UNIVERSIDADE FEDERAL DO CEARÁ

CURSO DE GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

JOSÉ EDIBERTO DO NASCIMENTO JÚNIOR

DETECÇÃO DE *DEEPFAKES* UTILIZANDO REDES NEURAIS CONVOLUCIONAIS

SOBRAL

2023

JOSÉ EDIBERTO DO NASCIMENTO JÚNIOR

DETECÇÃO DE *DEEPFAKES* UTILIZANDO REDES NEURAIS CONVOLUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Campus de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Orientador: Prof. Dr. Iális Cavalcante de Paula Júnior

SOBRAL

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

N195d Nascimento Júnior, José Ediberto do.
DETECÇÃO DE DEEPFAKES UTILIZANDO REDES NEURAIS CONVOLUCION / José Ediberto / José Ediberto do Nascimento Júnior. – 2023.
46 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Sobral, Curso de Engenharia da Computação, Sobral, 2023.
Orientação: Prof. Dr. Iális Cavalcante de Paula Júnior.

1. Deepfakes. 2. Redes Neurais. 3. Inteligência artificial. I. Título.

CDD 621.39

JOSÉ EDIBERTO DO NASCIMENTO JÚNIOR

DETECÇÃO DE *DEEPFAKES* UTILIZANDO REDES NEURAIS CONVOLUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Campus de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Iális Cavalcante de Paula Júnior (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Jarbas Joaci de Mesquita Sá Junior
Universidade Federal do Ceará (UFC)

Andressa Gomes Moreira
Universidade Federal do Ceará (UFC)

Aos meus pais, amigos e professores por sempre acreditarem em mim e me incentivarem a seguir, por estarem ao meu lado, me dando força e me motivando a não desistir, mesmo nos momentos mais difíceis;

AGRADECIMENTOS

Expresso minha gratidão a todos que acompanharam minha jornada acadêmica. O apoio de todos com certeza fez essa jornada valer a pena, sem vocês eu não teria conseguido chegar aqui.

Agradeço acima de tudo a minha família, minha mãe e meu pai por sempre me apoiarem, incentivarem e sempre sendo um dos principais pilares e exemplos na minha vida. Sou grato também pelos amigos, tanto os que me acompanham desde sempre quanto os que fiz ao longo do caminho, agradeço por me ajudarem a manter a motivação mesmo quando as coisas pareciam difíceis.

Agradeço especialmente ao meu orientador e professores, que me guiaram e ensinaram ao longo deste caminho. Seus conselhos e conhecimento foram inestimáveis.

Por fim, deixo minha gratidão também a todos que me ajudaram a concluir e melhorar este trabalho, sei que não é fácil tirar de seu tempo para ler um documento tão extenso, muitos inclusive leram diversas vezes. Suas críticas e comentários foram indispensáveis e me ajudaram a melhorar.

Novamente, muito obrigado a todos por seu apoio e incentivo ao longo desta jornada.

“O mestre falhou mais vezes do que o iniciante sequer tentou.”

(Stephen McCranie)

RESUMO

Explorando o uso de redes neurais convolucionais (CNNs) para detectar imagens e vídeos *deepfake*, mídias geradas usando técnicas de aprendizado de máquina para manipular ou substituir o conteúdo original. O estudo se concentra no desenvolvimento de um sistema de detecção de *deepfake* baseado em CNN que pode identificar e distinguir entre imagens e vídeos autênticos e manipulados com alta precisão. A pesquisa envolve treinar e testar diferentes arquiteturas usando conjuntos de dados de imagens e vídeos reais e sintéticos. O trabalho também discute o impacto de vários fatores no desempenho do sistema de detecção de *deepfake* baseado em redes convolucionais, como o tipo de método de geração de *deepfakes*, a qualidade e resolução da mídia de origem e o tamanho do conjunto de dados de treinamento. Os resultados mostram que os modelos baseados em redes neurais convolucionais podem efetivamente identificar *deepfakes* com alta precisão, e o sistema proposto pode ser aplicado a vários cenários do mundo real, incluindo mídia social e jornalismo, para combater a disseminação de notícias falsas e desinformação.

Palavras-chave: Informação, fakenews, Inteligência Artificial

ABSTRACT

Exploring the use of convolutional neural networks (CNNs) to detect deepfake images and videos, media generated using machine learning techniques to manipulate or replace the original content. The study focuses on developing a CNN-based deepfake detection system that can accurately identify and distinguish between authentic and manipulated images and videos. The research involves training and testing different architectures using datasets of real and synthetic images and videos. The paper also discusses the impact of various factors on the performance of the CNN-based deepfake detection system, such as the type of deepfake generation method, the quality and resolution of the source media, and the size of the training dataset. The results demonstrate that CNN-based models can effectively identify deepfakes with high accuracy, and the proposed system can be applied in various real-world scenarios, including social media and journalism, to combat the spread of fake news and misinformation.

Keywords: Information, video, fakenews

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AMLA	*Adversarial Machine Learning Attacks* /
AUC	*Area Under Curve* / Área sob a curva
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
GPU	*Graphics Processing Unit* / Unidade de processamento Gráfico
RNA	Rede Neural Artificial
ROC	*Receiver operating characteristic* / Característica de Operação do Receptor

LISTA DE SÍMBOLOS

P	Precisão do modelo
S	Sensibilidade do modelo
T_p	Verdadeiros Positivos
F_p	Falso Positivos
F_n	Falso Negativos
F_s	F-score do modelo

SUMÁRIO

1 INTRODUÇÃO

Um fenômeno vem se destacando e mostrando ser motivo de preocupação é o crescimento dos *deepfakes*, mídias geradas por algoritmos de aprendizagem profundo, conhecidos como *Deep Learning*. Esses algoritmos conseguem criar, alterar e trocar rostos automaticamente em imagens e vídeos. Os resultados obtidos através desse tipo de procedimento vêm se tornando cada vez mais impressionantes e quase impossíveis para um olho destreinado distinguir conteúdo real ou falso.

Atualmente, é comum se deparar com fotos e vídeos que foram digitalmente manipulados de alguma forma. A utilização de ferramentas para criação de *deepfake* tem vários propósitos, entretanto alguns desses levantam preocupações pelo impacto que podem causar. Por exemplo, existem casos em que *deepfakes* foram utilizados para produzir vídeos falsos de personalidades proeminentes. Um exemplo foi o caso onde um vídeo supostamente do Mark Zuckeberg, onde ele é retratado alegando que o Facebook rouba dados de seus usuários. Esses vídeos, mesmo sendo falsos, podem repercutir significantemente, ameaçando a imagem dos envolvidos, como também comprometendo a credibilidade das informações. Muitas vezes, mídias desse gênero são compartilhadas e difundidas nas redes sociais antes que as informações sejam averiguadas, levando a que várias pessoas acreditem no que foi exposto como sendo verdade.

Figura 1 – *Deepfake* com rosto do Mark Zuckerberg

Fonte: (POSTERS, 2019).

Nos últimos anos, com o avanço na área de *deepfakes*, a facilidade com que estes

podem ser criados representa um grande perigo. As ferramentas gratuitas, como o Google Colab ou o Kaggle (ALARCON, 2020), que incorporam máquinas virtuais equipadas com *Graphics Processing Unit* / Unidade de processamento Gráfico (GPU), têm simplificado significativamente o processo de treinamento e criação de *deepfakes*. Com tamanha facilidade, qualquer pessoa mal intencionada pode não apenas difamar, mas aproveitar da influência das vítimas para manipular grupos de pessoas com seus próprios objetivos.

Os *deepfakes* andam cada vez mais presentes na sociedade, tendo um impacto significativo no cotidiano. A facilidade de criação e do compartilhamento dessas mídias manipuladas levantam questões sobre a confiabilidade das informações. Com a grande facilidade e velocidade das redes sociais de proliferar informações, é fundamental perceber a influência que os *deepfakes* podem exercer na percepção da realidade. Imagine, por exemplo, o impacto que um *deepfake* sobre uma figura política durante uma campanha eleitoral. Esse tipo de mídia falsa pode ser facilmente utilizada para difamação de um candidato, ou então para manipulação da opinião pública. Esse tipo de situação afeta diretamente a democracia e a confiança nos agentes políticos. Além de que os *deepfakes* também podem ser usados para enganar indivíduos e obter informações confidenciais, o que compromete não só privacidade como também segurança pessoal.

Outro ponto importante é o impacto emocional e psicológico dos *deepfakes*. Muitas das vítimas, com seu rosto substituído em um vídeo ou imagem, têm não só a reputação como também a sua saúde mental destruída. Adicionalmente, existe o risco de que os "deepfakes"possam minar a autenticidade e a confiança nas provas visuais, resultando em uma sociedade onde a verdade se torna ambígua, tornando-se assim difícil confiar nesse tipo de evidência. Um infeliz exemplo é QTCinderella, renomada streamer na Twitch. A situação veio à tona quando o também streamer Atrioc, numa ação inescrupulosa, pagou para que fossem criados deepfakes de cunho pornográfico utilizando o rosto de QTCinderella. Essa violação grosseira da privacidade de QTCinderella causou um enorme abalo emocional e psicológico, afetando significativamente sua reputação e saúde mental. A experiência reforça a preocupação sobre como os *deepfakes* podem ser usados para fins mal-intencionados, incluindo a manipulação de imagens e vídeos, intensificando a insegurança sobre a veracidade das evidências visuais. Desta forma, o caso QTCinderella evidencia a necessidade urgente de regulação e medidas preventivas para lidar com essa forma emergente de abuso. Mesmo com o pedido de desculpas (TWITCH. . . , 2023) a situação escalou muito e inclusive outras streamers descobriram existirem *deepfakes* seus espalhados pela internet, sem consentimento nenhum.

É fundamental que o desenvolvimento e pesquisa de métodos de detecção e prevenção do *deepfakes* sejam priorizados. Apenas com abordagens eficazes podemos identificar e combater a disseminação de *deepfakes*, para proteger e preservar tanto a integridade das informações como a reputação das pessoas, garantindo assim que as informações na nossa sociedade sejam mais confiáveis.

1.1 Trabalhos Relacionados

A detecção de falsificações de mídia tem sido um tópico ativo de pesquisa à medida que a capacidade de produzir deepfakes realistas aumenta. Esta seção destaca dois trabalhos importantes que contribuíram significativamente para a detecção de deepfakes.

Em 2018, Korshuno e Marcel publicaram um estudo inovador intitulado "*DeepFakes: A New Threat to Face Recognition? Assessment and Detection*"(KORSHUNOV; MARCEL, 2018). Neste trabalho, os autores destacam a ameaça que as *deepfakes* representam para as tecnologias de reconhecimento facial e propõem um novo método para detectar essas falsificações. Eles desenvolveram um classificador baseado em aprendizado profundo usando uma rede neural convolucional (CNN) para distinguir entre imagens originais e manipuladas. A metodologia proposta foi treinada e testada em um conjunto de dados de deepfakes, demonstrando uma capacidade significativa de detecção.

Em um estudo subsequente, Afchar et al. exploraram a detecção de falsificação de vídeo facial com o trabalho "*Mesonet: a Compact Facial vídeo Forgery Detection Network*"(AFCHAR *et al.*, 2018). Neste estudo, eles introduziram a MesoNet, uma rede neural convolucional compacta projetada especificamente para a tarefa de detecção de *deepfakes*. A MesoNet foi projetada para ser leve e rápida, tornando-a adequada para uso em dispositivos com recursos limitados. O desempenho do MesoNet foi avaliado em uma variedade de *benchmarks* de *deepfakes*, mostrando resultados promissores.

Como também apresentado em *Deepfake Detection: A Systematic Literature Review* (RANA *et al.*, 2022) que analisou mais de 100 artigos publicados entre 2018 e 2020, comparando os resultados entre diversas técnicas de detecção de *deepfakes* entre elas: *deep learning*, *machine learning* convencionais, estatísticas e *blockchain*. Chegou-se a conclusão de que as técnicas baseadas em *deep learning* superaram as demais técnicas.

Esses estudos demonstram a utilidade dos métodos de aprendizado profundo para a detecção de *deepfakes*. No entanto, com a constante evolução das técnicas de criação de

deepfakes, a detecção continua a ser um desafio significativo que requer pesquisa contínua.

1.2 Objetivos

(NADA, 2023)

1.2.1 Objetivo Geral

Desenvolver um modelo para a detecção de *deepfakes*, utilizando técnicas de aprendizado de máquina, contribuindo para a preservação da integridade da informação, protegendo a privacidade e reputação das pessoas afetadas por conteúdos falsificados.

1.2.2 Objetivos Específicos

- Desenvolver modelos convolucionais para a detecção de *deepfakes*, utilizando um conjunto de dados de imagens autênticas e manipuladas.
- Avaliar o desempenho dos modelos a fim de identificar problemas e melhorar a acurácia dos mesmos.
- Propor recomendações para futuros desenvolvimentos na área de detecção de *deepfakes*, considerando os avanços tecnológicos e a evolução das técnicas de falsificação de conteúdo.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Rede Neural Artificial

Rede Neural Artificial (RNA) ou simplesmente Rede Neural é um tipo de modelo computacional que é baseado em como o sistema nervoso de animais, em especial o cérebro, funciona e aprende (WANG, 2003) por isso é utilizado em aprendizado de máquina e reconhecimento de padrões. Rede neurais são apresentadas como conjuntos de neurônios que são interconectados e capaz de computar valores de acordo com uma entrada, simulando uma rede neural biológica.

2.1.1 Aprendizado Supervisionado

O aprendizado supervisionado em redes neurais é uma das estratégias utilizadas em aprendizado de máquina, é uma técnica onde as entradas são pré-classificadas para o treinamento e a rede constantemente tenta prever a qual grupo a sua entrada pertence. Por exemplo, na etapa de treinamento pode se ter imagens classificadas como "A" ou "B" e sempre que a rede prever uma classificação para a entrada é calculado o erro dessa previsão, visando minimizar o erro, (que neste trabalho nos será referido como *loss*) o máximo possível.

2.1.2 Convolutional neural network

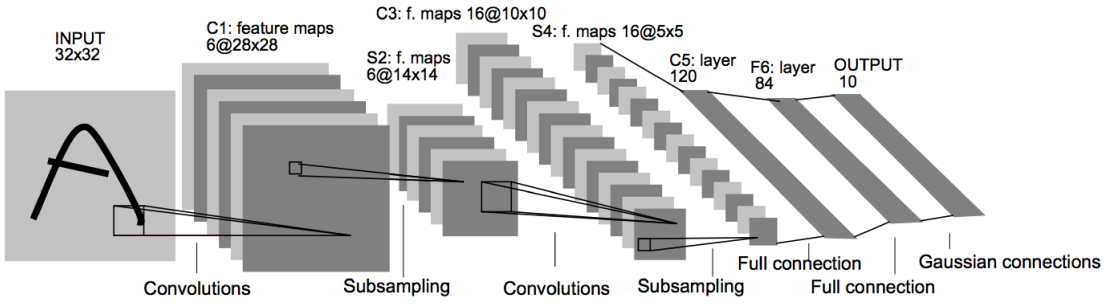
Convolutional Neural Network (CNN) (ou ConvNet) é uma categoria de rede neural amplamente aplicada para a análise de imagens. Essas redes foram inspiradas pela organização do córtex visual de animais, o que lhes permite desenvolver filtros e aprender a classificar determinadas características presentes nas imagens, tornando-as especialmente eficazes para tarefas de classificação de imagens.

As CNNs utilizam uma arquitetura multicamada de perceptrons, mas com uma característica distintiva: a aplicação de camadas convolucionais e camadas de po*o*ling em conjunto. As camadas convolucionais são responsáveis por extrair características relevantes das imagens, como linhas, formas geométricas, bordas e cores. Essas características são representadas pelos chamados *feature maps* na 2.

O processo de convolução envolve a aplicação de filtros a uma região da imagem para identificar padrões específicos. Em seguida, as camadas de *pooling* são aplicadas para reduzir o

tamanho das representações, diminuindo a quantidade de parâmetros e, assim, reduzindo o custo computacional e o tempo de treinamento.

Figura 2 – Arquitetura tipica de uma CNN



Fonte: (LECUN *et al.*, 1998).

Na 2, podemos ver a arquitetura típica de uma CNN, onde as camadas convolucionais e de *pooling* são intercaladas para identificar e reduzir gradualmente as características relevantes. A técnica de *flatten* é então utilizada para transformar os *arrays* multidimensionais gerados pelas camadas convolucionais em um *array* de uma dimensão, alimentando assim uma rede neural profunda totalmente conectada, representada pelas camadas C5 e F6 na 2.

As camadas de *pooling* também desempenham um papel importante na extração de características ao reduzir a dimensionalidade da representação, garantindo que características importantes sejam preservadas e permitindo que a rede foque em detalhes essenciais para a classificação.

Em resumo, as CNNs aproveitam as camadas convolucionais e de *pooling* para identificar características relevantes em imagens, tornando-as altamente eficazes na classificação de objetos, rostos e outros elementos visuais complexos. Essa abordagem eficiente reduz a complexidade do modelo e facilita o treinamento, tornando as CNNs um poderoso instrumento para tarefas de visão computacional.

2.1.3 Generative adversarial Network

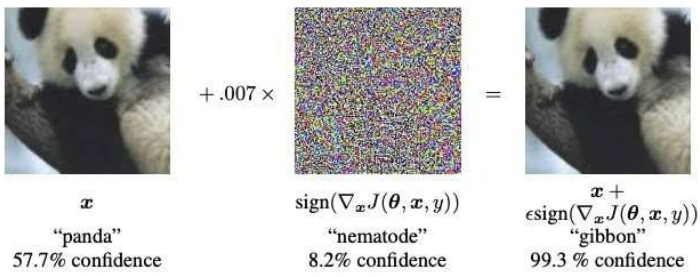
Generative Adversarial Network (GAN) é uma forma de aprendizagem de máquina onde duas redes neurais competem entre si, onde o ganho de uma é a perda da outra. Dado um conjunto para treino, essa técnica tenta usar uma rede geradora para criar dados baseados no conjunto de entrada. Por exemplo, uma GAN treinada em fotografias, pode gerar novas fotos que sejam autênticas a um olhar desatento. A ideia central de uma GAN é o treinamento indireto pelo

discriminador, outra rede neural que pode entender o quão "realista" a saída da rede geradora, onde as duas redes são atualizadas dinamicamente aumentando a rigorosidade. Isso significa que a rede geradora não tem o objetivo de criar uma imagem realista ou mesmo similar com o conjunto de entrada, mas sim de enganar a rede discriminadora. Isso possibilita obter um treinamento não supervisionado.

2.1.4 Adversarial Machine Learning Attacks

Adversarial Machine Learning Attacks / (AMLA) é o estudo de ataques em algoritmos de aprendizagem de máquina, e a defesa sobre esses ataques. Os algoritmos de aprendizagem de máquina são projetados para funcionar em condições específicas, assumindo que os dados de treinamento e de testes são gerados com a mesma distribuição estatística. Porém, assumir isso é perigoso, uma vez que fabricar entradas que violem o proposito dos algoritmos é uma tarefa não tão complexa quando se sabe a distribuição estatística esperada pelo algoritmo. A figura 3 mostra como um ataque desse pode mudar a classificação de um *panda* para um *gibbon* apenas adicionando "ruído" na entrada da rede.

Figura 3 – Exemplo de um *Adversarial Attack*

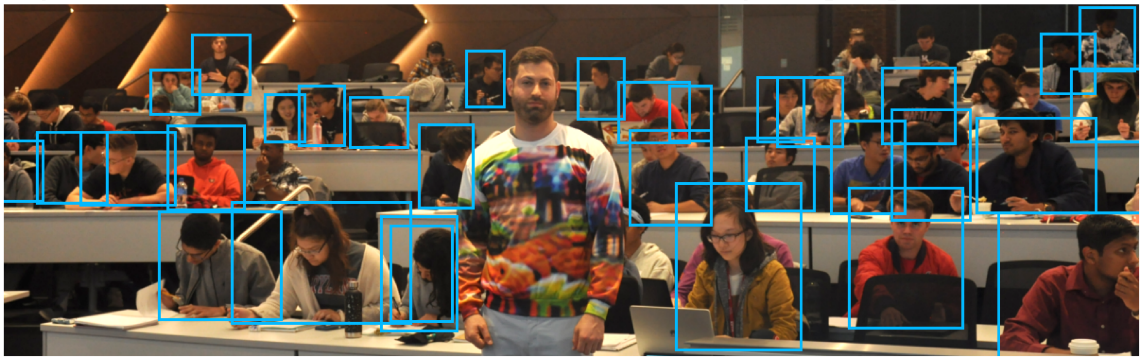


Fonte: (GOODFELLOW *et al.*, 2014).

É importante notar que esse tipo de ataque pode acontecer em qualquer momento antes da entrada ser processada pela rede, como, por exemplo, na figura 4, onde temos um padrão utilizado para enganar o detector YOLOV2 (REDMON; FARHADI, 2016) treinado no *dataset* COCO (LIN *et al.*, 2014).

Lidar com ataques dessa natureza representa um desafio significativo, especialmente diante do uso de ferramentas que empregam métodos semelhantes para criar *deepfakes*, conforme mencionado na seção 2.1.3. A complexidade desses ataques requer uma abordagem abrangente e sofisticada para mitigar seus efeitos adversos e preservar a autenticidade das evidências visuais em diferentes contextos.

Figura 4 – demonstração do detector YOLOV2 sendo engando pelo padrão na blusa



Fonte: (WU *et al.*, 2019).

2.2 Espaço de cores

Colorspace ou espaço de cores, é uma maneira específica de organizar cores para serem representadas por algum meio físico, o RGB (Red, Green, Blue - Vermelho, Verde, Azul) é um dos mais comuns no processamento de imagens digitais. Este sistema de cores aditivas cria cores pela combinação de luz vermelha, verde e azul em diferentes intensidades. Cada um destes canais de cores pode variar de intensidade de 0 a 255, permitindo a representação de mais de 16 milhões de cores distintas. A combinação destes três canais primários de cor em suas máximas intensidades resulta na cor branca, enquanto a ausência de luz em todos os canais resulta no preto (GONZALEZ; WOODS, 2008).

Nos modelos de redes neurais convolucionais, a entrada para processamento de imagens é tipicamente no formato de um tensor 3D, onde a altura e largura representam as dimensões da imagem, e a profundidade corresponde aos canais de cor. No caso do espaço de cores RGB, existem três canais correspondentes, permitindo que a rede processe informações de cor de forma independente em cada canal. Esta capacidade permite que a rede aprenda características específicas de cor que são importantes para a tarefa em questão, tais como a detecção de objetos ou a classificação de imagens (HEATON, 2018).

Estudos anteriores mostraram que a inclusão de informações de cor, tal como o espaço de cores RGB, pode melhorar significativamente o desempenho das redes neurais convolucionais em tarefas de processamento de imagens (KRIZHEVSKY *et al.*, 2012). Assim, o uso de espaços de cores, como o RGB, continua sendo uma prática padrão em muitas aplicações de aprendizado profundo baseadas em imagens.

2.3 Receiver operating characteristics

A curva ROC é uma ferramenta de diagnóstico amplamente utilizada em análises de classificação binária para determinar a qualidade de um modelo. Ela é um gráfico que mostra o desempenho de um modelo de classificação em todas as configurações de classificação possíveis. A curva ROC é gerada traçando a taxa de verdadeiros positivos (TPR) contra a taxa de falsos positivos (FPR) para diferentes pontos de corte de um modelo (FAWCETT, 2006).

A taxa de verdadeiros positivos (TPR, também conhecida como sensibilidade) é a proporção de positivos corretamente identificados como tal. Em contraste, a taxa de falsos positivos (FPR, também conhecida como 1-especificidade) é a proporção de negativos erroneamente identificados como positivos. Portanto, um modelo ideal teria uma curva ROC que passa pelo canto superior esquerdo do gráfico, indicando uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos.

Uma métrica relacionada, a Área sob a Curva ROC (AUC-ROC), proporciona uma medida agregada de desempenho em todos os limites de classificação possíveis. A AUC-ROC varia de 0 a 1, onde um valor de 1 indica um modelo de classificação perfeito e um valor de 0,5 sugere que o modelo não tem capacidade de discriminação (BRADLEY, 1997).

Na aprendizagem profunda, e especificamente em redes neurais convolucionais para classificação de imagens, a curva ROC e a AUC-ROC são comumente usadas para avaliar o desempenho de um modelo (LITJENS *et al.*, 2017).

2.4 Métrica F-score

O F-Score, também conhecido como medida F1 ou score F1, é uma métrica de avaliação que combina precisão e revocação (recuperabilidade) para fornecer uma única medida de qualidade para um sistema de classificação. Essa métrica é particularmente útil em situações onde as classes estão desequilibradas (SASAKI, 2007).

A precisão é a proporção de verdadeiros positivos (TP) em relação à soma de verdadeiros positivos e falsos positivos (FP), enquanto a revocação é a proporção de verdadeiros positivos em relação à soma de verdadeiros positivos e falsos negativos (FN). A fórmula para o F-Score é dada pela equação 2.1:

$$F_s = 2 \times \frac{P * S}{P + S} \tag{2.1}$$

Em outras palavras, o F-Score é a média harmônica de precisão e revocação. Esta métrica varia de 0 a 1, em que 1 é a melhor pontuação possível e 0 é a pior. A média harmônica é utilizada porque dá pesos iguais para precisão e revocação e penaliza sistemas de classificação com um desempenho extremo em uma métrica à custa da outra.

Assim como a curva ROC, o F-Score é uma métrica importante para avaliar a qualidade dos modelos de aprendizado profundo, como as redes neurais convolucionais, especialmente em contextos com classes desequilibradas ou onde a revocação é particularmente importante (GOUTTE; GAUSSIER, 2005).

2.5 *Deepfakes*

Deepfake é o nome dado aos vídeos e imagens onde rostos de pessoas são gerados, criados ou alterados por algorítimos de aprendizado profundo. São várias as aplicações para *deepfakes*, como na educação, por conseguir trazer figuras históricas resultando em aulas dinâmicas e interessantes. Na acessibilidade, onde voz pode ser sintetizada para dezenas de idiomas, na reconstituição de crimes para auxílio da polícia em alguns casos na recriação de cenas, etc. Porém, apesar de tantas utilizações benéficas, *deepfakes* vêm sendo utilizados de forma maléfica para disseminação de notícias falsas, chantagem, pornografia, entre outros. É difícil mensurar os danos que os *deepfakes* possam causar ao longo do tempo, com a tecnologia avançando rápido e o número de maneiras de criar vídeos e imagens falsas crescendo ainda mais rápido que maneiras de identificá-los. Um relatório publicado pela Deeptrace (DEEPTRACE, 2019) mostra que, em 2019, 96% dos vídeos *deepfakes* na internet eram pornográficos, o que torna obter dados sobre esse tipo de vídeo muito enviesado fazendo com que as redes neurais tenham dificuldade em classificar esse tipo de material. Além do problema da dificuldade da criação de uma base de dados tendenciosa, ainda existe o problema com os diferentes métodos de gerar *deepfakes*. Essas mídias sintéticas podem ser criadas de diversas maneiras utilizando estratégias diferentes para objetivos distintos, alguns exemplos são:

- *Face synthesis*, onde o objetivo é criar imagens ultra-realísticas de faces não existentes, normalmente para esse tipo de aplicação utiliza-se *Generative Adversarial Network* (GAN)
- *Face Swap*, trata-se de *deepfakes* onde ocorre a troca da face do vídeo base (*source*)

com a de uma referência. Esse tipo de *deepfake* é mais comum para ataques de pessoas famosas, notícias falsas e chantagem. Técnicas utilizadas nesse tipo de *deepfake* englobam localizar faces existentes e compor a imagem de referência utilizando alinhamento de face, otimização Newton-Gaussiana e mesclagem de imagens para o resultado.

- *Facial attributes and expression* consistem de modificar atributos da face como a cor do cabelo ou da pele, a idade, o gênero, expressões faciais, transformando rostos fazendo-os parecerem alegres, tristes ou com raiva, por exemplo. Um exemplo deste tipo de *deepfake* é o aplicativo *FaceApp* que foi amplamente utilizado por milhares de pessoas. Esse aplicativo utiliza um dos métodos mais poderosos para tal, o *StarGan*, que consiste em treinar um único modelo em diversos atributos simultaneamente, para cada domínio, essencialmente fazendo um modelo ser bom em identificar várias estruturas e modificá-las de modo a gerar o resultado desejado.

Existe uma gama considerável de *deepfakes* com atributos específicos para cada tipo de ferramenta com o qual foi criado, dessa maneira dificultando que um único modelo de rede neural consiga identificar e classificar com confiança se uma mídia foi manipulada digitalmente. Existem trabalhos que se demonstram capazes de identificar qual a técnica utilizada para criar o *deepfake*, embora o método demonstrado não consiga classificar como real ou falso geralmente.

De longe o método mais comum para se criar *deepfakes* é a utilização de uma GAN, já que ela não exige *hardware* especial e existem vários modelos prontos disponíveis na internet. Uma GAN funciona onde duas redes neurais batalham entre si, onde uma rede é responsável por gerar imagens e outra por discriminá-las. O objetivo normalmente desse tipo de rede é criar imagens mais realistas, onde a rede discriminadora determina o nível de realismo conforme a capacidade de classificação da mesmas (tendo em mente que o limite teórico desse realismo é uma junção da classificação com a capacidade generativa da outra rede). O resultado desse tipo de rede, onde as redes neurais treinam juntas e batalham entre si, é extraordinário e produzem geralmente resultados excelentes:

A Imagem 5 é uma evidência impressionante do avanço significativo na criação de *deepfakes*. Por meio da rede GAN StyleGAN2, foi possível gerar de forma artificial uma pessoa que não existe na realidade. Esses resultados destacam o quão avançada e sofisticada se tornou a tecnologia de *deepfake*.

Figura 5 – Imagem gerada pelo site thispersondoesnotexist.com

Fonte: Wang (2022).

3 METODOLOGIA

A metodologia utilizada neste trabalho baseia-se em aprendizado de máquina profundo (*deep learning*), especificamente com a utilização de aprendizado supervisionado. Este enfoque metodológico é essencial dada a natureza do problema — a detecção de *deepfakes*. Utilizando o *deep learning* podemos criar modelos capazes de reconhecer padrões nas imagens de maneira que esses modelos possam adequadamente detectar *deepfakes* com eficácia.

Os modelos de aprendizado de máquina são treinados utilizando um conjunto de dados, composto por dados e seus respectivos rótulos de saída. Neste caso, as entradas de dados serão imagens e os rótulos de saída serão binários, indicando 'verdadeiro' para mídias autênticas e 'falso' para *deepfakes*. O modelo aprende a partir desses exemplos e utiliza o conhecimento adquirido para classificar novas imagens em verdadeiras e falsas.

Um fator crucial para o sucesso desta abordagem é a qualidade e a diversidade do conjunto de dados utilizado para o treinamento. A representatividade e a diversidade do conjunto de dados são fundamentais para o desempenho do modelo na classificação correta das imagens como reais ou falsas. Por exemplo, um conjunto de dados com uma variedade de rostos, iluminação, qualidade de imagem e técnicas de *deepfake* aumentará a robustez do nosso modelo em situações reais.

Nesse sentido, a seleção apropriada do conjunto de dados e o uso cuidadoso de técnicas de aprendizado profundo são partes vitais da nossa metodologia para a detecção de *deepfakes*.

3.1 Conjunto de dados

Para o treinamento e teste dos modelos, foi empregado o conjunto de dados DFGC-21 (PENG *et al.*, 2021). Este conjunto de dados ou *dataset* é derivado do Celeb-DF (LI *et al.*, 2020), uma base ampla e diversificada de vídeos *deepfake* que contém mais de 5000 amostras envolvendo indivíduos de diferentes grupos étnicos e faixas etárias. A subseção escolhida do DFGC-21 consiste em cerca de 1000 imagens extraídas desses vídeos, proporcionando um bom conjunto de exemplos para o aprendizado do modelo.

É importante ressaltar, contudo, que a diversidade de *deepfakes* pode ser um desafio para a eficácia dos modelos finais. A variação nas técnicas de criação de *deepfakes*, nas características faciais dos indivíduos, na qualidade da imagem e em outros aspectos pode impactar

Figura 6 – Comparação entre uma imagem real e uma imagem gerada por *deepfake*

(a) Fake (b) Real

Fonte: (PENG *et al.*, 2021).

a capacidade do modelo de generalizar com eficácia para *deepfakes* não vistos anteriormente. Além disso, a quantidade limitada de *deepfakes* disponíveis pode representar uma restrição à robustez do modelo.

Para tentar contornar este problema, utilizou-se de uma técnica chamada *data augmentation*, essa estratégia nos permite aumentar a diversidade do conjunto de treinamento, de modo a melhorar a robustez dos modelos. Esta técnica envolve a criação de novas amostras aplicando uma série de transformações variadas nas imagens existentes, como rotação, deslocamento, *zoom*, *flip*, etc. Isso permite que aumentemos artificialmente o conjunto de dados aumentando a variedade e o número de exemplos para aprendizagem do modelo.

Mesmo com as limitações existentes, entre as opções de conjuntos de dados publicamente disponíveis na internet, o DFGC-21 se mostrou o mais adequado e completo para as necessidades do nosso trabalho, oferecendo uma combinação relevante de diversidade e complexidade.

O *dataset* utiliza-se majoritariamente o método de *FaceSwap* para criação do *deepfakes*, com alguns pós-processamentos para dificultar a classificação. O *dataset* é separado em vários subconjuntos criados aplicando métodos diferentes como pode-se observar na tabela 1.

3.2 Preparação dos dados

A preparação dos dados é uma etapa crucial para a eficácia de qualquer modelo de aprendizado de máquina. No caso deste trabalho, foi adotado um procedimento em várias etapas para preparar as imagens do conjunto de dados DFGC-21 para o treinamento do nosso modelo.

O conjunto de dados fornecido foi pré-processado com o uso da *Multi-task Cascaded Convolutional Networks (MTCNN)* (ZHANG *et al.*, 2016), um método popular para a detecção de rostos em imagens. Foram testadas outras maneiras como o algoritmo *Face Extractor* (FE)

Tabela 1 – Subconjuntos do *dataset*

Subconjunto	**Método**
real_fulls	Dados reais do conjunto original
fake_baseline	Dados falsos do conjunto original
DFGC_SYSU_852924	Adversarial Attacks com pós-processamento
jerryHUST_853638	FaceShifter+Adversarial Attacks+pós-processamento
miaotao_853000	FaceShifter
seanseattle_853068	FaceController + Adversarial Attacks
yZzzzzz_849853	MegaFS
DFischerHDA_852673	FaceMorpher + dlib landmarks
joshhu_853266	Adversarial Attacks
nbhh_853436	FaceShifter + Adversarial Attacks
smartz_849705	Anonimização facial
yangquanwei_852303	Troca facial baseando-se em áreas chave
zhaobh_852336	Modelo Adversarial para adicionar ruído
ctmiu_853213	FaceShifter + Adversarial Attacks
lowtec_853184	FacceShifter com pós-processamento
wany_853175	Faceshifter
yuejiang_852934	Recortar e colar

Fonte: (PENG *et al.*, 2021).

Figura 7 – Comparação entre os métodos de extração de face

(a) CV2 **(b)** MTCNN **(c)** FE

Fonte: Autor

e o OpenCV (BRADSKI, 2000), porém os melhores resultados foram obtidos pela MTCNN como pode-se observar comparação da figura 7. O objetivo desta etapa foi isolar e extrair as faces presentes em cada imagem. Isto é especialmente importante para o nosso problema, já que os *deepfakes* geralmente manipulam rostos em imagens ou vídeos.

Após a extração dos rostos, como dito anteriorment, recorremos à técnica de *data augmentation* para aumentar a diversidade e o tamanho do nosso conjunto de dados. Esta etapa permitiu que nosso modelo aprendesse a partir de uma variedade maior de exemplos, aumentando sua robustez.

Finalmente, o conjunto de dados foi dividido em três partes: treinamento, teste e validação. Seguimos a proporção de 70/20/10 para este fim, resultando em 6893 imagens para

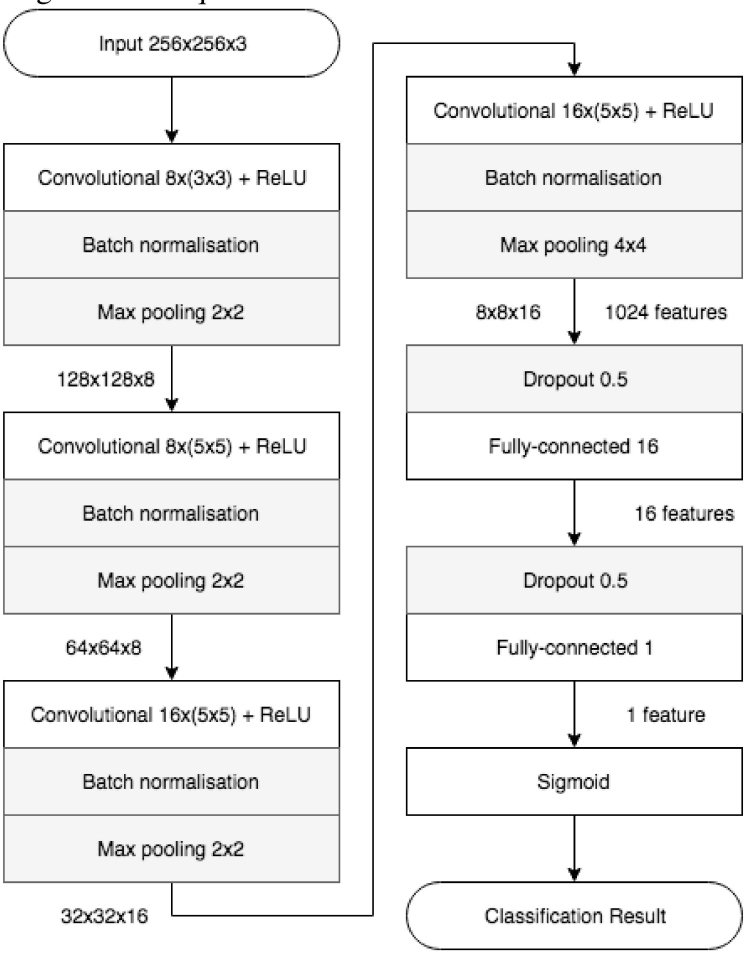
treinamento, e 2954 imagens para teste e validação.

Em suma, a preparação dos dados envolveu a extração de rostos, *data augmentation* e a separação dos dados em conjuntos de treinamento, teste e validação. Através destes passos, asseguramos que o modelo fosse treinado em um conjunto de dados diversificado e representativo, permitindo uma avaliação precisa de seu desempenho.

3.3 Desenvolvimento do Modelo

No desenvolvimento da nossa rede neural convolucional (CNN), diversos modelos foram experimentados e analisados para a detecção de *deepfakes*. O ponto de partida foi a arquitetura proposta pelo modelo MesoNet, que foi implementado e testado conforme descrito no trabalho (AFCHAR *et al.*, 2018). No entanto, os resultados obtidos com essa arquitetura inicial não foram satisfatórios para o conjunto de dados. Mesmo utilizando bibliotecas diferentes para a extração de faces, não houve melhoras substanciais.

Figura 8 – Arquitetura da MesoNet



Fonte: (AFCHAR *et al.*, 2018).

Na arquitetura demonstrada pela figura 8 foram utilizadas as camadas da biblioteca Keras (CHOLLET *et al.*, 2015). O modelo utilizado consta da utilização de uma camada Conv2D que realiza uma multiplicação na entrada, tendo como resultado diferentes matrizes juntado os pixels de cada canal de cor da imagem de entrada em um único pixel. A camada *BatchNormalization* é responsável por normalizar os dados, agora modificados pela operação convolucional da camada anterior. *Flatten*, a próxima camada, é uma operação que transforma as matrizes de entrada em um único *array* unidimensional, que no que lhe concerne é a entrada da parte totalmente conectada do modelo. Na parte totalmente conectada da rede, existem camadas *Dropouts*, responsáveis por evitar *overfitting* (evitando que o modelo se ajuste muito bem ao conjunto específico de treinamento, se tornando ineficaz em produzir novos resultados). A camada *Leaky* realiza um limite utilizando ReLU, qualquer neurônio que o resultado da ativação seja zero é multiplicado por um escalar (no caso 0.01).

Em busca de melhorar o desempenho, tentativas de integrar a MesoNet com outras CNNs pré-treinadas foram realizadas. A ideia era aproveitar a capacidade destas redes em identificar particularidades e reconhecer padrões, uma vez que são amplamente utilizadas para esses propósitos em muitos outros problemas de visão computacional. As redes pré-treinadas consideradas incluíram a VGG16, VGG19, a EfficientNetB0, a EfficientNetB7 e a ResNet.

Apesar de essas redes serem notáveis na literatura e amplamente utilizadas para uma variedade de tarefas de classificação, os resultados ainda não foram satisfatórios quando aplicados ao nosso problema específico inicialmente. Os modelos modificados apresentaram desempenho semelhante ao do modelo original da MesoNet, com acurácias de validação em torno de 40%.

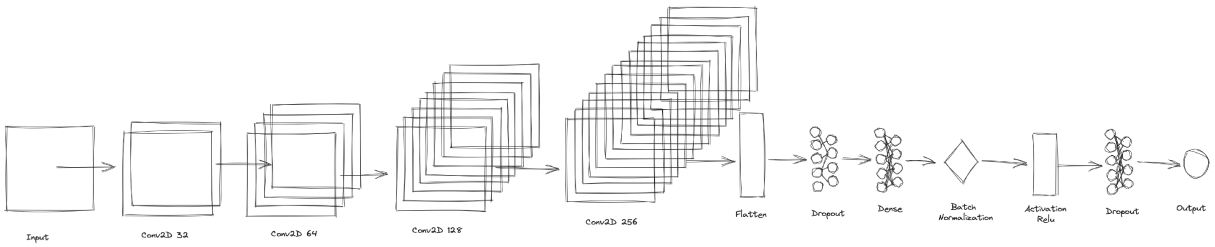
Com base na arquitetura inicial, foram adicionadas novas camadas convolucionais com o intuito de aprimorar a capacidade do modelo em capturar características mais complexas e sutis presentes nas imagens de *deepfakes*. Este processo de refinamento da arquitetura foi iterativo, com cada alteração sendo rigorosamente testada para avaliar seu impacto no desempenho do modelo.

A nova arquitetura, como mostrada na figura 9, resultante desse processo, ainda é fortemente influenciada pela MesoNet, mas possui mais profundidade e complexidade para lidar com a intrincada tarefa de detecção de *deepfakes*. A primeira modificação feita foi aumentar o número de camadas convolucionais com o propósito de capturar *features* mais complexas das entradas, adicionando também funções de ativação entre cada camada convolucional e

ainda aumentando a quantidade de unidades nas camadas Denses para melhor representação do aprendizado, este modelo foi batizado com o nome de Eilhart. Juntamente com as modificações nas arquiteturas utilizaram-se técnicas como *Early Stoping* e F-Score para não só ter uma ideia melhor sobre o desempenho das novas redes como otimizar o tempo gasto em modelos ruins quando a perda se mantém a mesma depois de várias épocas do treinamento.

O F-Score é a média harmônica de precisão e sensibilidade. A média harmônica é apropriada nesse caso porque ela é mais sensível a valores extremos. Portanto, para obter um F-Score alto, tanto a precisão quanto a sensibilidade devem ser altas. Se qualquer uma delas for baixa, o F-Score também será baixo.

Figura 9 – Nova Arquitetura



Fonte: O autor.

A partir dessa arquitetura utilizamos *Transfer learning* numa tentativa de que com mais camadas convolucionais poderíamos ter um uso mais eficaz de redes como a Efficientnet e VGG, redes pré-treinadas com os *wieghts* da ImageNet que já possui uma habilidade notável de reconhecer características úteis de imagens, o que é bem útil para obtermos melhores resultados.

4 RESULTADOS

Aqui são apresentados os resultados obtidos durante a fase experimental da pesquisa. A finalidade do estudo é avaliar a eficácia dos modelos desenvolvidos na tarefa de detecção de *deepfakes*. As métricas escolhidas para classificar foram acurácias, precisão, sensibilidade, especificidade e o valor F1 que são métricas comuns para avaliar modelos de classificação binários.

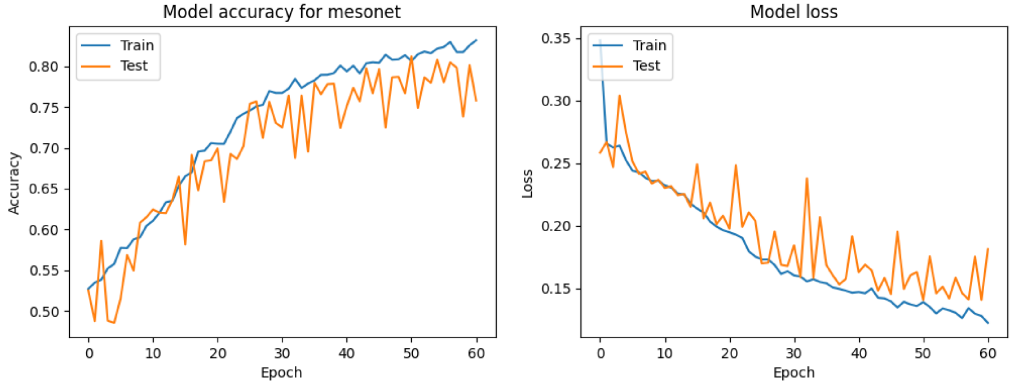
Para cada modelo desenvolvido, será apresentado inicialmente os resultados obtidos em termos das métricas de desempenho. Em seguida, esses resultados são discutidos e comparados para identificar o modelo ou modelos que se mostraram mais eficazes na detecção de *deepfakes* em nosso conjunto de dados. Além disso, também é apresentado as *Receiver operating characteristic* / Característica de Operação do Receptor (ROC) dos modelos para fornecer uma visão mais detalhada de seu desempenho.

4.1 MesoNet

O modelo MesoNet (AFCHAR *et al.*, 2018) foi implementado conforme descrito no trabalho original, e treinado com a mesma base de dados descrita na metodologia deste trabalho.

A figura 10 mostra a acurácia e a perda ao longo das épocas tanto no treinamento quanto no teste. A acurácia do modelo melhora progressivamente ao longo das épocas com o pico de 83% de acurácia no treinamento e 82% no teste demonstrando efetivamente que o modelo está melhorando e aprendendo a diferenciar *deepfakes*. Simultaneamente, a *loss* do modelo, também apresentada na Figura 10, diminui ao longo das épocas, sinalizando que o MesoNet está se tornando cada vez mais preciso em suas previsões.

Figura 10 – Acurácia e *Loss* do modelo MesoNet



Fonte: O autor.

Na tabela 2 é mostrado um panorama mais detalhado sobre as métricas do modelo MesoNet.

Tabela 2 – Estatísticas de desempenho do modelo MesoNet

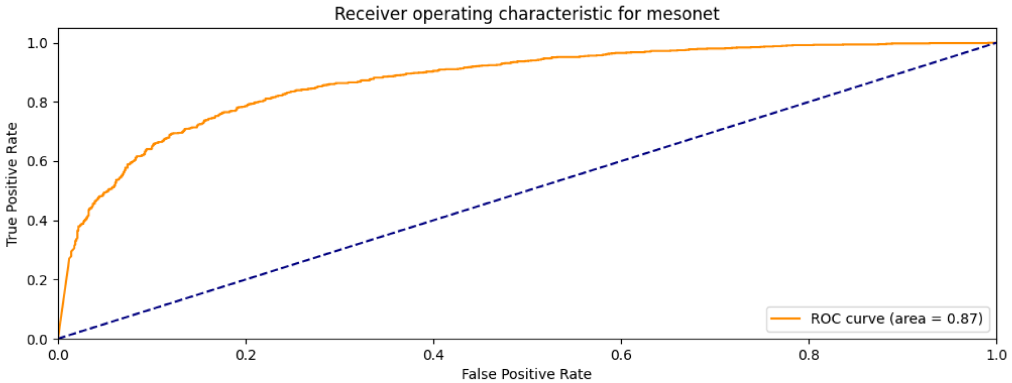
Item	**Max**	**Mean**	**Min**	**Std**
Loss	0.294028	0.173960	0.120012	0.041895
Accuracy	0.837555	0.741364	0.531507	0.083605
True Positives	2712	2367.261905	1634	304.378372
True Negatives	2583	2315.190476	1664	232.924292
False Positives	1490	838.809524	571	232.924292
False Negatives	1528	794.738095	450	304.378372
Val Loss	0.512160	0.212585	0.128619	0.081458
Val Accuracy	0.821429	0.691991	0.487723	0.102268
Val True Positives	874	649.619048	22	203.811766
Val True Negatives	921	590.428571	0	253.330214
Val False Positives	918	326.785714	2	253.291902
Val False Negatives	847	225.166667	0	202.833607

Fonte: O autor.

Observando as médias principalmente de validação os resultados estão abaixo do esperado, por exemplo, a média da acurácia da validação é de 69% sugerindo um desempenho mediano do modelo.

A figura 11 confirma que o modelo teve um bom desempenho, com uma *Area Under Curve* / Área sob a curva (AUC) de 0.87. A análise da ROC revela o *trade-off* entre a sensibilidade e a especificidade do MesoNet, indicando que o modelo alcançou resultados positivos. A curva próxima do canto superior esquerdo demonstra que o modelo possui uma boa capacidade de distinguir entre as classes, garantindo tanto uma alta sensibilidade quanto especificidade. Esses resultados apontam que o MesoNet é um ponto de partida sólido para a tarefa em questão.

Figura 11 – ROC MesoNet



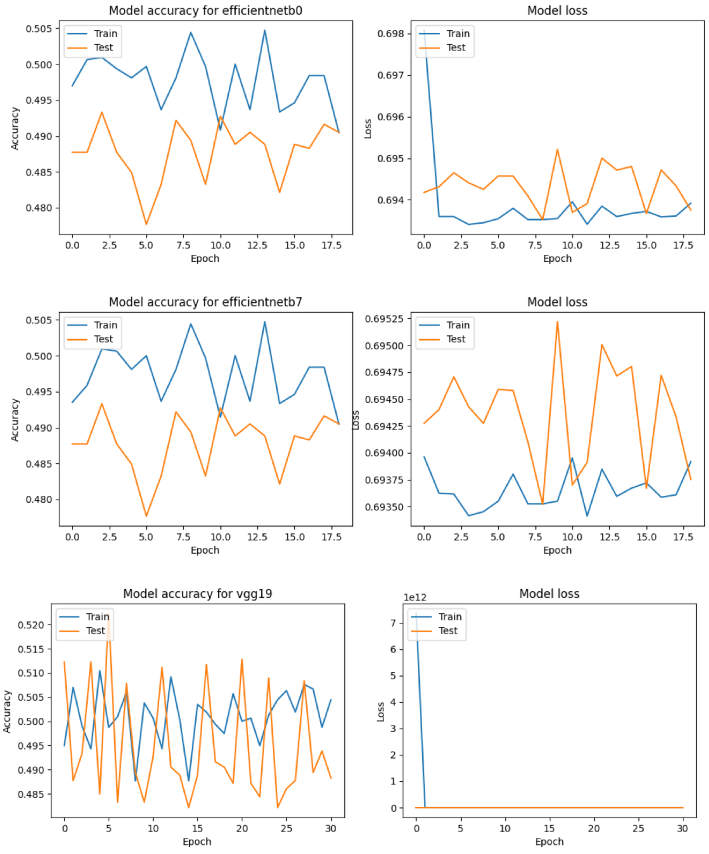
Fonte: O autor.

4.2 EfficientNet e VGG19

Numa tentativa de verificar o quão bom seriam modelos que utilizassem puramente redes pré-treinadas para classificar as redes em *deepfake*, foram treinados dois modelos com arquiteturas equivalentes, onde as camadas iniciais são as EfficientNetB0, EfficientNetB7 ou VGG19, ambos com pesos da ImageNet.

A figura 12 mostra que os modelos apresentam resultados inferiores à MesoNet em termos de acurácia e *loss*. Além disso, é observado que o número de épocas é baixo devido ao critério de parada, onde o treinamento é encerrado se o *loss* não melhora por 5 épocas consecutivas. Embora isso possa não ser uma regra absoluta, nesse experimento específico, era esperado encerrar o treinamento para modelos que ficaram estagnados por muito tempo. Isso indica que a MesoNet foi mais eficaz em alcançar melhores resultados para a tarefa em questão.

Figura 12 – Acurácia e *Loss* dos modelos utilizando EfficientNet e VGG19

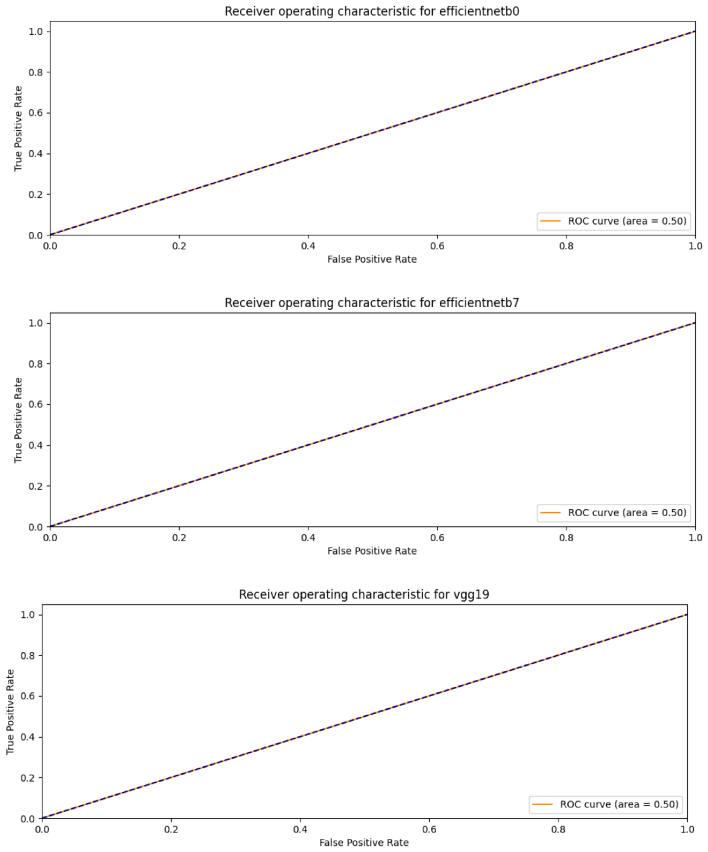


Fonte: O autor.

Podemos observar também nas figura 13, a AUC é de 0,5, o que nos indica que nossos modelos têm um desempenho tão bom quanto um classificador aleatório. Embora tanto a EfficientNet quanto a VGG sejam redes pré-treinadas muito utilizadas, por si só não conseguiram

realizar bem a tarefa de identificar os *deepfakes*.

Figura 13 – ROC dos modelos utilizando EfficientNet e VGG19



Fonte: O autor.

4.3 Eilhart

Partindo da arquitetura da MesoNet que teve bons resultados, aumentamos o número de camadas convolucionais de modo a melhorar a capacidade do modelo de entender as *features* do *dataset* e assim poder obter resultados ainda melhores, essa nova rede será chamada de Eilhart.

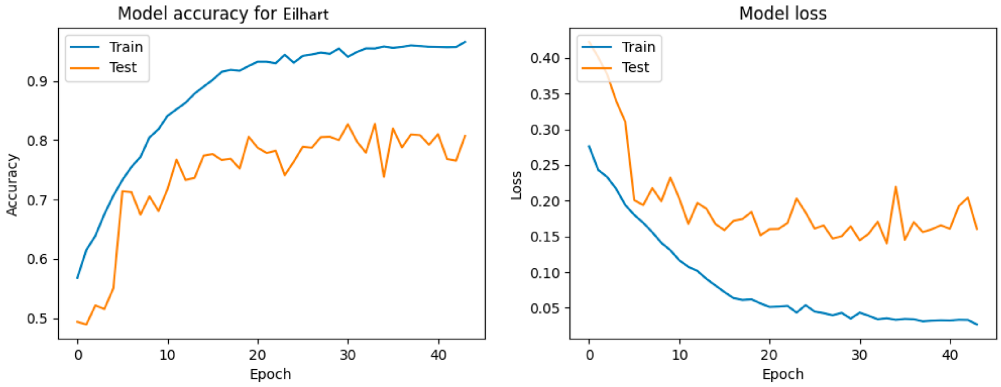
Conforme apresentado na figura 14, os gráficos são similares ao modelo MesoNet comum, como era de se esperar, mas com resultados um pouco melhores, com picos de 85% e 71% de acurácia para teste e validação respectivamente.

A tabela 3 mostra um panorama mais detalhado sobre as métricas do modelo Eilhart.

Outro indicativo da melhora do modelo é a AUC, que foi de 0,87 para 0,90 como mostra a figura 15.

Como aumentar as camadas convolucionais funcionou muito bem, incrementamos o

Figura 14 – Acurácia e *Loss* do modelo Eilhart



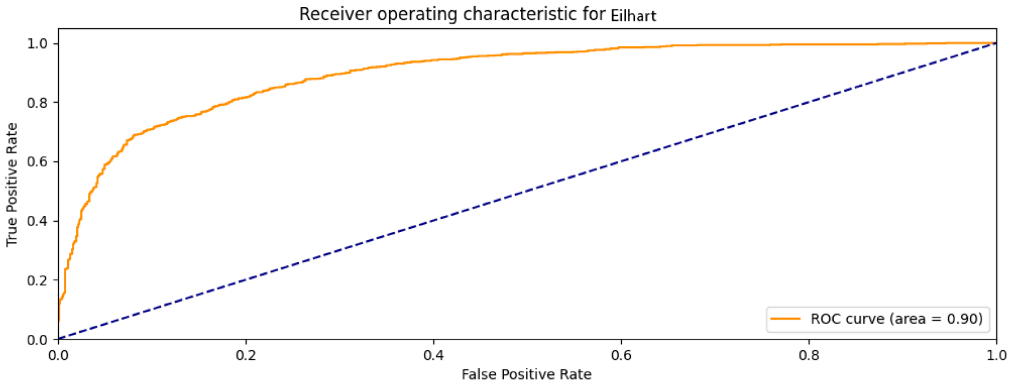
Fonte: O autor.

Tabela 3 – Estatísticas de desempenho do modelo Eilhart

item	max	mean	min	std
Loss	0.273774	0.104148	0.039163	0.066080
Accuracy	0.948543	0.850915	0.569031	0.105738
True Positives	2997	2663.724138	1668	359.796293
True Negatives	2994	2710.655172	1926	308.487218
False Positives	1228	443.344828	160	308.487218
False Negatives	1494	498.275862	165	359.796293
Val Loss	0.510903	0.224806	0.153502	0.089261
Val Acuracy	0.791853	0.710437	0.487165	0.084873
Val True Positives	876	698.275862	438	98.128310
Val True Negatives	852	574.827586	0	220.042170
Val False Positives	918	343.206897	67	220.317855
Val False Negatives	435	175.689655	0	98.063703

Fonte: O autor.

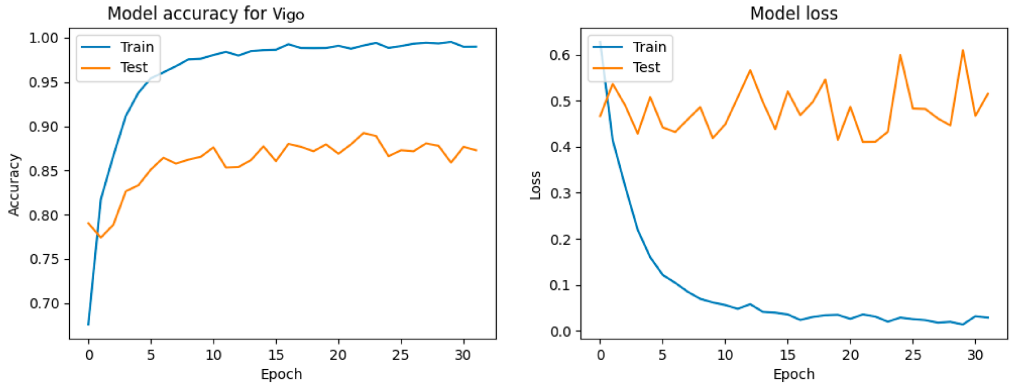
Figura 15 – ROC Eilhart



Fonte: O autor.

modelo com ainda mais camadas convolucionais com camadas *dense* entre elas, o que auxiliou na classificação das imagens no *dataset* como é possível visualizar na figura 16, este novo modelo recebeu o nome de Vigo.

Figura 16 – Acurácia e *Loss* do modelo Vigo



Fonte: O autor.

4.4 Transfer Learning

Mesmo com os resultados promissores do modelo Vigo, podemos aprimorar ainda mais os resultados aplicando a técnica de *transfer learning*. Utilizando redes pré-treinadas como ponto de partida, é possível melhorar o desempenho. Embora as redes por si só possam não ser excelentes em classificação de *deepfakes*, como vimos na sessão sobre EfficientNet e VGG19, podemos aproveitar a capacidade de extração de características dessas redes e incorporá-las aos nossos modelos. Dessa forma, criamos versões do nosso modelo que utilizam EfficientNet e VGG19 como ponto de entrada, resultando em melhorias no desempenho global.

Aplicando a técnica de *transfer learning* damos origem a dois modelos: Vengerberg que é o modelo criado a partir da junção do modelo Vigo com a EfficientNetB7, e o modelo Merigold que é a junção do modelo Vigo com a rede VGG19.

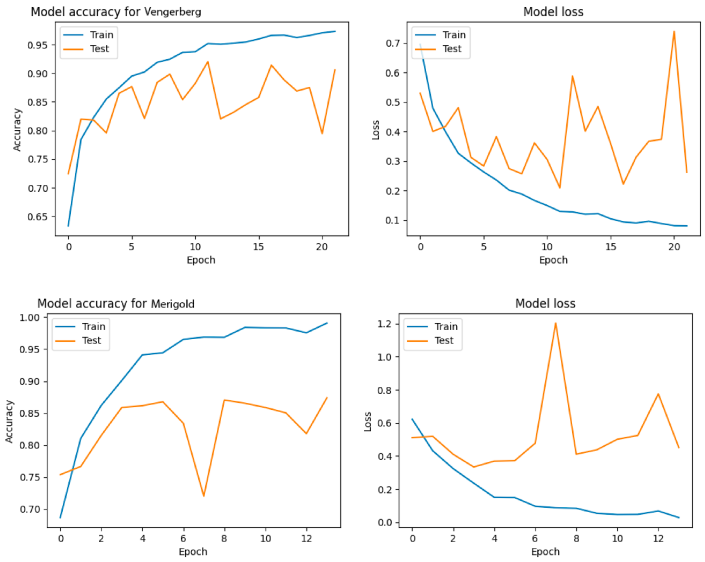
Observa-se ainda com mais detalhes nas tabelas 4 e 5 todas as métricas dos modelos com *transfer learning*.

Tabela 4 – Métricas do modelo Merigold

Item	**Max**	**Mean**	**Min**	**Std**
Loss	0.641125	0.178238	0.033856	0.172716
Accuracy	0.989234	0.922973	0.668778	0.087546
True Positives	3128	2893.142857	2002	309.076703
True Negatives	3120	2936.357143	2222	243.938987
False Positives	932	217.642857	34	243.938987
False Negatives	1160	268.857143	34	309.076703
Val Loss	0.604396	0.464197	0.352537	0.064229
Val Accuracy	0.875000	0.839086	0.756138	0.036825
Val True Positives	793	681.071429	484	83.070077
Val True Negatives	888	822.571429	715	56.783693
Val False Positives	210	95.428571	25	56.909344
Val False Negatives	400	192.928571	73	83.390188

Fonte: O autor.

Figura 17 – Acurácia e *Loss* dos modelos Vengerberg e Merigold



Fonte: O autor.

Tabela 5 – Métricas do modelo Vengerberg

Item	**Max**	**Mean**	**Min**	**Std**
Loss	0.706902	0.178226	0.052863	0.149192
Accuracy	0.981951	0.925023	0.627771	0.075929
True Positives	3104	2915	1946	256.007004
True Negatives	3098	2927.448276	2019	224.084588
False Positives	1135	226.551724	56	224.084588
False Negatives	1216	247	58	256.007004
Val Loss	1.141355	0.362103	0.192734	0.202537
Val Accuracy	0.925781	0.863801	0.713170	0.058161
Val True Positives	872	754.965517	523	97.255257
Val True Negatives	900	792.965517	408	120.611798
Val False Positives	508	125.068966	10	121.279367
Val False Negatives	351	119	5	98.950702

Fonte: O autor.

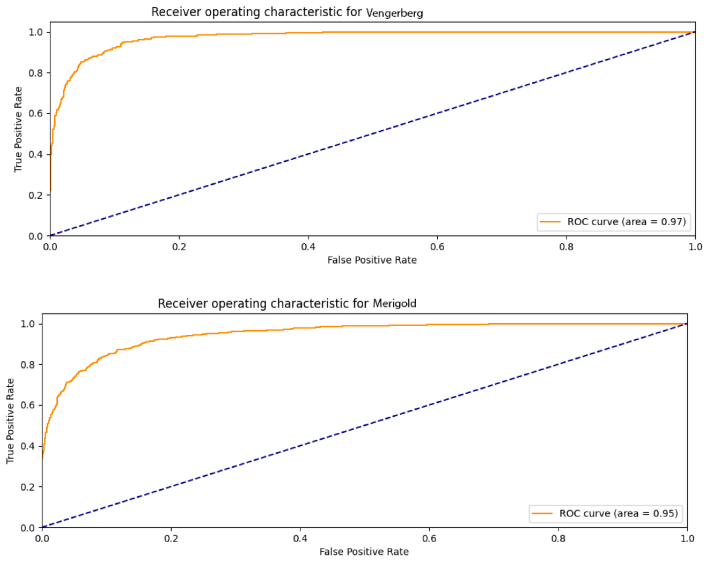
A figura 18 representa a ROC dos dois modelos, onde se observa um resultado ainda melhor, bem próximo do canto superior esquerdo do gráfico.

Analisando mais profundamente os resultados, observando o modelo Vengerberg, percebemos que a perda média é de 0,362. Acompanhada pela taxa de verdadeiros positivos com média de 755 e média de 793 de verdadeiros negativos. O que é um fator indicativo da boa capacidade de classificação do modelo. Além disso, a média de 125 falsos positivos e 119 falsos negativos nos permite afirmar que existe um bom equilíbrio entre a sensibilidade e especificidade do modelo.

Nas figuras 19 e 20 é exemplificado comparações entre os melhores modelos, tanto de acurácia como a comparação das F-scores (positivos e negativos verdadeiros e falsos)

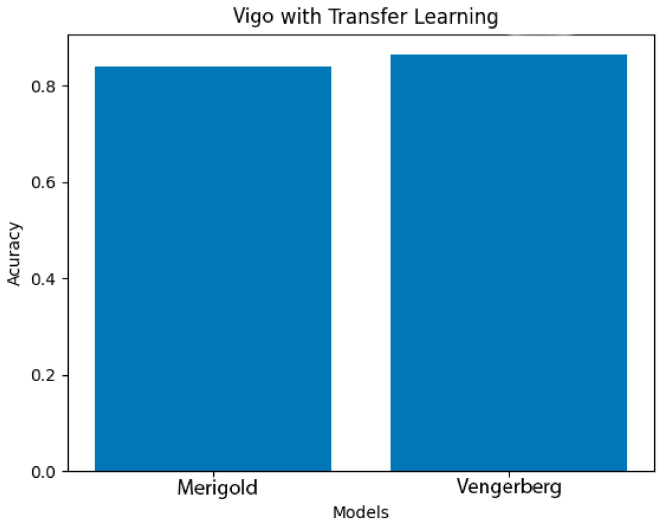
Utilizando os valores da tabela 6 chegamos então nos valores de 0.857 para precisão

Figura 18 – Acurácia e *Loss* dos modelos Vengerberg e Merigold



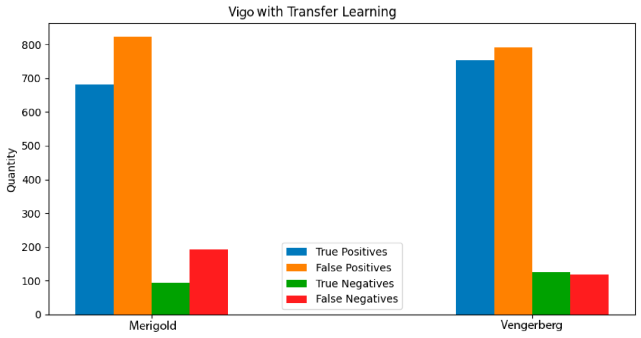
Fonte: O autor.

Figura 19 – Comparação da acurácia dos melhores modelos



Fonte: O autor.

Figura 20 – F-Score dos modelos



Fonte: O autor.

Tabela 6 – verdadeiros/falsos positivos e negativos dos melhores modelos

Model	True Positives	True Negatives	False Positives	False Negatives
Merigold	681.071429	22.571429	95.428571	192.928571
Vengerberg	754.965517	792.965517	125.068966	119.000000

Fonte: O autor.

e 0.864 para sensibilidade do modelo Vengerberg e 0.877 e 0.779 para a precisão e a sensibilidade do modelo Merigold, respectivamente.

Com base nesses cálculos, o modelo Merigold tem maior precisão, enquanto o modelo Vengerberg tem maior sensibilidade. No entanto, ao olhar para a média harmônica desses dois valores (F-Score), que representa uma medida de desempenho geral que equilibra precisão e sensibilidade, pode-se determinar qual modelo tem um desempenho geral melhor. Percebe-se que o modelo Vengerberg se destaca com um F-Score de 0.860, em comparação com o F-Score de 0.824 do modelo Merigold. Esta análise sugere que o modelo Vengerberg tem um desempenho melhor no equilíbrio entre a precisão e a sensibilidade. Contudo, é fundamental lembrar que o F-Score é apenas uma métrica. O desempenho real do modelo deve ser avaliado considerando outros fatores, como a relevância relativa da precisão e sensibilidade para o problema específico, assim como o desempenho do modelo em diferentes conjuntos de dados.

5 CONCLUSÕES E TRABALHOS FUTUROS

A detecção de *deepfakes* é uma tarefa importante e desafiadora na área de processamento de imagens e vídeos. A utilização de redes neurais convolucionais tem se mostrado uma abordagem promissora para enfrentar esse desafio, oferecendo resultados precisos e eficientes em termos de tempo de processamento.

Ao longo deste trabalho, foi possível observar que a detecção de *deepfakes* com redes neurais convolucionais envolve várias etapas, desde a seleção e pré-processamento dos dados, passando pelo treinamento e ajuste de parâmetros da rede, até a validação e teste do modelo.

Além disso, o desempenho da detecção de *deepfakes* depende muito da qualidade e quantidade dos dados utilizados no treinamento da rede, assim como da escolha adequada dos parâmetros dos modelos.

Embora os resultados dos testes realizados nos modelos: 'MesoNet Custom 512 with EfficientNet' e 'MesoNet Custom 512 with VGG19', tenham sido promissores, com uma *f-score* de 86,44 e 72,38 e acurácia de 86,3% e 83,9%, respectivamente, com dados reservados para esse teste, é importante salientar que essa alta precisão não garante necessariamente que os modelos sejam completamente confiáveis para a detecção de *deepfakes* em todas as situações possíveis.

Primeiramente, vale lembrar que o desempenho do modelo depende fortemente da qualidade e quantidade dos dados utilizados para o treinamento. A quantidade de dados disponíveis para o treinamento dos modelos em nosso estudo, embora significativa, está longe de cobrir toda a gama de possíveis *deepfakes* que podem ser encontrados em situações reais. Portanto, a capacidade dos modelos de generalizar a partir dos dados de treinamento para novos exemplos de *deepfakes* ainda é uma questão em aberto.

Além disso, as limitações de hardware também impõem restrições ao treinamento dos modelos. Para alcançar uma precisão de detecção próxima a 100%, provavelmente seriam necessários modelos de rede neural muito mais complexos e computacionalmente exigentes do que os que foram possíveis de implementar com o hardware disponível em nosso estudo.

Um exemplo ilustrativo dessas limitações pode ser encontrado nos resultados do Deepfake Detection Challenge (DFGC), promovido pelo Facebook. Nesse desafio, o modelo de melhor desempenho no conjunto de dados público alcançou 82.56% de precisão média, uma medida comum de precisão para tarefas de visão computacional. No entanto, ao avaliar os participantes contra o conjunto de dados *black box* (dados privados e não disponíveis ao público), a classificação dos modelos de melhor desempenho mudou significativamente. O modelo com

melhor desempenho foi um modelo apresentado por Selim Seferbekov, que alcançou uma precisão média de apenas 65.18% (DEEPFAKE. . . , 2023). Isso demonstra que, mesmo com uma alta precisão em um conjunto de dados de teste, um modelo pode não ser necessariamente eficaz ao se deparar com *deepfakes* desconhecidos.

Em suma, embora os modelos de detecção de *deepfakes* apresentados neste trabalho mostrem resultados encorajadores, é preciso ter cautela ao interpretar esses resultados. A detecção confiável de *deepfakes* é um problema complexo que vai além do treinamento de um modelo em um conjunto de dados específico. Mais pesquisa, mais dados e mais recursos de hardware são necessários para o desenvolvimento de modelos que possam efetivamente lidar com a variedade e a complexidade dos *deepfakes* no mundo real.

Portanto, concluímos que o uso de redes neurais convolucionais para detecção de *deepfakes* é uma abordagem promissora que requer cautela, tanto na seleção e pré-processamento dos dados quanto no ajuste dos parâmetros do modelo, para obter resultados precisos e confiáveis na identificação de *deepfakes*.

5.1 Trabalhos Futuros

Embora tenham sido alcançados progressos significativos na detecção de *deepfakes*, os avanços contínuos nas técnicas de geração de imagens artificiais destacam a necessidade de pesquisa contínua e desenvolvimento de ferramentas de detecção mais robustas. Recentemente, modelos baseados em difusão estável (stable diffusion) começaram a ser utilizados para gerar imagens sintéticas de alta qualidade, o que coloca novos desafios para a comunidade de pesquisa.

Os modelos de difusão estável, tais como aqueles apresentados por Jonathan Ho, Ajay Jain e Pieter Abbeel (HO *et al.*, 2020), têm a capacidade de criar imagens altamente realistas como também arte baseado em qualquer estilo que se treine estes modelos. Estes modelos representam um avanço significativo em relação aos métodos anteriores de geração de imagens, tais como as GAN. Portanto, há uma necessidade urgente de desenvolver métodos de detecção de *deepfakes* que sejam capazes de identificar imagens geradas por modelos de difusão estável.

Em 2023, Alessandro Coccomini e colaboradores publicaram um trabalho destacando a criação de modelos capazes de identificar imagens geradas por modelos de difusão estável (COCCOMINI *et al.*, 2023). Além disso, em 2022, Riccardo Corvi e sua equipe também publicaram um trabalho abordando a detecção dessas imagens (CORVI *et al.*, 2022). Esses

avanços demonstram um potencial significativo no desenvolvimento de modelos cada vez mais eficazes na identificação de conteúdos criados por meio de técnicas avançadas de geração, como os modelos de difusão estável.

Em trabalhos futuros, espera-se investigar a aplicabilidade dos métodos de detecção de *deepfakes* existentes a imagens geradas por modelos de difusão estável. Além disso, a exploração da concepção de novos métodos de detecção, especificamente adaptados para enfrentar este novo desafio, poderia ser uma área de pesquisa relevante. Esta é uma área extremamente importante, considerando a evolução contínua das técnicas de geração de *deepfakes* e o impacto potencialmente disruptivo que essas imagens sintéticas podem ter em diversos campos, incluindo segurança, mídia e política.

REFERÊNCIAS

RICCARDO Corvi and Davide Cozzolino and Giada Zingarini and Giovanni Poggi and Koki Nagano and Luisa Verdoliva. Globo, 2023. Disponível em: <https://g1.globo.com/politica/noticia/2023/08/10/silvinei-diz-a-pf-que-blitze-eram-para-coibir-compra-de-votos-e-nao-para-impedir-lulistas-de-votar.ghtml>.

AFCHAR, D.; NOZICK, V.; YAMAGISHI, J.; ECHIZEN, I. MesoNet: a compact facial video forgery detection network. In: **2018 IEEE International Workshop on Information Forensics and Security (WIFS)**. IEEE, 2018. Disponível em: <https://doi.org/10.1109%2Fwifs.2018.8630761>.

ALARCON, N. **How Kaggle Makes GPUs Accessible to 5 Million Data Scientists**. 2020. Disponível em: <https://developer.nvidia.com/blog/how-kaggle-makes-gpus-accessible-to-5-million-data-scientists/>.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. **Pattern Recognition**, v. 30, n. 7, p. 1145–1159, 1997. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320396001422>.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

CHOLLET, F. *et al.* **Keras**. GitHub, 2015. Disponível em: <https://github.com/fchollet/keras>.

COCCOMINI, D. A.; ESULI, A.; FALCHI, F.; GENNARO, C.; AMATO, G. **Detecting Images Generated by Diffusers**. 2023.

CORVI, R.; COZZOLINO, D.; ZINGARINI, G.; POGGI, G.; NAGANO, K.; VERDOLIVA, L. **On the detection of synthetic images generated by diffusion models**. 2022.

DEEPFAKE Detection Challenge Dataset — ai.facebook.com. 2023. <https://ai.facebook.com/datasets/dfdc/>. [Accessed 01-Jun-2023].

DEEPTRACE. **The State of Deepfake - Landscape, Threats, and Impact**. 2019. Disponível em: <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf>.

FAWCETT, T. An introduction to roc analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>.

GONZALEZ, R.; WOODS, R. **Digital Image Processing**. Prentice Hall, 2008. ISBN 9780131687288. Disponível em: <https://books.google.com.br/books?id=8uGOnjRGEzoC>.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. **Explaining and Harnessing Adversarial Examples**. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1412.6572>.

GOUTTE, C.; GAUSSIER, É. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: **European Conference on Information Retrieval**. [S.l.: s.n.], 2005.

HEATON, J. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning. **Genetic Programming and Evolvable Machines**, v. 19, n. 1, p. 305–307, Jun 2018. ISSN 1573-7632. Disponível em: <https://doi.org/10.1007/s10710-017-9314-z>.

HO, J.; JAIN, A.; ABBEEL, P. **Denoising Diffusion Probabilistic Models**. 2020.

KORSHUNOV, P.; MARCEL, S. **DeepFakes: a New Threat to Face Recognition? Assessment and Detection**. 2018.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, p. 84 – 90, 2012.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LI, Y.; YANG, X.; SUN, P.; QI, H.; LYU, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2020.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; BOURDEV, L.; GIRSHICK, R.; HAYS, J.; PERONA, P.; RAMANAN, D.; ZITNICK, C. L.; DOLLáR, P. **Microsoft COCO: Common Objects in Context**. arXiv, 2014. Disponível em: <https://arxiv.org/abs/1405.0312>.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. van der; GINNEKEN, B. van; SÁNCHEZ, C. I. A survey on deep learning in medical image analysis. **Medical Image Analysis**, Elsevier BV, v. 42, p. 60–88, dec 2017. Disponível em: <https://doi.org/10.1016\%2Fj.media.2017.07.005>.

PENG, B.; FAN, H.; WANG, W.; DONG, J.; LI, Y.; LYU, S.; LI, Q.; SUN, Z.; CHEN, H.; CHEN, B.; HU, Y.; LUO, S.; HUANG, J.; YAO, Y.; LIU, B.; LING, H.; ZHANG, G.; XU, Z.; MIAO, C.; LU, C.; HE, S.; WU, X.; ZHUANG, W. **DFGC 2021: A DeepFake Game Competition**. 2021.

POSTERS, B. **Imagine this...** 2019. Disponível em: <https://www.instagram.com/p/ByaVigGFP2U/?utm_source=ig_embed&ig_rid=3e977738-ff29-4ee9-9669-943008378708>.

RANA, M. S.; NOBI, M. N.; MURALI, B.; SUNG, A. H. Deepfake detection: A systematic literature review. **IEEE Access**, v. 10, p. 25494–25513, 2022.

REDMON, J.; FARHADI, A. Yolo9000: Better, faster, stronger. **arXiv preprint arXiv:1612.08242**, 2016.

SASAKI, Y. The truth of the f-measure. **Teach Tutor Mater**, 01 2007.

TWITCH streamer Atrioc gives tearful apology after paying for deepfakes of female streamers - Dexerto — dexerto.com. 2023. <https://www.dexerto.com/entertainment/twitch-streamer-atrioc-gives-tearful-apology-after-paying-for-deepfakes-of-female-streamers-2047162/>. [Accessed 31-May-2023].

WANG, P. **This Person Does Not Exist**. 2022. Disponível em: <https://thispersondoesnotexist.com>.

WANG, S.-C. Artificial neural network. In: ____. **Interdisciplinary Computing in Java Programming**. Boston, MA: Springer US, 2003. p. 81–100. ISBN 978-1-4615-0377-4. Disponível em: <https://doi.org/10.1007/978-1-4615-0377-4_5>.

WU, Z.; LIM, S.-N.; DAVIS, L.; GOLDSTEIN, T. **Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors**. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1910.14667>.

ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y. Joint face detection and alignment using multitask cascaded convolutional networks. **IEEE Signal Processing Letters**, Institute of Electrical and Electronics Engineers (IEEE), v. 23, n. 10, p. 1499–1503, oct 2016. Disponível em: <https://doi.org/10.1109%2Flsp.2016.2603342>.