Computer Vision for Automated FAMACHA Classification in Sheep and Goats: A Comparison of Color-Based and Deep Learning Strategies

Mauricio Benjamin da Rocha[1], **Luciano Sousa Barbosa**[1], **Júlio Vitor Monteiro Marques**[1], **Mateus Assis Veloso** [1], **José Lindenberg Rocha Sarmento** [1], **Flávio Henrique Duarte de Araújo** [1]

[1]Universidade Federal do Piauí (UFPI), Teresina, PI, Brasil

mauriciobenjamin@ufpi.edu.br, luciano.barbosa@ufpi.edu.br, juliomonteiro@ufpi.edu.br

mateusveloso@ufpi.edu.br, sarmento@ufpi.edu.br, flavio86@ufpi.edu.br

Abstract. *The FAMACHA© method selects anemic sheep and goats for treatment through ocular mucosa color, but relies on subjective human assessment. This work compares three strategies for automating it under a single 5-fold stratified cross-validation protocol, with an identical partition and the same segmenter-derived region of interest, over a purpose-built dataset of 843 expert-annotated images: classical machine learning on color features, transfer learning with CNNs, and YOLO classification. The YOLO classifier performed best (F1 = 0.840, Cohen's κ = 0.635), above the other two by a paired McNemar test ($p < 0.02$), and its errors concentrate on the decision boundary: none of the images scored 5 were missed. The color-based approach (F1 = 0.799) remained competitive and interpretable, supported by a LASSO regression.*

Resumo. *O método FAMACHA© seleciona caprinos e ovinos anêmicos para tratamento pela coloração da mucosa ocular, mas depende de avaliação humana subjetiva. Este trabalho compara três estratégias para automatizá-lo sob um único protocolo de validação cruzada estratificada de 5 folds, com particionamento idêntico e a mesma região de interesse fornecida pelo segmentador, em uma base própria de 843 imagens anotadas por especialista: aprendizado de máquina clássico sobre features de cor, transfer learning com CNNs e classificação YOLO. O classificador YOLO obteve o melhor desempenho (F1 = 0,840; κ = 0,635), superior às outras duas pelo teste pareado de McNemar ($p < 0{,}02$), e seus erros concentram-se na fronteira de decisão: nenhuma imagem de escore 5 deixou de ser indicada para tratamento. A abordagem baseada em cor (F1 = 0,799) manteve-se competitiva e interpretável, apoiada por uma regressão LASSO.*

1. Introdução

A caprinocultura e a ovinocultura desempenham papéis estratégicos na agropecuária brasileira, sobretudo em regiões semiáridas, onde são fonte de renda e segurança alimentar. O rebanho nacional, de cerca de 12,6 milhões de caprinos e 20,5 milhões de ovinos, concentra-se na Região Nordeste, com forte presença na agricultura familiar [IBGE 2022a, IBGE 2022b]. Apesar do potencial produtivo, a verminose gastrintestinal, em especial as infecções por *Haemonchus contortus*, helminto hematófago que

causa anemia, perda de peso e mortalidade, limita a eficiência desses sistemas, agravada pelo clima quente e úmido que favorece as larvas infectantes [Teixeira et al. 2023, Nascimento 2022]. O controle por vermifugação em massa acelerou a resistência anti-helmíntica, comprometendo a eficácia dos princípios ativos e elevando custos.

Como alternativa sustentável, o Tratamento Seletivo Direcionado (TSD) trata apenas os animais que necessitam, destacando-se o método FAMACHA©, que identifica animais anêmicos pela coloração da mucosa ocular [Molento et al. 2004]. O método reduziu em 75,6% o uso de anti-helmínticos em ovinos e teve aplicabilidade confirmada em diferentes contextos [de Oliveira Diniz et al. 2022, Senoamadi et al. 2022]. Sua aplicação, porém, depende de avaliação visual humana, cuja acurácia varia com o treinamento do avaliador e com a iluminação em campo, limitação recorrente na literatura que aplica e automatiza o método [Molento et al. 2004, Senoamadi et al. 2022, Freitas et al. 2023].

Paralelamente, a Inteligência Artificial tem transformado a pecuária de precisão [Arshad et al. 2024]. Na visão computacional aplicada à saúde animal a classificação é supervisionada em todos os casos, e o que distingue as abordagens é quem define a representação da imagem: descritores explícitos e interpretáveis, como estatísticas de cor, entregues a classificadores como SVM, KNN e Random Forest; ou redes convolucionais, incluindo a família YOLO [Redmon et al. 2016], que aprendem a representação dos *pixels* ao custo de menor transparência. A escolha envolve um compromisso entre desempenho, custo e interpretabilidade, cuja avaliação depende de métricas adequadas [Strauss et al. 2022].

A literatura de automação do FAMACHA©, porém, explora essas abordagens de forma majoritariamente isolada, cada qual sob protocolo próprio, e carece de bases públicas anotadas segundo o escore, o que dificulta comparações justas e a reprodutibilidade. Diante disso, o trabalho aborda a pergunta: dentre as estratégias de visão computacional disponíveis, qual oferece o melhor compromisso entre desempenho, complexidade e interpretabilidade para a seleção automatizada de caprinos e ovinos anêmicos? O objetivo é comparar sistematicamente três estratégias de classificação, aprendizado de máquina sobre *features* de cor, *transfer learning* com CNNs e classificação por modelos YOLO, sob validação cruzada estratificada idêntica e sobre a mesma região de interesse, em uma base construída para este fim.

As contribuições deste trabalho são: (i) uma base própria de 843 imagens da mucosa ocular anotada por especialista, diante da inexistência de conjuntos públicos; (ii) a comparação entre ML clássico, *transfer learning* e YOLO sob protocolo unificado, com particionamento e ROI idênticos, de modo que a diferença observada seja atribuível ao classificador; (iii) a avaliação do impacto da variante de ROI, dos espaços de cor e do balanceamento, com o custo computacional de cada rota; (iv) uma análise de interpretabilidade por regressão LASSO; e (v) uma análise de erros por escore FAMACHA© original, com teste pareado entre as estratégias, que mostra onde cada rota falha e que seus erros só coincidem em parte.

2. Trabalhos Relacionados

A automação do método FAMACHA© busca reduzir a subjetividade da avaliação clínica e ampliar a adoção do TSD, e organiza-se em três eixos: aplicações móveis de campo, delimitação da mucosa na imagem e classificação por aprendizado de máquina e pro-

fundo. Os dois últimos são interdependentes, pois classificar por descritores de cor exige a mucosa isolada do restante da cena, e é no grau de automação dessa delimitação e na representação entregue ao classificador que os trabalhos mais se distinguem (Tabela 1).

Molento et al. [Molento et al. 2004] fundamentam clinicamente o método no Brasil, correlacionando-o a parâmetros hematológicos. Entre as soluções de processamento clássico de imagens, Demoliner e Alves [Demoliner and Alves 2017] informatizaram o método no aplicativo *Anemímetro*, que decide por *Naive Bayes* sobre as médias HSV de 414 fotos recortadas manualmente; Ribeiro [Ribeiro 2018] isolou a conjuntiva por limiarização no canal H e classificou 114 imagens em três faixas do escore com KNN e SVM; Souza et al. [de Souza et al. 2023, de Souza et al. 2021] desenvolveram um aplicativo Android semiautomático em que SVM e rede neural decidem entre saudável e anêmico, com 87% e 83% de acurácia; e Freitas et al. [Freitas et al. 2023] extraíram quantis de cor de 1.156 imagens, segmentadas por um classificador de *pixels*, e as classificaram por Random Forest. No aprendizado profundo, Almeida [Almeida 2021] comparou U-Net, ERFNet e DeepLabV3 na segmentação de 106 imagens, seguidas de CNNs; Freitas et al. [Freitas et al. 2024] serviram U-Net e VGG19 em uma aplicação *web* sobre 3.441 imagens; Silva [Silva 2025] integrou YOLOv5 e CNN para os cinco níveis em *smartphone* com 165 imagens; e Souza Filho [Souza Filho 2022] treinou três CNNs sobre as 128 imagens cedidas por Almeida e as serviu por API.

Tabela 1. Síntese comparativa dos trabalhos relacionados de visão computacional. *Delimitação*: como a mucosa é isolada da cena. *Representação*: descritores de cor definidos por especialista ou aprendidos dos *pixels*. *Classes*: formulação do alvo. *Famílias*: quantas representações o trabalho mede entre si, e apenas este compara mais de uma sobre a mesma base, partição e ROI.

Trabalho	Classificadores	Delimitação	Representação	Classes	Famílias
Demoliner & Alves	*Naive Bayes*	Manual	Cor	5	1
Ribeiro	KNN e SVM	Limiarização	Cor	3	1
Souza et al.	SVM e rede neural	Semiautomática	Cor	2	1
Freitas et al. (2023)	Random Forest e regressão multinomial	Automática	Cor	5	1
Freitas et al. (2024)	VGG19	Automática	Aprendida	2	1
Almeida	CNN	Automática	Aprendida	2	1
Silva	YOLOv5 e CNN	Automática	Aprendida	5	1
Souza Filho	AlexNet, LeNet-5 e ZFNet	Não realiza	Aprendida	2	1
Este trabalho	**RF, MobileNet e YOLO**	**Automática**	**Cor e aprendida**	**2**	**3**

Há, portanto, uma trajetória de amadurecimento rumo a *pipelines* automatizados, mas dois obstáculos persistem. Primeiro, cada trabalho adota uma única família de representação sob protocolo próprio, com bases de 106 a 3.441 imagens, formulações de duas, três ou cinco classes e partições distintas, de modo que os desempenhos publicados não são comparáveis entre si nem permitem dizer qual família é preferível. Silva [Silva 2025], por exemplo, reporta F1 = 0,903 $\pm$ 0,052 nos cinco escores, acima do que aqui se obtém na tarefa binária, mais simples, mas sobre 165 imagens equilibradas em 33 por escore, ao passo que uma base de campo é dominada pelos escores intermediários: equilibrar a amostra remove justamente a dificuldade que a decisão clínica enfrenta. Segundo, nenhuma dessas bases é pública, o que impede a reprodução e a agregação dos resultados. É nessa lacuna que este trabalho se insere: mede as três famílias sobre o

mesmo conjunto de dados, o mesmo particionamento e a mesma região de interesse, isolando o classificador como única variável, e acrescenta a análise de erros e o custo de cada rota.

3. Metodologia

A metodologia organiza-se em um *pipeline* experimental unificado (Figura 1) e esta seção o percorre na ordem de execução: aquisição e anotação da base (Seção 3.1), segmentação da mucosa (Seção 3.2), extração das variantes de ROI e dos descritores de cor (Seção 3.3), as três estratégias de classificação (Seção 3.4), a interpretabilidade por regressão LASSO (Seção 3.5) e o protocolo de avaliação comum a todas (Seção 3.6).

A partir da base própria, um modelo de segmentação localiza a mucosa e define a região de interesse (ROI), consumida pelas três estratégias, que diferem apenas no classificador e na representação sobre a qual ele opera. Com particionamento idêntico, as diferenças de desempenho decorrem dos modelos, não do protocolo. A Seção 4.2 acrescenta a ablação em que o YOLO recebe a imagem inteira, sem a ROI, para separar o efeito da delimitação do efeito do classificador.

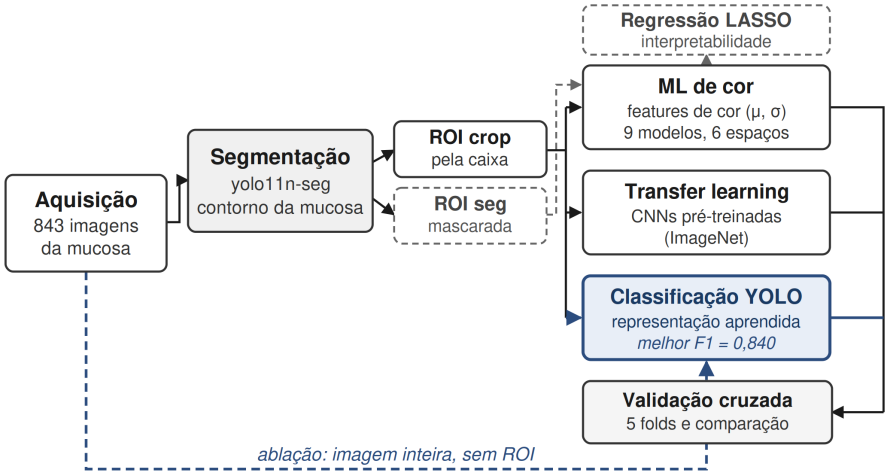


Figura 1. Visão geral do *pipeline*. A ROI tem duas variantes, o recorte pela caixa (*crop*) e o mascarado pelo polígono (*seg*), esta avaliada apenas na rota de cor. As três estratégias consomem a mesma ROI e partição; a linha tracejada é a ablação sem ROI.

3.1. Aquisição e anotação dos dados

Diante da inexistência de bases públicas anotadas segundo o escore FAMACHA©, construiu-se uma base própria por coletas em campo, em propriedades rurais de Picos e Teresina, no Piauí, usando o cartão FAMACHA© como referência de classificação. As imagens passaram por curadoria e anotação supervisionada por um especialista em Zootecnia. A anotação foi feita exclusivamente sobre a fotografia da mucosa, sem acesso ao animal, à sua condição corporal ou a exames hematológicos, de modo que o rótulo não carrega informação indisponível ao modelo. Em contrapartida, cada imagem recebeu o escore de um único avaliador, o que impede estimar a concordância entre especialistas e, com ela, o teto de desempenho da tarefa, limitação retomada na Seção 5.

A captura coube aos coletores, com seus *smartphones*, sem padronizar modelo, sensor ou aplicativo de câmera e sem controlar nem registrar a iluminação. Em cada animal o operador fez a contenção, everteu a pálpebra inferior e fotografou a mucosa em plano aproximado, o mesmo gesto que o método exige do avaliador humano. Na anotação as fotos foram padronizadas em 640×640 *pixels*, etapa que descartou os metadados de aquisição e impede estratificar os resultados por aparelho. A heterogeneidade de dispositivo e de luz fica, assim, embutida na base, o que a aproxima do uso real e responde por parte do ruído fotométrico.

Para alinhar o problema à decisão de manejo, as cinco classes foram agrupadas em duas: *Vermifuga* (graus 3 a 5) e *Não Vermifuga* (graus 1 e 2, cuja vermifugação pode ser adiada). Como a decisão clínica opera sobre um limiar (escore ≥ 3), a formulação binária captura a informação operacionalmente relevante; os cinco escores ficam como trabalho futuro. Cada imagem corresponde a um animal distinto, o que evita o mesmo indivíduo em partições diferentes. Das 843 imagens, 279 (33,1%) são *Não Vermifuga* e 564 (66,9%) *Vermifuga*, distribuídas pelos escores 1 a 5 como 86, 193, 294, 189 e 81: o desbalanceamento é de cerca de 2:1 e os escores extremos são os menos representados, o que motiva avaliar o balanceamento e usar métricas robustas (F1 *weighted* e κ de Cohen). Não é possível dizer, a partir desta amostra, se o predomínio de animais que requerem tratamento é próprio do método ou específico das propriedades visitadas, pois a coleta foi por conveniência em duas localidades. Base e código não são disponibilizados; para viabilizar a replicação, descrevem-se a base, a grade experimental e os hiperparâmetros.

3.2. Segmentação da mucosa ocular

A segmentação localiza a mucosa e delimita a ROI consumida pelas três estratégias. O alvo supervisionado foi construído por anotação manual: em cada imagem o contorno visível da mucosa ocular foi delimitado por um polígono, produzindo máscaras de referência de classe única (*membrane*). Sobre essas 849 máscaras treinou-se um modelo de segmentação de instância yolo11n-seg, com partição fixa de treino, validação e teste (595/169/85 imagens) em vez de validação cruzada, por duas razões: treinar o segmentador custa uma ordem de grandeza mais que os classificadores, e a métrica da tarefa já é uma média sobre limiares de sobreposição e sobre todas as instâncias de teste. Reporta-se o mAP da máscara em mAP@0,5, que mede se a mucosa foi localizada sob sobreposição mínima de 50%, e mAP@0,5:0,95, que penaliza contornos imprecisos, aspecto crítico aqui porque é o contorno que define os *pixels* dos descritores de cor. Das 849 imagens anotadas, seis saíram da base de classificação porque o recorte resultante não isolava a mucosa de forma confiável, o que explica as 843 imagens usadas na classificação: uma máscara imprecisa contamina diretamente os descritores calculados sobre ela. O segmentador foi treinado por 100 épocas e submetido a uma ablação de um fator por vez, alterando a resolução de entrada (640 e 768), desativando o *mosaic*, aplicando `copy_paste` = 0,3 e trocando o modelo pela variante maior yolo11s-seg, com os demais hiperparâmetros nos valores padrão da implementação [Jocher and Qiu 2024].

3.3. Pré-processamento e extração de *features* de cor

Da localização fornecida pela segmentação geraram-se duas variantes de ROI: *crop*, o recorte retangular pela caixa delimitadora; e *seg*, o recorte mascarado, no qual apenas os *pixels* internos ao polígono são preservados e todos os externos recebem valor zero em todos os canais, resultando em fundo preto. Ambas foram redimensionadas para 224×224

pixels, resolução canônica de entrada das redes convolucionais pré-treinadas no ImageNet; padronizar a ROI nesse tamanho permite que a mesma entrada alimente as três estratégias sem reprocessamento, preservando o protocolo unificado. Como a coloração da mucosa reflete o volume globular e a hemoglobina, descritores estatísticos de cor são uma representação compacta e interpretável do grau de anemia. Para o ML clássico, calcularam-se, para cada canal $X = \{x_1, \ldots, x_n\}$ com n pixels, a média $\mu = \frac{1}{n}\sum_i x_i$ e o desvio-padrão $\sigma = \sqrt{\frac{1}{n}\sum_i (x_i - \mu)^2}$. Preferiram-se esses dois momentos a histogramas completos ou descritores de textura por três motivos: o vetor fica com seis dimensões por espaço de cor, proporção adequada a uma base de centenas de imagens, contra dezenas ou centenas de dimensões de um histograma; a média descreve o deslocamento da coloração de avermelhada para pálida, que é o que o cartão FAMACHA© compara; e cada coeficiente permanece legível na análise de interpretabilidade (Seção 3.5). Descritores mais ricos ficam como trabalho futuro. Foram avaliados seis espaços de cor (RGB, HSV, HLS, Lab, LUV e YCrCb), que vão dos que misturam informação cromática e luminância (RGB) aos que as separam (Lab, LUV, YCrCb). O espaço é eixo experimental sob a hipótese de que, com a luminância em canal próprio, a variação de iluminação inevitável em campo recai sobretudo nela e deixa os canais cromáticos mais estáveis para descrever a palidez, que é como a anemia se manifesta.

Às *features* agregadas não se aplicou *data augmentation*: transformações geométricas não alteram a média nem o desvio-padrão dos *pixels* da região, e a fotométrica alteraria a própria cor que constitui o sinal, configurando vazamento de rótulo. A *augmentation* só é pertinente às abordagens convolucionais, que aprendem das imagens.

3.4. Estratégias de classificação

Três estratégias foram comparadas sob o protocolo unificado (Seção 3.6), todas na tarefa binária (*Vermifuga* vs. *Não Vermifuga*) e todas sobre a mesma ROI, o recorte por caixa. A seleção das famílias segue o eixo da comparação, o grau de autonomia da representação: descritores definidos por especialista, representação pré-treinada em domínio genérico e representação aprendida para a tarefa. Dentro de cada família tomaram-se os modelos compatíveis com o alvo de uso, um dispositivo móvel em campo, o que exclui arquiteturas de grande porte.

ML clássico sobre *features* de cor. Nove classificadores (AdaBoost, Árvore de Decisão, *Gradient Boosting*, k-NN, Regressão Logística, MLP, *Perceptron*, Random Forest e SVM) treinados sobre os vetores de cor, com hiperparâmetros por busca em grade e *class_weight* nos estimadores que o suportam. O conjunto cobre propositalmente fronteiras de decisão distintas, lineares, por distância, por margem, por árvore isolada, por *ensemble* e por rede rasa, o que permite verificar se o sinal de cor é linearmente separável. O espaço de busca contemplou as duas variantes de ROI, os seis espaços de cor e quatro estratégias de balanceamento (nenhuma, sobreamostragem, subamostragem e ponderação por classe).

***Transfer learning* com CNNs.** Arquiteturas pré-treinadas no ImageNet e ajustadas à tarefa, das famílias MobileNet (v2, v3-small, v3-large) e EfficientNet (b0, b1), escolhidas por serem referências de classificação em dispositivos móveis, todas com menos de oito milhões de parâmetros, sobre o recorte da ROI. Avaliaram-se balanceamento

e intensidades de *augmentation*.

Classificação YOLO. Modelos de classificação da família YOLO [Redmon et al. 2016, Jocher and Qiu 2024] (variantes *nano*, *small* e *medium* das versões 8, 11 e 26), que aprendem a representação a partir dos *pixels* do recorte, explorando cor, textura e organização espacial. A família entra por suas variantes *nano* serem o padrão de fato em borda e por ser a mesma do segmentador, o que permitiria um único *runtime*. Avaliaram-se balanceamento e *augmentation*.

Nas três famílias os hiperparâmetros foram definidos sem acesso ao teste de cada *fold*: busca em grade no ML clássico e, nas redes, uma etapa prévia de *tuning* reutilizada em todos os *folds*, com 100 épocas de treino. O modelo YOLO vencedor foi ainda retreinado sob a mesma configuração e partição recebendo a imagem inteira (Seção 4.2).

3.5. Regressão LASSO para interpretabilidade

Como análise complementar, modelou-se o escore FAMACHA© como variável contínua (escala ordinal de 1 a 5) e treinou-se uma regressão linear com regularização L_1 (LASSO). A troca de variável alvo é deliberada: a tarefa do restante do trabalho é binária porque a decisão de manejo é binária, enquanto esta análise não decide, e sim descreve como a cor varia ao longo da escala clínica, o que um alvo binário achataria em dois patamares. Os dois modelos não competem: o classificador diz se tratar, e a regressão explicita o sinal de cor por trás disso; nenhum resultado de classificação da Seção 4 provém do LASSO. O modelo estima os coeficientes β minimizando o erro quadrático penalizado:

$$\hat{\beta} = \arg\min_{\beta} \; \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - x_i^{\top}\beta\right)^2 + \alpha \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

em que y_i é o escore, x_i o vetor de *features* de cor e α o parâmetro de regularização. Diferentemente da penalização L_2, o termo L_1 zera os coeficientes irrelevantes, o que importa aqui porque os canais de cor são correlacionados e a seleção embutida devolve um subconjunto parcimonioso. As 18 *features* (média e desvio-padrão dos canais em RGB, HSV e Lab) foram padronizadas, de modo que os coeficientes sejam comparáveis em magnitude. O α e o desempenho foram estimados por validação cruzada aninhada de 5 *folds*, com o laço interno selecionando α apenas no treino.

3.6. Protocolo de avaliação

Para comparabilidade, todas as estratégias foram avaliadas sob validação cruzada estratificada de 5 *folds* (*StratifiedKFold*), com semente fixa e o mesmo particionamento, o que preserva a proporção das classes e garante treino e teste sobre os mesmos subconjuntos. Reporta-se a média e o desvio-padrão das métricas sobre as cinco iterações. O balanceamento incidiu apenas sobre o treino de cada *fold*.

A validação cruzada aninhada da Seção 3.5 aplica-se apenas ao LASSO, cujo α é selecionado dentro de cada partição de treino. Nas três estratégias de classificação, os hiperparâmetros foram fixados antes da validação cruzada (Seção 3.4), de modo que existe um único laço de 5 *folds* e nenhuma decisão de modelo foi tomada sobre as partições de teste.

A grade completa compreende 605 execuções na tarefa binária, cada uma avaliada nos mesmos 5 *folds*, o que totaliza 3.025 treinos: 432 no ML clássico (9 classificadores $\times$ 6 espaços de cor $\times$ 4 balanceamentos $\times$ 2 variantes de ROI), 65 no *transfer learning* (5 arquiteturas $\times$ 13 combinações) e 108 no YOLO (9 modelos $\times$ 12 combinações). Tudo rodou em uma estação com GPU de 16 GB, 12 *threads* de CPU e 62 GB de RAM; os cinco *folds* de uma configuração levam de 16 s, no ML de cor, a 886 s na melhor CNN.

O desempenho foi quantificado por acurácia, precisão, *recall* e F1-*score* [Strauss et al. 2022], sendo o F1 o mais adequado a conjuntos desbalanceados, e pelo κ de Cohen, que corrige a concordância pelo acaso. A significância entre os classificadores clássicos foi avaliada pelo teste de Friedman, complementado pelo teste pareado de Wilcoxon [Zimmerman and Zumbo 1993] entre os modelos de topo. A comparação entre estratégias usou o teste exato de McNemar [McNemar 1947] sobre as predições fora de amostra: como o particionamento é idêntico, cada imagem tem uma predição por estratégia, e o teste mede as discordâncias par a par sem depender da média por *fold*.

4. Resultados

Os resultados seguem o *pipeline* apresentado na Figura 1: valida-se a segmentação, comparam-se as estratégias de classificação e explora-se a interpretabilidade do sinal de cor. Salvo indicação contrária, as métricas são média $\pm$ desvio-padrão sobre os 5 *folds*, com F1 *weighted* e κ de Cohen.

4.1. Segmentação da mucosa ocular

O segmentador yolo11n-seg localiza a mucosa de forma confiável na partição de teste fixa: mAP@0,5 = 0,945 e mAP@0,5:0,95 = 0,659 para a máscara, com precisão e *recall* de 0,918. Na ablação de um fator por vez, apenas `copy_paste` = 0,3 melhora o resultado base (0,670 contra 0,659), por agregar variedade de instâncias sem distorcer a cor; a entrada de 768 px (0,636) e a desativação do *mosaic* (0,634) pioram, e o modelo maior yolo11s-seg fica bem atrás (0,574), evidência de *overfitting* em base reduzida.

A segmentação cumpre dois papéis: garante que a região analisada é de fato a mucosa, e não pele, pelo ou fundo, e define a ROI que alimenta as três estratégias. Uma ressalva antecipa as seções seguintes: a máscara estrita não melhora a classificação frente ao recorte por caixa (Seção 4.3), o que indica que o contorno pixel a pixel não é pré-requisito para classificar, ainda que prescindir da delimitação custe 3,4 pontos de F1 (Seção 4.2).

4.2. Comparação entre estratégias de classificação

A Tabela 2 compara as abordagens na tarefa binária, todas sob o mesmo protocolo de 5 *folds* e sobre a mesma ROI. O classificador YOLO (yolo11n-cls) obtém o melhor desempenho (F1 = 0,840; κ = 0,635), superando o *transfer learning* (MobileNet-V3-Small, F1 = 0,808; κ = 0,566) e o ML de cor (RF, F1 = 0,799; κ = 0,541). A diferença não vem da variação entre *folds*: no teste exato de McNemar sobre as 843 predições fora de amostra, o YOLO supera o ML de cor ($p = 0{,}009$; erra sozinho em 59 imagens contra 92) e o *transfer learning* ($p = 0{,}014$; 47 contra 75), enquanto estes dois não diferem entre si ($p = 0{,}75$). A ordem da tabela é, portanto, estatisticamente sustentada apenas quanto à liderança do YOLO.

Pelo κ, o YOLO é o único na faixa *substancial* (0,635), contra a *moderada* do *transfer learning* (0,566) e do ML de cor (0,541). Como as três recebem a mesma ROI e

Tabela 2. Comparação consolidada entre as estratégias na tarefa binária, sob protocolo único de 5 *folds*, e a ablação sem ROI. Acurácia e F1 em % e κ de Cohen, todos como média $\pm$ desvio sobre os *folds*; o *recall* das classes *Vermifuga* (V) e *Não Vermifuga* (NV) é calculado sobre as 843 predições fora de amostra. Sob o desbalanceamento de 2:1 a acurácia é otimista, por isso a discussão prioriza F1 e κ.

Abordagem	Configuração	Acur.	F1	κ	Rec. V	Rec. NV
YOLO (ROI)	yolo11n-cls	**84,2 ± 1,9**	**84,0 ± 2,0**	**0,635 ± 0,046**	**0,906**	**0,713**
Transfer learn.	MobileNet-V3-Small	80,9 ± 0,7	80,8 ± 0,9	0,566 ± 0,026	0,862	0,703
ML de cor	RF / Lab / *crop*	80,3 ± 1,8	79,9 ± 2,4	0,541 ± 0,063	0,878	0,652
YOLO (imagem inteira)	yolo11n-cls, ablação	80,9 ± 2,9	80,5 ± 3,0	0,556 ± 0,073	0,883	0,659

a mesma partição, o ganho é atribuível à representação: as estatísticas agregadas de cor descartam a organização espacial, os filtros pré-treinados vêm de um domínio genérico ajustado a uma base pequena, e o YOLO aprende a representação para a tarefa a partir dos *pixels* do recorte. Modelos maiores não ajudam nesta base: o melhor F1 é 0,840 no *nano*, 0,839 no *medium* e 0,834 no *small*.

Para verificar se essa vantagem viria de operar sobre a imagem inteira, e não do classificador, o modelo vencedor foi retreinado com a imagem completa no lugar do recorte, sob a mesma configuração, semente e *folds*: o desempenho cai a F1 = 0,805 $\pm$ 0,030 e κ = 0,556, perda de 3,4 pontos confirmada por McNemar ($p = 0{,}037$). A delimitação da mucosa não é, portanto, dispensável, e sem ela o YOLO cai ao patamar do *transfer learning* com ROI (0,805 contra 0,808): recorte e capacidade de representação são vias parcialmente substituíveis para o mesmo acerto.

Por escore original, os erros concentram-se na fronteira da decisão. No YOLO com ROI, a taxa de erro é de 11,6% no escore 1, 36,3% no 2, 16,0% no 3, 3,2% no 4 e 0,0% no 5, e o perfil se repete nas outras três configurações (erro máximo de 44,6% no escore 2, no ML de cor, e de 4,8% no escore 4, no *transfer learning*). Em nenhuma delas um animal de escore 5 deixou de ser indicado para tratamento: o custo clínico recai sobre os casos menos graves, não sobre os severamente anêmicos. Isso também explica o *recall* mais baixo da classe *Não Vermifuga* (0,652 a 0,713), pois cerca de metade dos erros está no escore 2, logo abaixo do limiar, onde a diferença de coloração é a menor da escala.

Os erros são apenas parcialmente coincidentes: das 266 imagens erradas por alguma estratégia, só 53 o são pelas três, com Jaccard par a par entre 0,33 e 0,41 e erros exclusivos de 26 (YOLO), 41 (*transfer learning*) e 58 (ML de cor). Sob ROI e partição idênticas, as representações capturam aspectos distintos do sinal, o que abre espaço para combinação por comitê.

O custo separa as estratégias mais do que o desempenho. Na mesma estação, em CPU, os classificadores custam por imagem 0,19 ms (Random Forest sobre cor), 3,0 ms (MobileNet-V3-Small, 2,54 M de parâmetros) e 3,8 ms (yolo11n-cls, 2,81 M), enquanto a localização da mucosa, comum às três, custa 35,3 ms, mais de 90% do orçamento. Otimizar o classificador pouco altera o tempo total: é a delimitação que decide a viabilidade em campo, e a ablação precifica sua remoção em 3,4 pontos de F1.

4.3. Análise das abordagens baseadas em cor

O *Random Forest* no espaço Lab atinge F1 = 0,799, 1 ponto abaixo do melhor *backbone* CNN e 4 abaixo do YOLO. O intervalo estreito indica que a cor, isoladamente, já contém grande parte do sinal que discrimina os grupos; o mérito da rota está na interpretabilidade aliada ao baixo custo (Seção 3.5), o que as CNNs não oferecem.

O espaço de cor influencia o resultado. Os três que separam luminância de crominância lideram, Lab (RF, 0,799), LUV (MLP, 0,795) e YCrCb (RF, 0,793), contra HSV (MLP, 0,771), RGB (MLP, 0,757) e HLS (SVM, 0,754), vantagem de cerca de 4 pontos que sustenta a hipótese da Seção 3.3: com a iluminação em canal próprio, os canais cromáticos descrevem a palidez de forma mais estável. Já a escolha do classificador foi avaliada no espaço Lab sobre o recorte, sem reamostragem, pelo teste de Friedman ($\chi^2_F = 23{,}20$; $p = 0{,}003$; 8 graus de liberdade), que confirma diferença significativa entre os nove modelos: os *ensembles* RF e GB lideram (rank médio 2,8), seguidos do MLP (3,4), com a árvore isolada e o *Perceptron* em último (8,2 e 7,6), sinal de que a fronteira de decisão não é linear. O Wilcoxon pareado não separa os modelos de topo entre si ($p \geq 0{,}63$) e, com apenas cinco pares, não alcança significância contra os últimos colocados, ainda que RF e GB superem a árvore isolada nos cinco *folds* ($p = 0{,}0625$, o menor valor possível nesse desenho).

Por fim, duas escolhas de pré-processamento. Nenhuma reamostragem melhorou o F1 (0,791 com sobreamostragem e 0,774 com subamostragem, contra 0,795 sem tratamento), de modo que o desbalanceamento de 2:1 dispensa tratamento agressivo. E o recorte por caixa (F1 = 0,799) superou levemente a máscara estrita (0,776), pois a borda do recorte carrega informação de cor útil que a máscara descarta. Dispensável é o contorno pixel a pixel, não a etapa de delimitação: na prática, uma aplicação móvel pode trocar a segmentação por uma detecção de caixa, mais simples e barata, mas não pode entregar a imagem inteira ao classificador sem perder acerto.

4.4. Interpretabilidade do sinal de cor via regressão LASSO

Para explicitar *como* a cor da mucosa se relaciona com o grau de anemia, ajustou-se uma regressão LASSO que estima o escore a partir das 18 *features* de cor. Em dados não usados no treino, o modelo erra em média meia categoria (MAE $= 0{,}554 \pm 0{,}029$) e explica cerca de 62% da variação ($R^2 = 0{,}625 \pm 0{,}010$), praticamente o mesmo ajuste obtido *in-sample* ($R^2 = 0{,}655$), o que indica boa generalização e qualifica a equação como indicador auxiliar, não como classificador isolado.

Mais importante que o desempenho é a leitura dos pesos. Dos 18 descritores o modelo manteve 14, e os cinco de maior magnitude são `mean_G` ($-1{,}108$), `std_L` ($+1{,}061$), `mean_V` ($+0{,}931$), `std_a` ($-0{,}867$) e `std_G` ($-0{,}552$). O escore estimado cresce com o brilho e com a dispersão de luminância, e cai com o nível de verde e com a dispersão cromática: em conjunto, esses termos descrevem a transição de uma mucosa vermelha e homogênea para uma mucosa clara, que é o que o cartão FAMACHA© compara. Como os canais de cor são correlacionados, cada coeficiente é um efeito parcial, condicionado aos demais, e não deve ser lido isoladamente. Para o profissional em campo, a equação rende duas utilidades que a saída binária descarta e que as redes convolucionais não oferecem: a estimativa contínua, pela qual um escore previsto de 2,7 aponta um animal a reavaliar antes de outro de 1,3 mesmo dispensando tratamento agora; e a auditabilidade, pois é

possível mostrar quais descritores sustentaram a indicação e confrontá-los com o que o avaliador vê no cartão. Nenhuma substitui a avaliação clínica; ambas a instrumentam.

5. Conclusão

Este trabalho investigou a automação do método FAMACHA© para a seleção de caprinos e ovinos anêmicos, comparando ML clássico sobre *features* de cor, *transfer learning* com CNNs e classificação por modelos YOLO sob um protocolo único de validação cruzada, com particionamento e ROI idênticos, em uma base própria de 843 imagens anotadas por especialista.

O classificador YOLO (yolo11n-cls) foi a abordagem mais adequada, com o melhor F1 (0,840) e κ (0,635) e superioridade confirmada por McNemar sobre as duas alternativas. Sob ROI idêntica, isso indica que aprender a representação para a tarefa recupera informação espacial e de textura que os descritores de cor descartam e que filtros pré-treinados em domínio genérico não reconstroem em uma base pequena. O *transfer learning* (MobileNet-V3-Small, F1 = 0,808) ficou em segundo, e o ML de cor com Random Forest, embora inferior (F1 = 0,799), mostrou-se competitivo, com inferência cerca de vinte vezes mais barata, e é a única rota interpretável: a equação do LASSO ($R^2 = 0{,}625$) reproduz o critério de cor do método. A análise de erros localizou as falhas na fronteira da decisão, e a ablação com a imagem inteira mostrou que a delimitação da ROI vale 3,4 pontos de F1 e concentra mais de 90% do custo de inferência.

Quatro limitações delimitam a leitura dos resultados. A principal está no *ground truth*: cada imagem recebeu o escore de um único avaliador, sem medida de concordância entre especialistas, de modo que o teto da tarefa é desconhecido e não se pode dizer se $\kappa = 0{,}635$ está próximo dele. A base, de tamanho moderado e restrita a duas localidades e poucas raças, limita a generalização para outros dispositivos, raças, espécies e contextos, ainda que a coleta em campo dê validade prática aos resultados; base e código não são disponibilizados. A comparação isola a ROI e a partição, mas não a configuração de treino, pois cada família manteve seu espaço de *augmentation* e de hiperparâmetros, e só uma delas foi avaliada sobre a imagem inteira. Por fim, o ganho prático não foi medido: se o gargalo for conter o animal e obter uma foto utilizável, e não classificar, um sistema acurado pode não reduzir o tempo de decisão. Como continuidade, pretende-se medir a concordância entre avaliadores, estender a classificação aos cinco escores, ampliar a base, combinar as estratégias por comitê, dado que seus erros são só parcialmente coincidentes, investigar outras formas de captura, como a face completa ou a região pupilar, e avaliar a implantação em dispositivos móveis para uso *offline* no local de manejo.

Referências

Almeida, A. M. A. (2021). Detecção de anemia em ovinos através de aprendizagem profunda em imagens de mucosa ocular. Dissertação (mestrado em engenharia elétrica e de computação), Universidade Federal do Ceará, Sobral.

Arshad, M. F., Burrai, G. P., Varcasia, A., Sini, M. F., Ahmed, F., Lai, G., Polinas, M., Antuofermo, E., Tamponi, C., Cocco, R., et al. (2024). The groundbreaking impact of digitalization and artificial intelligence in sheep farming. *Research in Veterinary Science*, page 105197.

de Oliveira Diniz, V. A., Teixeira, L. F., de Almeida, G. D., Meirelles, A. C. F., Junior, E. G., and Figueira, P. T. (2022). Utilização do método famacha como auxílio no controle de verminose em ovinos. *Pubvet*, 16:195.

de Souza, L. F., Costa, M. H., and Riet-Correa, B. (2023). Mobile app for targeted selective treatment of haemonchosis in sheep. *Veterinary Parasitology*, 316:109902.

de Souza, L. F., Mior, L. B., Costa, M. H., and Riet-Correa, B. (2021). Sistema para classificação de infestação parasitária em pequenos ruminantes. *Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, pages 1–5.

Demoliner, G. and Alves, R. J. F. (2017). Anemímetro: app móvel para implementação do método famacha. *Unoesc & Ciência - ACET*, 8(1):25–32.

Freitas, L. A., da Rocha, N. C., Barbosa, A. M., Dorea, J. R., Paz, C. C., and Rosa, G. J. (2024). Sheepeye: a based-web app for real-time diagnosis of sheep anemia. *Translational Animal Science*, 8:txae144.

Freitas, L. A., Ferreira, R. E., Savegnago, R. P., Dórea, J. R., Stafuzza, N. B., Rosa, G. J., and Paz, C. C. (2023). Image analysis to automatically classify anemia based on famacha score in sheep using ocular conjunctiva images. *Translational Animal Science*, 7(1):txad118.

IBGE (2022a). Rebanho de caprinos (bodes e cabras).

IBGE (2022b). Rebanho de ovinos (ovelhas e carneiros).

Jocher, G. and Qiu, J. (2024). Ultralytics yolo11.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Molento, M. B., Tasca, C., Gallo, A., Ferreira, M., Bononi, R., and Stecca, E. (2004). Método famacha como parâmetro clínico individual de infecção por haemonchus contortus em pequenos ruminantes. *Ciência Rural*, 34(4):1139–1145.

Nascimento, M. I. d. S. S. (2022). *Produção, comercialização e consumo dos produtos e subprodutos caprinos e ovinos no Nordeste do Brasil*. PhD thesis, Universidade Federal Rural de Pernambuco.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Ribeiro, W. F. (2018). Uma metodologia para detectar anemia em pequenos ruminantes. Trabalho de conclusão de curso (graduação), Universidade Federal do Maranhão.

Senoamadi, M., Louis, T., and Chitura, T. (2022). Application of the famacha system for the evaluation of haemonchosis in small ruminants reared in a communal system of the molemole municipality, limpopo province, south africa. *Adv Anim Vet Sci*, 10(5):1101–1109.

Silva, J. V. d. A. (2025). Aplicativo para detecção de anemia em ovinos. Trabalho de conclusão de curso (graduação), Universidade Federal do Ceará.

Souza Filho, J. L. d. (2022). Desenvolvimento de api para aprendizado profundo de imagens de pequenos ruminantes. Trabalho de conclusão de curso (graduação), Universidade Federal do Ceará.

Strauss, E., Júnior, M. V. B., and Ferreira, W. L. L. (2022). A importância de utilizar métricas adequadas de avaliação de performance em modelos preditivos de machine learning. *Projectus*, 7(2):52–62.

Teixeira, M., Matos, A. F. I. M. d., Oliveira, T. N. d., and Souza, C. M. (2023). Cartilha controle de verminoses em caprinos e ovinos: recomendações para o controle integrado de verminose em caprinos e ovinos no sertão do são francisco. *PE*.

Zimmerman, D. W. and Zumbo, B. D. (1993). Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks. *The Journal of Experimental Education*, 62(1):75–86.