Detecção e diagnóstico automático de patologias na retina utilizando arquitetura baseada em Transformers

Thalisson J. C. Silva, Saulo E. R. Fernandes, João D. S. de Almeida, Darlan B. P. Quintanilha, Geraldo Braz Junior

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

{thalissoncutrim, saulorodrigues, jdallyson}@nca.ufma.br,

{dquintanilha, geraldo}@nca.ufma.br

Abstract. *Globally, more than 2.2 billion people are visually impaired, with around one billion of these cases being preventable. Early detection of eye diseases is crucial to prevent the progression of irreversible conditions such as blindness. Therefore, this study presents a new method for detecting multiple eye pathologies in fundus images, using a neural network architecture based on* transformers*, called Query2Label. The experiments were carried out on the RFMiD public dataset, revealing promising results, with an average accuracy of 99.8% in the "D. Risk" category. Compared to the state of the art, the method effectively detected the "ODP" class. It surpassed accuracy in other specific categories, such as "CSR", "LS", highlighting its feasibility and effectiveness in classifying ophthalmic pathologies.*

Resumo. *Globalmente, mais de 2,2 bilhões de pessoas têm deficiência visual, com cerca de um bilhão desses casos podendo ser evitados. A detecção precoce de doenças oculares é crucial para prevenir a evolução para estados irreversíveis, como a cegueira. Assim, este estudo apresenta um novo método para detectar múltiplas patologias oculares em imagens de fundo de olho, com a utilização de uma arquitetura de rede neural baseada em* transformers*, denominada Query2Label. Os experimentos foram realizados no conjunto de dados público RFMiD, revelando resultados promissores, com destaque para uma precisão média de 99,8% na categoria "D. Risk". Em comparação com o estado da arte, o método mostrou excelente desempenho na detecção da classe "ODP" e superou a precisão em outras categorias específicas, como "CSR","LS", destacando sua viabilidade e eficácia na classificação de patologias oftalmológicas.*

1. Introdução

Globalmente, pelo menos 2,2 bilhões de pessoas têm deficiência visual. Dessas, em pelo menos um bilhão, ou quase metade desses casos, a deficiência visual poderia ter sido evitada [Organization et al. 2019]. No Brasil, cerca de 3,5% têm deficiência visual [IBGE 2010]. Segundo o Conselho Brasileiro de Oftalmologia (CBO), 90% dos casos de deficiência visual são preveníveis ou tratáveis [Umbelino and Ávila 2023]. Assim, é de extrema importância que aconteça avanços tecnológicos relacionados a diagnósticos preventivos e/ou automáticos, principalmente em doenças que atingem todos os anos uma larga escala de pessoas, como a Degeneração Macular relacionada à idade

(DMRI), Retinopatia Diabética (RD) e Glaucoma, que causam cegueira em mais de 10 milhões de pessoas em todo o mundo [Mittal and Rajam 2020]. O exame de fundo de olho é realizado com a visualização da região da retina, por meio de fotos coloridas do fundo de olho, oferecendo um exame não invasivo da microcirculação sistêmica da retina [Pachade et al. 2021].

Já que múltiplas doenças podem acometer a retina de um único paciente, o trabalho em questão foi tratado como um caso de classificação Multirrótulo. Neste tipo de aprendizado temos um conjunto de treinamento composto por instâncias, nas quais uma única instância está associada a vários rótulos de diferentes classes simultaneamente [Zhang et al. 2018]. Tal método é de suma importância para o diagnóstico de doenças oculares, já que pacientes que, por exemplo, sofrem de RD podem também sofrer de outras doenças, como Glaucoma e DMRI. Portanto, a detecção de múltiplas doenças é essencial caso aja o risco de presença de mais de uma patologia em um paciente [Pachade et al. 2021].

A literatura apresenta alguns trabalhos relacionados a esta pesquisa, as quais propuseram métodos e técnicas computacionais para a detecção de patologias de fundo de olho que contribuíram para a realização deste trabalho.

Estudos como [Araújo et al. 2017, Silva et al. 2018, Ceschini et al. 2022] visam auxiliar no diagnóstico do glaucoma. [Araújo et al. 2017] propõe um método de diagnóstico do glaucoma em imagens de fundo de olho, usando os índices de diversidade de Shannon e McIntosh como descritores de textura e SVM para classificação. Os índices de textura alcançaram uma acurácia média de 88,35%, sensibilidade média de 84,50% e especificidade média de 91,37%. [Silva et al. 2018] apresenta um método computacional para detecção automática do glaucoma em retinografias, usando descritores de textura como LBP, LQP, CS-LBP e CLBP, e SVM para classificação. O método obteve uma acurácia de 90,70%. Por fim, [Ceschini et al. 2022] unifica duas redes de segmentação, reduzindo o tempo de processamento em 24,24%. Além disso, adicionaram uma segunda rede de classificação direta, aumentando a sensibilidade do modelo em 3% em comparação com o método base.

No estudo apresentado em [Dominik Müller and Kramer 2021], foi empregado o método de Aprendizado em Conjunto após um *Up-Sampling* em toda base de imagens. Este método utiliza dois modelos distintos: um para detectar a doença de fundo de olho e outro para classificar em caso de patologia detectada. Ambos os modelos foram pré-treinados no conjunto de imagens ImageNet [Deng et al. 2009]. Utilizando o conjunto de imagens RFMiD, foram empregados quatro *backbones* diferentes: DenseNet201, ResNet152, InceptionV3 e EfficientNetB4.

Um método de classificação utilizando *Transformers*, com uma solução que emprega *Transformer Decoders* para questionar a presença de uma categoria foi apresentado em [Liu et al. 2021]. Também é utilizado um método de Atenção Cruzada para o *Adaptively Feature Pooling*, visando detectar diferentes partes importantes em uma imagem. A entrada passa por um *backbone* que extrai a localização das características encontradas, antes de ser processada pela arquitetura proposta.

Em [Rodriguez et al. 2022], foi adotado um método baseado em *transformers*, utilizando a arquitetura C-Tran proposta em [Farnell et al. 2008]. Essa arquitetura consiste

em um *transformer encoder* alimentado por características visuais extraídas de uma CNN e por máscaras das categorias. Cada imagem é processada por uma Rede Neural Convolucional *backbone*, e um conjunto de representações de rótulos é gerado para realizar as predições.

Um método de *Ensemble Learning* para combinar as predições de vários modelos de redes convolucionais em [Dominik Müller and Kramer 2021]. Por outro lado, os trabalhos [Liu et al. 2021, Rodriguez et al. 2022] demonstram o diferencial do uso de *Transformers* para a tarefa de classificação. O método de [Rodriguez et al. 2022] em específico consegue obter um ótimo resultado mesmo sem conhecimento prévio, ou seja, pré-treinamento, utilizando o método LMT (*label mask training*), separando parte das categorias para aprender combinações entre as *labels*.

Dessa forma, o objetivo deste estudo é propor um modelo de rede neural baseado em aprendizado profundo para classificar automaticamente tanto doenças oculares frequentes como patologias raras na retina. O método proposto neste trabalho é principalmente baseado no Query2Label [Liu et al. 2021], que oferece flexibilidade ao usar diferentes *backbones* para a extração de características da imagem, possibilitando o uso de modelos pré-treinados para melhores resultados. Sua fácil adaptabilidade a diferentes *backbones* e foco na classificação de múltiplos rótulos são os principais motivos para sua escolha como base. Até o momento, o Query2Label não foi avaliado ou aplicado no contexto específico deste estudo, o que representa um fator relevante para sua escolha como base.

Destaca-se como principal contribuição deste trabalho a aplicação da arquitetura Query2Label com distintos *backbones*, aplicada no problema de classificação de retinografias com múltiplos rótulos. Para tanto, utilizou-se do conjunto de imagens RFMiD juntamente com técnicas de tratamento de dados.

2. Materiais e método

Esta seção apresenta o método proposto para detecção e classificação de patologias na retina. Além do conjunto de dados utilizado, é apresentado: o método proposto, arquitetura utilizada e métricas aplicadas.

2.1. Base de Imagens

O método proposto foi validado utilizando o conjunto de dados RFMiD [Pachade et al. 2021], uma base de imagens de retina de acesso público. Esta versão do conjunto de dados abrange 29 diferentes patologias. As imagens são divididas em subconjuntos de treinamento (1.920) e teste (640). A distribuição de amostras para cada classe no subconjunto de treinamento é apresentada na Tabela 1.

Para melhor entendimento das siglas correspondentes às patologias oculares presentes no conjunto de dados, as mesmas serão listadas a seguir: D.Risk (Disease Risk): Presença de doença/anormalidade, DR: Retinopatia Diabética, ARMD: Degeneração Macular Relacionada à Idade, MH: Neblina na mídia, DN: Drusas, MYA: Miopia, BRVO: Oclusão de Veia Retiniana de Ramo, TSLN: Tesselação, ERM: Membrana Epirretiniana, LS: Cicatrizes de Laser, MS: Cicatriz Macular, CSR: Retinopatia Serosa Central, ODC: Escavação do Disco Óptico, CRVO: Oclusão de Veia Retiniana Central, TV: Vasos Tortuosos, AH: Hialose Asteróide, ODP: Palidez do Disco Óptico, ODE: Edema do Disco

Tabela 1. Frequência das patologias presentes no subconjunto de dados de treinamento do Dataset RFMiD.

Patologia	Amostra	Patologia	Amostra	Patologia	Amostra
D.Risk	1519	MS	15	PT	11
DR	376	CSR	37	RT	14
ARMD	100	ODC	282	RS	43
MH	317	CRVO	28	CRS	32
DN	138	TV	6	EDN	15
MYA	101	AH	16	RPEC	22
BRVO	73	ODP	65	MHL	11
TSLN	186	ODE	58	RP	6
ERM	14	ST	5	OTHER	34
LS	47	AION	17		

Óptico, ST: Derivação Optociliar, AION: Neuropatia Óptica Isquêmica Anterior, PT: Telangiectasia Parafoveal, RT: Tração Retiniana, RS: Retinite, CRS: Coriorretinite, EDN: Exsudação, RPEC: Alterações no Epitélio Pigmentado da Retina, MHL: Buraco Macular, RP: Retinite Pigmentosa, Other: Outras patologias oculares.

2.2. Método proposto

O método proposto é delineado nas etapas mostradas na Figura 1. Após a coleta de dados, ocorre a etapa de pré-processamento do modelo, onde o subconjunto é dividido em 85% das imagens para treinamento e 15% para validação. Apenas as imagens de treinamento passam pelo processo de corte da retina e *Up-Sampling*, resultando em um subconjunto com um maior balanceamento entre os rótulos. Assim, o subconjunto de treinamento contém 3.354 imagens após o pré-processamento, enquanto o de validação possui 640 imagens. Posteriormente, o modelo é treinado, com a aplicação de *augmentations* para aprimoramento da generalização do modelo.

2.2.1. Pré-processamento

A seguir serão mostrados os passos realizados na fase de pré-processamento do conjunto de dados, sendo realizada uma técnica de corte da imagem, *up-sampling* em todo o conjunto de imagens e aplicações de *augmentation* em todo esses dados alterados durante o treinamento.

A técnica de corte de imagens tem o objetivo de manter a parte central da imagem do fundo do olho e preservar sua proporção após o redimensionamento. O corte foi adaptado para diferentes resoluções de imagem, incluindo 1424x1424, 1536x1536 e 3464x3464 pixels. Posteriormente, as imagens foram redimensionadas para o treinamento da rede, atingindo o tamanho de 384x384 pixels, conforme necessário pelo modelo. A Figura 2 ilustra essa transformação.

Treinar um classificador multirrótulo eficaz é desafiador devido ao desequilíbrio entre classes, o que é um obstáculo conhecido na construção de modelos [Kaur and Gosain 2018]. Para mitigar esse desequilíbrio, foi empregado o método de *up-sampling*. Este método, juntamente com a técnica de balanceamento de pesos das classes, ajuda a equilibrar o conjunto de dados, aumentando o número de amostras em patologias

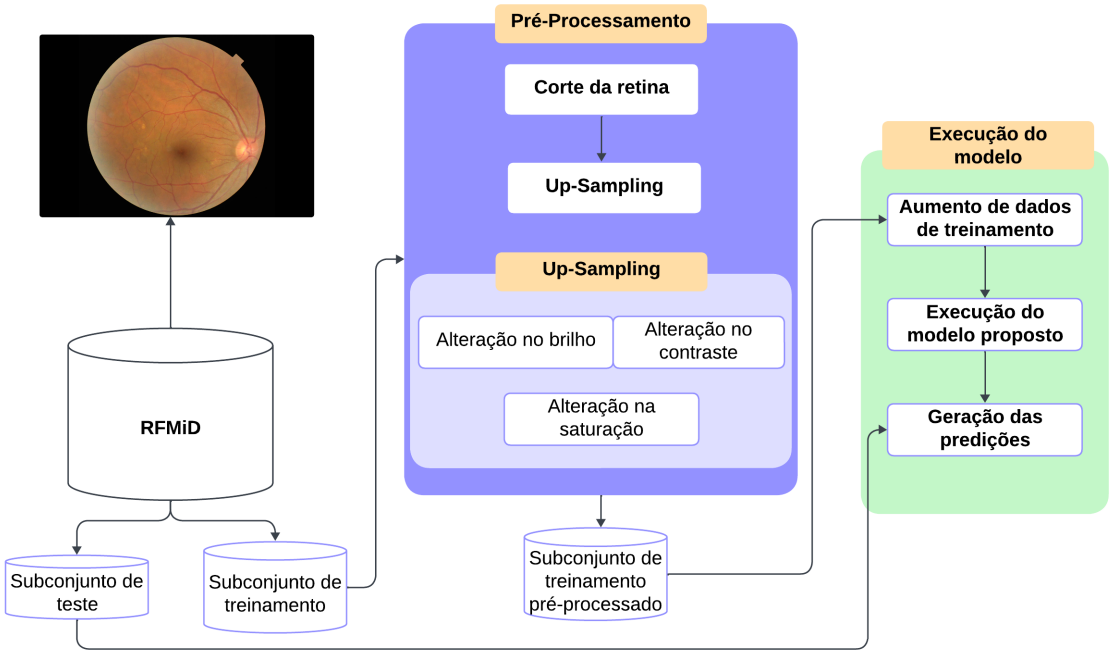


Figura 1. Etapas do método proposto.

Figura 2. Imagem após passar pelo Up-Sampling e função de corte.

menos frequentes. Apesar disso, algumas classes específicas ainda apresentam desafios devido ao baixo número de amostras.

O *upsampling* garantiu que cada categoria tivesse pelo menos 100 amostras no conjunto de dados, sendo realizado apenas no subconjunto de treinamento. Foram aplicadas técnicas como alteração no brilho, tonalidade, contraste e saturação. Essas transformações foram escolhidas devido às variações naturais nas propriedades visuais das imagens, como luminosidade, comuns em conjuntos de dados reais. A Tabela 2 apresenta a frequência das patologias após a aplicação desta técnica.

Para aumentar a diversidade do conjunto de dados de treinamento, aprimorando a capacidade de generalização do modelo e evitando *overfitting*, foi empregada a técnica de *augmentation*. Esta técnica consiste na aplicação de transformações aleatórias às imagens durante o treinamento de modelos de aprendizado de máquina. Foram aplicadas transformações de rotação dentro do intervalo [-90; 90] e espelhamento vertical e horizontal, bem como ajustes na saturação, brilho, contraste e tonalidade da imagem dentro dos intervalos [-0,2; 0,2].

Tabela 2. Frequência das patologias presentes no subconjunto de dados de treinamento do Dataset RFMiD após o up-sampling.

Patologia	Amostra	Patologia	Amostra	Patologia	Amostra
D.Risk	2953	MS	105	PT	104
DR	566	CSR	110	RT	103
ARMD	168	ODC	571	RS	103
MH	455	CRVO	100	CRS	104
DN	300	TV	102	EDN	105
MYA	124	AH	101	RPEC	102
BRVO	124	ODP	272	MHL	101
TSLN	345	ODE	109	RP	100
ERM	108	ST	101	OTHER	188
LS	139	AION	101		

2.2.2. Arquitetura utilizada

A arquitetura utilizada foi a Query2Label [Liu et al. 2021], desenvolvida especificamente para tarefas de classificação multirrótulo com o uso de *transformers*. Sua arquitetura é exibida na Figura 3.

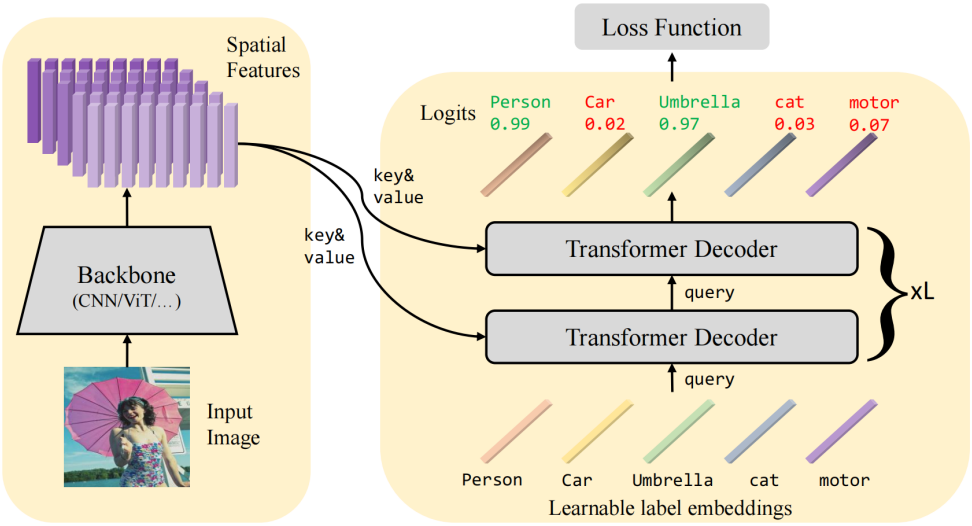


Figura 3. Esquema do modelo Query2Label [Liu et al. 2021].

Na primeira etapa, um *backbone* é empregado para extrair características das imagens. Esse *backbone* pode ser constituído por modelos convencionais de CNN, como ResNet, DenseNet, entre outros, ou por modelos de Vision Transformer.

Na segunda etapa, as características espaciais das imagens extraídas são direcionadas para um decodificador de *transformer* com várias camadas, onde *label embeddings* - representações numéricas das classes - são utilizados como consultas para determinar a presença de cada rótulo. Isso é realizado por meio de uma atenção cruzada (*Cross-Attention*) de várias cabeças, adaptativamente agrupando as características dos objetos, permitindo então a classificação binária subsequente. Esse processo é seguido por uma camada de projeção linear que calcula logits de previsão para cada rótulo correspondente.

Para avaliar o método proposto, foram utilizadas a métrica de Precisão e o gráfico

da Área Sob a Curva ROC (AUC-ROC). A precisão mede a porcentagem de exemplos classificados corretamente em uma categoria específica dentro do conjunto de dados. Por outro lado, a AUC-ROC mede a capacidade de um modelo em distinguir entre classes positivas e negativas. Quanto mais próximo de 1 os valores das métricas, melhor é a capacidade do modelo em classificar corretamente as instâncias positivas e negativas.

3. Resultados

Esta seção apresenta e discute os resultados que foram alcançados realizando os experimentos no *dataset* RFMiD.

3.1. *Backbone* ResNet

Neste experimento, a ResNet foi utilizada como backbone, com as variantes ResNet18, ResNet34, ResNet50 e ResNet101. As técnicas aplicadas em cada backbone da ResNet foram as mesmas utilizadas na ResNet101, incluindo os mesmos parâmetros. No entanto, a ResNet101 demonstrou melhor desempenho em comparação com as outras arquiteturas.

Os dois primeiros experimentos foram conduzidos com 250 épocas de treinamento, usando um tamanho de lote de 64 e uma taxa de aprendizado inicial de 1×10^{-4}, com o otimizador "AdamW". No último teste mencionado, foi incorporada uma taxa de decaimento de peso (*weight decay*) e implementada a técnica de busca em grade (*Grid Search*) para otimização de hiperparâmetros. Isso implicou em ajustar a taxa de aprendizado e o decaimento de peso para encontrar os valores ideais de cada parâmetro. A validação foi realizada utilizando a técnica *Holdout*, com uma proporção de 85:15, onde 85% dos dados foram usados para o treinamento e os 15% restantes para a validação.

Os resultados mostrados na Tabela 3 foram alcançados utilizando a arquitetura base apresentada juntamente com o backbone da ResNet101.

Tabela 3. Precisão de cada categoria (Q2L-ResNet101).

Patologia	Precisão	Patologia	Precisão	Patologia	Precisão
D.Risk	98,7	MS	8,5	PT	2,6
DR	88,7	CSR	28,1	RT	73,8
ARMD	60,8	ODC	55,2	RS	82,5
MH	86	CRVO	73,3	CRS	27,1
DN	46	TV	0,6	EDN	16,6
MYA	87,3	AH	47,5	RPEC	7,1
BRVO	57,7	ODP	37	MHL	43,1
TSLN	64,9	ODE	72,9	RP	8
ERM	6	ST	1	OTHER	17,5
LS	53,9	AION	20,6		

A classe D.Risk alcançou uma precisão notavelmente alta de 98,7%. No entanto, outras classes, como "ST", "RP", "ERM", "TV" e "MS", registraram valores de precisão muito baixos. Isso se deve principalmente ao reduzido número de amostras disponíveis no conjunto de treinamento para essas classes, o que, combinado com a complexidade em detectar as características específicas dessas patologias, contribui para o desempenho inferior. Apesar do *Up-Sampling* realizado para abordar a baixa precisão em classes com poucas amostras, ainda houve dificuldades em melhorar a precisão dessas categorias.

Para dados mais específicos, a classe “TV” possui apenas 6 amostras iniciais, enquanto a classe “ST” tem ainda menos, com apenas 5 amostras. Essas classes, com um número extremamente reduzido de imagens, provaram ser um desafio significativo para alcançar uma precisão satisfatória.

3.2. *Backbone* ViT

Um segundo experimento foi realizado utilizando a arquitetura *Vision Transformers* como *backbone*, previamente treinada. Após o pré-processamento, as imagens foram redimensionadas para 384x384 para serem usadas como entrada na arquitetura modificada.

Na arquitetura modificada, foram introduzidas diversas alterações na etapa de pré-processamento visando uma melhor utilização da memória disponível, o que possibilitou um aumento no tamanho do lote durante a execução da rede nos testes subsequentes. Além disso, foi empregado um modelo pré-treinado denominado “CVT_W24” como extrator de características, originalmente treinado na base de dados ImageNet. Também foi aplicada uma função de decaimento de peso (*weight decay*) às camadas do modelo. Os hiperparâmetros foram otimizados utilizando a técnica de *Grid Search*. Os parâmetros inseridos nessa grade foram o *learning rate* (taxa de aprendizado) e o *weight decay* (decaimento de peso). Por consequência dessas modificações, essa arquitetura foi denominada “Q2L-CVT_W24”.

A Tabela 4 apresenta a precisão de cada categoria do conjunto de dados RFMiD após o treinamento do método em questão.

Tabela 4. Precisão de cada categoria (Q2L-CVT_W24).

Patologia	Precisão	Patologia	Precisão	Patologia	Precisão
D.Risk	99,8	MS	17	PT	2,4
DR	89,5	CSR	65,4	RT	78,7
ARMD	67,1	ODC	35,8	RS	85,9
MH	86,2	CRVO	82,4	CRS	29,7
DN	29,2	TV	0,6	EDN	21,9
MYA	83,2	AH	35,6	RPEC	4,5
BRVO	65,3	ODP	20,7	MHL	7,4
TSLN	75,2	ODE	72,8	RP	51,5
ERM	10,3	ST	0,6	OTHER	20,7
LS	65,9	AION	35,6		

Apesar das alterações implementadas, não houve avanços significativos em comparação com experimento anterior, exceto em casos específicos, como nas patologias “CSR” e “RP”, onde houve um aumento notável na precisão.

Como estado da arte, foi utilizado o estudo [Rodriguez et al. 2022], que também utilizou *transformers* no *dataset* RFMiD. Apesar de usar o mesmo conjunto de dados, o trabalho em questão não utilizou todas as classes do *dataset*, já que ele usou apenas classes que também apresentavam em outro conjunto de dados estudado no trabalho em questão. Comparando apenas a detecção da patologia, foi obtido um resultado superior: no trabalho em questão, foi utilizado a categoria dos normais, que basicamente seria a mesma categoria do “D.Risk”, porém havendo falsos positivos ao invés de verdadeiros positivos, ou seja: a mesma categoria apenas utilizando os casos normais como métrica ao invés de casos com doença. Na classe dos normais foi obtido uma precisão de 85,9% no

trabalho comparado, enquanto no método proposto se obteve 99,8%. A Tabela 5 mostra a comparação dos experimentos com o estados da arte.

Tabela 5. Comparação da precisão (%) entre os experimentos Q2L-ResNet101 e Q2L-CVT_W24 com o estado da arte [Rodriguez et al. 2022].

Patologia	Q2L-ResNet101	Q2L-CVT_W24	Estado da arte
D.Risk	98,7	99,8	85,9
ARMD	60,8	67,1	80,0
MH	86,0	86,2	87,5
DN	46,0	29,2	70,8
MYA	87,3	83,2	81,0
BRVO	57,7	65,3	92,9
TSLN	64,9	75,2	80,0
LS	53,9	65,9	50,0
CSR	28,1	65,4	44,4
ODC	55,2	35,8	66,1
CRVO	73,3	82,4	60,0
ODP	37,0	20,7	0,0
ODE	72,9	72,8	83,3
RS	82,5	85,9	100,0
CRS	27,1	29,7	40,0
OTHER	17,5	20,7	58,7

Ainda comparando com o método proposto em [Rodriguez et al. 2022], em algumas categorias, como o caso da classe "ODP", o estado da arte não adquiriu nenhum acerto, obtendo 0% de precisão. Por outro lado, este trabalho obteve 37% de precisão no primeiro experimento apresentado e 20,7% no segundo. Outras patologias apresentaram resultados semelhantes e outras pequenas melhoras, como "MYA", com uma pequena superioridade de 6,3% utilizando o primeiro experimento e com 2,2% no segundo.

Apesar disso, houve também classes com resultados inferiores em comparação com a arquitetura comparada, como a classe "DN" com uma piora de 24,8% se comparada com o primeiro experimento; "RS", o qual foi obtido 100% e "BRVO" com uma piora de 35,2% comparado com o primeiro experimento.

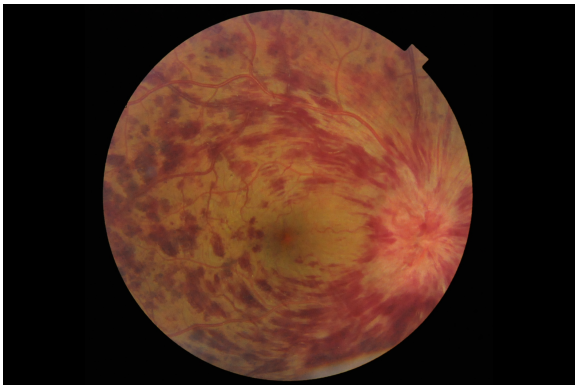
3.3. Estudos de Casos

Algumas categorias apresentaram valores de precisão significativamente abaixo do esperado, como a categoria "TV". Esta patologia é caracterizada por um padrão de grande tortuosidade nos vasos sanguíneos, que parecem dilatados e seguem um caminho sinuoso, conforme ilustrado na Figura 4. Além da complexidade na detecção desse padrão, o conjunto de dados inclui apenas seis amostras dessa doença ocular, o que contribui para a baixa precisão nesta patologia.

Por outro lado, a categoria que apresentou maior discrepância positiva em termos de precisão em relação ao método de [Rodriguez et al. 2022] foi a categoria "CRVO". Essa patologia ocorre quando a veia principal da retina fica bloqueada perto do início ou no nervo óptico. Os sinais clínicos incluem hemorragias em formato de chama, conforme mostrado na Figura 5. Apesar de ter sido treinado com um conjunto pequeno de amostras, apenas 28 no total, o método demonstrou habilidade em aprender os padrões presentes nessas imagens. Isso se deve à sua capacidade de identificar facilmente a característica principal, que se destaca consideravelmente em relação às demais amostras.

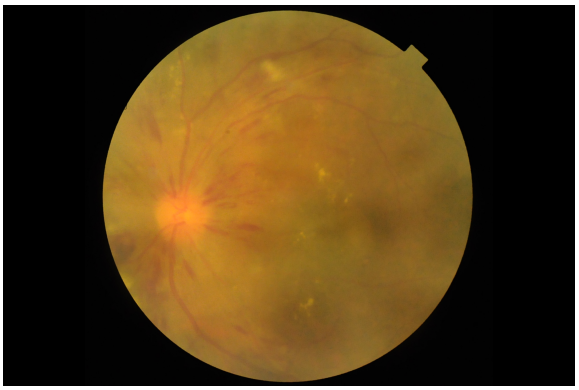
Figura 4. Vasos tortuosos presentes na categoria “TV”.

Figura 5. Hemorragias em formato de chama da categoria “CRVO”.



A amostra apresentada na Figura 6 exibe as patologias “CRVO”, mencionada anteriormente, e “MH”. A opacidade visual observada na imagem com “MH” pode indicar a presença de catarata, opacidades no vítreo, edema de córnea ou pupilas pequenas. A patologia “MH” também alcançou uma precisão significativa na rede, atingindo 86,2%, aproximando-se do desempenho do método proposto em [Rodriguez et al. 2022].

Figura 6. Amostra apresentando a patologia “MH” e “CRVO”.



4. Conclusão

Este estudo apresentou uma abordagem alternativa para a classificação multirrótulo de patologias na retina, utilizando o conjunto de dados RFMiD. Destacou-se a eficácia na

detecção de patologias (Disease Risk) e na classificação de diversas categorias, como "ODP", "CRVO" e outras presentes no conjunto de dados.

Nos testes realizados, o método proposto alcançou uma precisão de 99,8% na categoria de D.Risk, indicando uma detecção quase ideal da patologia ocular, superior ao estado da arte o qual possui uma precisão de 99,8% na mesma categoria. Destaca-se também outras categorias, como "CRVO", que alcançou 82,4% de precisão com apenas 28 imagens no subconjunto de treinamento; "LS", com 65,9% de precisão com apenas 47 imagens, e "CSR", com 65,4% de precisão com 37 imagens no mesmo subconjunto. Em comparação com o método de [Rodriguez et al. 2022], utilizado como estado da arte, o método proposto superou algumas classes e se equiparou em outras, sem apresentar saltos expressivos em categorias nas quais não superou.

As contribuições deste estudo se destacam pela introdução de uma nova arquitetura de rede baseada em *Transformers*, utilizada para a classificação de doenças em retinografias do conjunto de dados RFMiD, aliada às técnicas de pré-processamento adotadas. Um avanço significativo é observado na detecção de patologias, evidenciando uma precisão notável.

Por fim, propõe-se, para trabalhos futuros, a investigação de outras arquiteturas de *backbone*, como a DenseNet [Huang et al. 2017] ou a TResNet [Ridnik et al. 2021]. Além disso, sugere-se a combinação do método proposto com abordagens adicionais, como a apresentada em [Oh and Park 2022], utilizando um extrator de características com dois classificadores distintos, conforme a performance em classes específicas. Por exemplo, a aplicação do método proposto para a detecção de patologias, como "D.Risk" e "CRVO", combinada com o método do estado da arte para classes como "BRVO" e "RS". Essa abordagem, utilizando o melhor de dois classificadores diferentes, pode resultar em um desempenho aprimorado. Espera-se que a fusão do método proposto com o trabalho citado possa elevar o nível de desempenho no conjunto de dados utilizado neste estudo.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 , Fundação de Amparo a Pesquisa do Maranhão (FAPEMA) (Termo: 000527/2024), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Empresa Brasileira de Serviços Hospitalares (Ebserh) Brazil (Proc. 409593/2021-4).

Referências

Araújo, J., de Paiva, A., de Almeida, J., Neto, O. P. S., de Sousa, J., Silva, A., and Júnior, G. B. (2017). Diagnóstico de glaucoma em imagens de fundo de olho utilizando os Índices de diversidade de shannon e mcintosh. In *Anais do XVII Workshop de Informática Médica*, Porto Alegre, RS, Brasil. SBC.

Ceschini, L., Policarpo, L., Rodrigues, V., Righi, R., and Ramos, G. (2022). Otimizando o diagnóstico automatizado de glaucoma a partir de imagens de fundo de olho. In *Anais da XXII Escola Regional de Alto Desempenho da Região Sul*, pages 9–12, Porto Alegre, RS, Brasil. SBC.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dominik Müller, I. S.-R. and Kramer, F. (2021). Multi-disease detection in retinal imaging based on ensembling heterogeneous deep learning models.

Farnell, D. J., Hatfield, F. N., Knox, P., Reakes, M., Spencer, S., Parry, D., and Harding, S. P. (2008). Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. *Journal of the Franklin institute*, 345(7):748–765.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

IBGE (2010). Censo demográfico.

Kaur, P. and Gosain, A. (2018). Issues and challenges of class imbalance problem in classification. *International Journal of Information Technology*, pages 1–7.

Liu, S., Zhang, L., Yang, X., Su, H., and Zhu, J. (2021). Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*.

Mittal, K. and Rajam, V. (2020). Computerized retinal image analysis-a survey. *Multimedia Tools and Applications*, 79(31):22389–22421.

Oh, Y.-t. and Park, H. (2022). End-to-end two-branch classifier for retinal imaging analysis. In *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–3. IEEE.

Organization, W. H. et al. (2019). World report on vision.

Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., and Mériaudeau, F. (2021). Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. *Data*, 6(2):14.

Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., and Friedman, I. (2021). Tresnet: High performance gpu-dedicated architecture. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1400–1409.

Rodriguez, M., AlMarzouqi, H., and Liatsis, P. (2022). Multi-label retinal disease classification using transformers. *arXiv preprint arXiv:2207.02335*.

Silva, M., Pessoa, A., de Almeida, J., Júnior, G. B., and de Paiva, A. (2018). Diagnóstico do glaucoma em imagens de retinografia usando variantes de padrões locais binários. In *Anais do XVIII Simpósio Brasileiro de Computação Aplicada à Saúde*, Porto Alegre, RS, Brasil. SBC.

Umbelino, C. C. and Ávila, M. P. (2023). As condições de saúde ocular no brasil. *São Paulo: Conselho Brasileiro de Oftalmologia*.

Zhang, M.-L., Li, Y.-K., Liu, X.-Y., and Geng, X. (2018). Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.