* e *ProGAN*
Este Trabalho	CNN, LSTM, GRU, CvT e Regra de Benford	97,32%	*Celeb-DF v2*

Tabela 1 – Análise comparativa dos trabalhos relacionados.

Os estudos revisados agrupam-se em duas vertentes: modelos de aprendizado profundo, que demonstram alta acurácia com maior custo computacional, e a Lei de Benford, que oferece eficiência, mas cuja aplicação requer métricas de correlação para a validação dos resultados. Observa-se, no entanto, a ausência de uma avaliação comparativa direta destas arquiteturas e do método estatístico empregados no mesmo dataset de alta fidelidade (Celeb-DF V2) e sob idênticas condições de teste. O objetivo deste trabalho é preencher esta lacuna, estabelecendo um *benchmark* de desempenho entre estas metodologias na tarefa de classificação de deepfakes.

4. Análise Do Celeb-DF v2 e De Outros Datasets Disponíveis

A eficácia na detecção de deepfakes depende fortemente de datasets com alta qualidade e diversidade. Entre as bases mais relevantes, o Celeb-DF v2 se destaca pelo realismo

nas trocas faciais de celebridades, preservando expressões, iluminação e movimentos naturais. Como ilustrado na Figura 1, essas características tornam o dataset especialmente desafiador para os modelos de detecção, sendo ideal para simular cenários do mundo real.

Além de sua qualidade visual, o Celeb-DF v2 apresenta variedade de rostos, gêneros e contextos, o que contribui para o treinamento de modelos mais robustos e generalistas. Ao todo, a base contém 5.639 vídeos manipulados e 890 vídeos reais, extraídos de vídeos públicos do YouTube, totalizando mais de 2 milhões de frames com cerca de 10 segundos cada, no formato MP4.

Comparativamente, o DFDC, criado para o desafio promovido pelo Facebook, oferece grande volume e diversidade de manipulações, mas com qualidade facial inferior ao Celeb-DF v2, como também demonstrado na Figura 1. Já o FaceForensics++ inclui manipulações por autoencoders e GANs em diversos níveis de compressão, sendo útil para testar robustez, embora com realismo inferior, conforme observado na Figura 2.

Figura 1 –Comparativo da base de dados Celeb-DF v2 a esquerda, com a base de dados DFDC a direita. Adaptado de Li et al. (2020) e Dolhansky et al. (2020).

Figura 2 – Comparativo visual entre o FaceForensics++ e o Celeb-DF v2. Reimpresso de *Visual attention-based deepfake video forgery detection* por Ganguly et al., 2022.

Diante dessas considerações, o Celeb-DF v2 foi escolhido como principal base deste estudo por oferecer o melhor equilíbrio entre fidelidade visual e diversidade, sendo mais adequado para treinar e avaliar modelos de detecção em contextos realistas e complexos.

4.1. Pré-processamento e Treinamento

Após a escolha da base de dados, iniciou-se o pré-processamento. Os vídeos foram convertidos em trechos de quatro segundos, com resolução padronizada de 224x224 pixels, por meio de recorte facial (*face crop*) utilizando a biblioteca *face recognition*. Essa etapa garantiu foco nos rostos, facilitando o treinamento dos modelos.

Vídeos corrompidos e com áudio foram removidos, por não contribuírem para os objetivos do estudo. Com os dados limpos, os vídeos foram carregados na memória do ambiente de execução para início do treinamento.

O dataset Celeb-DF V2 apresentava desequilíbrio entre classes, com predominância de vídeos manipulados. Para evitar viés de classificação, foi aplicado balanceamento, igualando-se a quantidade de vídeos reais e sintéticos, totalizando 1.869 vídeos (50% reais e 50% falsos), conforme recomendações de Krishna et al. (2022).

O treinamento com os modelos ResNext50 (LSTM), Xception (CNN) e GRU foi realizado no Google Colab, com GPU NVIDIA Tesla T4. Os vídeos foram extraídos do Google Drive e processados com OpenCV, resultando em média de 148 frames por vídeo. As classes foram definidas com base em um arquivo CSV contendo os nomes e rótulos dos vídeos (REAL ou FAKE).

A divisão do conjunto de dados foi feita com a biblioteca *Scikit-learn*: 80% para treinamento, 10% para teste e 10% para validação, totalizando 1.495 vídeos para treino e 374 para teste e validação. Os modelos foram treinados por 20 épocas, com taxa de aprendizado de 0.00001. Em geral, os treinamentos duraram entre duas e quatro horas, considerando o tempo de carregamento dos vídeos (cerca de 40 minutos).

Para o modelo Convolutional Vision Transformer (CvT), houve uma diferença metodológica: a extração dos frames e a separação dos dados foram feitas manualmente, mantendo as proporções de 80/10/10. Isso resultou em um total de 276.460 imagens. O treinamento foi realizado no Google Colab com GPU NVIDIA Tesla A100, com duração aproximada de oito horas, sendo o processo mais demorado entre os modelos avaliados.

5. Resultados

Conforme especificado na seção 4.1, quatro diferentes algoritmos de DL foram treinados com uma distribuição de dados de 80% para treinamento, 10% para validação e 10% para teste. Os modelos treinados consistiram no modelo LSTM com a arquitetura ResNext50, o modelo CNN com a arquitetura Xception2, que incorpora uma variante da arquitetura Xception, o modelo GRU com a arquitetura ResNet-50, e o modelo CvT.

A principal métrica avaliada foi a acurácia, que consiste na divisão das previsões corretas pelo número total de previsões conforme observado em (2). Em outras palavras, é a divisão da soma dos verdadeiros positivos (VP) e verdadeiros negativos (VN) pela soma dos verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN).

$$Acurácia\ AC = \frac{VP + VN}{\mathrm{VP+VN+FP+FN}} \tag{2}$$

Os modelos foram avaliados com base em sua matriz de confusão e acurácia, além de serem submetidos a um teste utilizando 10 vídeos reais e 10 vídeos sintéticos da base *Celeb-DF V2*, os quais não fizeram parte da etapa de treinamento. Nesta etapa de teste, é feita uma extração de um dos frames desses vídeos, retornando-o com um mapa de calor, e um valor de confiança de predição com o rótulo categorizado pelo modelo.

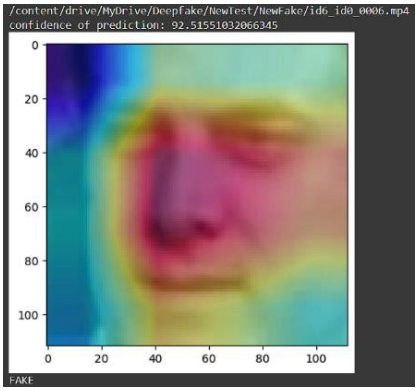

Figura 3 – Etapa de avaliação do modelo com vídeos separados da base.

No treinamento do modelo LSTM com a arquitetura ResNext50 por 20 épocas, foi alcançada uma acurácia máxima de 97,32% na fase de validação, com uma queda consistente nos valores de perda de treinamento e validação conforme a Figura 4. A matriz de confusão na Tabela 2 revelou que, entre 200 vídeos sintéticos e 174 vídeos reais, houve uma maior ocorrência de falsos negativos do que de falsos positivos. Na etapa de avaliação, o modelo demonstrou bom desempenho na classificação de frames, acertando todos os 10 vídeos sintéticos testados. No entanto, ao classificar os 10 vídeos reais, dois foram incorretamente rotulados como falsos positivos, conforme a Tabela 3.

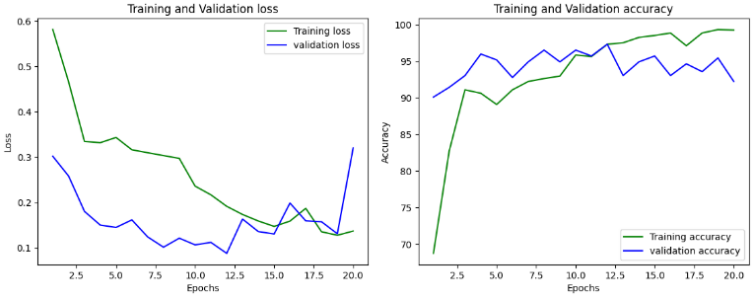

Figura 4 – Gráficos de Perda e de Acurácia para Treinamento e Validação da arquitetura LSTM.

Já no treinamento do modelo CNN com a arquitetura Xception2 por 20 épocas, foi registrada na Figura 5 uma acurácia máxima de 78,34% na validação, com picos de perda

variando entre 0,47 e 0,59. Na Tabela2, a matriz de confusão indicou uma quantidade maior de falsos negativos em comparação aos falsos positivos. Na fase de avaliação, o desempenho do modelo CNN foi menos consistente em relação ao modelo anterior, apresentando menor confiança nas predições, além da ocorrência de um verdadeiro negativo e dois falsos positivos, conforme a Tabela 3.

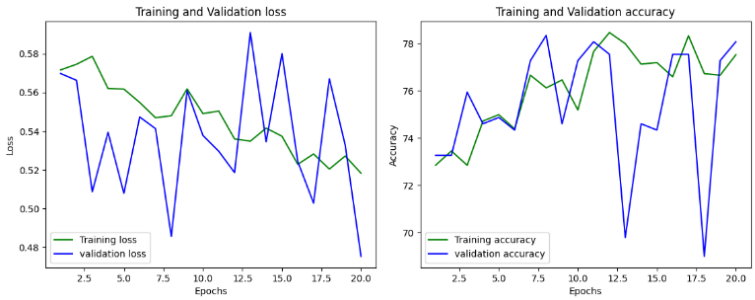

Figura 5 – Gráficos de Perda e de Acurácia para Treinamento e Validação da arquitetura CNN.

O modelo GRU, treinado por 20 épocas, alcançou uma acurácia máxima de 96,52% na validação e apresentou na Figura 6, uma redução consistente nos valores de perda de treinamento e validação, comportamento semelhante ao modelo LSTM. Sua matriz de confusão na Tabela 2 indicou uma leve predominância de falsos negativos sobre falsos positivos. Na etapa de avaliação, o GRU demonstrou desempenho destacado tanto nos níveis de confiança das predições quanto na taxa de acertos, registrando apenas um falso positivo na classificação de vídeos reais, Conforme a Tabela 3.

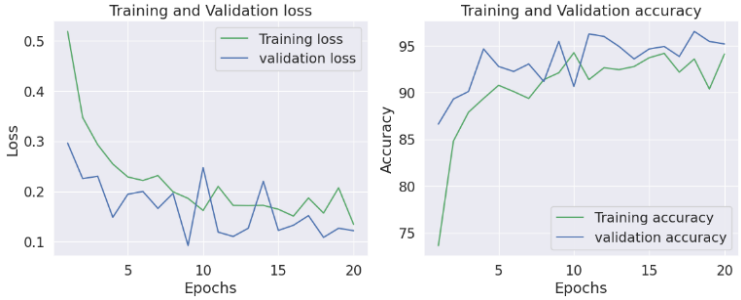

Figura 6 – Gráficos de Perda e de Acurácia para Treinamento e Validação da arquitetura GRU.

O modelo CvT, apesar de exigir mais tempo de treinamento em comparação aos demais, apresentou o pior desempenho entre os modelos avaliados, com uma acurácia de apenas 50,53% na classificação de deepfakes e perdas estagnadas durante o treinamento e validação. Conforme a Tabela 3 na etapa de avaliação com 10 vídeos reais e 10 sintéticos, observou-se que o modelo classificou todos os vídeos como deepfakes, utilizando uma abordagem baseada em valores de predição — abaixo de 0,5 para REAL e igual ou

acima de 0,5 para FAKE. Todos os valores de predição foram idênticos, indicando um comportamento inconsistente.

Classe	LSTM	CNN	GRU
Verdadeiro Positivo (VP)	186	189	198
Falso Positivo (FP)	3	11	2
Falso Negativo (FN)	26	71	16
Verdadeiro Negativo (VN)	159	103	158

Tabela 2 – Matriz de Confusão das arquiteturas LSTM, CNN e GRU.

Classe	Classe	Acertos	Erros	Tipo de Erro	Média da Confiança de Predição (%)
LSTM	FAKE	10	0	-	97,57%
LSTM	REAL	8	2	Falso Positivo	95,74%
CNN	FAKE	9	1	Falso Negativo	75,42%
CNN	REAL	8	2	Falso Positivo	71,10%
GRU	FAKE	10	0	-	98,49%
GRU	REAL	9	1	Falso Positivo	93,56%
CvT	FAKE	10	0	-	51,04%
CvT	REAL	0	10	Falso Positivo	51,04%

Tabela 3 – Desempenho dos modelos para classificação de vídeos.

Adicionalmente, foi realizada uma análise em frames de vídeos do dataset Celeb-DF v2 utilizando a Regra do Primeiro Dígito, ou Regra de Benford, com o objetivo de identificar possíveis manipulações. O processo envolveu a extração dos valores RGB das imagens, normalização e aplicação da Transformada Discreta de Cosseno (DCT), que converte os dados para o domínio da frequência, separando componentes de baixa e alta frequência. A partir dos coeficientes da DCT, foram extraídos os primeiros dígitos e analisada sua distribuição, comparando-a com o padrão teórico da Regra de Benford.

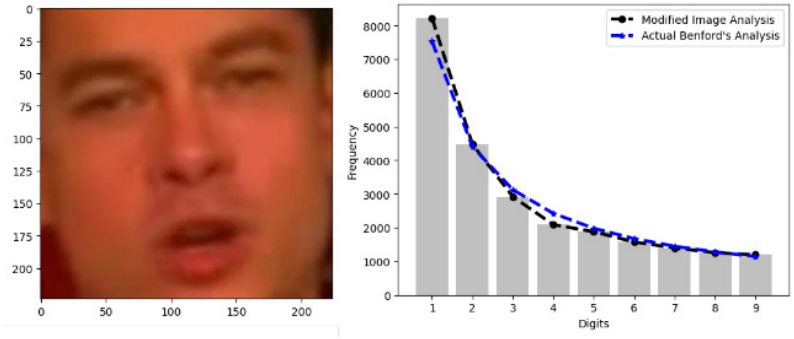

Figura 7 – Análise de um frame pertencente à classe FAKE, com variação à Regra de Benford.

A conformidade com a regra foi visualizada por meio de gráficos que comparam a frequência dos dígitos observados com a distribuição esperada. Como observado na Figura 7, imagens de vídeos manipulados apresentaram, em alguns casos, menor aderência à lei do que as imagens reais. No entanto, também ocorreram casos contrários, em que deepfakes se aproximaram mais da lei do que imagens reais, o que indica que os resultados ainda não são conclusivos sem o cálculo das distâncias entre as linhas nos gráficos. Além disso, foram realizadas análises com frames completos dos vídeos, sem focar apenas nos rostos.

6. Conclusões

Este trabalho abordou a crescente relevância da detecção de deepfakes avaliando a eficácia de diferentes modelos de DL e a aplicação da Regra de Benford no desafiador dataset Celeb-DF V2. Os resultados demonstram a alta eficácia das arquiteturas recorrentes no contexto de alta fidelidade do Celeb-DF V2, com o modelo LSTM (97,32% de acurácia) e o GRU (96,52%) superando o objetivo mínimo de 80%. Este desempenho é notavelmente superior ao encontrado em trabalhos relacionados que utilizaram arquiteturas similares em outros datasets (por exemplo, Kularkar et al., 2023, e Mogulkoc e Yuce, 2024) e, em comparação direta, superou a acurácia de modelos mais complexos como o GenConViT de Wodajo e Atnafu (2023) no mesmo dataset. Esta superioridade sugere que modelos DL especializados em sequências temporais, quando otimizados com backbones CNN, fornecem o desempenho mais robusto para a detecção de manipulações sutis em vídeos de alta qualidade.

Em contraste, o modelo CvT apresentou baixa acurácia (50,53%) e alto custo computacional, falhando em generalizar para a classe REAL. Essa queda abrupta de desempenho, contrastando com o resultado de 91,5% obtido por Wodajo e Atnafu (2021) no dataset DFDC, sugere que, embora os Vision Transformers sejam promissores, eles são particularmente sensíveis às características de alta complexidade do Celeb-DF V2, indicando que podem exigir maior volume de dados ou ajustes mais finos no treinamento para estabilização. Observou-se também uma tendência dos modelos recorrentes a encontrar mais falsos positivos (classificar real como falso) do que verdadeiros negativos.

Adicionalmente, a análise com a Regra de Benford mostrou-se promissora, mas ainda inconclusiva sem métricas estatísticas formais. A inconsistência observada, com deepfakes se aproximando da Lei em alguns casos, ressalta a advertência de Hill (1998) sobre a aplicabilidade limitada da regra. Embora outros trabalhos (Fernandes e Antunes, 2023; Bonettini et al., 2020) tenham obtido alta performance com o método, a ausência de um parâmetro de correlação clara aqui reforça a necessidade de integrar a análise com testes estatísticos robustos para validar o desvio.

Como trabalhos futuros, propõe-se testar os modelos com vídeos produzidos pelos próprios autores — respeitando aspectos éticos e legais — gerados via *Roop-Unleashed,* e aplicar métodos como a correlação de Pearson para aprimorar a análise estatística dessas manipulações.

Referências

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, 551-572.

Bonettini, N., Bestagini, P., Milani, S., & Tubaro, S. (2021). On the use of Benford's law to detect GAN-generated images. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5495-5502). IEEE.

Chen, H., & Magramo, K. (2025). *Funcionário de multinacional paga US$ 25 mi a golpista que usou deepfake para simular reunião*. CNN Brasil. https://www.cnnbrasil.com.br/internacional/funcionario-de-multinacional-paga-us-25-mi-a-golpista-que-usou-deepfake-para-simular-reuniao/

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, *5*(1), 17-34.

Fernandes, P., & Antunes, M. (2023). Benford's law applied to digital forensic analysis. *Forensic Science International: Digital Investigation*, *45*, 301515.

Ganguly, S., Ganguly, A., Mohiuddin, S., Malakar, S., & Sarkar, R. (2022). ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection. *Expert Systems with Applications*, *210*, 118423.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), 2451-2471.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks* (Vol. 385). Studies in Computational Intelligence.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical science*, 354-363.

Hill, T. P. (1998). The first digit phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist*, *86*(4), 358-363.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hui, J. (2018). How deep learning fakes videos (Deepfake) and how to detect it?. *Medium Corporation*, *28*.

Krishna, P., Arukala, S., Battula, V. R., & Sri, M. B. (2022). Deepfake Detection using LSTM and ResNext. *Journal of Engineering Sciences*, *13*(07).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kularkar, T., Jikar, T., Rewaskar, V., Dhawale, K., Thomas, A., & Madankar, M. (2023). Deepfake detection using LSTM and ResNext.

Lai, G., Chang, W. C., Yang, Y., & Liu, H. (2018). Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 95-104).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Lee, H., Lee, C., Farhat, K., Qiu, L., Geluso, S., Kim, A., & Etzioni, O. (2024). The Tug-of-War Between Deepfake Generation and Detection. *arXiv preprint arXiv:2407.06174*.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2019). Celeb-df (v2): a new dataset for deepfake forensics [j]. *arXiv preprint arXiv*.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).

Machado, S. (2024). *'Eram meu rosto e minha voz, mas era golpe': Como criminosos 'clonam pessoas' com inteligência artificial*. BBC News Brasil. https://www.bbc.com/portuguese/articles/cd1jv45dq3go

Mogulkoc, Z., & Yuce, B. (2024, 25 de março). *A comparative analysis of state-of-the-art algorithms for robust deep fake detection*. Abdullah Gul University. Acessado em 15 de fevereiro de 2025.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association*, *18*(1), 72.

Nigrini, M. J., & Mittermaier, L. J. (1997). The use of Benford's Law as an Aid in Analytical Procedures. *Auditing: A journal of practice & theory*, *16*(2).

Oberoi, G. (2018). *Exploring DeepF*akes. https://goberoi.com/exploring-deepfakes-20c9947c22d9

Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., & Tao, D. (2024). Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE access, 10, 25494-25513.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). Lstm neural networks for language modeling. In *Interspeech* (Vol. 2012, pp. 194-197).

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., & Dosovitskiy, A. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, *34*, 24261-24272.

Vishnu, U. (2021). *Deepfake detection using Benford's Law and Distribution Variance Statistic. International Research Journal of Engineering and Technology (IRJET)*, 8(10), 712–719. Acessado em 13 de fevereiro de 2025.

Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*.

Wodajo, D., Atnafu, S., & Akhtar, Z. (2023). Deepfake video detection using generative convolutional vision transformer. *arXiv preprint arXiv:2307.07036*.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22-31).

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*, 611-629.