Lorrayne Somerlatte dos Santos

O Impacto das Configurações de Hiperparâmetros no Reconhecimento Facial: Uma Avaliação do Dataset FairFace

Belo Horizonte

2024

Lorrayne Somerlatte dos Santos

O Impacto das Configurações de Hiperparâmetros no Reconhecimento Facial: Uma Avaliação do Dataset FairFace

Pesquisa Científica

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas – ICEx
Departamento de Ciência da Computação

Orientadora: Ana Paula Couto da Silva

Belo Horizonte

2024

Sumário

1 Introdução

A análise de bases de dados faciais que incluem uma ampla variedade de informações sobre pessoas, é fundamental para o desenvolvimento de sistemas de reconhecimento facial. Esses dados oferecem uma visão completa e detalhada das características e comportamentos dos usuários, incluindo histórico de interações, preferências e demografia. No entanto, a complexidade dos modelos analíticos utilizados nesse processo pode apresentar desafios significativos. Em (SANTOS et al., 2023) há a discussão sobre como as tecnologias de reconhecimento facial têm se espalhado pelo mundo e causado controvérsia em seu uso e eficácia, principalmente em populações de diferentes etnias. Nesse contexto, torna-se relevante discutir a configuração e o viés presentes nos algoritmos de *Deep Learning* para reconhecimento facial. Esses modelos, embora poderosos, podem apresentar desafios relacionados à interpretação e à equidade, especialmente quando consideramos os hiperparâmetros envolvidos.

Ao avaliar modelos de reconhecimento facial, a necessidade de interpretabilidade e previsões explicáveis é evidente. A utilização de modelos opacos, que não permitem uma interpretação clara, pode gerar desconfiança e preocupações sobre possíveis práticas discriminatórias, que são justificadas por estudos como o de (BUOLAMWINI; GEBRU, 2018) que mostra que os sistemas comerciais de verificação facial têm um desempenho significativamente inferior na identificação de pessoas negras, especialmente mulheres negras. Analisar o impacto da variação de hiperparâmetros em modelos de *Deep Learning* se torna uma tarefa crucial para verificar a eficácia preditiva e a transparência dos algoritmos.

Neste contexto, o objetivo deste trabalho é investigar de forma mais detalhada como a variação dos hiperparâmetros em modelos de *Deep Learning* afeta o desempenho do reconhecimento facial. Ao analisar essas variações, não apenas buscamos compreender os efeitos dos hiperparâmetros na eficácia preditiva dos modelos, mas também entender como essas mudanças impactam a interpretabilidade e a transparência das decisões de reconhecimento facial. A partir das relações encontradas entre hiperparâmetros e o desempemho dos modelos, podemos desenvolver estratégias mais eficazes para garantir a equidade e a justiça nas aplicações de reconhecimento facial. Além disso, a compreensão desses aspectos contribui para o avanço no desenvolvimento de modelos de *Deep Learning* mais responsáveis e éticos, promovendo uma tomada de decisão mais transparente e alinhada com os princípios de equidade e não discriminação.

1.1 Objetivo Geral

O objetivo geral deste estudo é investigar como a modificação de hiperparâmetros específicos influencia o desempenho de modelos de reconhecimento facial baseados em *Deep Learning*, especialmente no contexto de variabilidade racial. Além disso, visa-se avaliar o impacto da escolha desses hiperparâmetros na precisão e robustez dos modelos em reconhecer diferentes raças. Com base nos resultados obtidos, pretende-se propor diretrizes para a seleção de hiperparâmetros que otimizem o desempenho dos modelos de reconhecimento facial, levando em consideração diferentes configurações de parâmetros e cenários de aplicação, como ambientes com iluminação variável, diferentes ângulos de captura e diversidade de características faciais.

1.2 Objetivos Específicos

Objetivo 1 - Seleção do Modelo:

- Realizar uma análise detalhada de modelos de *Deep Learning* para reconhecimento facial, considerando suas características e desempenho em diferentes raças.
- Implementar e conduzir testes dos modelos selecionados, variando funções de ativação, métodos de otimização e esquemas de inicialização de pesos.

Objetivo 2 - Experimentação e Análise:

- Preparar o conjunto de dados FairFace (KARKKAINEN; JOO, 2021) para treinar e testar os modelos finais selecionados, incluindo etapas de pré-processamento e balanceamento dos dados.
- Treinar os modelos finais com o conjunto de dados preparado e realizar uma análise detalhada do impacto dos hiperparâmetros no desempenho dos modelos, especialmente em relação à precisão e robustez para diferentes raças.

Objetivo 3 - Elaboração dos Resultados:

- Analisar os resultados obtidos na experimentação, considerando o impacto dos hiperparâmetros na precisão, robustez e equidade dos modelos de reconhecimento facial.
- Discutir as conclusões encontradas e propor diretrizes para a seleção de hiperparâmetros que otimizem o desempenho dos modelos de reconhecimento facial, levando em conta a variabilidade racial e diferentes configurações de dados.

2 Referencial Teórico

Na literatura, o *Deep Learning* tem sido amplamente explorado em diversas áreas, como saúde, finanças e justiça criminal. Em (DOSHI-VELEZ; KIM, 2017), discute-se o desafio da interpretabilidade em modelos de *Deep Learning*, especialmente devido à sua complexidade e natureza opaca.

Os autores em (SANTOS et al., 2023) abordam a interseção entre racismo e tecnologia, destacando o reconhecimento facial e explorando como essa tecnologia pode criar invisibilidades ou reforçar visibilidades, particularmente em relação aos corpos negros e à realidade brasileira. Também discute a politização da gestão algorítmica e enfatiza a necessidade de ampliar as vozes que denunciam o racismo na produção de técnicas supostamente imparciais.

O estudo de (MAGNO; BEZERRA, 2020) revela que 90% das 151 pessoas detidas com base em câmeras de reconhecimento facial são negras. Ele destaca como essa tecnologia pode se tornar uma ameaça para populações socialmente vulneráveis.

A pesquisa de (BRITO; COLAVOLPE, 2023) examina como a diferença de raças afeta o desempenho dos sistemas de reconhecimento facial ao identificar suspeitos de crimes. O viés racial nos algoritmos resulta em taxas de erro desiguais para diferentes grupos raciais, com maior impacto em pessoas negras e outras minorias étnicas. O artigo destaca a necessidade de regulamentação rigorosa e consideração ética ao implementar essa tecnologia.

Com base nas conclusões e metodologias apresentadas nos estudos mencionados, este trabalho pretende investigar como a alteração dos hiperparâmetros afeta o desempenho dos modelos de Deep Learning no reconhecimento facial. Em particular, o estudo focará no impacto dessas alterações na precisão e na acurácia dos modelos, além de examinar como esses ajustes podem influenciar a presença de vieses raciais. A pesquisa busca entender como a configuração dos hiperparâmetros pode otimizar o desempenho dos modelos, garantindo uma maior equidade na identificação facial de diferentes grupos raciais.

3 Metodologia

Para atingir os objetivos propostos neste estudo, foi adotada uma abordagem metodológica composta por várias etapas distintas. O fluxo de trabalho mostrado a seguir descreve as principais fases da metodologia.

3.1 Entendimento e Caracterização dos Dados

A primeira etapa consistiu em adquirir um entendimento completo dos dados da base de dados FairFace (KARKKAINEN; JOO, 2021). Isso envolveu a análise preliminar dos dados para compreender sua natureza, estrutura e conteúdo. Foi realizado a visualização das imagens e rótulos, seguido de uma análise exploratória dos dados para entender a distribuição das classes raciais. Além disso, identificou-se possíveis desbalanceamentos e peculiaridades dos dados para uma caracterização inicial mais precisa da base de dados, permitindo identificar as variáveis-chave e seus atributos.

3.2 Filtragem e Preparação dos Dados

Após o entendimento e caracterização dos dados, foi iniciada a fase de filtragem e preparação dos dados. Nessa etapa, foram aplicados critérios de seleção para reter apenas os dados que eram relevantes e adequados para as análises de desempenho e robustez dos modelos de reconhecimento facial. A preparação dos dados envolveu o redimensionamento e normalização das imagens, codificação dos rótulos e balanceamento dos conjuntos de treinamento e validação para garantir uma distribuição equitativa das classes raciais.

3.3 Implementação e Treinamento dos Modelos

Com a base de dados refinada e preparada, foram implementados e treinados diferentes modelos de *Deep Learning*, variando hiperparâmetros específicos. As atividades realizadas incluíram a definição e implementação das arquiteturas de redes neurais convolucionais (CNNs) (LECUN et al., 1998), o treinamento dos modelos com diferentes combinações de funções de ativação, métodos de otimização e esquemas de inicialização de pesos, além do monitoramento e registro do desempenho dos modelos durante o treinamento.

3.3.1 Definição e Implementação das Arquiteturas de Redes Neurais Convolucionais (CNNs)

As redes neurais convolucionais (CNNs) foram escolhidas para a construção dos modelos de *Deep Learning* devido à sua eficácia comprovada em tarefas de reconhecimento de padrões em imagens. As CNNs são especialmente adequadas para processamento de dados visuais por várias razões: as camadas convolucionais capturam automaticamente características espaciais hierárquicas das imagens; o pooling reduz a dimensionalidade, preservando as informações mais relevantes; e os parâmetros compartilhados reduzem a complexidade do modelo, facilitando o treinamento e melhorando a generalização.

3.3.2 Escolha das Funções de Ativação

As funções de ativação são fundamentais para determinar o comportamento dos neurônios em redes neurais. Este estudo analisa três funções de ativação específicas:

- **ReLU (Rectified Linear Unit)** (NAIR; HINTON, 2010):
  - **Motivação**: Amplamente utilizada devido à sua simplicidade e eficácia em mitigar o problema do gradiente desaparecendo, permitindo que redes profundas sejam treinadas de forma mais eficiente.
  - **Comportamento**: Ativa apenas valores positivos, introduzindo não-linearidade ao modelo.
- **Sigmoid** (Rumelhart; Hinton; Williams, 1986):
  - **Motivação**: Comumente usada em camadas de saída para problemas de classificação binária, sendo analisada aqui para fins de comparação.
  - **Comportamento**: Mapeia valores de entrada para um intervalo entre 0 e 1, útil para interpretar probabilidades de classes.
- **Tanh (Tangente Hiperbólica)** (LECUN et al., 1998):
  - **Motivação**: Ideal para normalizar dados de entrada, gerando valores entre -1 e 1, o que pode levar a um aprendizado mais rápido em comparação com a função Sigmoid.
  - **Comportamento**: Comparada à função Sigmoid, a Tanh é zero-centralizada, o que pode levar a um aprendizado mais rápido.

3.3.3 Métodos de Otimização

Os métodos de otimização são responsáveis por ajustar os pesos do modelo para minimizar a função de perda. A função de perda quantifica a diferença entre os valores previstos por um modelo e os valores reais observados, medindo o custo associado a um evento ou a um conjunto de valores, representando o quão bem ou mal o modelo está performando (ZHOU et al., 2019). Dois métodos foram explorados:

- **SGD (Stochastic Gradient Descent)** (SOYDANER, 2020):
  - **Motivação**: O SGD é a base para muitos métodos de otimização mais avançados e fornece uma comparação direta.
  - **Características**: Atualiza os parâmetros iterativamente, usando gradientes calculados a partir de mini-lotes, o que pode levar a uma convergência mais estável mas potencialmente mais lenta.
- **Adam (Adaptive Moment Estimation)** (KINGMA; BA, 2017):
  - **Motivação**: Combina as vantagens de dois outros métodos de otimização: RMSProp e Stochastic Gradient Descent (SGD) com momentum.
  - **Características**: Adaptativamente ajusta a taxa de aprendizado para cada parâmetro, convergindo mais rápido e com mais estabilidade em problemas ruidosos.

3.3.4 Esquemas de Inicialização de Pesos

A inicialização dos pesos influencia significativamente a velocidade de convergência e o desempenho do modelo. Dois esquemas foram avaliados:

- **Inicialização Normal** (THIMM; FIESLER, 1995):
  - **Motivação**: Utiliza uma distribuição normal com média zero e um pequeno desvio padrão para inicializar os pesos, garantindo que os valores sejam pequenos.
  - **Características**: Ajuda a evitar a saturação de neurônios na função de ativação Sigmoid ou Tanh.
- **Inicialização He** (DATTA, 2020):
  - **Motivação**: Desenvolvida especificamente para redes com funções de ativação ReLU, esta inicialização usa uma distribuição normal com variância escalada.
  - **Características**: Proporciona uma melhor propagação dos sinais nas camadas iniciais, evitando a diminuição do gradiente.

3.3.5 Monitoramento e Registro do Desempenho dos Modelos

Durante o treinamento dos modelos de reconhecimento facial, o desempenho foi monitorado e registrado continuamente. Foram utilizadas várias métricas e métodos para essa avaliação, descritos a seguir:

- **Acurácia** (MONICO JOÃO FRANCISCO DAL PÓZ, 2009): A acurácia mede a proporção de previsões corretas em relação ao total de previsões feitas. É calculada pela fórmula:

$$\text{Acurácia} = \frac{\text{Número de Previsões Corretas}}{\text{Total de Previsões}} \tag{3.1}$$

  Ela fornece uma visão geral de quão bem o modelo está performando na tarefa de reconhecimento facial, mas pode ser enganosa em conjuntos de dados desbalanceados.

- **Precisão (Precision)** (ARORA; KANJILAL; VARSHNEY, 2016): A precisão mede a proporção de verdadeiros positivos em relação ao total de positivos preditos, refletindo a exatidão das previsões positivas do modelo. É calculada pela fórmula:

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Positivos}} \tag{3.2}$$

  Esta métrica é particularmente útil quando o custo de falsos positivos é alto, como em sistemas de segurança que utilizam reconhecimento facial.

- **Revocação (Recall)** (ARORA; KANJILAL; VARSHNEY, 2016): A revocação, também conhecida como sensibilidade, mede a proporção de verdadeiros positivos em relação ao total de positivos reais. É calculada pela fórmula:

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos + Falsos Negativos}} \tag{3.3}$$

  Esta métrica é importante quando o custo de falsos negativos é alto, como na identificação de suspeitos em vigilância por reconhecimento facial.

- **F1-Score** (YACOUBY; AXMAN, 2020): O F1-Score é a média harmônica entre precisão e revocação, proporcionando um equilíbrio entre as duas. É particularmente útil quando há uma necessidade de balancear precisão e revocação. É calculado pela fórmula:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão + Revocação}} \tag{3.4}$$

  Esta métrica oferece uma única pontuação que leva em consideração tanto falsos positivos quanto falsos negativos, sendo útil para avaliar o desempenho global do modelo de reconhecimento facial.

A acurácia de treinamento e validação foi medida para avaliar o desempenho geral do modelo, enquanto a perda de treinamento e validação foi monitorada para verificar o ajuste do modelo aos dados. Esses registros foram cruciais para uma análise detalhada do impacto dos hiperparâmetros e para identificar as melhores configurações, visando otimizar tanto o desempenho quanto a equidade do modelo de reconhecimento facial.

4 Resultados e Análises

Nesta seção, são apresentados os resultados das avaliações de diferentes modelos de *Deep Learning* implementados e treinados com variações específicas de hiperparâmetros. Abaixo, seguem os resultados obtidos para cada modelo em relação às diferentes raças presentes no dataset FairFace.

4.1 Modelo 1: Relu + Adam + Normal Initialization

O primeiro modelo foi treinado utilizando a função de ativação *Relu*, o otimizador *Adam* e a inicialização de pesos normal com média 0 e desvio padrão de 0.01. Abaixo estão as métricas de precisão, recall e f1-score e acurácia para cada classe racial:

Tabela 1 – Avaliação do Modelo 1: Relu + Adam + Normal Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.49	0.66	0.49
Leste Asiático	1	1.00	0.56	0.72	0.56
Indiano	2	1.00	0.34	0.51	0.34
Latino/Hispânico	3	1.00	0.34	0.51	0.34
Oriente Médio	4	1.00	0.12	0.21	0.12
Sudeste Asiático	5	1.00	0.20	0.34	0.20
Branco	6	1.00	0.53	0.69	0.53

O Modelo 1, utilizando *Relu*, *Adam* e inicialização normal, apresentou desempenho variado entre as diferentes classes raciais. A precisão foi alta (1.00) para todas as classes, mas o recall e o F1-Score mostraram deficiências significativas, especialmente em classes como Negro, Indiano, Latino/Hispânico, Oriente Médio e Sudeste Asiático, onde a acurácia foi abaixo de 0.56. Isso sugere que o modelo teve dificuldades em identificar corretamente a maioria dos exemplos nessas classes, resultando em baixas acurácias e recalls.

A raça Branco apresentou o melhor desempenho com uma acurácia de 0.53, recall de 0.53 e F1-Score de 0.69. Apesar de ser o melhor resultado entre as classes, ainda há considerável espaço para melhorias. Essas variações destacam a necessidade de ajustes no modelo para melhorar a identificação e a acurácia geral para todas as classes raciais.

4.2 Modelo 2: Relu + SGD + Normal Initialization

O segundo modelo foi treinado utilizando a função de ativação *Relu*, o otimizador *SGD* e a inicialização de pesos normal com média 0 e desvio padrão de 0.01. Abaixo estão as métricas de precisão, recall e f1-score para cada classe racial, além da acurácia:

Tabela 2 – Avaliação do Modelo 2: Relu + SGD + Normal Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.47	0.64	0.47
Leste Asiático	1	1.00	0.50	0.66	0.50
Indiano	2	1.00	0.21	0.35	0.21
Latino/Hispânico	3	1.00	0.38	0.55	0.38
Oriente Médio	4	0.00	0.00	0.00	0.00
Sudeste Asiático	5	0.00	0.00	0.00	0.00
Branco	6	1.00	0.19	0.32	0.19

A Tabela 2 indica que a precisão para cada raça foi consistentemente 1.00, mas o recall e o F1-Score apresentaram diferenças significativas. A raça Negro teve um recall de 0.47 e um F1-Score de 0.64, mostrando um desempenho razoável em comparação com outras raças. A raça Leste Asiático obteve um recall de 0.50 e um F1-Score de 0.66, destacando-se ligeiramente em relação à Negro. Em contraste, as raças Oriente Médio e Sudeste Asiático não foram detectadas nesse modelo.

A acurácia geral do modelo também reflete a disparidade no desempenho entre as raças. A acurácia para Negro e Leste Asiático foi de 0.47 e 0.50, respectivamente, sugerindo que o modelo tem um desempenho razoável nessas categorias. No entanto, a acurácia caiu drasticamente para raças como Oriente Médio e Sudeste Asiático, onde foi registrada como 0.00, indicando que o modelo não conseguiu identificar essas classes de forma eficaz. A raça Branco teve a menor acurácia, com 0.19, demonstrando dificuldades na identificação correta. Esses resultados indicam que o modelo enfrenta desafios significativos na identificação de algumas raças e sugerem a necessidade de ajustes no treinamento ou na arquitetura do modelo para melhorar o desempenho geral.

4.3 Modelo 3: Tanh + Adam + Normal Initialization

O terceiro modelo foi treinado utilizando a função de ativação *Tanh*, o otimizador *Adam* e a inicialização de pesos normal com média 0 e desvio padrão de 0.01. Abaixo estão as métricas dos resultados:

Tabela 3 – Avaliação do Modelo 3: Tanh + Adam + Normal Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.66	0.80	0.66
Leste Asiático	1	1.00	0.47	0.64	0.47
Indiano	2	1.00	0.32	0.48	0.32
Latino/Hispânico	3	1.00	0.12	0.21	0.12
Oriente Médio	4	1.00	0.01	0.03	0.01
Sudeste Asiático	5	1.00	0.16	0.28	0.16
Branco	6	1.00	0.66	0.80	0.66

A Tabela 3 demonstra que o modelo teve um desempenho particularmente alto para as raças Negro e Branco, com uma precisão e F1-Score de 1.00 e um recall de 0.66 para ambas as classes. No entanto, o desempenho foi muito inferior para outras classes, com as métricas de precisão e recall de 1.00, mas F1-Scores muito baixos, indicando que o modelo teve dificuldades.

O modelo mostrou um desempenho muito baixo para as classes Oriente Médio e Latino/Hispânico, com precisão de 1.00 mas recall de apenas 0.01 e 0.12, respectivamente. Isso sugere uma capacidade limitada do modelo em generalizar para essas classes. A análise indica que, apesar de alcançar altos F1-Scores para algumas classes, a variabilidade na identificação de outras classes pode ser um desafio significativo para o modelo, especialmente com a função de ativação *Tanh*.

4.4 Modelo 4: Tanh + SGD + Normal Initialization

O quarto modelo foi treinado com a função de ativação *Tanh*, o otimizador *SGD* e a inicialização de pesos normal com média 0 e desvio padrão de 0.01. A seguir, apresentamos as métricas de precisão, recall e F1-Score para cada classe racial, bem como a acurácia do modelo:

Tabela 4 – Avaliação do Modelo 4: Tanh + SGD + Normal Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.35	0.52	0.35
Leste Asiático	1	1.00	0.22	0.36	0.22
Indiano	2	1.00	0.29	0.45	0.29
Latino/Hispânico	3	1.00	0.30	0.46	0.30
Oriente Médio	4	1.00	0.06	0.11	0.06
Sudeste Asiático	5	1.00	0.28	0.44	0.28
Branco	6	1.00	0.54	0.70	0.54

Esse modelo mostrou um desempenho razoavelmente melhor para a classe Branco, com precisão de 1.00, recall de 0.54 e F1-Score de 0.70. No entanto, para outras classes, como Negro, Leste Asiático, Indiano e Sudeste Asiático, a precisão foi de 1.00, mas os recalls foram significativamente mais baixos, resultando em F1-Scores variáveis e indicando uma baixa taxa de identificação correta para essas classes.

O desempenho foi particularmente ruim para a classe Oriente Médio, com recall de 0.06 e F1-Score de 0.11, sugerindo dificuldades significativas na classificação correta dessas imagens. Em resumo, o Modelo 4 mostrou uma capacidade limitada de generalização para a maioria das classes, apesar da alta precisão para algumas.

4.5 Modelo 5: Relu + Adam + He Initialization

O quinto modelo foi treinado utilizando a função de ativação *Relu*, o otimizador *Adam* e a inicialização de pesos de He. Abaixo, seguem as métricas de precisão, recall e F1-Score, assim como a acurácia do modelo:

Tabela 5 – Avaliação do Modelo 5: Relu + Adam + He Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.62	0.77	0.62
Leste Asiático	1	1.00	0.50	0.66	0.50
Indiano	2	1.00	0.23	0.38	0.23
Latino/Hispânico	3	1.00	0.35	0.51	0.35
Oriente Médio	4	1.00	0.06	0.11	0.06
Sudeste Asiático	5	1.00	0.12	0.21	0.12
Branco	6	1.00	0.46	0.63	0.46

O Modelo 5 apresentou uma precisão de 1.00 para todas as classes raciais, mas mostrou variações significativas no recall e no F1-Score. Para as classes Negro e Leste Asiático, o modelo teve um desempenho relativamente bom, com F1-Scores de 0.77 e 0.66, respectivamente. No entanto, para as classes Indiano, Latino/Hispânico, Oriente Médio e Sudeste Asiático, o desempenho foi consideravelmente inferior, com F1-Scores variando de 0.11 a 0.51.

Esses resultados indicam que, embora o modelo seja preciso, ele enfrenta dificuldades em termos de recall e F1-Score para a maioria das classes menos representadas. A combinação de *Relu*, *Adam* e a inicialização de He parece ser eficaz apenas para algumas classes, sugerindo a necessidade de ajustes adicionais para melhorar a generalização e o desempenho do modelo para todas as classes raciais.

4.6 Modelo 6: Tanh + Adam + He Initialization

O sexto modelo foi treinado utilizando a função de ativação *Tanh*, o otimizador *Adam* e a inicialização de pesos de He. Seguem as métricas e a acurácia para cada classe racial:

Tabela 6 – Avaliação do Modelo 6: Tanh + Adam + He Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.53	0.69	0.53
Leste Asiático	1	1.00	0.45	0.62	0.45
Indiano	2	1.00	0.25	0.40	0.25
Latino/Hispânico	3	1.00	0.23	0.38	0.23
Oriente Médio	4	1.00	0.10	0.19	0.10
Sudeste Asiático	5	1.00	0.16	0.27	0.16
Branco	6	1.00	0.49	0.66	0.49

Nesse modelo, há uma precisão perfeita de 1.00 para todas as classes raciais, mas mostrou variações significativas no recall e no F1-Score. Para as classes Negro e Leste Asiático, o modelo teve um desempenho razoável, com F1-Scores de 0.69 e 0.62, respectivamente. No entanto, para as classes Indiano, Latino/Hispânico, Oriente Médio e Sudeste Asiático, o desempenho foi consideravelmente inferior, com F1-Scores variando de 0.19 a 0.40.

Os resultados mostram que, apesar da precisão do modelo, ele tem dificuldades com recall e F1-Score nas classes menos representadas. A combinação de *Tanh*, *Adam* e a inicialização de He parece funcionar bem apenas para algumas classes, indicando que são necessários ajustes adicionais para melhorar a generalização e o desempenho geral do modelo.

4.7 Modelo 7: Tanh + SGD + He Initialization

O sétimo modelo foi treinado utilizando a função de ativação *Tanh*, o otimizador *SGD* e a inicialização de pesos de He. Seguem os dados de sáida do modelo:

Tabela 7 – Avaliação do Modelo 7: Tanh + SGD + He Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.46	0.63	0.46
Leste Asiático	1	1.00	0.26	0.41	0.26
Indiano	2	1.00	0.19	0.33	0.19
Latino/Hispânico	3	1.00	0.23	0.38	0.23
Oriente Médio	4	1.00	0.14	0.25	0.14
Sudeste Asiático	5	1.00	0.05	0.10	0.05
Branco	6	1.00	0.45	0.62	0.45

O Modelo 7 apresentou uma precisão de 1.00 para todas as classes raciais, mas mostrou variações significativas no recall e no F1-Score. Para as classes Negro e Leste Asiático, o modelo teve um desempenho razoável, com F1-Scores de 0.63 e 0.41, respectivamente. Já para as classes Indiano, Latino/Hispânico, Oriente Médio e Sudeste Asiático, o desempenho foi consideravelmente inferior, com F1-Scores variando de 0.10 a 0.38.

Aqui, o modelo enfrenta desafios significativos em termos de recall e F1-Score para a maioria das classes menos representadas. A combinação de *Tanh*, *SGD* e a inicialização de He parece funcionar bem apenas para algumas classes, indicando a necessidade de ajustes adicionais.

4.8 Modelo 8: Sigmoid + SGD + He Initialization

O oitavo modelo foi treinado utilizando a função de ativação *Sigmoid*, o otimizador *SGD* e a inicialização de pesos de He. Seguem os resultados:

Tabela 8 – Avaliação do Modelo 8: Sigmoid + SGD + He Initialization

Raça	**Classe**	**Precisão**	**Recall**	**F1-Score**	**Acurácia**
Negro	0	1.00	0.05	0.10	0.05
Leste Asiático	1	1.00	0.57	0.73	0.57
Indiano	2	1.00	0.01	0.01	0.01
Latino/Hispânico	3	0.00	0.00	0.00	0.00
Oriente Médio	4	0.00	0.00	0.00	0.00
Sudeste Asiático	5	1.00	0.58	0.74	0.58
Branco	6	1.00	0.04	0.07	0.04

O Modelo 8 apresenta um desempenho bastante desigual entre as diferentes classes raciais. Para a classe Negro, embora a precisão seja perfeita (1.00), o recall é extremamente baixo (0.05), resultando em um F1-Score de apenas 0.10. A classe Leste Asiático, por outro lado, obteve uma precisão de 1.00, um recall de 0.57 e um F1-Score de 0.73, indicando um desempenho mais equilibrado. No entanto, o modelo enfrentou grandes dificuldades com as classes Indiano e Latino/Hispânico, apresentando um recall muito baixo (0.01) e um F1-Score de 0.01 para a classe Indiano, e desempenho nulo em todas as métricas para a classe Latino/Hispânico.

A classe Sudeste Asiático destacou-se com um recall de 0.58 e um F1-Score de 0.74, demonstrando o melhor desempenho entre todas as classes. Em contraste, a classe Branco, apesar de uma precisão de 1.00, teve um recall muito baixo (0.04), resultando em um F1-Score de 0.07. Em resumo, o Modelo 9 mostra uma precisão alta para todas as classes, mas enfrenta sérios problemas de recall e F1-Score, especialmente para as classes Latino/Hispânico, Oriente Médio e Branco.

5 Considerações finais e Trabalhos Futuros

Neste trabalho, foi investigado o impacto de diferentes configurações de hiperparâmetros em modelos de reconhecimento facial, com foco particular no desempenho do modelo em diversas etnias. O dataset FairFace foi utilizado para avaliar o desempenho de modelos de rede neural profunda com diferentes combinações de funções de ativação, otimizadores e métodos de inicialização de pesos. Comparando as funções de ativação ReLU, Tanh e Sigmoid, observou-se que ReLU apresentou desempenho mais consistente e melhor para a maioria das classes raciais. A função Tanh mostrou um desempenho variável, e a Sigmoid frequentemente resultou em baixa acurácia e F1-Score. A utilização do SGD (Stochastic Gradient Descent) se mostrou menos eficiente do que o Adam em termos de desempenho geral, frequentemente resultando em baixa acurácia e F1-Score, especialmente para classes menos representadas. A inicialização de He demonstrou um desempenho superior em comparação com a inicialização de Xavier e aleatória, principalmente quando combinada com funções de ativação não saturantes como ReLU.

A análise das métricas revelou que o desempenho do modelo variou significativamente entre as diferentes etnias. Em particular, as classes menos representadas, como Latino/Hispânico, Oriente Médio e Sudeste Asiático, apresentaram desafios notáveis, com baixa acurácia e F1-Score, sugerindo que o modelo pode estar enfrentando problemas de viés e generalização. Por outro lado, algumas classes como Leste Asiático e Negro obtiveram resultados mais consistentes, mas ainda apresentam margem para melhorias.

Para possíveis trabalhos futuros é possível investigar técnicas avançadas de balanceamento de dados e aumentar a diversidade do dataset pode melhorar a representação das classes. O uso de técnicas de aumento de dados e estratégias de amostragem pode ajudar a mitigar o viés. Testar diferentes arquiteturas de redes neurais, como redes convolucionais profundas (CNNs) mais sofisticadas e redes neurais profundas baseadas em atenção (transformers), pode oferecer melhorias no reconhecimento facial, especialmente para classes etnicamente diversas. Continuar a experimentação com diferentes combinações de funções de ativação, otimizadores e inicializações de pesos, realizando uma busca mais abrangente por hiperparâmetros, pode ajudar a encontrar configurações que melhoram o desempenho geral e reduzem o viés. Explorar métodos de regularização, como dropout e técnicas de regularização baseadas em dados, pode melhorar a robustez do modelo e a capacidade de generalização. Realizar validações extensivas dos modelos em cenários do mundo real e com dados de teste mais variados pode assegurar que o desempenho seja consistente em diferentes condições e contextos.

Em conclusão, embora os modelos investigados tenham demonstrado um desempenho promissor, há uma necessidade clara de mais refinamento e ajustes para alcançar uma precisão e

equidade melhores no reconhecimento facial. O contínuo desenvolvimento e inovação nesta área são essenciais para criar sistemas de reconhecimento facial mais robustos e justos.

Referências

ARORA, M.; KANJILAL, U.; VARSHNEY, D. Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, v. 12, n. 2, p. 224–236, 2016. PMID: 74482. Disponível em: <https://www.inderscienceonline.com/doi/abs/10.1504/IJICBM.2016.074482>. Citado na página 10.

BRITO, G. R. G.; COLAVOLPE, L. E. L. S. O impacto da diferença de raças no reconhecimento facial de suspeitos de crimes. *Boletim IBCCRIM*, v. 31, n. 365, p. 20–22, abr. 2023. Disponível em: <https://publicacoes.ibccrim.org.br/index.php/boletim_1993/article/view/469>. Citado na página 6.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: <https://proceedings.mlr.press/v81/buolamwini18a.html>. Citado na página 4.

DATTA, L. A survey on activation functions and their relation with xavier and he normal initialization. *CoRR*, abs/2004.06632, 2020. Disponível em: <https://arxiv.org/abs/2004.06632>. Citado na página 9.

DOSHI-VELEZ, F.; KIM, B. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. Citado na página 6.

KARKKAINEN, K.; JOO, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2021. p. 1548–1558. Citado 2 vezes nas páginas 5 e 7.

KINGMA, D. P.; BA, J. *Adam: A Method for Stochastic Optimization*. 2017. Disponível em: <https://arxiv.org/abs/1412.6980>. Citado na página 9.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 7.

LECUN, Y. et al. Efficient backprop. In: ____. *Neural Networks: Tricks of the Trade*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 9–50. ISBN 978-3-540-49430-0. Disponível em: <https://doi.org/10.1007/3-540-49430-8_2>. Citado na página 8.

MAGNO, M. E. d. S. P.; BEZERRA, J. S. Vigilância negra: O dispositivo de reconhecimento facial e a disciplinaridade dos corpos. *Novos Olhares*, v. 9, n. 2, p. 45–52, dez. 2020. Disponível em: <https://www.revistas.usp.br/novosolhares/article/view/165698>. Citado na página 6.

MONICO JOÃO FRANCISCO DAL PÓZ, A. P. G. M. C. D. S. M. C. D. O. L. G. AcurÁcia e precisÃo: Revendo os conceitos de forma acurada. *Boletim de Ciências Geodésicas*, 2009. ISSN 1413-4853. Disponível em: <https://www.redalyc.org/articulo.oa?id=393937709010>. Citado na página 10.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Madison, WI, USA: Omnipress, 2010. (ICML'10), p. 807–814. ISBN 9781605589077. Citado na página 8.

Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. , v. 323, n. 6088, p. 533–536, out. 1986. Citado na página 8.

SANTOS, L. G. d. M. et al. Reconhecimento facial: Tecnologia, racismo e construÇÃo de mundos possÍveis. *Psicologia Sociedade*, Associação Brasileira de Psicologia Social, v. 35, p. e277141, 2023. ISSN 0102-7182. Disponível em: <https://doi.org/10.1590/1807-0310/2023v35e277141>. Citado 2 vezes nas páginas 4 e 6.

SOYDANER, D. A comparison of optimization algorithms for deep learning. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 34, n. 13, p. 2052013, 2020. Disponível em: <https://doi.org/10.1142/S0218001420520138>. Citado na página 9.

THIMM, G.; FIESLER, E. Neural network initialization. In: MIRA, J.; SANDOVAL, F. (Ed.). *From Natural to Artificial Neural Computation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995. p. 535–542. ISBN 978-3-540-49288-7. Citado na página 9.

YACOUBY, R.; AXMAN, D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In: EGER, S. et al. (Ed.). *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. Online: Association for Computational Linguistics, 2020. p. 79–91. Disponível em: <https://aclanthology.org/2020.eval4nlp-1.9>. Citado na página 10.

ZHOU, Y. et al. Mpce: A maximum probability based cross entropy loss function for neural network classification. *IEEE Access*, v. 7, p. 146331–146341, 2019. Citado na página 9.