UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Adaptação de modelos de reconhecimento de expressões faciais em cenários com poucos dados rotulados

Gustavo Villela Guimarães

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP
SÃO CARLOS

Gustavo Villela Guimarães

Adaptação de modelos de reconhecimento de expressões faciais em cenários com poucos dados rotulados

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Dr. Brucce Neves dos Santos

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP, com os
dados inseridos pelo(a) autor(a)

G963a

Guimarães, Gustavo Villela . Adaptação de modelos de reconhecimento de expressões faciais em cenários com poucos dados rotulados / Gustavo Villela Guimarães; orientador Brucce Neves dos Santos. -- São Carlos, 2025.
52 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. FER. 2. Reconhecimento de Expressões Faciais. 3. Redes Neurais. 4. Reconhecimento de Emoções. 5. Modelos de Inteligência Artificial. I. dos Santos, Brucce Neves, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Gustavo Villela Guimarães

Adaptation of facial expression recognition models in scenarios with limited labeled data

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Dr. Brucce Neves dos Santos

Original version

São Carlos

2025

Dedico este trabalho primeiramente a Deus, Inteligência Suprema e causa primária de todas as coisas, que me permitiu trilhar a grandiosa experiência da vida.
Dedico igualmente à minha mãe, Sônia Maria Leal Villela (in memoriam), inspiração fundamental deste estudo. Ela me ensinou que as expressões faciais são o verdadeiro espelho da alma, uma linguagem que expressa os sentimentos mais genuínos, muito além do que as palavras podem descrever.

AGRADECIMENTOS

Aos meus alicerces, minha família. Agradeço de todo coração à minha esposa e aos meus filhos pelo amor, suporte incondicional e pela imensa paciência durante minha ausência, compreendendo as longas horas dedicadas aos estudos deste MBA. Vocês foram minha maior motivação.

Expresso minha profunda gratidão ao meu Diretor, amigo e incentivador, Luis Casuscelli. Seu apoio foi fundamental neste desafio, tanto no aspecto motivacional quanto pelo decisivo incentivo financeiro, viabilizado pela bolsa de estudos do programa de desenvolvimento profissional da empresa Bull Ltda.

No âmbito acadêmico, agradeço à Profa. Dra. Solange Rezende por todo o suporte e incentivo ao longo do MBA em Inteligência Artificial e *Big Data* da USP. Sou especialmente grato por sua indicação que me levou ao meu orientador, Dr. Brucce Neves dos Santos, a quem dedico um agradecimento especial. Sua orientação precisa, didática e apoio foram essenciais para que eu chegasse com sucesso ao final desta jornada.

O texto deste trabalho foi corrigido com o auxílio de inteligência artificial, usando o ChatGPT para refinar a linguagem e aprimorar a clareza ao longo deste estudo, sem alterar seu significado original, propósito acadêmico ou autoria.

"Nossa maior fraqueza está em desistir.
A maneira mais certa de ter sucesso é sempre tentar mais uma vez."
(Thomas Edison)

RESUMO

GUIMARAES, G.V. **Adaptação de modelos de reconhecimento de expressões faciais em cenários com poucos dados rotulados**. 2025. 52 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Esse estudo tem como objetivo analisar o desempenho de modelos de inteligência artifical no campo do reconhecimento de emoções faciais (FER) a partir de imagens estáticas. A pesquisa aborda o desafio central da escassez de dados anotados, um cenário comum em aplicações práticas da inteligência artificial, especialmente no contexto de interações humano-computador. O trabalho investiga o desempenho de modelos de redes neurais com arquiteturas distintas, buscando compreender sua capacidade de adaptação e generalização sob condições de dados limitados. Os resultados obtidos demonstram a viabilidade de alcançar otimizações relevantes em modelos de FER, mesmo diante de restrições significativas na disponibilidade de dados e de considerações de privacidade. Ao discutir as implicações práticas dos achados, este estudo ressalta o potencial da tecnologia de reconhecimento facial em diversas áreas, desde o monitoramento de bem-estar até aplicações de segurança e diagnóstico. Por fim, delineia as limitações inerentes à pesquisa e sugere direções para investigações futuras, consolidando a contribuição deste trabalho para o avanço do aprendizado de máquina em cenários de dados desafiadores.

Palavras-chave: FER. Reconhecimento de Expressões Faciais. Redes Neurais. Reconhecimento de Emoções. Modelos de Inteligência Artificial.

ABSTRACT

GUIMARAES, G.V. **Adaptation of facial expression recognition models in scenarios with limited labeled data**. 2025. 52 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This study has the objective to analyze the performance of artificial intelligence models in the field of facial emotion recognition (FER) from static images. The research addresses the central challenge of the scarcity of annotated data, a common scenario in practical applications of artificial intelligence, especially in the context of human-computer interactions. The work investigates the performance of neural network models with distinct architectures, seeking to understand their capacity for adaptation and generalization under conditions of limited data. The results obtained demonstrate the feasibility of achieving relevant optimizations in FER models, even in the face of significant constraints on data availability and privacy considerations. In discussing the practical implications of the findings, this study highlights the potential of facial recognition technology in various areas, ranging from well-being monitoring to security and diagnostic applications. Finally, it outlines the inherent limitations of the research and suggests directions for future investigations, consolidating this work's contribution to the advancement of machine learning in challenging data scenarios.

Keywords: FER. Facial Expression Recognition. Neural Networks. Emotion Recognition. Artificial Intelligence Models.

LISTA DE FIGURAS

Figura 1 – CNN de 4 camadas - Fonte: (Pallavi; Chavan, 2024) 26
Figura 2 – Visão de conexões residuais na arquitetura ResNet50. Fonte: (Kim, 2018) 27
Figura 3 – Arquitetura *Vision Transformer* (ViT) - Fonte: (Koyyada; Rawat; Singh, 2022) 28
Figura 4 – Fluxograma das Etapas de Processamento 31
Figura 5 – Matriz de confusão - ResNet *Fold* 2 com 20 amostras por classe 36
Figura 6 – Matriz de confusão - ViT *Fold* 1 com 20 amostras por classe 38
Figura 7 – Matriz de confusão - ViT *Fold* 2 com 10 amostras por classe 47
Figura 8 – Matriz de confusão - ResNet *Fold* 2 com 10 amostras por classe 48
Figura 9 – Matriz de confusão - ViT *Fold* 1 com 40 amostras por classe 49
Figura 10 – Matriz de confusão - ResNet *Fold* 2 com 40 amostras por classe 50
Figura 11 – Matriz de confusão - ViT *Fold* 5 com 100 amostras por classe 51
Figura 12 – Matriz de confusão - ResNet *Fold* 5 com 100 amostras por classe 52

LISTA DE TABELAS

Tabela 1 – Distribuição das amostras por classes 34
Tabela 2 – Métricas dos melhores modelos em cada execução para cada número de amostras. Entre parenteses está o desvio padrão. 35

SUMÁRIO

1 INTRODUÇÃO 23
1.1 Objetivo 24
1.2 Organização do Texto 24

2 FUNDAMENTAÇÃO TEÓRICA 25
2.1 Reconhecimento de Expressões Faciais 25
2.2 Redes Neurais Convolucionais 26
2.3 ResNet - Residual Network 27
2.4 Vision Transformer 28
2.5 Trabalhos Relacionados 29

3 METODOLOGIA 31

4 AVALIAÇÃO EXPERIMENTAL 33
4.1 Conjunto de Dados - AffectNet 33
4.2 Configurações Experimentais 34
4.3 Métricas de Avaliação e Seleção dos Modelos 35

5 CONCLUSÕES 41

REFERÊNCIAS 43

APÊNDICES 45

APÊNDICE A – MATRIZES DE CONFUSÃO DE TODAS AS AMOSTRAS 47

1 INTRODUÇÃO

As expressões faciais constituem um elemento fundamental na comunicação não verbal. Um sorriso caloroso pode irradiar simpatia, enquanto uma testa franzida pode comunicar desaprovação de maneira inequívoca. Segundo Ekman; Friesen (1978), as expressões faciais representam um componente universal da linguagem emocional, sendo essenciais para o entendimento e resposta social apropriada. Por esse motivo a análise de expressões faciais é um estudo de grande relevância na identificação de emoções em seres humanos.

O reconhecimento de emoções faciais (FER – *Facial Emotion Recognition*) é uma área de pesquisa interdisciplinar que une ciência da computação, psicologia, neurociência e engenharia, com aplicações crescentes em diversos setores como saúde, educação, segurança, marketing e interação humano-computador. O estudo de FER visa habilitar sistemas computacionais a identificar e interpretar emoções humanas a partir de expressões faciais, promovendo uma interação mais natural e empática entre humanos e computadores. No contexto tecnológico, sistemas de FER possibilitam avanços em áreas como diagnósticos médicos, apoio a pessoas com transtornos do espectro autista, monitoramento do bem-estar emocional, análise de sentimentos em ambientes virtuais e aprimoramento de assistentes virtuais e robôs sociais (Li; Deng, 2022).

Além disso, expressões sutis ou mascaradas, que se manifestam, por exemplo, em condições clínicas como a depressão, demandam modelos mais sensíveis e específicos. Embora este trabalho não se concentre diretamente nesse tipo de aplicação, tais cenários reforçam a importância de desenvolver abordagens robustas e generalizáveis para o reconhecimento de expressões faciais. Assim, compreender e mitigar esses obstáculos é essencial para a evolução do FER em contextos sensíveis e de alto impacto, como o diagnóstico assistido de transtornos emocionais.

Desafios técnicos, como variações culturais, iluminação, posturas e oclusões, motivam desenvolvimentos constantes em algoritmos de *machine learning* e *deep learning* para FER, conforme destaca Zhao; Liu; Zhou (2021). Diante desse cenário, torna-se necessário investigar e avaliar o desempenho de modelos de reconhecimento de expressões faciais em imagens estáticas, considerando suas limitações práticas e técnicas. Embora muitos estudos estejam concentrados em aplicações específicas ou utilizem conjuntos de dados idealizados. Ainda há carência de análises comparativas que explorem a capacidade de generalização desses modelos em condições mais adversas, como subdomínios de imagens em que há escassez ou completa ausência de dados rotulados sobre as emoções expressas.

1.1 Objetivo

O objetivo deste trabalho é avaliar o desempenho de modelos pré-treinados para o reconhecimento de expressões faciais (FER) em imagens estáticas. A análise será conduzida em um cenário com escassez de dados rotulados, representando contextos onde a anotação de imagens é limitada ou onerosa. Ao considerar essas duas condições, busca-se investigar a capacidade de adaptação e generalização dos modelos em contextos realistas, contribuindo para a compreensão de seus limites e potencialidades em aplicações práticas.

1.2 Organização do Texto

Este trabalho foi organizado conforme a estrutura a seguir:

- **Capítulo 1 - Introdução:** estabelece a relevância do reconhecimento de expressões faciais, delineia os desafios existentes, como a escassez de dados rotulados, e apresenta o objetivo central do estudo: avaliar modelos pré-treinados para FER em imagens estáticas sob condições de dados limitados.
- **Capítulo 2 - Fundamentação Teórica:** explora os conceitos essenciais do FER, detalhando as redes neurais convolucionais, redes residuais e os vision transformers fechando com uma breve explicação sobre trabalhos relacionados ao tema.
- **Capítulo 2 - Metodologia:** descreve o processo experimental em quatro etapas principais: seleção de amostras, treinamento dos modelos, seleção do melhor modelo por *fold* e, por fim, a avaliação final do modelo.
- **Capítulo 4 - Avaliação Experimental:** apresenta os resultados obtidos no experimento, detalhando o conjunto de dados AffectNet, as configurações dos experimentos e as métricas de avaliação empregadas nos modelos.
- **Capítulo 5 - Conclusões:** nesse último capítulo está sumarizado os achados do estudo, destacando o modelo que foi o mais eficiente nas condições testadas, discutem as implicações práticas da pesquisa, apresenta limitações e propõem direções para investigações futuras no campo do reconhecimento de expressões faciais.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os fundamentos teóricos necessários para a compreensão do reconhecimento automático de expressões faciais, bem como uma revisão dos principais trabalhos relacionados à área. Na Seção 2.1, são introduzidos os conceitos básicos sobre o Reconhecimento de Expressões Faciais (FER), incluindo suas etapas, desafios e importância prática. A Seção 2.2 aborda as Redes Neurais Convolucionais (CNNs - *Convolutional Neural Network*), que constituem a base de diversas arquiteturas modernas utilizadas em tarefas de visão computacional. Na Seção 2.3, é detalhada a arquitetura ResNet, destacando suas características estruturais e relevância para o problema em estudo. A Seção 2.4 apresenta os *Vision Transformers* (ViTs), discutindo seu funcionamento e aplicação em FER. Por fim, a Seção 2.5 revisa trabalhos relacionados, com foco em abordagens baseadas em CNNs e *Transformers*, além de destacar lacunas presentes na literatura atual.

2.1 Reconhecimento de Expressões Faciais

O reconhecimento de expressões faciais é um processo que envolve identificar e interpretar os movimentos e padrões faciais com base em alterações nos músculos do rosto, que estão associados a diferentes estados emocionais. Psicologicamente, estudos realizados por Ekman; Friesen (1978) estabeleceram que as expressões faciais são resultados de emoções universais e inatas, presentes em todas as culturas. Essas expressões, conhecidas como “emoções básicas”, por exemplo felicidade, tristeza, nojo, medo, raiva e surpresa, são marcadores visuais que permitem a comunicação interpessoal e desempenham um papel essencial na regulação social (Ekman; Friesen, 1978).

Na ciência da computação, o reconhecimento de expressões faciais é uma área central no domínio da visão computacional, permitindo que computadores interpretem informações visuais para identificar emoções com base em traços faciais. Essa tarefa é intrinsecamente desafiadora devido à variabilidade nos dados de entrada, que inclui diferenças culturais, idades, gêneros, condições de iluminação e ângulos de captura. Avanços em aprendizado profundo, particularmente no uso de redes neurais convolucionais (CNNs), possibilitaram ganhos substanciais na precisão desses sistemas, tornando viável sua aplicação em diagnóstico médico, sistemas de vigilância e experiência do usuário (He *et al.*, 2016).

O reconhecimento de expressões faciais também é um componente-chave na área da Computação Afetiva, campo que visa desenvolver computadores capazes de processar e responder a emoções humanas (Picard, 2000). Para isso, são utilizados algoritmos alimentados por grandes bases de dados, como o AffectNet, que oferecem imagens anotadas

para diferentes classes emocionais, auxiliando no treinamento e validação dos modelos.

2.2 Redes Neurais Convolucionais

Redes Neurais Artificiais são modelos computacionais inspirados no funcionamento do cérebro humano, compostos por camadas de neurônios artificiais que processam e aprendem padrões a partir de dados (Goodfellow *et al.*, 2016). Dentre as arquiteturas existentes, a mais comum para visão computacional é a CNN, que utiliza camadas de convolução para extrair automaticamente características relevantes das imagens. Essas redes neurais superaram abordagens mais tradicionais que exigiam extração manual de características, permitindo, por exemplo, reconhecer faces, identificar placas de trânsito e analisar exames médicos com elevada precisão (LeCun; Bengio; Hinton, 2015).

As CNNs destacam-se pela sua capacidade de trabalhar diretamente com dados em formato de imagem, explorando propriedades espaciais e reduzindo drasticamente o número de parâmetros em comparação com redes tradicionais completamente conectadas. A principal característica das CNNs é o uso de camadas de convolução, onde pequenos filtros (ou *kernels*) processam a imagem, extraindo padrões locais como bordas, texturas e formas básicas. Essas características extraídas são, posteriormente, combinadas em níveis mais profundos da rede para detectar padrões cada vez mais complexos (Goodfellow *et al.*, 2016).

Além das camadas de convolução, as CNNs frequentemente utilizam camadas de *pooling* (subamostragem), que reduzem a dimensionalidade dos mapas de ativação, tornando o processamento mais eficiente e contribuindo para a invariância espacial. Essa estrutura hierárquica permite que as CNNs aprendam representações robustas e generalizáveis a partir das imagens, sendo especialmente eficaz em tarefas de reconhecimento visual, como a classificação de objetos no ImageNet ou a identificação automática de anomalias em exames médicos (LeCun; Bengio; Hinton, 2015). Na Figura 1 estão ilustradas essas camadas, como também as etapas do treinamento de um modelo de reconhecimento facial de emoções utilizando uma CNN de 4 camadas.

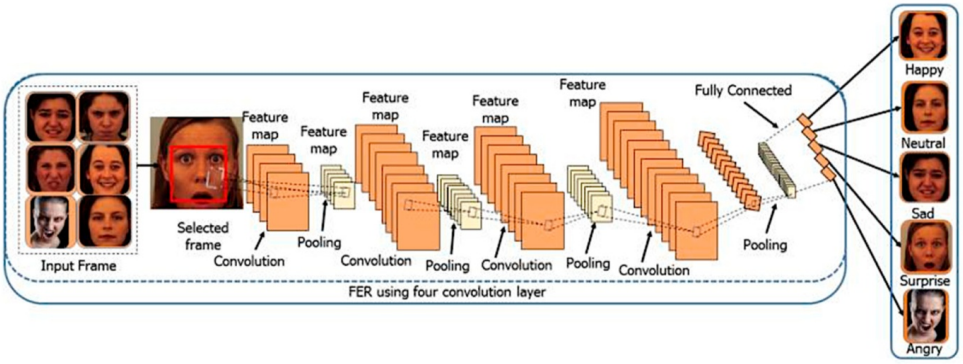

Figura 1 – CNN de 4 camadas - Fonte: (Pallavi; Chavan, 2024)

Com o avanço das pesquisas, arquiteturas de CNN mais profundas e sofisticadas foram desenvolvidas, como a ResNet e a Inception. Essas redes conseguem explorar centenas de camadas graças a inovações como conexões residuais e módulos multi-escala, tornando as CNNs a espinha dorsal dos sistemas de visão computacional modernos. O impacto dessas redes pode ser observado em aplicações que vão desde carros autônomos até sistemas de vigilância inteligente e análise de dados biomédicos (He *et al.*, 2016).

2.3 ResNet - Residual Network

A *Residual Network* (ResNet) é uma arquitetura de rede neural profunda proposta por He *et al.* , (2016), amplamente utilizada em tarefas de visão computacional, como classificação de imagens e reconhecimento facial. Seu diferencial está na introdução dos blocos residuais, que implementam conexões de atalho (*skip connections*) entre camadas não adjacentes. O funcionamento da ResNet baseia-se na hipótese de que é mais fácil otimizar o resíduo (diferença entre entrada e saída) do que o mapeamento original de uma camada para outra. Assim, cada bloco residual aprende uma função de resíduo (F(x) = H(x) - x), onde (H(x)) é o objetivo desejado e (x) é a entrada do bloco. Por meio da soma direta da entrada com a saída processada, facilitando o treinamento de redes com centenas ou até milhares de camadas, tornando-as menos suscetíveis à degradação de performance conforme a arquitetura se aprofunda.

Isso significa que enquanto as CNNs convencionais transmitem a informação de forma sequencial, camada por camada, a ResNet permite a passagem de informações de camadas anteriores diretamente para camadas posteriores, "saltando" etapas intermediárias. Esse mecanismo inovador diminui a perda de informações importantes e evita que camadas muito profundas prejudiquem a acurácia da rede. Em razão dessas características, a ResNet representou um marco no desenvolvimento de redes neurais profundas, influenciando diretamente o avanço e o sucesso de modelos modernos utilizados atualmente. Na Figura 2 está ilustrado o funcionamento da ResNet comparado ao de uma CNN convencional.

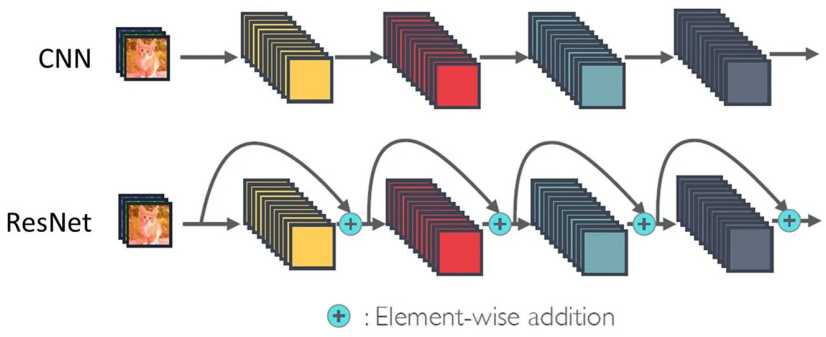

Figura 2 – Visão de conexões residuais na arquitetura ResNet50. Fonte: (Kim, 2018)

2.4 Vision Transformer

Dentre os novos paradigmas em visão computacional, o *Vision Transformer* (ViT) se destaca por utilizar a arquitetura do *Transformers*, originalmente desenvolvida para processamento de linguagem natural. Diferente das CNNs, o ViT divide a imagem em pequenos *patches* (blocos) bidimensionais, que são achatados e tratados como uma sequência, similar ao que ocorre em textos. Cada bloco é transformado em um vetor de características e alimentado no modelo *Transformer,* permitindo a captura de relações globais entre diferentes regiões da imagem (Dosovitskiy *et al.*, 2021)

Essa abordagem inovadora trouxe vantagens consideráveis em tarefas de classificação de imagens, especialmente ao lidar com conjuntos de dados muito grandes. A arquitetura ViT demonstrou desempenho competitivo ou superior ao de redes convolucionais profundas, sendo capaz de aprender relações de longo alcance entre partes distantes da imagem de forma mais eficiente. O uso de mecanismos de atenção do *Transformer* permite que o modelo foque seletivamente em regiões relevantes, o que potencializa o reconhecimento em contextos complexos de visão computacional (Touvron *et al.*, 2021).

Além disso, o sucesso do *Vision Transformer* fomentou uma rápida expansão de variantes e adaptações para diferentes aplicações visuais, incluindo detecção de objetos e segmentação semântica. A versatilidade, escalabilidade e facilidade de pré-treinamento com grandes volumes de dados posicionam o ViT como uma referência emergente em visão computacional, abrindo caminho para novos avanços tanto em pesquisa quanto em implementações práticas (Khan *et al.*, 2022). Na Figura 3 é apresentado de forma esquemática o desenho de uma arquitetura ViT.

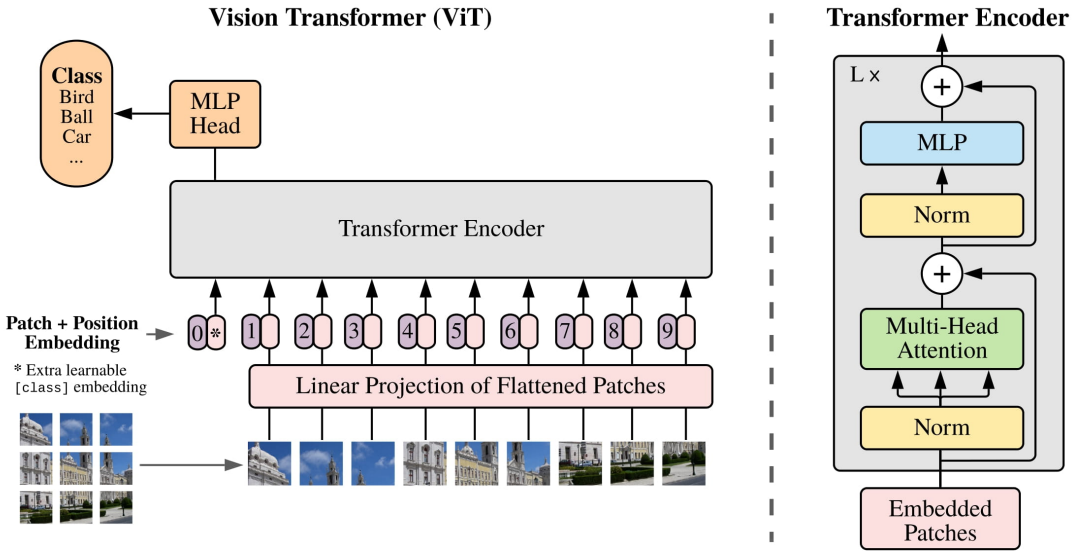

Figura 3 – Arquitetura *Vision Transformer* (ViT) - Fonte: (Koyyada; Rawat; Singh, 2022)

2.5 Trabalhos Relacionados

Nesta seção, são apresentados e discutidos estudos relevantes que abordam o FER com enfoque em modelos baseados em CNNs, *Transformers* e abordagens voltadas para cenários com escassez de dados rotulados. Os trabalhos selecionados contribuem para a compreensão dos avanços recentes na área e oferecem subsídios para a definição da abordagem adotada neste estudo.

Wang; Wang; Cui (2023) propoem um modelo baseado no ResNet com melhorias denominado Y-Net. Este modelo aprimora a arquitetura original ao incorporar conceitos do DenseNet, otimizando a propagação de características e a robustez. A pesquisa destaca a eficácia de técnicas como *transfer learning* e validação cruzada *K-fold* para acelerar o treinamento e garantir a estabilidade. Os resultados experimentais, testados nos conjuntos de dados FER2013, FERPlus e RAF-DB, demonstraram um desempenho eficiente, atingindo precisões de 79,9%, 87,5% e 85,2% respectivamente. Um aspecto crucial para o estudo de FER é a influência direta da distribuição de amostras na acurácia do reconhecimento, com categorias como felicidade apresentando maior acerto devido ao volume de dados, e a importância de balancear os conjuntos de dados para otimizar o reconhecimento de expressões menos representadas, como nojo e medo. O estudo aponta ainda para futuras aplicações em cenários do mundo real e dispositivos embarcados.

Já Kim; Kim (2023) abordadam a aprendizagem com poucas amostras - *Few-Shot Learning* (FSL) para um contexto de FER. A pesquisa aborda os desafios inerentes ao FER, como a escassez de dados anotados e a dificuldade de generalização em ambientes não-controlados, posicionando a FSL como uma solução promissora. O estudo classifica e discute diversos conjuntos de dados de FER: CK+, CFEE, FER2013, FER+, EmotioNet, AffectNet, RAF-DB e AFEW, detalhando problemas relacionados a dados e métodos, como classes desbalanceadas e *overfitting*. A FSL é apresentada como uma técnica capaz de otimizar a eficiência de tempo tanto no treinamento quanto na previsão. O trabalho analisa modelos como CRN, FedAffect, EGS-Net e MERAU, comparando suas arquiteturas, número de parâmetros e complexidade temporal. Dentre eles, o CRN se destaca como o modelo mais leve (0.2 M parâmetros) e mais rápido (0.9 ms na GPU) para inferência. Com base nesse estudo, conclui-se que a FSL permite a inferência em categorias não-treinadas de forma eficiente, apresentando diversas melhorias oferecendo caminhos para otimizações e extração de características mais robustas.

Por fim, Shen (2024) investigam a eficácia de modelos híbridos para FER. Eles destacam a utilização das CNNs na extração de características locais e dos ViT na captura de relações globais. Três arquiteturas híbridas são propostas: CNN-before-ViT, ViT-before-CNN e Parallel CNN-ViT. Os resultados, obtidos no conjunto de dados FER2013, demonstram que os modelos híbridos, em geral, superam o ViT isolado. O Parallel CNN-ViT, ao combinar as capacidades de CNN e ViT em paralelo (com ResNet-50 como

base), alcançou a maior precisão geral (75,14% na categoria *happy*). Notavelmente, este modelo mostrou eficácia em categorias com menor número de amostras, como *disgust*. Em contraste, a arquitetura ViT-before-CNN consistentemente reduziu a acurácia, sugerindo que a sequência de extração de características é crucial para o desempenho.

3 METODOLOGIA

O objetivo deste trabalho é avaliar o desempenho de modelos pré-treinados para FER em cenários com disponibilidade limitada de dados rotulados para treinamento. A metodologia adotada, ilustrada de forma esquemática na Figura 4, é composta por quatro etapas principais. Sendo que a primeira etapa realiza a seleção de um subconjunto de imagens representativas de cada classe. Em seguida na segunda etapa, os modelos são treinados ao longo de múltiplas épocas utilizando essas amostras, sendo selecionado na terceira etapa, para cada *fold*, a época que apresenta o melhor desempenho no conjunto de validação. Por fim, o modelo correspondente à melhor época de cada *fold* é avaliado sobre o conjunto de teste.

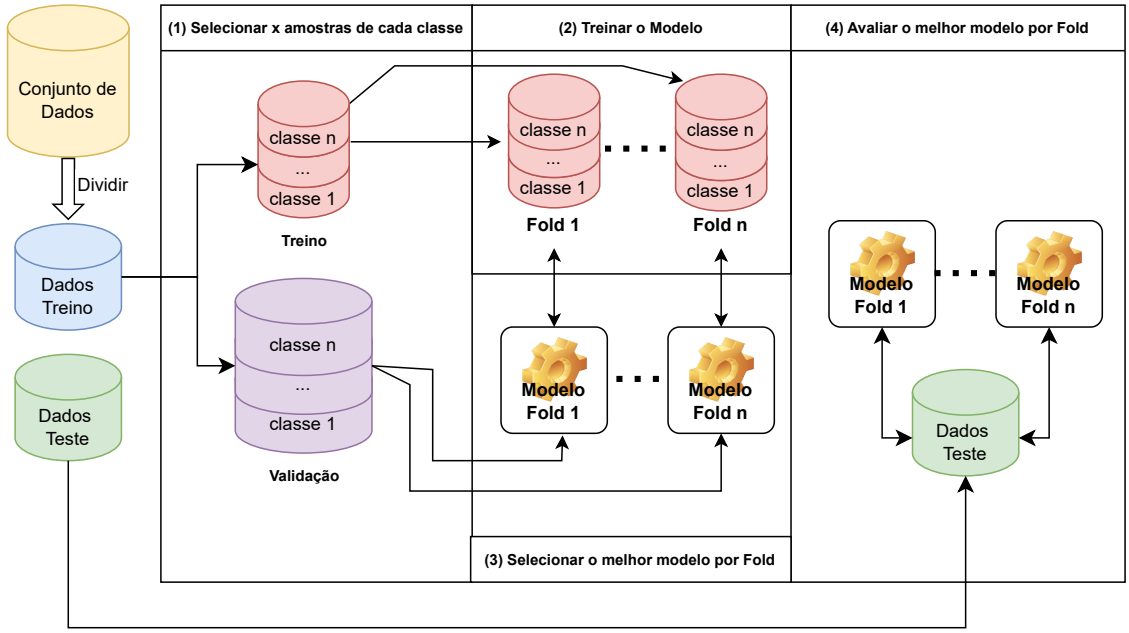

Figura 4 – Fluxograma das Etapas de Processamento

Antes do início dessas etapas, é realizada uma fase preliminar de preparação dos dados, na qual o conjunto é dividido em duas partes: dados de teste, mantidos isolados até a avaliação final; dados de treinamento, utilizados para ajuste dos modelos.

Etapa 1 - Selecionar X amostras de cada classe: nesta etapa, realizada exclusivamente sobre a partição de treinamento previamente definida, seleciona-se um número reduzido e balanceado de amostras por classe, de modo a simular cenários com poucos rótulos disponíveis. Após a seleção, o remanescente da partição de treinamento é reservado para validação, garantindo que nenhuma

informação do conjunto de teste seja utilizada e permitindo avaliar a capacidade de generalização dos modelos a partir de poucos rótulos. Nessa etapa, para cada execução do experimento, o número "X" de amostras por classe foi sendo modificado, primeira execução com 20, depois 40 e por último com 100 amostras por classe.

Etapa 2 - Treinar o modelo: os dados de treinamento são divididos em *folds*. Cada *fold* é utilizado para treinar o modelo ao longo de um número pré-definido de épocas, sendo cada época avaliada no conjunto de validação. Em cada época, são coletadas as métricas função de perda (*loss*), acurácia e F1-*Macro* (média harmônica entre precisão e recall). A definição e a interpretação dessas métricas serão apresentadas no capítulo 4.

Etapa 3 - Selecionar o melhor modelo por *fold*: após o treinamento, para cada *fold* é selecionado o modelo que apresenta o maior valor de F1-*Macro* no conjunto de validação. O modelo escolhido é armazenado juntamente com a indicação da época em que obteve esse desempenho, possibilitando sua posterior utilização na etapa de avaliação.

Etapa 4 - Avaliar o melhor modelo por *fold*: com os melhores modelos selecionados, realiza-se a avaliação final utilizando o conjunto de teste, que permaneceu isolado durante todas as etapas anteriores. O objetivo é comparar o desempenho obtido por cada modelo e identificar aquele que apresenta o maior valor de F1-*Macro* no cenário de teste, garantindo assim uma análise consistente e imparcial do desempenho em dados não vistos.

4 AVALIAÇÃO EXPERIMENTAL

Neste capítulo serão apresentados os resultados que tem por objetivo avaliar os modelos treinados com poucos dados rotulados. Na Seção 4.1 será explicado o conjunto de dados utilizado chamado AffectNet. Na Seção 4.2 será detalhado o ambiente utilizado e as configurações que foram utilizados na execução dos programas e parâmetros de configurações. Na Seção 4.3 serão explicadas as métricas utilizadas para avaliação e seleção dos modelos. Na Seção 4.4 apresentamos os dados obtidos e a análise sobre os resultados alcançados.

4.1 Conjunto de Dados - AffectNet

O AffectNet[1] (Mollahosseini; Hasani; Mahoor, 2017) é um dos conjuntos de dados mais completos e amplos voltados para o reconhecimento de emoções faciais. Foi criado com o objetivo de fornecer um recurso rico para o treinamento e avaliação de algoritmos de reconhecimento de expressões faciais. O conjunto de dados contém mais de 1 milhão de imagens de rostos humanos, coletadas automaticamente por meio de motores de busca usando 1250 palavras-chave relacionadas a emoções em seis idiomas diferentes. Cada imagem no conjunto de dados é anotada manualmente com rótulos que indicam uma dentre as seguintes emoções: Neutro, Feliz, Triste, Surpreso, Medo, Nojo, Raiva e Desdém. Além disso, as imagens também são marcadas com pontos de referência faciais que ajudam na localização precisa de características faciais. O tamanho e diversidade do AffectNet permitem que ele seja uma ferramenta crucial no desenvolvimento de modelos de aprendizado profundo que requerem grandes quantidades de dados para generalizar bem em tarefas de detecção e classificação de emoções. Ele é aplicado em áreas como interação humano-computador, monitoramento de comportamento humano, entre outras.

Entretanto, para este experimento, em razão das restrições de recursos computacionais e do tempo de processamento, optou-se por utilizar uma versão reduzida[2] do conjunto de dados. A distribuição das amostras por classe nessa versão está ilustrado na Tabela 1.

[1] O conjunto em sua versão original pode ser encontrado em: https://mohammadmahoor.com/pages/databases/affectnet/. Acessado em 03-mar-2025.

[2] O conjunto de dados utilizado foi obtido através do site Kaggle em: https://www.kaggle.com/datasets/mstjebashazida/affectnet. Acessado em 04-mai-2025

	Amostras	
Emoções	**Treinamento**	**Teste**
anger	1500	1718
contempt	1559	1312
disgust	1229	1248
fear	1512	1664
happy	2360	2704
neutral	2758	2368
sad	3091	1594
surprise	2119	1920
TOTAL	16128	14528

Tabela 1 – Distribuição das amostras por classes

Foi selecionado o número de amostras por classe para o treinamento em cada *fold*. Esse valor foi testado em dois cenários: 20 amostras por classe e 40 amostras por classe. Tal configuração é central para o estudo, pois simula ambientes com poucos dados rotulados, permitindo avaliar a capacidade de generalização dos modelos nessas condições.

Para melhorar a capacidade de generalização dos modelos e reduzir o risco de *overfitting*, foram aplicadas técnicas de *data augmentation* com hiperparâmetros específicos, conforme descrito a seguir. Inicialmente, todas as imagens foram redimensionadas para 224×224 pixels, tamanho padrão esperado por arquiteturas pré-treinadas como ResNet e *Vision Transformer*. Em seguida, aplicou-se espelhamento horizontal aleatório, com probabilidade de 50%, de modo a introduzir variações na orientação das faces. Essa técnica força o modelo a aprender características faciais independentes da pose, evitando que ele associe direções específicas do olhar a determinadas classes.

Adicionalmente, foi utilizada rotação aleatória de até 10 graus, à esquerda ou à direita, introduzindo variação na orientação angular das imagens. Também aplicou-se a variação de brilho e contraste, permitindo alterações de até 20% nesses atributos, o que favorece a robustez do modelo frente a diferentes condições de iluminação. Por fim, todas as imagens foram normalizadas, utilizando as médias ([0.485, 0.456, 0.406]) e desvios padrão ([0.229, 0.224, 0.225]) por canal do conjunto ImageNet. Essa normalização garante que os dados de entrada estejam na mesma escala esperada pelos modelos pré-treinados, preservando a compatibilidade e o aproveitamento do conhecimento previamente adquirido.

4.2 Configurações Experimentais

Para a condução dos experimentos, foram utilizados modelos pré-treinados ResNet e *Vision Transformer*, adaptados para a tarefa de reconhecimento de expressões faciais. Os principais hiperparâmetros empregados foram definidos de forma a equilibrar viabilidade computacional e qualidade dos resultados. O tamanho do lote (*batch size*) foi fixado em 64 amostras, valor que permite um treinamento estável por iteração, ao mesmo tempo

em que aproveita a capacidade de processamento disponível. O número de épocas de treinamento foi estabelecido em cinco, em função de restrições de recursos computacionais, buscando evitar sobreajuste e reduzir o tempo de processamento. Para avaliação dos modelos, adotou-se validação cruzada com cinco partições (*5-fold cross-validation*), o que garante maior robustez na estimativa de desempenho e aproveitamento equilibrado das amostras de todas as classes. A taxa de aprendizado (*learning rate*) foi fixada em 0,001, utilizando-se o otimizador Adam. Esse valor busca um equilíbrio entre velocidade de convergência e estabilidade do treinamento, evitando oscilações excessivas ou lentidão no ajuste dos parâmetros.

4.3 Métricas de Avaliação e Seleção dos Modelos

As métricas que foram coletadas durante as execuções foram: função de perda, acuracidade e F1-*Macro*. Para determinar o melhor modelo de cada *fold*, foi utilizado a métrica de F1-*Macro* que representa a média do F1 de todas as classes, usando os valores reais e previstos. Na Tabela 2 está ilustrado os resultado obtidos em cada execução, para cada modelo, onde cada linha da tabela representa o resultado obtido a partir do número de amostras para cada classe. As colunas Acurácia Média e F1-*Macro* Média, representam respectivamente a média aritimética por *fold* das métricas de acurácia e F1-*Macro* e entre parenteses está o desvio padrão.

Tabela 2 – Métricas dos melhores modelos em cada execução para cada número de amostras. Entre parenteses está o desvio padrão.

Amostras por Classe	**Modelo**	**Acurácia Média**	**F1-*Macro* Média**
10	ResNet	0.1349 (0.0448)	0.0834 (0.0533)
	ViT	0.1131 (0.0429)	0.0289 (0.0123)
20	ResNet	0.1883 (0.0511)	0.1174 (0.0359)
	ViT	0.1236 (0.0368)	0.0294 (0.0112)
40	ResNet	0.1530 (0.0288)	0.1058 (0.0514)
	ViT	0.1234 (0.0272)	0.0310 (0.0051)
100	ResNet	0.2809 (0.0450)	0.1937 (0.0405)
	ViT	0.1108 (0.0165)	0.0369 (0.0114)

É possível observar que o modelo ResNet consistentemente supera o ViT nos resultados. Porém essa análise foi baseada apenas em média e desvio padrão de F1-*Macro*, sendo necessário uma análise mais detalhada das distribuições das classes e a natura de cada classe. Uma vez que elas representam emoções identificadas através de expressões faciais o que pode influenciar diretamente em como os modelos estão acertando ou errando suas predições. Enquanto a Tabela 2 fornece uma visão agregada do desempenho por meio de métricas como F1-*Macro* e Acurácia, as matrizes de confusão oferecem um panorama detalhado dos acertos e erros por classe. Permitindo entender quais classes cada modelo

está tendo mais dificuldade em classificar. Para isso foi selecionado os resultados contendo 20 amostras pois apesar de não ser o melhor resultado dos modelos, rotular 20 amostras por classe já seria um esforço considerável. As demais matrizes de confusão foram incluidas no Apêndice A.

Em relação ao desempenho do Modelo ResNet, ao analisar a matriz de confusão ilustrada na Figura 5, referente ao *Fold* 2 que obteve o melhor resultado. É possível observar um comportamento de classificação mais equilibrado, embora com desafios notáveis.

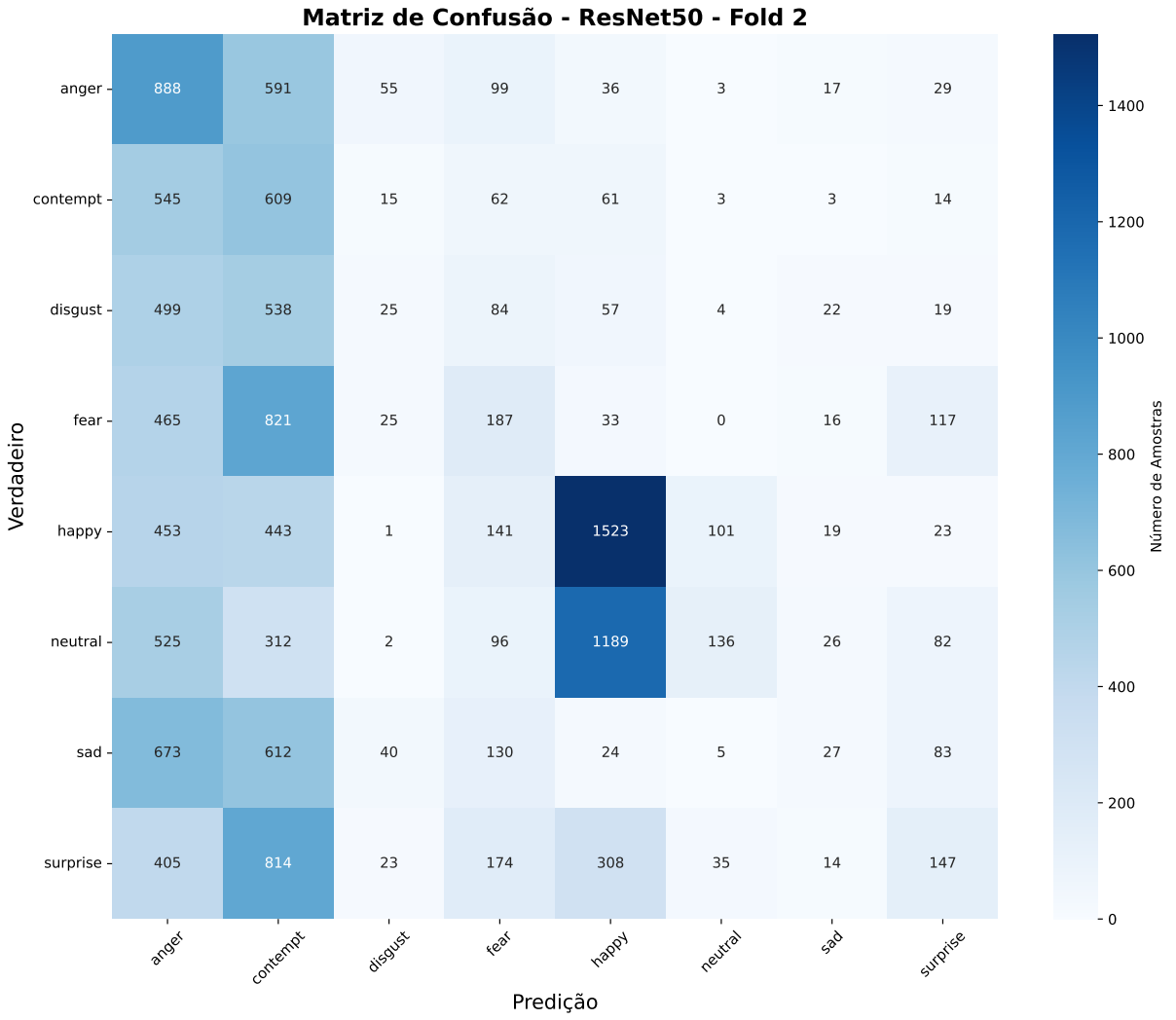

Figura 5 – Matriz de confusão - ResNet *Fold* 2 com 20 amostras por classe

- **Pontos Fortes Relativos:** a emoção *"happy"* (felicidade) foi a mais consistentemente bem classificada, com 1523 verdadeiros positivos. Este resultado sugere que as características faciais associadas à felicidade são suficientemente distintas para serem aprendidas pelo modelo mesmo com um conjunto limitado de dados. O recall para *"happy"* foi de 56.3%, indicando que mais da metade das ocorrências de felicidade foram corretamente identificadas.

- **Desempenho Moderado e Confusões Comuns:** as emoções *"anger"* (raiva) e *"contempt"* (desprezo) demonstraram recalls moderados (51.7% e 46.4%, respectivamente), com 888 e 609 verdadeiros positivos. No entanto, uma confusão significativa foi observada entre elas. Por exemplo, 591 amostras de *"anger"* foram classificadas como *"contempt"* e 545 amostras de *"contempt"* foram classificadas como *"anger"*. Esta mútua confusão é um desafio comum em sistemas de reconhecimento de expressões faciais, dadas as sobreposições sutis em suas manifestações faciais.

- **Pontos Fracos Críticos:** o modelo apresentou extrema dificuldade em classificar corretamente as emoções de *"surprise"* (nojo), *"sad"* (tristeza), *"neutral"* (neutro), *"fear"* (medo) e *"surprise"* (surpresa). Para *"surprise"* e *"sad"*, os recalls foram alarmantemente baixos (2.0% e 1.7%, respectivamente), com apenas 25 e 27 verdadeiros positivos. A maioria das amostras dessas classes foi erroneamente atribuída a outras categorias, predominantemente *"anger"* e *"contempt"*. Por exemplo, 673 amostras de *"sad"* foram classificadas como *"anger"* e 612 como *"contempt"*. Similarmente, *"neutral"* (5.7% recall) foi amplamente confundida com *"happy"* (1189 casos) e *"anger"* (525 casos), enquanto *"fear"* (11.2% recall) e *"surprise"* (7.7% recall) frequentemente foram classificadas como *"contempt"* (821 e 814 casos, respectivamente).

- **Padrões de Confusão Gerais:** a ResNet demonstrou um viés em classificar erroneamente muitas emoções como *"anger"*, *"contempt"* ou *"happy"*, sugerindo que, em situações de incerteza, o modelo tendia a se basear nas características mais aprendidas ou nas classes com maior representatividade percebida.

Em síntese, a ResNet, embora não perfeita, exibiu um desempenho que indica uma tentativa de diferenciar todas as classes, mesmo que com sucesso variável. Suas falhas são mais distribuídas e refletem desafios esperados na complexidade do reconhecimento de emoções. Por outro lado o desempenho do Modelo ViT ilustrado na Figura 6, referente ao *Fold* 1 que obteve melhores resultados. revela um padrão de classificação drasticamente diferente e mais problemático:

- **Desempenho Concentrado na Classe *"Happy"*:** o ViT demonstrou uma capacidade excepcional para identificar a emoção *"happy"*, registrando 2591 verdadeiros positivos e um recall impressionante de 95.8%. Este é, de longe, o ponto mais forte e quase exclusivo do modelo.

- **Incapacidade de Classificação para Outras Emoções:** em contraste gritante, o ViT falhou quase completamente em classificar as demais emoções. Para *"anger"*, *"contempt"*, *"surprise"*, *"fear"* e *"surprise"*, o modelo registrou 0 (zero) verdadeiros positivos e um recall de 0%. Isso significa que o ViT nunca classificou corretamente uma amostra dessas emoções quando elas eram a classe verdadeira.

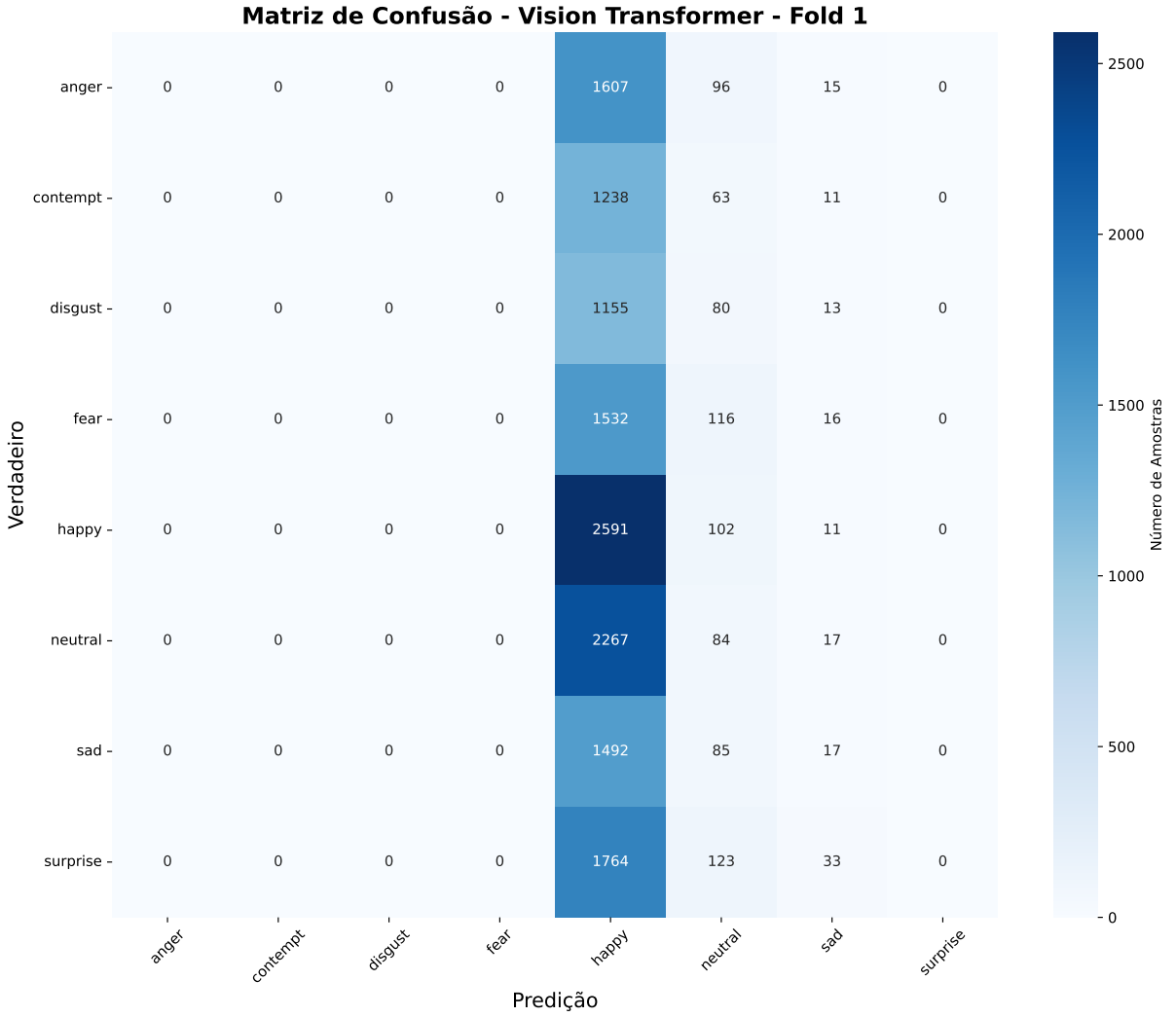

Figura 6 – Matriz de confusão - ViT *Fold* 1 com 20 amostras por classe

- **Desempenho Negligenciável para *"neutral"* e *"sad"*:** as emoções *"neutral"* e *"sad"* tiveram um desempenho igualmente pobre, com recalls de 3.5% (84 acertos) e 1.1% (17 acertos), respectivamente.

- **Viés Extremo na Predição:** o padrão mais notável do ViT foi seu viés extremo em classificar a grande maioria das amostras como *"happy"*. Por exemplo, 1607 amostras de *"anger"*, 1238 de *"contempt"*, 1155 de *"surprise"*, 1532 de *"fear"*, 2267 de *"neutral"*, 1492 de *"sad"* e 1764 de *"surprise"* foram erroneamente classificadas como *"happy"*. Pequenas porções dessas amostras também foram classificadas como *"neutral"* ou *"sad"*, mas a predominância da classe *"happy"* nas predições foi avassaladora.

Este comportamento sugere que o ViT, neste contexto de poucos dados rotulados, não conseguiu aprender as características discriminatórias para a maioria das emoções, desenvolvendo, em vez disso, um forte viés para prever a classe *"happy"*. Isso pode ser

atribuído a questões como a predominância da classe *"happy"* no conjunto de dados, que pode ter levado o modelo a otimizar sua função de custo pela classificação massiva nesta categoria.

A comparação entre o desempenho da ResNet e do ViT, com base em suas respectivas matrizes de confusão e considerando as métricas da Tabela 2, revela diferenças fundamentais:

- **Robusteza e Equilíbrio:** a ResNet, apesar de suas limitações para certas classes, demonstrou ser um classificador mais robusto e equilibrado. Ela tenta diferenciar todas as 8 emoções, apresentando verdadeiros positivos (mesmo que poucos) para todas elas. Embora apresente dificuldades para classificar para algumas classes, ainda sim apresenta indicativos de um aprendizado parcial das características de cada emoção. O F1-*Macro* médio da ResNet (variando entre 0.0834 a 0.1937) consistentemente superou o do ViT.
- **Viés e Ineficácia do ViT:** o *Vision Transformer*, neste experimento, exibiu um viés de classificação extremo, tornando-o praticamente ineficaz para o reconhecimento multi-classe de emoções. Sua incapacidade de classificar corretamente a maioria das emoções, aliada à sua forte tendência de prever *"happy"* indiscriminadamente, indica que o modelo não conseguiu generalizar adequadamente as características das diferentes expressões faciais sob condições de dados limitados. Embora seja possivel ver na Tabela 2 que o F1-*Macro* do ViT aumenta conforme a quantidade de amostras aumenta, seus valores baixos confirmam uma falha na diferenciação das classes.

Em conclusão, com base na análise aprofundada das matrizes de confusão, observou-se que a ResNet apresentou um desempenho superior e mais funcional para a tarefa de reconhecimento de expressões faciais em cenários com poucos dados rotulados. Embora ambas as arquiteturas enfrentem desafios inerentes à escassez de dados, a ResNet demonstrou maior capacidade de aprender as distinções entre as diferentes emoções. O comportamento enviesado do ViT neste contexto ressalta a necessidade de estratégias de treinamento mais avançadas ou adaptações arquitetônicas específicas para que modelos como o *Vision Transformer* possam operar eficazmente em cenários de dados limitados para tarefas de classificação multi-classe como o FER.

5 CONCLUSÕES

Este trabalho investigou o desempenho de modelos na tarefa de FER a partir de imagens estáticas em cenários de escassez de dados rotulados. Foram analisadas duas arquiteturas de redes neurais, ResNet e *Vision Transformer*, treinadas com poucas amostras do conjunto de dados AffectNet. Os resultados indicaram que a ResNet foi o modelo mais eficiente, demonstrando robustez e um desempenho superior ao ViT, mesmo com um número restrito de dados e poucas épocas de treinamento.

A principal contribuição deste estudo reside na demonstração da viabilidade de otimização de modelos de IA mesmo diante da escassez de dados anotados, um desafio comum em reconhecimento facial devido a questões de privacidade (como a LGPD) e ao alto custo/sensibilidade da coleta. Essa abordagem tem potencial de aplicação em diversas áreas, desde o monitoramento do bem-estar emocional e satisfação do cliente até diagnósticos assistidos. Incluindo a contribuição para a triagem de transtornos emocionais, monitoramento de níveis de satisfação e detecção de situações atípicas, como fadiga em operadores de máquinas e motoristas, ou indicadores de *bullying* em ambientes escolares.

Apesar das contribuições, esse trabalho apresenta algumas limitações que abrem caminhos para investigações futuras. A amostragem aleatória inicial poderia ser substituída por uma curadoria manual dos dados por classe para maior diversidade e qualidade. Além disso, as restrições computacionais limitaram os experimentos a quatro cenários de dados e cinco épocas de treinamento. Seria importante fazer testes aumentando o aumento do número de épocas (ex: para cem ou mais) pois isso pode elevar o desempenho dos modelos. A continuidade da pesquisa sobre o tema mostra-se, portanto, essencial para aprofundar o entendimento, robustecer a evidência empírica e fomentar inovações na área. Permitindo superar as limitações identificadas e explorar novas abordagens para o reconhecimento de expressões faciais em cenários desafiadores.

REFERÊNCIAS

DOSOVITSKIY, A. *et al.* **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021. Disponível em: https://arxiv.org/abs/2010.11929.

EKMAN, P.; FRIESEN, W. V. Facial action coding system. **Environmental Psychology & Nonverbal Behavior**, 1978.

GOODFELLOW, I. *et al.* **Deep learning**. [*S.l.: s.n.*]: MIT press Cambridge, 2016. v. 1.

HE, K. *et al.* Deep residual learning for image recognition. *In*: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [*S.l.: s.n.*], 2016. p. 770–778.

KHAN, S. *et al.* Transformers in vision: A survey. **ACM Computing Surveys**, Association for Computing Machinery (ACM), v. 54, n. 10s, p. 1–41, jan. 2022. ISSN 1557-7341. Disponível em: http://dx.doi.org/10.1145/3505244.

KIM, C.-L.; KIM, B.-G. Few-shot learning for facial expression recognition: a comprehensive survey. **Journal of Real-Time Image Processing**, v. 20, n. 3, p. 52, May 2023. ISSN 1861-8219. Disponível em: https://doi.org/10.1007/s11554-023-01310-x.

KIM, H. **DenseNet &x2014; Organize everything I know documentation — oi.readthedocs.io**. 2018. https://oi.readthedocs.io/en/latest/computer_vision/cnn/densenet.html. [Accessed 23-08-2025].

KOYYADA, S.; RAWAT, A.; SINGH, T. P. Lung infection detection using contemporary techniques of artificial intellligence. **Computology: Journal of Applied Computer Science and Intelligent Technologies**, v. 2, p. 14–22, 12 2022.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, May 2015. ISSN 1476-4687. Disponível em: https://doi.org/10.1038/nature14539.

LI, S.; DENG, W. Deep facial expression recognition: A survey. **IEEE Transactions on Affective Computing**, v. 13, n. 3, p. 1195–1215, 2022.

MOLLAHOSSEINI, A.; HASANI, B.; MAHOOR, M. Affectnet: A database for facial expression, valence, and arousal computing in the wild. **IEEE Transactions on Affective Computing**, PP, 08 2017. Disponível em: https://ieeexplore.ieee.org/document/8013713.

PALLAVI, M.; CHAVAN, P. Emotion detection using haar-cascade classifier and cnn. *In*: . [*S.l.: s.n.*], 2024. p. 1–8.

PICARD, R. W. **Affective computing**. [*S.l.: s.n.*]: MIT press, 2000.

SHEN, Z. A comparative study of hybrid cnn and vision transformer models for facial emotion recognition. *In*: **2024 11th International Conference on Dependable Systems and Their Applications (DSA)**. [*S.l.: s.n.*], 2024. p. 401–408.

TOUVRON, H. *et al.* **Training data-efficient image transformers distillation through attention**. 2021. Disponível em: https://arxiv.org/abs/2012.12877.

WANG, X.; WANG, G.; CUI, Y. Facial expression recognition based on improved resnet. **The Journal of China Universities of Posts and Telecommunications**, v. 30, n. 1, p. 28–38, February 2023. Disponível em: https://jcupt.bupt.edu.cn/EN/10.19682/j.cnki.1005-8885.2023.2003.

ZHAO, Z.; LIU, Q.; ZHOU, F. Robust lightweight facial expression recognition network with label distribution training. *In*: **Proceedings of the AAAI conference on artificial intelligence**. [*S.l.: s.n.*], 2021. v. 35, n. 4, p. 3510–3519. Disponível em: https://ojs.aaai.org/index.php/AAAI/article/view/16465.

APÊNDICES

APÊNDICE A – MATRIZES DE CONFUSÃO DE TODAS AS AMOSTRAS

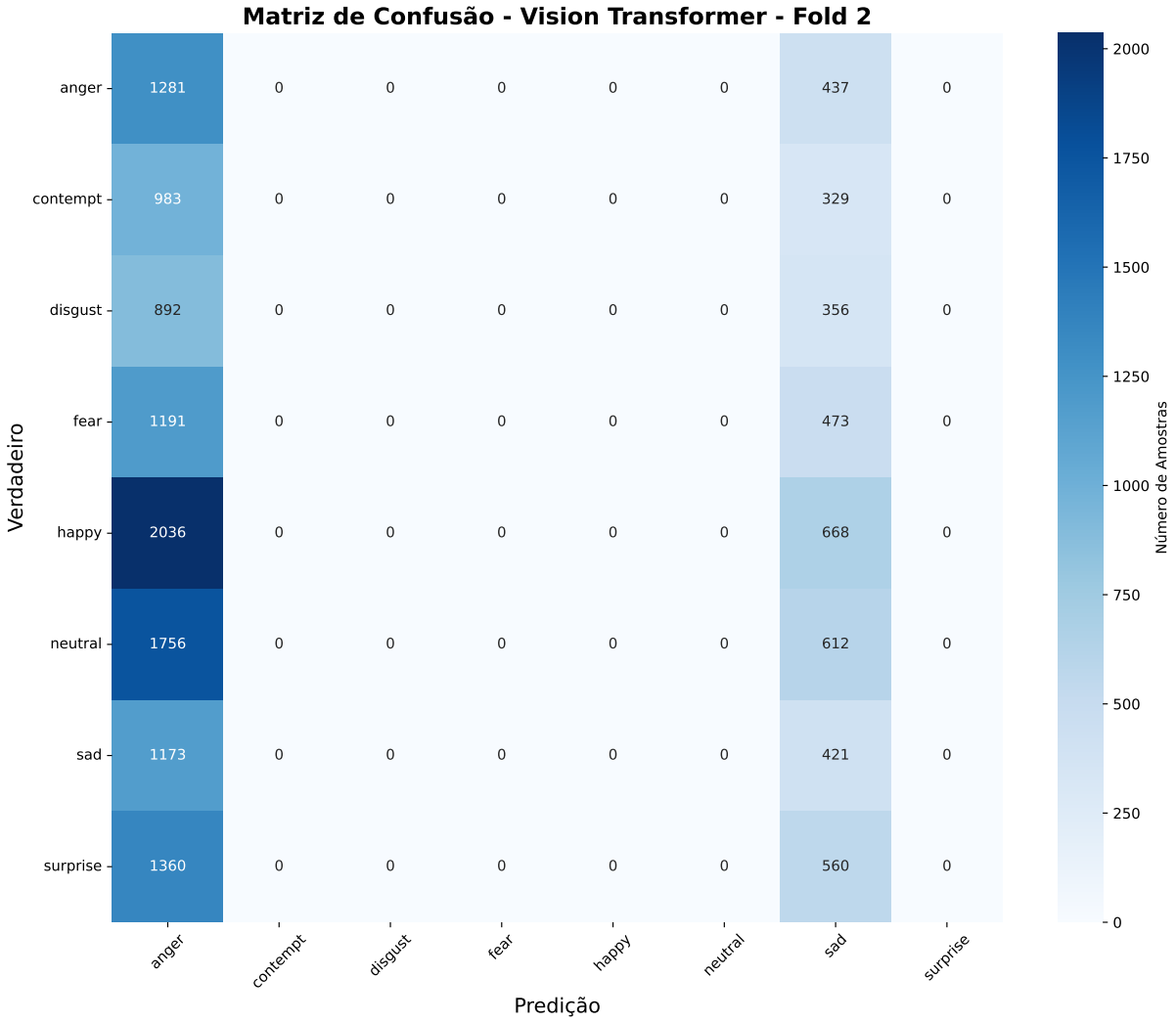

Figura 7 – Matriz de confusão - ViT *Fold* 2 com 10 amostras por classe

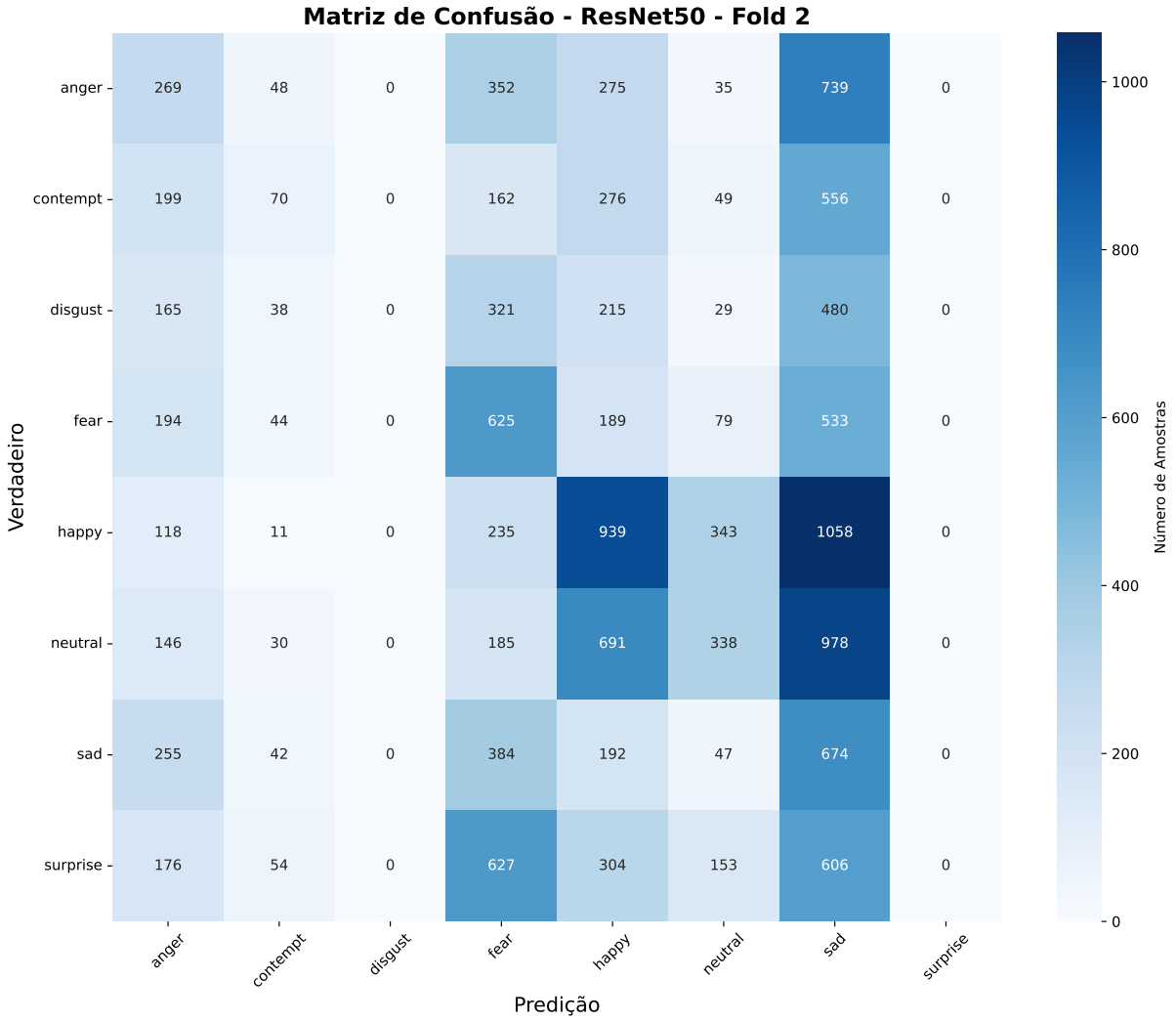

Figura 8 – Matriz de confusão - ResNet *Fold* 2 com 10 amostras por classe

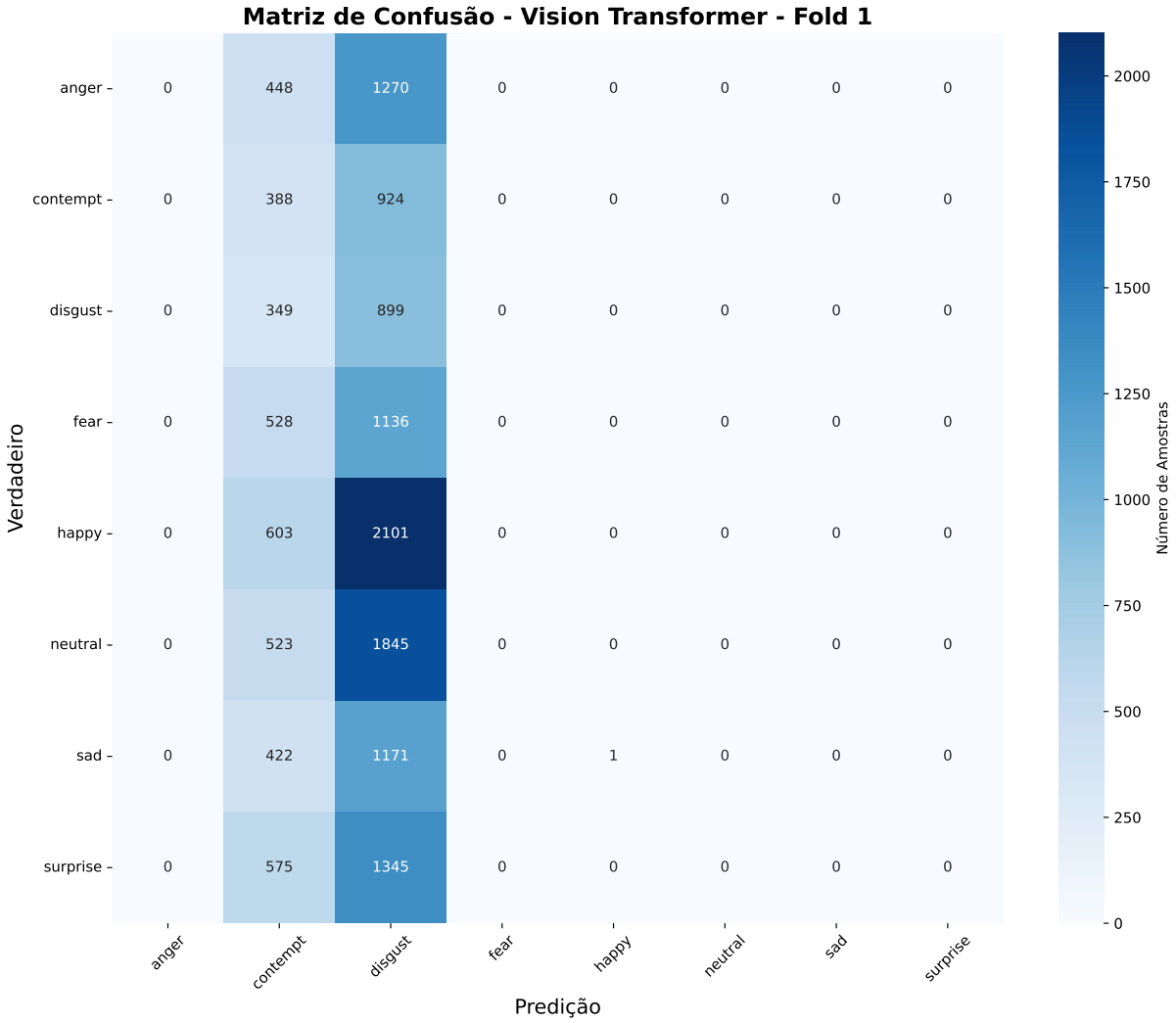

Figura 9 – Matriz de confusão - ViT *Fold* 1 com 40 amostras por classe

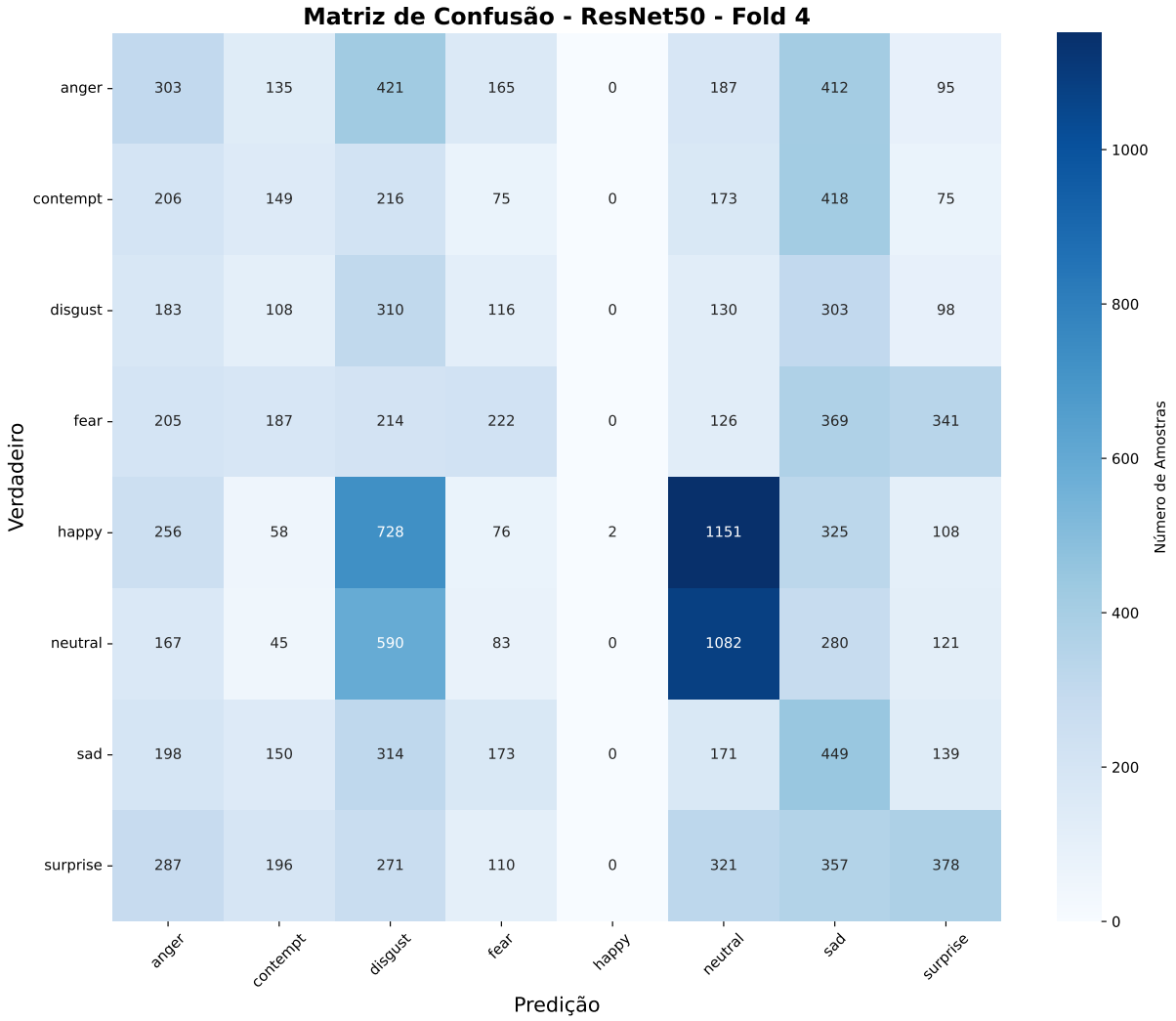

Figura 10 – Matriz de confusão - ResNet *Fold* 2 com 40 amostras por classe

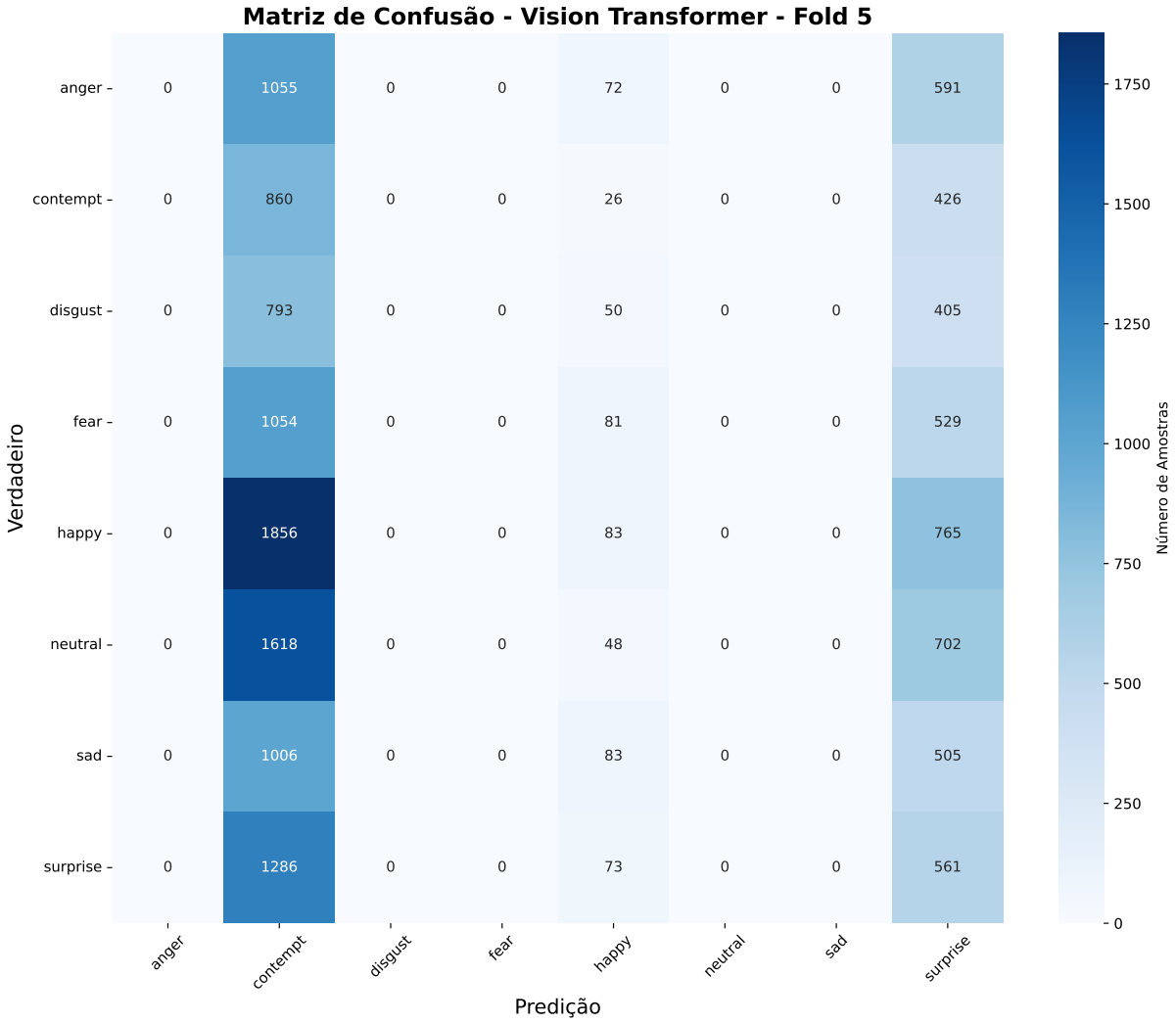

Figura 11 – Matriz de confusão - ViT *Fold* 5 com 100 amostras por classe

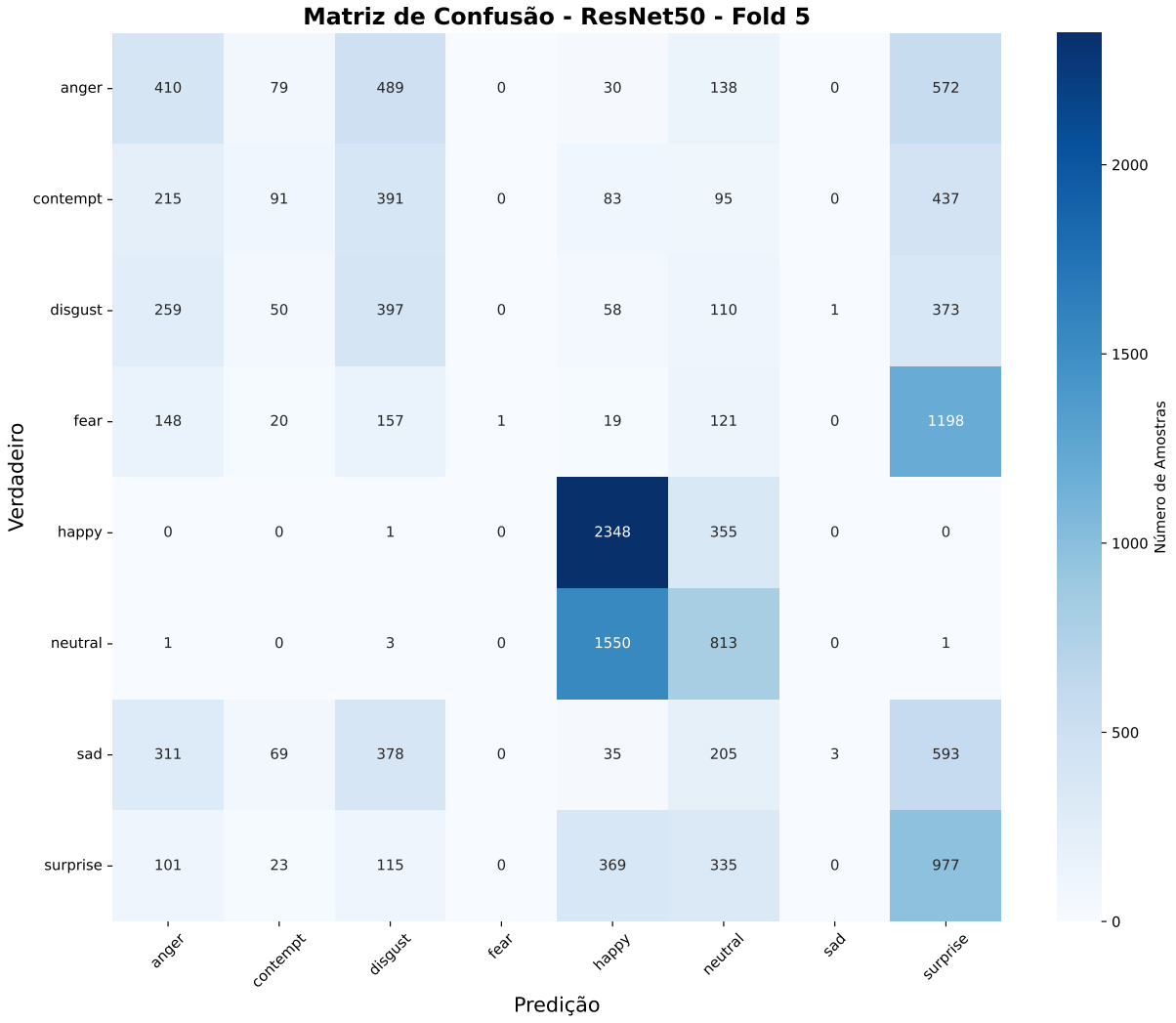

Figura 12 – Matriz de confusão - ResNet *Fold* 5 com 100 amostras por classe