Cross-Database in Deepfake Detection Based on a Convolutional Neural Network and Vision Transformer

Erikson Eler Ferreira[1], Jefferson Oliveira Andrade[2], Karin Satie Komati[3]
[123]Programa de Pós-graduação em Computação Aplicada (PPComp)
[123]Instituto Federal do Espírito Santo (IFES) Campus Serra
eriksonferreira12@gmail.com[1], jefferson.andrade@ifes.edu.br[2], kkomati@ifes.edu.br[3]

***Resumo*—The proliferation of Deepfake techniques has raised concerns due to their potential to generate misleading multimedia content, leading to ethical, social, and political implications. In response to this emerging issue, collaborative efforts between academia and leading technological entities have committed on developing robust detection methods. Initially, Convolutional Neural Networks (CNNs) were prominent, recently proposed methods, which combine features of CNNs with Vision Transformers (ViT) have shown improved performance. This research centers on evaluating the generalization capacity of these advanced models by subjecting them to cross-database tests with different datasets than those used in their training phases. Our analysis reveals that while both models perform well on known datasets, they face challenges related to overfitting when transitioning to new datasets. Consequently, this study underscores the need for further research in Deepfake detection, ensuring its adaptability and effectiveness in diverse scenarios.**

***Index Terms*—deepfakes, generalização, cnn, vit, overfitting**

I. Introdução

O termo *Deepfake* é uma amálgama de "deep learning" (DL, do inglês, aprendizado profunda) e "fake" (falso em inglês). *Deepfakes* é uma técnica de síntese de imagens ou sons humanos baseada em modelos de DL, que é muito usada para gerar conteúdos de multimídia falsos. Em vídeos, é muito usado para substituir a face de uma pessoa em um vídeo já existente, em termos de voz, o objetivo é clonar a voz de uma pessoa para criar um discurso inexistente [1].

Infelizmente, a técnica têm sido usada com o objetivo de enganar e manipular o público [2]. As aplicações maliciosas do *Deepfake* são evidentes no contexto de notícias falsas, em que a propagação rápida e ampla de conteúdos enganosos pode influenciar a opinião pública e prejudicar a estabilidade social [3]. Essa técnica levanta sérias preocupações éticas, políticas e sociais, uma vez que pode ser usada para difamar pessoas públicas, espalhar desinformação e influenciar processos democráticos. É imperativo desenvolver técnicas eficazes para a detecção de *Deepfakes*, a fim de combater a disseminação desse tipo de conteúdo enganoso.

Para combater esse problema, a comunidade científica e grandes empresas de tecnologia têm investido esforços no desenvolvimento de métodos de detecção de *Deepfake*. No início destes esforços, a partir de 2020, as abordagens mais promissoras indicavam o uso de CNNs (em inglês Convolutional Neural Networks, em português Redes Neurais Convolucionais), principalmente a EficientNet, chegando a alcançar a medida-F1 de 0,97 na base de dados DFDC (Deepfake Detection Challenge) [4].

Recentemente, no domínio da detecção de *Deepfakes*, têm sido empregadas abordagens híbridas que combinam redes CNN e Vision Transformers (ViT) [5], chamadas neste trabalho de CNN+ViT. O ViT é uma arquitetura de rede neural que aplica a abordagem Transformer [6], originalmente desenvolvida para processamento de linguagem natural, para tarefas de reconhecimento de imagem. Ele divide a imagem em *patches*, os transforma em sequências de entrada e utiliza mecanismos de auto-atenção para capturar as relações entre os *patches* [7].

Os resultados quantitativos apresentados em artigos sobre *DeepFake* em vídeo alcançam métricas promissoras. No entanto, como seria o resultado frente à vídeos de uma base de dados diferente daquela que o modelo foi treinado? Assim, a pergunta deste estudo é: "Qual é a capacidade de generalização do modelo treinado?".

A pergunta de pesquisa, avalia qual é o sobre-ajuste (do inglês *overfitting*) do modelo. O sobre-ajuste é um termo quando um modelo se ajusta muito bem ao conjunto de dados anteriormente observado, mas não se mostra eficaz para prever novos resultados [8].

Para responder à pergunta, a proposta é: (i) testar em uma base de dados C, um Modelo A baseado em CNN treinado em uma base de dados A e (ii) testar um Modelo B em uma base de dados C, baseado em CNN+ViT treinado em uma base de dados B. Com essa proposta de teste em base de dados cruzada, podemos avaliar se os modelos, A e/ou B, sofrem de sobre-ajuste. Ainda é possível verificar qual arquitetura, CNN ou CNN+ViT sofre mais com o sobre-ajuste. Nos experimentos deste artigo, o modelo baseado em CNN é

Os autores agradecem à FAPES e CAPES pelo PDPG (Programa de Desenvolvimento de Pós-Graduação - Parcerias Estratégicas nos Estados, processo 2021-2S6CD, FAPES nº132/2021). A professora Karin Komati agradece ao CNPq pela Bolsa de Produtividade DT-2 (308432/2020-7) e pelo projeto 407742/2022-0, também agradece à FAPES pelo Auxílio Taxa de Pesquisa (nº 293/2021) e pelo projeto nº1023/2022 P:2022-8TZV6.

a proposta de [9] que foi treinado usando duas bases de dados DFDC e FaceForencics++. O modelo baseado em CNN+ViT é a proposta Cross Efficient ViT de [10] que foi treinado no DFDC. A base de dados de testes, que não foi usada para o treinamento dos modelos é o Celeb-DF V2.

O texto está dividido em uma seção de trabalho correlatos, seguido por uma seção de materiais e métodos, com a apresentação dos resultados e análise dos resultados na quarta seção e por fim, finaliza-se com as conclusões.

II. Trabalhos Correlatos

Nos trabalhos correlatos, destacamos três artigos que abordam o desafio de desempenho e generalização entre bancos de dados de técnicas de aprendizado profundo em aplicações em faces humanas, tanto para reconhecimento de emoções e *Deepfake*. Estes estudos destacam a importância de modelos robustos e adaptáveis em aplicações do mundo real, onde os conjuntos de dados podem variar em termos de conteúdo e características.

No estudo de [11], os autores propõe a técnica de Deep Emo-transfer Network (DETN), para reconhecimento de expressões faciais em domínio cruzado (*cross-domain*). Durante os experimentos, a base de dados RAF-DB foi usada como domínio de origem, e extensas avaliações foram realizadas em diferentes bases de dados alvo bem estabelecidas na literatura: CK+, JAFFE, MMI, SFEW e FER2013. O DETN não só mostrou uma eficácia destacável, mas também superou vários métodos previamente estabelecidos, particularmente em ambientes e bases de dados não controlados, realçando sua robustez um cenários reais, alcançando 78,83% de acurácia no *dataset* CK. Através da experimentação, foi evidenciado que o hiper-parâmetro λ, que representa o grau de transferência de domínio, foi fundamental para otimizar o desempenho do sistema. O comportamento da acurácia em diferentes bases de dados a medida que λ varia, sugere que classes desbalanceadas de fato causam um gargalo no reconhecimento de expressões faciais, o autor ainda afirma que o DETN pode mitigar este problema efetivamente aprendendo a reamostrar os pesos do domínio fonte de forma adequada.

No estudo de [12], os autores propõem uma abordagem para reconhecimento de micro expressões usando a técnica de bases de dados cruzadas, o problema de CDMER (do inglês Cross-database micro-expression recognition). Os autores ainda afirmam que tal tarefa é difícil, onde as amostras de teste e treinamento são de diferentes bases de dados de micro-expressões (ME), resultando na inconsistência de distribuição de características entre eles e afetando o desempenho dos vários métodos de reconhecimento existentes. Para isso, os autores propuseram uma CNN de fluxo duplo (DSCNN, do inglês Dual-Stream Convolutional Neural Network). Esta rede foi projetada para extrair características espaço-temporais a partir de amostras de ME, visando uma representação mais robusta e adaptativa. Para avaliar o desempenho do DSCNN os autores conduziram testes Tipo 1 e Tipo 2. Os experimentos do Tipo 1 usam subconjuntos da base SMIC [13], treinando com um subconjunto e testando em outro. Os experimentos do Tipo 2 usam os subconjuntos da base SMIC e a base de dados CASME II [14], com variações de conjunto de treinando e testes. A métrica escolhida foi a média F1-*score* e acurácia. O método DSCNN obteve uma média F1-*score*/acurácia de 0,779/78,09% nos experimentos Tipo 1 e 0,695/70,77% nos experimentos Tipo 2, o que é significantemente maior que que a maioria dos métodos de adaptação de domínio existentes para tarefas de CDMER.

No trabalho de [15], os autores buscam desenvolver um detector de *Deepfakes* mais generalizável. De acordo com os autores, os métodos atuais de detecção de falsificação de rosto alcançam alta precisão no cenário dentro do banco de dados onde o treinamento e os testes de falsificações são sintetizados pelo mesmo algoritmo. No entanto, poucos deles obtêm desempenho satisfatório sob o cenário de banco de dados cruzado onde treinamento e teste são sintetizadas por diferentes algoritmos. Após uma análise detalhada dos modelos baseados em CNN, os autores afirmam que essa arquitetura aprende a capturar padrões de textura e cor específicos ao método de falsificação, desta forma tais modelos falham em generalizar devido a este viés de textura. Observando que os ruídos da imagem removem as texturas das cores e expõe discrepâncias entre regiões autênticas e adulteradas, os autores usaram ruídos de alta frequência para detecção de falsificação de rosto. O modelo apresentado é composto por três módulos: um para extração de ruídos de alta frequência em múltiplas escalas, um módulo de atenção espacial guiado por resíduos com o objetivo de enfatizar vestígios de alteração, e um módulo de atenção *cross-modality dual* para uma efetiva integração entre as duas modalidades propostas. O modelo obteve bom desempenho, utilizando o FF++ como conjunto de treinamento, alcançando 0,797 de acurácia no DFDC e 0,794 de acurácia no Celeb-DF V1 indicando a capacidade do modelo de generalizar.

III. Materiais e Métodos

A. Bases de Dados

1) DFDC: O DFDC [16] é uma base de vídeos gerada através de uma parceria entre empresas como Facebook, Microsoft e Amazon com o intuito de estimular o desenvolvimento de modelos de aprendizado de máquina com a capacidade de predizer se determinado vídeo é ou não um *Deepfake*. Foram utilizados oito diferentes métodos para a criação da base de vídeos, são eles: DF-128, DF-256, MM/NN, NTH, FSGAN, StyleGAN, *refinement*, e *audio swaps*. Com objetivo de reduzir o *overfitting* da base, técnicas de *data augmentation* foram aplicadas aleatoriamente nos vídeos criando novas cópias dos vídeos contendo alguns detalhes alterados (inclusão de figuras geométricas, de textos, alteração de cores, dentre outros). Um exemplo das técnicas de *data augmentation* é mostrado na Fig. 1, onde se apresentam várias versões da mesma imagem com técnicas diferentes para cada uma delas.

A base foi dividida em três grupos: treinamento, validação e testes. O primeiro grupo possui aproximadamente 120 mil vídeos, já o segundo 4 mil e o terceiro grupo é composto por 10 mil vídeos, sendo que esta última base não foi divulgada. Os vídeos possuem até 10 segundos de duração e foram

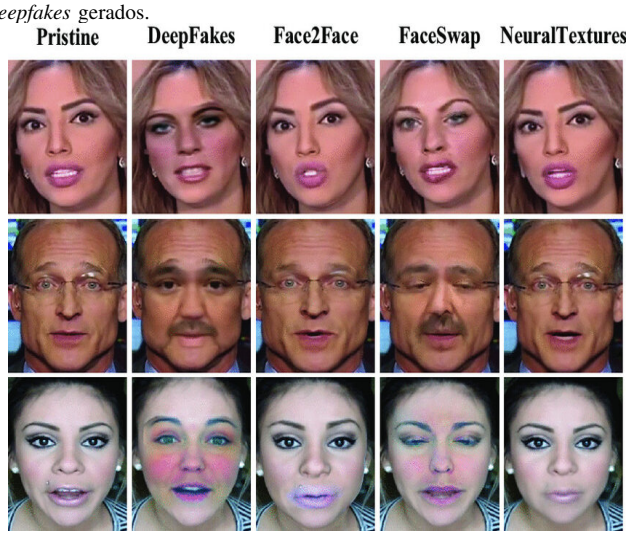
Figura 1. Exemplos de *frames* da base de vídeos DFDC com aplicações de técnicas de *data augumentation*.

gravados com a participação de cerca de 3 mil atores, que concordaram formalmente em terem o material divulgado com suas faces trocadas por de outras pessoas.

No ano de 2020, foi lançado uma competição utilizando a base do DFDC[1] como base para o treinamento dos modelos criados pelos competidores, a competição recebeu aproximadamente 2.000 subscrições, o modelo de [17] foi o vencedor.

2) FaceForensics++: O FaceForensics++ (ou apenas FF++) [18] foi criado como uma espécie de *benchmark* para pesquisadores treinarem métodos de detecção de *Deepfakes*, que são: Face2Face, FaceSwap, DeepFakes e NeuralTextures. Grande parte dos vídeos foram extraídos do YouTube. A Fig. 2 mostra como é o processo de criação das imagens do *dataset*.

Figura 2. Exemplos de quadros da base de vídeos FaceForensics++. A primeira coluna é o quadro real e as quatro colunas à direita são os quadros *Deepfakes* gerados.



Em conjunto, técnicas como FaceShifter [19] foram utilizadas também. FaceShifter é uma técnica de criação de *deepfakes* que utiliza uma abordagem de dois fatores para criar troca de faces de alta fidelidade. Uma técnica para aumentar a quantidade de dados é utilizar os *datasets* já disponíveis e aplicar técnicas diferentes de criação de *Deepfakes*. O FaceShifter foi utilizado por [5], [20] e [21] aplicado ao *dataset* FaceForensics++. No total, 1.000 vídeos foram selecionados contendo 509.914 imagens que foram utilizados como conjunto de dados.

3) Celeb-DF V2: O conjunto de dados Celeb-DF (v2) [22] contém vídeos reais e sintetizados por DeepFake com qualidade visual semelhante aos que circulam *online*. O conjunto de dados Celeb-DF (v2) é maior em relação ao Celeb-DF (v1) anterior, que contém apenas 795 vídeos DeepFake. Até o momento, o Celeb-DF inclui 590 vídeos originais coletados do YouTube com assuntos de diferentes idades, grupos étnicos e gêneros. A base de vídeos Celeb-DF V2, possui um total de 5.369 vídeos *Deepfake*, correspondendo a mais de 2 milhões de quadros. Todos os vídeos foram criados a partir de um algoritmo de síntese de *Deepfake* aprimorado, resultando uma perceptível qualidade das imagens. Na Fig. 3 é apresentada uma amostra da qualidade de quadros presentes na base, sendo a primeira coluna à esquerda a imagem real e as cinco outras colunas mais à direita versões alteradas da versão original [23].

No início de 2021, foi realizada outra edição da competição DFGC[2] (DeepFake Game Competition) [24], desta vez patrocinada pela empresa Alibaba, onde o Celeb-DF V2 foi utilizado como a base de vídeos da disputa, o sistema DFGC Detection Solution[3] foi o vencedor desta edição.

Figura 3. Exemplos de *frames* da base de vídeos Celeb-DF. A primeira coluna é o quadro real e as cinco colunas à direita são os quadros *Deepfakes* correspondentes.

B. Métodos

1) Modelo baseado exclusivamente em CNN: O trabalho de [9] propõe um *ensemble* de CNNs, especificamente da família da EfficientNet. A EfficientNetB4 foi escolhida como *baseline* por obter um melhor custo-benefício em termos de dimensões ou quantidade de parâmetros, tempo de execução

[1]https://www.kaggle.com/c/deepfake-detection-challenge

[2]https://competitions.codalab.org/competitions/29583

[3]https://github.com/beibuwandeluori/DFGC_Detection

e desempenho em classificação. O modelo foi treinado e testado utilizando duas bases de dados DFDC e FF++. Uma MTCNN (do inglês Multi-Task Cascaded Convolutional Neural Networks) [25] foi utilizada para extração da face em cada quadro.

A proposta do autor trás uma variação da arquitetura EfficientNetB4 padrão, fazendo uso de dois diferentes conceitos: (i) camadas de atenção; (ii) treinamento siamês. As camadas de atenção foram usadas em diversas contribuições no campo de processamento de linguagem natural e visão computacional. Um bloco de mecanismo de atenção é aplicado no modelo, similar ao já presente na própria EfficientNet (Figura 4). O modelo obteve bom desempenho alcançando no FF++ AUC de 0,94 e logloss de 0,32 e no DFDC AUC de 0,87 e logloss de 0,46.

Figura 4. Arquitetura proposta por [9].

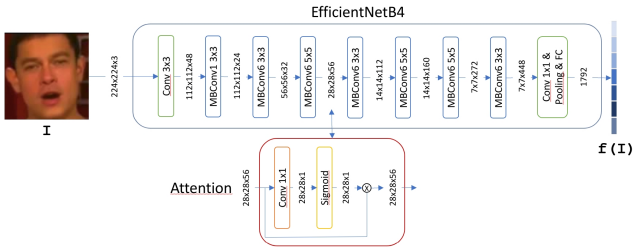
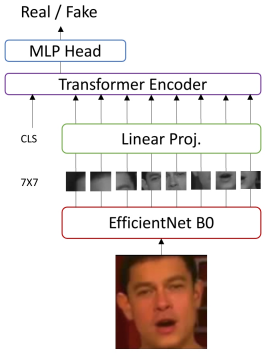


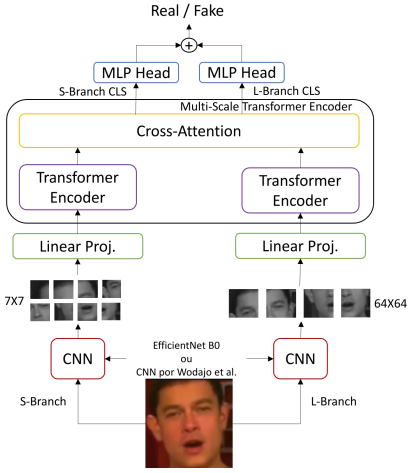
Figura 5. Efficient ViT proposta por [10].



2) Modelo baseado em CNN + ViT: O método proposto por [10] utiliza uma CNN EfficientNetB0 para extração de características em conjunto com um ViT conforme mostra a Figura 5. O *token* CLS é utilizado para classificação binária. A arquitetura utiliza pesos pré-treinados no EfficientNetB0 e é ajustada para uma extração de características específica para a tarefa. Essas características facilitam o treinamento do ViT devido aos seus detalhes de imagem de baixo nível embutidos, o modelo obteve medida-F1 de 0,838 e AUC de 0,91.

Uma segunda abordagem foi proposta, os autores aplicarem uma técnica que chamaram de Convolutional Cross ViT, onde afirmam que usar apenas *patches* pequenos pode não ser a escolha ideal, já que os artefatos introduzidos pelos métodos de geração de *deepfakes* podem aparecer tanto localmente quanto globalmente. Os autores criaram dois ramos no *pipeline* onde

Figura 6. Cross-Efficient ViT proposta por [10].



a mesma imagem passa por uma CNN EfficientNetB0 e um ramo tem saída de 7×7 e o outro de 64×64. Assim entram no *encoder* do *Transformer* e por fim são classificados por uma MLP (Multilayer Perceptron). A arquitetura detalhada é apresentada na Figura 6. Os autores também usaram a MTCNN para extração de face e *albumentations* para aumento de dados. A base de dados utilizada no treinamento e testes foi o DFDC. A abordagem dos autores obteve medida-F1 de 0,88 e AUC de 0,95.

C. Métricas

Em problemas de classificação, a matriz de confusão é uma tabela com duas linhas e duas colunas que relata o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Isso permite uma análise mais detalhada do que a mera proporção de classificações corretas (precisão) [26]. Onde:

- TP = True Positive (Verdadeiro Positivo), representam os casos em que o modelo previu corretamente a classe positiva, ou seja, o modelo classificou corretamente um exemplo como pertencente à classe positiva.
- FP = False Positive (Falso Positivo), representam os casos em que o modelo previu incorretamente a classe positiva, ou seja, o modelo classificou erroneamente um exemplo como pertencente à classe positiva, quando na verdade não era.
- TN = True Negative (Verdadeiro Negativo), representam os casos em que o modelo previu corretamente a classe negativa, ou seja, o modelo classificou corretamente um exemplo como não pertencente à classe positiva.
- FN = False Negative (Falso Negativo), representam os casos em que o modelo previu incorretamente a classe negativa, ou seja, o modelo classificou erroneamente um exemplo como não pertencente à classe positiva, quando na verdade era.

1) Medida-F1: A medida-F1 (ou em inglês, F1-score) avalia o desempenho de um modelo de classificação a partir da matriz de confusão, agregando as medidas de Precisão e

revocação sob o conceito de média harmônica. A fórmula da medida-F1 pode ser interpretada como uma média ponderada entre a precisão e o revocação, onde o F1-Score atinge seu melhor valor em 1 e o pior valor em 0. A contribuição relativa da precisão e do revocação é igual no medida-F1, e a média harmônica é útil para encontrar o melhor equilíbrio entre as duas quantidades. A equação (1) demonstra o cálculo.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

A precisão é a fração de elementos TP dividida pelo número total de unidades previstas como positivas (TP + TN). A precisão expressa a proporção de unidades que o modelo classifica como positivas e que realmente são positivas. Em outras palavras, a precisão nos diz o quanto podemos confiar no modelo quando ele prevê um indivíduo como positivo [27].

Revocação (ou recall em inglês) é a fração de elementos TP dividida pelo número total de unidades classificadas como positivas (soma da linha dos positivos reais). Revocação mede a precisão preditiva do modelo para a classe positiva: intuitivamente, mede a capacidade do modelo de encontrar todas as unidades positivas no conjunto de dados.

2) Área sob a curva (AUC): AUC é a área coberta sob a curva ROC, que é usada para medir a precisão do modelo, quanto mais perto de 1, melhor o resultado. A curva ROC é usada para visualizar o modelo de classificação. A abscissa da curva ROC é FPR (False Positive Rate) e a ordenada é TPR (True Positive Rate). Suas fórmulas são:

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$AUC = \int_a^b TPR(FPR) dFPR \tag{4}$$

3) Log loss: Log loss (ou perda logarítmica) é uma métrica de avaliação de modelos de classificação, sendo uma métrica negativa, ou seja, quanto menor, melhor. Ela varia entre 0 (previsões perfeitas) e infinito (previsões completamente erradas).

$$Log\ Loss(y, p) = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{5}$$

onde:

n é o número de amostras;

y é a classe real (0 ou 1);

p é a probabilidade prevista pelo modelo (entre 0 e 1).

IV. Experimentos, Resultados e Discussão

Para a realização dos experimentos, foram usados dois modelos para detecção de *deepfakes* disponíveis publicamente:

- O modelo de [9] é baseado em um *ensemble* de CNNs, disponível em [28] e
- o modelo de [10] é baseado em uma arquitetura que aplica CNN em conjunto de um ViT, disponível em [29].

Os modelos treinados propostos por [9] e [10] tiveram seus pesos carregados sem alterações.

A base de dados escolhida para os experimentos foi o CelebDF-V2 [22], esta é a base de dados que não foi vista pelos modelos durante o treinamento. O pré-processamento realizado foi a extração de faces utilizando uma MTCNN.

Tabela I
TABELA COMPARATIVA ENTRE OS MÉTODOS

modelo/base de dados	DFDC	FF++	Celeb-DF v2
modelo baseado em CNN	AUC = 0,94 logloss = 0,32	AUC = 0,87 logloss = 0,46	AUC = 0,32 Logloss=0,88
modelo baseado em CNN+ViT: "Cross-Efficient ViT"	AUC = 0,95 medida-F1=0,808		AUC = 0,07 medida-F1=0,19
modelo baseado em CNN+ViT: "Efficient ViT"	AUC = 0,95 medida-F1=0,808		AUC = 0,14 medida-F1=0,25

Na Tabela I, avaliamos o desempenho dos 3 modelos na base Celeb-DF v2 usando as métricas AUC, logloss e medida-F1, escolhidas com base nas utilizadas pelos autores originais dos modelos para uma comparação justa e consistente. Observando o AUC, que reflete a capacidade do modelo de distinguir entre classes verdadeiras e falsas, o modelo baseado em CNN obteve 0,32, enquanto os modelos "Cross-Efficient ViT"e "Efficient ViT"mostraram valores mais baixos. Esse desempenho baixo sugere dificuldade em discriminar corretamente *deepfakes* de vídeos reais em um contexto completamente novo. Em relação ao **logloss**, que quantifica a confiança das probabilidades preditivas, o modelo baseado em CNN teve um alto valor de 0,88, indicando que as previsões feitas podem não ser tão confiáveis. A medida-F1, um equilíbrio entre precisão e revocação, revelou eficácia inferior para os modelos CNN+ViT. Isso pode indicar que, enquanto o modelo pode estar fazendo previsões corretas, ele pode estar perdendo muitos verdadeiros positivos ou incluindo muitos falsos positivos. Tais resultados, juntamente com a evidente dificuldade de adaptação a novos conjuntos de dados, apontam para um cenário de *overfitting*.

Técnicas de regularização das CNNs são fundamentais para melhorar seus resultados, pois ajuda a evitar o *overfitting* nos dados de treinamento [30]. A regularização pode ser dividida em 3 categorias, sendo a regularização de entrada que atua antes da imagem ser alimentada na rede, podendo envolver técnicas como aumento de dados. A regularização interna aplica-se depois que a imagem é alimentada na rede, um exemplo é o *dropout* [31]. A regularização de rótulo atua na camada de saída. Santos e Papa [30] ainda pontuam que problemas como a falta do uso de arquiteturas mais simples e a falta de uma avaliação dos métodos de regularização em

dados mais complexos, *datasets* desbalanceados por exemplo, são alguns problemas encontrados na maioria dos trabalhos.

É importante ressaltar que algumas das técnicas de regularização foram usadas: a base de dados DFDC já contém técnicas de *data augmentation*, o modelo baseado em CNN usou duas bases de dados diferentes, e no modelo CNN+ViT foram usadas *albumentations* para aumento de dados. Mesmo com estas técnicas presentes nos trabalhos avaliados, ainda não foi o suficiente para evitar o sobre-ajuste dos modelos. E portanto, corroboramos com as afirmações do trabalho de [15] em que os autores identificam que falta de generalização em detectores de *deepfakes*.

V. Conclusões

Este estudo procurou abordar justamente essa problemática, examinando a capacidade de generalização de modelos baseados em CNN e CNN+ViT em um cenário de teste com base de dados cruzada. Os resultados preliminares indicam a necessidade de uma análise mais aprofundada sobre a adaptabilidade dos modelos frente a diferentes conjuntos de dados, a fim de garantir sua eficácia prática.

Em síntese, à medida que a ameaça dos *deepfakes* se intensifica, a busca por modelos de detecção que sejam não apenas precisos, mas também generalizáveis, torna-se cada vez mais crítica. Como trabalhos futuros, espera-se reproduzir o trabalho de [15], avaliando principalmente a questão do ruído de alta frequência, além disso estender tal trabalho para experimentos com outras bases de dados e investigar outros modelos. Outro trabalho futuro é a combinação de diferentes bases de dados durante o treinamento e diferentes formas de treinamento.

Referências

[1] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[2] J. Bakdash, C. Sample, M. Rankin, M. Kantarcioglu, J. Holmes, S. Kase, E. Zaroukian, and B. Szymanski, "The future of deception: Machine-generated and manipulated images, video, and audio?," in *2018 International Workshop on Social Sensing (SocialSens)*, pp. 2–2, IEEE, 2018.

[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[4] S. R. Ahmed, E. Sonuç, M. R. Ahmed, and A. D. Duru, "Analysis survey on deepfake detection and recognition with convolutional neural networks," in *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–7, 2022.

[5] L. Deng, J. Wang, and Z. Liu, "Cascaded network based on EfficientNet and transformer for deepfake video detection," *Neural Processing Letters*, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2021.

[8] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, pp. 1–48, 2021.

[9] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," *CoRR*, vol. abs/2004.07676, 2020.

[10] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," *CoRR*, vol. abs/2107.02612, 2021.

[11] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3092–3099, IEEE, 2018.

[12] B. Song, Y. Zong, K. Li, J. Zhu, J. Shi, and L. Zhao, "Cross-database micro-expression recognition based on a dual-stream convolutional neural network," *IEEE Access*, vol. 10, pp. 66227–66237, 2022.

[13] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, IEEE, 2013.

[14] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS ONE*, vol. 9, pp. 1–8, 01 2014.

[15] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

[16] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[17] S. Seferbekov, "Deepfake detection (dfdc) solution by @selimsef," 2020.

[18] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

[19] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *CoRR*, vol. abs/1912.13457, 2019.

[20] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, may 2023.

[21] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," 2021.

[22] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A new dataset for deepfake forensics," *CoRR*, vol. abs/1909.12962, 2019.

[23] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "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, 2020.

[24] B. Peng, H. Fan, W. Wang, J. Dong, Y. Li, S. Lyu, Q. Li, Z. Sun, H. Chen, B. Chen, Y. Hu, S. Luo, J. Huang, Y. Yao, B. Liu, H. Ling, G. Zhang, Z. Xu, C. Miao, C. Lu, S. He, X. Wu, and W. Zhuang, "DFGC 2021: A deepfake game competition," *CoRR*, vol. abs/2106.01217, 2021.

[25] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[26] S. Visa, B. Ramsay, A. Ralescu, and E. Knaap, "Confusion matrix-based feature selection.," vol. 710, pp. 120–127, 01 2011.

[27] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *ArXiv*, vol. abs/2008.05756, 2020.

[28] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Davide-coccomini/combining-efficientnet-and-vision-transformers-for-video-deepfake-detection: Code for video deepfake detection model from "combining EfficientNet and vision transformers for video deepfake detection" presented at ICIAP 2021.." https://github.com/davide-coccomini/Combining-EfficientNet-and-Vision-Transformers-for-Video-Deepfake-Detection, 2022.

[29] N. Bonettini, C. Bonettini, E. Daniele, Mandelli, Sara, Bondi, Luca, Bestagini, Paolo, Tubaro, and et al., "Polimi-ispl/icpr2020dfdc: Video face manipulation detection through ensemble of cnns." https://github.com/polimi-ispl/icpr2020dfdc, 2021.

[30] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–25, 2022.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of machine learning research*, vol. 15, pp. 1929–1958, 2014.