PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ
PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO

PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA – PIBIC

RELATÓRIO FINAL

IMPLEMENTAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO ULTRASSONOGRÁFICO EM CÃES

CURITIBA
2025

BRUNO BETIATTO ALVES
ANDRÉ GUSTAVO HOCHULI
CIÊNCIA DA COMPUTAÇÃO - EP
MODALIDADE FUND. ARAUCÁRIA – INCLUSÃO SOCIAL

IMPLEMENTAÇÃO DE INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO ULTRASSONOGRÁFICO EM CÃES

Relatório Final apresentado ao Programa Institucional de Bolsas de Iniciação Científica e Tecnológica, Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação da Pontifícia Universidade Católica do Paraná.

Orientador: Prof. Dr. André Gustavo Hochuli

CURITIBA
2025

RESUMO

O uso de inteligência artificial aplicada ao diagnóstico por imagem tem ganhado destaque na medicina veterinária, especialmente em cenários onde o acesso a profissionais especialistas é limitado. Neste contexto, este trabalho buscou desenvolver e avaliar um modelo de aprendizado profundo capaz de classificar imagens ultrassonográficas de rins caninos como normais ou alterados, com foco em explorar os desafios impostos pela escassez de dados, desbalanceamento de classes e variação de sondas utilizadas na aquisição das imagens. O principal objetivo foi investigar a eficácia de um modelo leve, baseado na arquitetura YOLOv11-nano, na detecção de alterações renais, avaliando diferentes estratégias de aumento de dados (data augmentation) e diferentes conjuntos de imagens, incluindo combinações de imagens obtidas com sondas convexas e lineares. Para isso, foram utilizados conjuntos de dados reais obtidos em ambiente clínico, segmentados por tipo de sonda e enriquecidos com técnicas de aumento sintético. As imagens foram organizadas em diferentes cenários experimentais, e o desempenho do modelo foi avaliado por meio de validação cruzada estratificada (k-fold), utilizando métricas como acurácia, precisão, recall, F1-score e mAP. Os resultados mostraram que, apesar do desbalanceamento de classes e da baixa quantidade de imagens lineares, o modelo apresentou desempenho satisfatório nos cenários com aumento de dados, especialmente no conjunto "Linear Ampliado", que obteve acurácia média de 86 %, superando os conjuntos com imagens convexas originais e ampliadas. Esse resultado indica que o modelo consegue se adaptar bem às imagens lineares quando exposto a variações sintéticas, demonstrando o potencial de aplicação clínica mesmo com recursos limitados. Por outro lado, a performance inferior nos cenários sem aumento de dados evidencia a importância da ampliação e balanceamento da base de dados. Conclui-se que o uso de modelos de deep learning é viável na detecção de alterações renais em cães por meio de imagens de ultrassom, desde que se adotem estratégias adequadas de pré-processamento e validação. Como recomendações para trabalhos futuros, destaca-se a necessidade de expandir a base de dados, explorar arquiteturas mais complexas, aplicar técnicas avançadas de oversampling e realizar validações externas em ambientes clínicos reais, além de investir em ferramentas de explicabilidade e interfaces amigáveis para aplicação prática por profissionais da área.

Palavras-chave: 1. Inteligência artificial; 2. Aprendizado profundo; 3. Ultrassonografia veterinária; 4. YOLO; 5 Rim canino.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

AI - Inteligência Artificial
CNN - Rede Neural Convolucional (Convolutional Neural Network)
IA - Inteligência Artificial
YOLO - You Only Look Once (Arquitetura de detecção de objetos em tempo real)
YOLOv11 - Versão 11 da arquitetura YOLO
YOLOv11-nano - Variante leve da arquitetura YOLOv11
mAP - Mean Average Precision
MAP@50 - Mean Average Precision com limiar de IoU ≥ 0,50
MAP@50-95 - Mean Average Precision com IoU variando de 0,50 a 0,95
F1-score - Média harmônica entre precisão e recall
K-fold - Validação cruzada com k partições
TP - True Positives (Verdadeiros Positivos)
FP - False Positives (Falsos Positivos)
FN - False Negatives (Falsos Negativos)
GPU - Graphics Processing Unit (Placa Gráfica)
GAN - Generative Adversarial Network
cGAN - Conditional Generative Adversarial Network
Grad-CAM - Gradient-weighted Class Activation Mapping
LIME - Local Interpretable Model-agnostic Explanations
IoU - Intersection over Union (Métrica de sobreposição entre caixas preditas e reais)
DICOM - Digital Imaging and Communications in Medicine
JPEG - Joint Photographic Experts Group
CEUA - Comitê de Ética no Uso de Animais
ABNT - Associação Brasileira de Normas Técnicas
PIBIC - Programa Institucional de Bolsas de Iniciação Científica
ROBOFLOW - Plataforma de anotação e treinamento em visão computacional

LISTA DE SÍMBOLOS

@ - arroba

IoU - Intersection over Union (métrica de sobreposição entre caixas preditas e reais)

SUMÁRIO

1 INTRODUÇÃO

A crescente evolução das técnicas de inteligência artificial (IA) tem transformado significativamente a análise de imagens médicas, oferecendo novas perspectivas para o diagnóstico em diferentes especialidades, inclusive na medicina veterinária. Nesse contexto, tecnologias como o deep learning têm se destacado por sua capacidade de identificar padrões complexos em dados visuais com alto grau de precisão, contribuindo para diagnósticos mais rápidos e confiáveis. Uma das áreas que pode se beneficiar enormemente dessas tecnologias é a ultrassonografia veterinária, especialmente no exame dos rins de cães, cuja interpretação ainda depende fortemente da experiência do operador e da qualidade da imagem capturada.

Apesar do potencial revolucionário da IA, sua aplicação na prática clínica veterinária enfrenta desafios importantes. Um dos principais obstáculos é a escassez de dados disponíveis para o treinamento de modelos robustos. Ao contrário da medicina humana, onde há maior disponibilidade de bancos de dados estruturados, a área veterinária ainda sofre com a fragmentação das informações e com a baixa quantidade de exames armazenados digitalmente com qualidade suficiente para análise automatizada. A maioria dos modelos de deep learning propostos na literatura depende de grandes volumes de imagens rotuladas para alcançar alto desempenho, o que não condiz com a realidade enfrentada por muitas clínicas e centros de pesquisa veterinária.

Essa limitação levanta uma questão central que orienta o presente estudo: seria possível desenvolver uma ferramenta de inteligência artificial eficaz para classificar imagens ultrassonográficas de rins caninos, mesmo com uma base de dados limitada? Essa problemática revela uma lacuna importante na literatura científica e propõe uma investigação relevante para ampliar a aplicabilidade de soluções baseadas em IA em contextos com recursos restritos. Diante disso, torna-se fundamental explorar estratégias que possibilitem o uso eficiente de conjuntos de dados pequenos, tais como técnicas de aumento de dados (data augmentation), transferência de aprendizado (transfer learning) e validação cruzada robusta.

Nesse sentido, a revisão sistemática de Soni e Rai (2024) destaca o desempenho promissor do algoritmo YOLO na detecção de objetos médicos em diferentes domínios, reforçando seu potencial também para aplicações veterinárias. No entanto, os autores apontam limitações importantes, como a exigência por grandes

volumes de dados e recursos computacionais elevados, o que reforça a relevância de pesquisas voltadas a contextos com menor disponibilidade de dados estruturados.

Outro fator crítico para o desenvolvimento de soluções de IA confiáveis é a qualidade da anotação dos dados. A rotulagem manual, que envolve a delimitação precisa das regiões de interesse nas imagens e a identificação de alterações patológicas, é um processo trabalhoso e suscetível à variabilidade Inter observador. No entanto, é esse processo que confere valor ao conjunto de dados, tornando possível o treinamento de modelos preditivos capazes de auxiliar o diagnóstico clínico.

Diante desse cenário, este trabalho propõe o desenvolvimento de uma solução baseada em IA para a análise de imagens ultrassonográficas de rins caninos, focando em superar os desafios impostos pela limitação de dados disponíveis. A hipótese central é que, mesmo com um número reduzido de imagens, é possível construir um modelo assertivo com desempenho clínico relevante, desde que aplicadas técnicas adequadas de processamento e modelagem. Tal avanço não só contribuiria para a prática veterinária como também abriria caminho para futuras aplicações em outras áreas da medicina veterinária que enfrentam desafios semelhantes.

A distinção entre rins caninos normais e alterados em imagens ultrassonográficas representa um desafio significativo, mesmo para profissionais experientes. Isso ocorre porque muitas vezes as alterações patológicas se manifestam de maneira sutil e sobreposta às variações anatômicas individuais. As Figuras 1 e 2 ilustram esse cenário:

FIGURA 1 – Rim canino normal

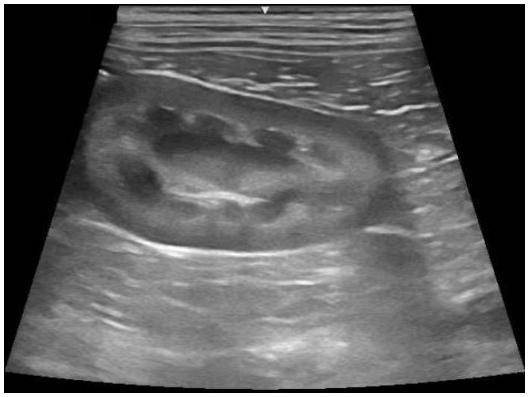
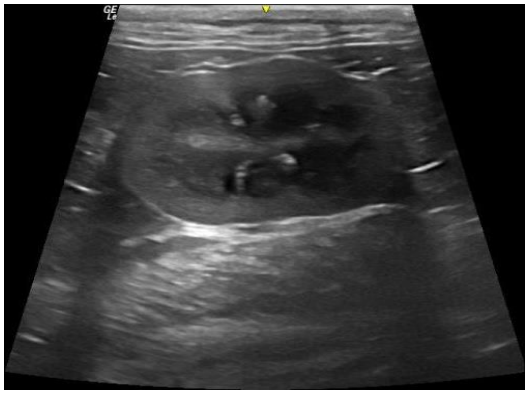


FIGURA 2 – Rim canino alterado



Embora uma represente um rim com padrão considerado normal e a outra exiba alterações morfológicas compatíveis com anormalidade, visualmente ambas apresentam características muito semelhantes, como ecogenicidade difusa e presença de estruturas internas pouco definidas. Essa similaridade pode comprometer a acurácia diagnóstica e reforça a necessidade de ferramentas automatizadas baseadas em inteligência artificial que auxiliem na detecção de padrões com maior sensibilidade e consistência, especialmente em contextos de baixa experiência clínica ou qualidade de imagem variável.

2 OBJETIVO

Desenvolver uma solução de inteligência artificial baseada em técnicas de *deep learning* capaz de identificar e classificar imagens ultrassonográficas de rins caninos e, potencialmente, felinos com precisão clínica satisfatória. O intuito é criar uma ferramenta computacional que auxilie profissionais da medicina veterinária no processo de diagnóstico por imagem, contribuindo para a tomada de decisões clínicas mais rápidas, seguras e padronizadas. A pesquisa também busca avaliar a viabilidade e a eficácia do uso de modelos de aprendizado profundo mesmo em cenários com bases de dados limitadas, promovendo a integração de tecnologias emergentes ao contexto veterinário.

2.1 OBJETIVOS ESPECÍFICOS

- Construir uma base de dados própria, composta por imagens ultrassonográficas de rins de cães e gatos, coletadas de maneira padronizada, com anotações precisas e metadados relevantes, assegurando qualidade e consistência para o treinamento dos modelos;
- Investigar e aplicar técnicas de aprendizado profundo (como transfer learning e data augmentation) adequadas à realidade de bases de dados reduzidas, visando obter desempenho aceitável mesmo com número limitado de amostras;
- Avaliar o desempenho dos modelos desenvolvidos por meio de métricas quantitativas (acurácia, mAP, recall e F1-score).
- Fornecer subsídios técnicos e científicos que possam apoiar futuras pesquisas no desenvolvimento de sistemas de apoio à decisão baseados em IA no campo da medicina veterinária.

3 MATERIAIS E MÉTODO

Este estudo empregou o modelo **YOLOv11** para detecção e segmentação de estruturas renais em imagens ultrassonográficas de rins caninos e felinos. Foram utilizados seis conjuntos de dados distintos, totalizando **680 imagens diferentes (Podendo repetir dependendo do conjunto)**, organizadas conforme descrito na Tabela 1:

TABELA 1 – Imagens Utilizadas

Conjunto	Total de imagens	Partição (Treino/Validação/Teste)
Convexo	566	338/118/113
Convexo Ampliado	1113	800/200/113
Linear	114	67/25/22
Linear Ampliado	522	400/100/22
Completo	680	428/117/135
Completo Ampliado	1835	1300/400/135

A Tabela 2 apresenta a distribuição total de imagens de acordo com o tipo de transdutor utilizado para cada conjunto na aquisição dos exames, sendo eles convexo e linear:

TABELA 2 – Distribuição entre Transdutores

Conjunto	Total de imagens	Partição (Convexo/Linear)
Convexo	566	566/0
Convexo Ampliado	1113	1113/0
Linear	114	0/114
Linear Ampliado	522	0/522
Completo	680	566/114
Completo Ampliado	1835	914/921

A tabela 3 apresenta a distribuição de classes para cada conjunto, sendo eles Normal e Alterado:

TABELA 3 – Distribuição entre as Classes

Conjunto	Total de imagens	Partição (Normal/Alterado)
Convexo	566	263/304
Convexo Ampliado	1113	560/553
Linear	114	32/82
Linear Ampliado	522	257/265
Completo	680	298/382
Completo Ampliado	1835	914/921

A tabela 4 apresenta a distribuição entre espécies em cada conjunto, sendo eles canino e felino:

TABELA 4 – Distribuição entre espécies

Conjunto	Total de imagens	Partição (Canino/Felino)
Convexo	566	129/454
Convexo Ampliado	1113	560/553
Linear	114	36/78
Linear Ampliado	522	257/265
Completo	680	147/497
Completo Ampliado	1835	914/921

3.1 ANOTAÇÃO E PRÉ-PROCESSAMENTO

Antes do treinamento, todas as imagens passaram por um protocolo de pré-processamento, que consistiu nas seguintes etapas: (i) conversão dos arquivos em formato DICOM para JPEG, utilizando o programa MicroDICOM; (ii) padronização da base, onde cada exame foi nomeado na planilha como “ORIGEM.ID do paciente” e detalhado com informações como frequência, ganho, distância e o transdutor utilizado;

(iii) censura dos dados sensíveis, removendo toda informação que possa identificar o paciente; e (iv) delimitação do parênquima renal utilizando o site ROBOFLOW. Os critérios de inclusão consideraram exames realizados em aparelhos GE Logic v2, Samsung e Sonoscape S2, com corte sagital, em espécies caninas e felinas, datados a partir de 2024, captados com transdutores linear, convexo e microconvexo. Devido à natureza dos arquivos DICOM e à necessidade de sigilo, algumas imagens podem apresentar recortes ou marcações internas. Além disso, o uso das imagens foi aprovado pelo Comitê de Ética para Uso de Animais em Pesquisa (CEUA). Um exemplo de diferença entre imagens **convexas** e **lineares** pode ser observado nas Figuras 2 e 3:

FIGURA 3 – Rim canino convexo

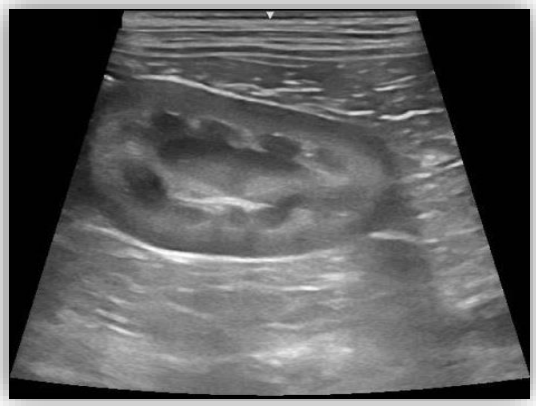
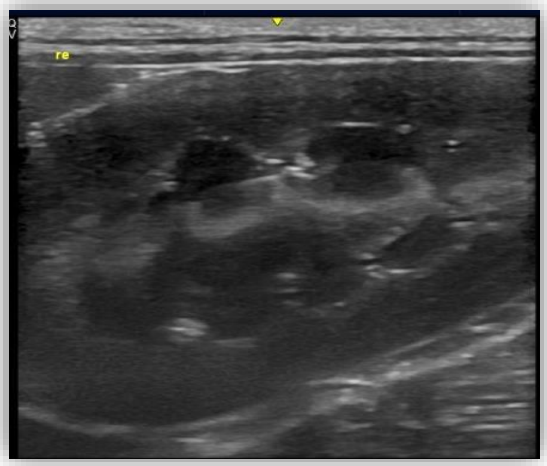


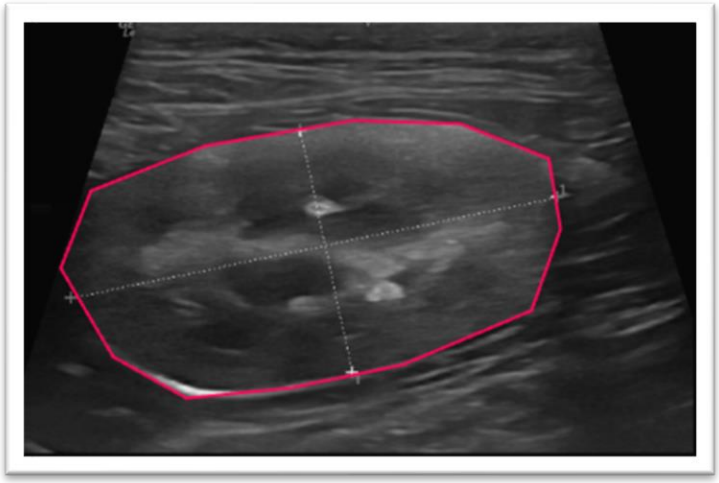
FIGURA 4 – Rim canino linear



As imagens foram previamente anotadas por especialistas, utilizando o formato de bounding boxes no padrão YOLO (arquivos .txt correspondentes a cada imagem). Para fins de classificação, foram definidas duas classes: Normais e

Alteradas, cujas etiquetas foram atribuídas de forma independente por duplas de biólogos especialistas. Um exemplo de imagem anotada pode ser visto na figura 3:

FIGURA 5 – Exemplo de rim anotado



Durante o particionamento para validação cruzada (k-fold), cada imagem passou por uma etapa adicional de crop automático do parênquima renal, utilizando uma máscara binária de fundo preto para isolar a região de interesse.

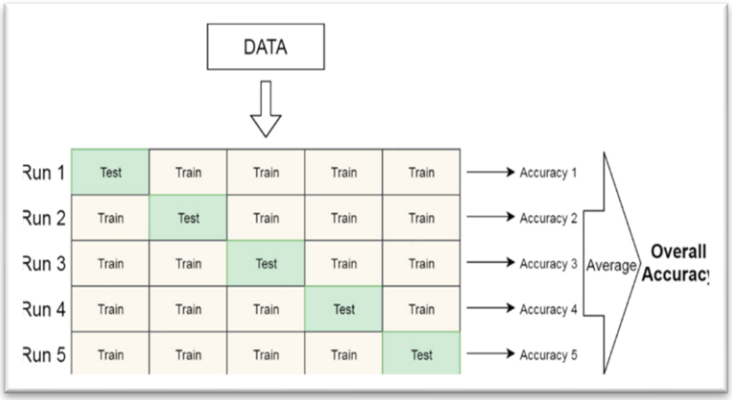
Além disso, o aumento de dados foi realizado de forma “limpa”, ou seja, cada imagem foi copiada no máximo três vezes, mantendo a distribuição uniforme entre os exemplos e sem aplicação de transformações visuais artificiais. Os efeitos de variação visual foram deixados exclusivamente para o momento do treinamento, utilizando o recurso RandAugment já incorporado no pipeline do YOLOv11.

3.1.1 Estratégia de Particionamento e Validação K-fold

Adotou-se a validação cruzada estratificada K-fold com K = 5 para avaliar de modo mais robusto o desempenho do modelo. Nesse procedimento, todo o conjunto de imagens é inicialmente dividido em cinco partes iguais (folds); em cada uma das cinco iterações, um fold diferente é reservado para teste, enquanto os quatro folds restantes são usados para treinamento e validação, de forma que, a cada rodada, aproximadamente 60 % dos dados servem para treinar o modelo, 20 % ficam para validar ajustes de hiperparâmetros e os 20 % finais compõem o conjunto de teste. Para evitar que o modelo se beneficie de características peculiares de um mesmo animal, todas as imagens de um mesmo indivíduo são mantidas juntas em apenas

uma das partições (treino, validação ou teste), garantindo que nenhum animal apareça simultaneamente nos conjuntos de treino e de teste. Ao término de cada rodada calcula-se a métrica de interesse (por exemplo, acurácia), e a média das cinco métricas resultantes fornece uma estimativa mais estável e confiável do desempenho real do classificador em novos animais. A figura 6 abaixo ilustra esse fluxo de divisão e avaliação:

FIGURA 6 – aplicação de K-fold



3.2 TREINAMENTO

O treinamento foi realizado utilizando o modelo **YOLOv11-nano**, uma das variantes leves da família YOLO desenvolvida no framework **Ultralytics YOLO**. Enquanto o YOLOv11-nano é otimizado para cenários com recursos computacionais restritos e bases de dados reduzidas, existem outras versões mais robustas (YOLOv11-small, -medium, -large) que empregam backbones e necks mais profundos, oferecendo maior capacidade de extração de características às custas de maior consumo de memória e processamento. A arquitetura do YOLOv11 apresenta aprimoramentos importantes sobre a eficiência computacional, bastante necessário para um hardware mais modesto no qual foi disponível.

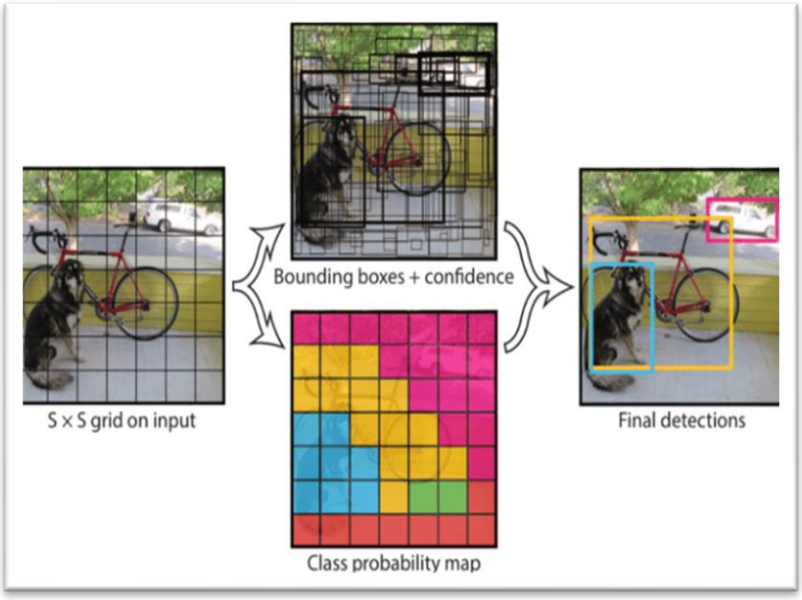
Segundo S Nikhileswara Rao (2024), autor de “YOLOv11 Explained: Next-Level Object Detection with Enhanced Speed and Accuracy”, essas inovações elevam o desempenho do modelo com melhor equilíbrio entre precisão e velocidade, especialmente em detecção de pequenos objetos e contextos complexos.

Importante destacar que, diferente de arquiteturas que finalizam com camadas densas (fully connected), o YOLOv11 não utiliza redes densas tradicionais,

operando exclusivamente com camadas convolucionais. Essa estrutura contribui para a alta velocidade de inferência e menor complexidade computacional do modelo.

Para facilitar a compreensão do funcionamento do YOLO (JOCHER, 2024), o pipeline do modelo pode ser resumido da seguinte forma: a imagem de entrada é dividida em uma grade, onde cada célula é responsável por prever caixas delimitadoras (bounding boxes) e as probabilidades das classes. Em uma única passagem, o modelo extrai as características da imagem, identifica os objetos (neste caso, as regiões dos rins) e classifica cada detecção. A Figura 6 ilustra de forma simplificada esse fluxo de detecção e classificação.

FIGURA 7 – Pipeline YOLO



Além disso, adotou-se o **RandAugment**, uma técnica de aumento de dados que automatiza e simplifica o processo de criar variações sintéticas nas imagens sem necessidade de buscar políticas complexas. Em vez de definir probabilidades e magnitudes individuais para cada transformação, o RandAugment trabalha com apenas dois parâmetros:

- N: número de operações de aumento a serem aplicadas sequencialmente em cada imagem;
- M: magnitude (intensidade) com que todas essas operações são executadas.

O método dispõe de um conjunto fixo de 14 possíveis transformações (como rotações, ajustes de brilho e contraste, posterização, cisalhamento e deslocamentos).

A cada iteração, ele seleciona aleatoriamente N dessas transformações e as aplica com intensidade M. Essa abordagem reduz drasticamente o espaço de busca em comparação a técnicas anteriores pois só precisamos ajustar N e M e, ainda assim, garante um nível elevado de diversidade visual no conjunto de treinamento. Com isso, o modelo exposto a uma maior variedade de versões de cada imagem tende a generalizar melhor, principalmente quando a base de dados é pequena ou desbalanceada.

No nosso caso, **optou-se por não usar efeitos que introduzem misturas de imagens** (como Mosaic), cortes regionais ou deformações intensas (como um alto), pois esses efeitos poderiam remover ou sobrepor partes críticas do parênquima renal, prejudicando a capacidade de classificação das imagens. Exemplos de transformações usadas podem ser vistos na imagem 9 e 10, em comparação a uma imagem sem alterações na imagem 7.

FIGURA 8 – Exemplo de imagem inalterada de rim

FIGURA 9 – Imagem com transformação 1

FIGURA 10 – Imagem com transformação 2

3.3 AVALIAÇÃO DO MODELO

Para cada modelo avaliado, foram realizados cinco testes independentes, correspondentes às cinco partições de teste definidas pelo procedimento de K-fold (K = 5). Em cada teste, foi utilizado um conjunto de imagens exclusivo, que não participou das etapas de treinamento nem de validação, sendo reservado apenas para a fase de teste. A avaliação do desempenho dos modelos foi realizada por meio das seguintes métricas: acurácia, recall, F1-score, mAP@50 e mAP@50–95.

3.3.1 Acurácia

A acurácia representa a proporção de previsões corretas sobre o total de amostras testadas

3.3.2 Precisão

A **precisão (precision)** é uma métrica que quantifica a exatidão das predições positivas do modelo, ou seja, de todas as vezes em que o modelo apontou “alterado”, quantas vezes ele realmente acertou. Matematicamente, ela é definida como:

$$\frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos(TP)\ +\ Falso\ Positivos(FP)}$$

- Verdadeiros Positivos (TP): número de imagens realmente alteradas que o modelo classificou como alteradas.
- Falsos Positivos (FP): número de imagens normais que o modelo classificou incorretamente como alteradas.

A precisão, portanto, indica a proporção de detecções positivas que são de fato corretas. Valores de precisão elevados (próximos de 1, mas aqui transformados em porcentagem) significam que há poucos falsos positivos, mas não dizem respeito aos falsos negativos para isso, utilizamos a sensibilidade (recall).

3.3.3 Recall

O **recall** (também chamado de sensibilidade) mede a capacidade do modelo de identificar corretamente todas as ocorrências da classe positiva, ou seja, de todas as imagens que são realmente “alteradas”, quantas foram efetivamente detectadas como tal. Formalmente, é definido por:

$$\frac{Verdadeiros\ Positivos\ (TP)}{Verdadeiros\ Positivos(TP)\ +\ Falso\ Negativos(FN)}$$

Onde Verdadeiros Positivos (TP) corresponde ao número de imagens alteradas corretamente classificadas como alteradas e Falsos Negativos (FN) representa o número de imagens alteradas classificadas incorretamente como normais.

O recall, portanto, indica a fração de casos positivos reais que o modelo consegue capturar; valores próximos de 1 significam que há poucos casos positivos perdidos, mas não refletem quantos falsos positivos podem existir (o que é avaliado pela precisão).

3.3.4 F1-score

O **F1-score** combina precisão e recall em uma única métrica para refletir o equilíbrio entre exatidão e completude das predições, especialmente útil em cenários de classes desbalanceadas. Ele é definido como a média harmônica de precisão (P) e recall (R):

$$F1 - score\ =\ 2\ \times\ \frac{P\ \times\ R}{P\ +\ R}$$

Valores altos de F1-score indicam que o modelo apresenta simultaneamente alta precisão (erra pouco ao classificar exemplos positivos) e alto recall (é capaz de encontrar a maior parte dos exemplos positivos). Por ser uma média harmônica, o F1 penaliza situações em que há desequilíbrio entre essas duas métricas — por exemplo, se a precisão for alta mas o recall for baixo (ou vice-versa), o F1-score resultante será significativamente reduzido. Assim, essa métrica é especialmente útil para avaliar o

desempenho geral de modelos em contextos onde existe desbalanceamento entre as classes ou quando os falsos negativos e falsos positivos têm custos relevantes.

3.3.5 mAP@50

O mAP@50 (mean average precision com IoU ≥ 0,5) calcula a média da precisão em diferentes pontos de recall considerando predições com sobreposição mínima de 50% com as anotações reais

3.3.6 mAP@50-95

O mAP@50-95 é a média da mAP obtida em múltiplos limiares de IoU variando de 0,5 a 0,95 em intervalos de 0,05, oferecendo uma visão detalhada do desempenho do modelo em diferentes níveis de exigência de sobreposição.

4 RESULTADOS

Os resultados obtidos para cada cenário experimental são apresentados de forma comparativa por meio das seis principais métricas de avaliação mAP@50, mAP@50–95, acurácia, precisão, recall e F1-score e exibidos em gráficos de barras com valores percentuais. Para cada cenário, o pipeline de validação seguiu o mesmo procedimento: durante a etapa de validação, definiu-se um limiar de confiança de 0,25 e um limiar de IoU de 0,6, de modo que apenas detecções com probabilidade de classe acima de 25% e que apresentassem sobreposição mínima de 60% fossem consideradas. Em seguida, para cada imagem de teste, selecionou-se a predição (bounding box) de maior confiança, que foi então comparada ao rótulo real para cálculo das métricas de classificação (acurácia, precisão, recall e F1-score).

- Convexo: Dataset original com imagens exclusivamente convexas
- Conv. Ampliado: Dataset convexo ampliado com técnicas de aumento de dados
- Linear: Dataset original com imagens exclusivamente lineares
- Linear Ampliado: Dataset linear ampliado com técnicas de aumento de dados

- Completo: Dataset com imagens convexas e lineares
- Comp. Ampliado: Dataset semi-completo ampliado com técnicas de aumento de dados

TABELA 5 – Distribuição das métricas médias entre todos os cenários

Cenários	Acurácia	Precision	Recall	F1 Score	mAP@50	mAP@50-95
Convexo	0.61	0.64	0.66	0.64	0.65	0.56
Convexo Aumentado	0.68	0.71	0.71	0.71	0.79	0.68
Linear	0.58	0.59	0.73	0.64	0.57	0.50
Linear Ampliado	0.86	0.91	0.89	0.89	0.91	0.84
Completo	0.69	0.73	0.72	0.72	0.62	0.53
Completo Ampliado	0.69	0.76	0.66	0.70	0.84	0.74
Convexo	0.71	0.72	0.74	0.73	0.75	0.65
Convexo Ampliado	0.69	0.75	0.64	0.68	0.79	0.66

5 DISCUSSÃO

Os resultados obtidos confirmam apenas parcialmente a hipótese de que técnicas de transfer learning combinadas com data augmentation podem levar a uma precisão clínica satisfatória em bases de dados restritas. No melhor dos cenários experimentais (“Linear Ampliado”), o modelo alcançou 86% de acurácia, um resultado consideravelmente interessante. Entretanto, em todos os demais cenários, a acurácia manteve-se abaixo de 75%, o que indica que, com o volume e a heterogeneidade atuais dos dados, a performance não é robusta o suficiente para aplicação clínica ampla.

Entre as limitações metodológicas, destaca-se o reduzido número de imagens lineares, que compõem o subconjunto de menor representatividade e provavelmente explicam o desempenho mais baixo observado no cenário sem aumento. Curiosamente, após a aplicação de data augmentation, o cenário “Linear Ampliado” não só superou a acurácia obtida com as imagens convexas originais, mas também ficou à frente dos conjuntos convexos com e sem aumento, sugerindo que o modelo se adapta especialmente bem às características das imagens lineares quando exposto a variações sintéticas.

Ao mesmo tempo, manteve-se um desbalanço de classes com predominância de exemplos “Normais” em relação a “Alterados”, o que pode ter influenciado negativamente o aprendizado e a generalização nos cenários “Convexo” e “Semi-Completo”. A falta de controle rigoroso sobre parâmetros de aquisição ultrassonográfica (ganho, ângulo e equipamento) acrescentou ruído aos dados, dificultando o treinamento e a avaliação. Além disso, a opção pela arquitetura YOLOv11-nano, embora eficiente em termos de velocidade e consumo de recursos, pode ter limitado a extração de características mais sutis, sugerindo que versões mais profundas ou técnicas de pruning e quantização poderiam melhorar o equilíbrio entre precisão e leveza.

A Figura 8 exemplifica um caso de predição correta, no qual a imagem rotulada como “normal” foi corretamente classificada pelo modelo com alta confiança (98%). Em contrapartida, a Figura 9 apresenta um caso de erro: a imagem, cuja rotulagem real é “normal”, foi classificada como “alterada”, evidenciando limitações do modelo em certos contextos visuais. Esses exemplos ilustram na prática tanto o potencial quanto os desafios enfrentados pela abordagem utilizada, sendo fundamentais para a compreensão crítica da performance observada.

FIGURA 11 – resultado com acerto

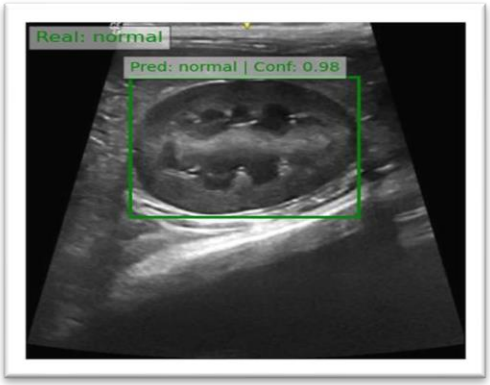
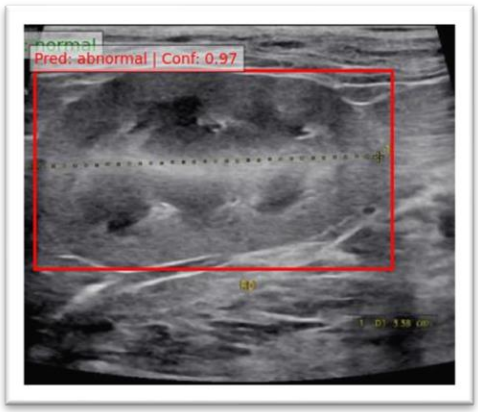


FIGURA 12 – Resultado com erro



Em termos de implicações, este estudo demonstra que, embora o data augmentation seja capaz de compensar parcialmente a escassez e o desbalanceamento de amostras — sobretudo em imagens lineares —, a simples adição de exemplos sintéticos não elimina por completo as restrições impostas por bases pequenas e desbalanceadas. Para avançar na consolidação de uma ferramenta de auxílio diagnóstico veterinário, recomenda-se ampliar o dataset por meio de coletas multicêntricas, garantindo maior diversidade de equipamentos e protocolos, e empregar estratégias de oversampling ou geração sintética direcionada para reforçar as classes menos representadas. Técnicas de normalização de domínio, como histogram matching, podem reduzir variações de aquisição, enquanto abordagens de fine-tuning semi-supervisionado ou por contraste poderiam ajudar a extrair melhor as características latentes dos dados.

6 CONCLUSÃO

Os resultados obtidos confirmam apenas parcialmente a hipótese de que técnicas de *transfer learning* combinadas com *data augmentation* poderiam resultar em uma precisão clínica satisfatória, especialmente considerando as limitações impostas por um banco de dados restrito. Ao longo do desenvolvimento do projeto, o número de imagens disponíveis foi gradualmente ampliado, permitindo o aprimoramento contínuo dos métodos de treinamento e validação. Inicialmente, as limitações quantitativas impunham severas barreiras à construção de modelos robustos, mas, com o tempo, a incorporação de novos dados possibilitou a experimentação de abordagens mais refinadas.

Durante essa trajetória, o projeto enfrentou diversos entraves relacionados à curadoria do banco de dados. O processo de *data cleaning* foi especialmente desafiador, exigindo um olhar atento sobre inconsistências nos dados e erros de anotação. O principal problema observado estava na rotulagem incorreta das imagens entre as classes “normal” e “alterado”, o que afetou diretamente a capacidade do modelo de aprender padrões consistentes. Como se trata de um trabalho que depende da colaboração entre áreas distintas — biologia e ciência da computação —, nem sempre houve consenso na interpretação clínica das imagens ultrassonográficas, o que gerou um nível considerável de ruído nos dados de entrada.

Entre os experimentos realizados, o cenário “Linear Ampliado” destacou-se com 86% de acurácia, superando todos os outros contextos testados. Isso se mostrou particularmente interessante, considerando que as imagens lineares estavam entre as menos representadas no conjunto inicial. A aplicação de *data augmentation* teve papel fundamental nessa melhora, mostrando que o modelo conseguiu adaptar-se bem às variações sintéticas das imagens lineares. Por outro lado, os conjuntos “Convexo” e “Semi-Completo” mantiveram-se com acurácia inferior a 75%, sugerindo que a combinação de heterogeneidade de imagens e desbalanceamento de classes comprometeu o desempenho nesses cenários.

A predominância de imagens “Normais” em relação às “Alteradas” também influenciou negativamente a generalização do modelo. Adicionalmente, a falta de padronização nos parâmetros de aquisição — como frequência, ganho e tipo de transdutor — introduziu variações de textura e ruído que tornaram o treinamento mais difícil. A escolha pela arquitetura YOLOv11-nano, embora eficiente e leve, limitou a capacidade do modelo de extrair padrões mais sutis. Em diversos casos, a sensibilidade excessiva a ruídos ou artefatos visuais comprometeu a acurácia das predições.

Mesmo com os avanços obtidos, a análise conduzida ainda não atinge a profundidade que o problema exige. A complexidade das imagens ultrassonográficas renais e a variabilidade anatômica entre os pacientes indicam que há muito a ser explorado e compreendido. O projeto representa uma etapa inicial importante, que conseguiu estruturar uma base experimental promissora, mas que ainda carece de investigações mais detalhadas para alcançar uma solução clinicamente aplicável. O equilíbrio entre rigor técnico, conhecimento biomédico e qualidade dos dados será determinante para a evolução dessa linha de pesquisa.

6.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Como recomendações para trabalhos futuros, sugere-se, inicialmente, ampliar e equilibrar o conjunto de dados por meio de coletas multicêntricas que incorporem maior variedade de casos, incluindo diferentes raças, idades e condições clínicas, com ênfase na obtenção de mais imagens lineares originais, de modo a reduzir o viés de classe e avaliar o impacto do data augmentation em cenários de maior diversidade. Além disso, recomenda-se experimentar arquiteturas e técnicas avançadas, testando modelos mais profundos ou híbridos (como variantes de YOLOv11, como YOLOV11-M etc., EfficientDet ou detectores baseados em Transformers) e comparando seu desempenho ao YOLOv11-nano, bem como avaliar estratégias de módulos de atenção e ensemble learning para capturar características mais sutis. No tocante à geração de dados, é importante aprimorar técnicas de oversampling e normalização de domínio, investigando métodos como GANs condicionais (cGANs) e MixUp avançado, e aplicar adaptações de domínio para tornar o modelo mais robusto a diferentes aparelhos de ultrassom e protocolos de aquisição. Para validar a generalização, recomenda-se realizar estudos prospectivos e testes em bases independentes, idealmente conduzidos por médicos veterinários em campo, comparando o desempenho do modelo com avaliações humanas e mensurando sua aceitabilidade clínica. A interpretação e explicabilidade devem ser abordadas com técnicas de explainable AI (por exemplo, Grad-CAM e LIME), a fim de visualizar as regiões mais influentes nas decisões do modelo e orientar ajustes finos na arquitetura.

7 USO DE INTELIGÊNCIA ARTICIFIAL GENERATIVA

PERGUNTAS SOBRE O USO DE IA GENERATIVA
1) Para escrita deste relatório, alguma ferramenta de inteligência artificial generativa foi utilizada? Sim (☒) Não (☐)
2) Qual(is) ferramenta(s) de IA generativa você utilizou? Não se aplica (☐) Se sim, cite quais: ChatGPT
3) Indique quais os usos de IA generativa foram aplicadas no neste relatório. Não se aplica (☐) Correção gramatical (ortografia e concordância) (☐) Formatação das referências (☒) Gerar partes do texto escrito (ex: frases, parágrafos, conceitos) (☐) Gerar a totalidade do texto escrito (☐) Gerar citações (☐) Criação/edição das imagens e gráficos (☐) Correção/auxílio na formatação final dos códigos estatísticos (ou outro software) (☒) Outros usos – especificar:
4) Declaração do uso de qualquer ferramenta de IA: ☒ Durante a preparação deste Relatório Final, o(s) autor(es) usaram CHAT-GPT, O4-MINI para a adicionar todas as referências as normas ABNT. Após usar essa ferramenta, o(s) autor(es) revisaram e editaram o conteúdo conforme necessário e assumem total responsabilidade pelo conteúdo. ☐ Durante a preparação deste Relatório Final, o(s) autor(es) afirmam que não utilizaram nenhuma inteligência artificial (IA).

REFERÊNCIAS

SCHARRE, Annabel; SCHOLLER, Dominik; GESSELL-MAY, Stefan; MÜLLER, Tobias; ZABLOTSKI, Yury; ERTEL, Wolfgang; MAY, Anna. **Comparison of veterinarians and a deep learning tool in the diagnosis of equine ophthalmic diseases.** 2024. Disponível em: https://beva.onlinelibrary.wiley.com/doi/full/10.1111/evj.14087. Acesso em: 6 jan. 2025.

LAURA, De Rosa; L'ABBATE, Serena; KUSMIC, Claudia; FAITA, Francesco. **Applications of Deep Learning Algorithms to Ultrasound Imaging Analysis in Preclinical Studies on In Vivo Animals**. 2023. Disponível em: https://www.researchgate.net/publication/373207099_Applications_of_Deep_Learning_Algorithms_to_Ultrasound_Imaging_Analysis_in_Preclinical_Studies_on_In_Vivo_Animals. Acesso em: 13 jan. 2025.

JOCHER, Glenn. Ultralytics YOLO11. 2024. Disponível em: https://docs.ultralytics.com/pt/models/yolo11/. Acesso em: 13 jan. 2025.

SONI, Akanksha; RAI, Avinash. **YOLO for Medical Object Detection (2018–2024).** 2024. Disponível em: https://ieeexplore.ieee.org/document/10653506. Acesso em: 4 jul. 2025.

RAO, S Nikhileswara. **YOLOv11 Explained: Next-Level Object Detection with Enhanced Speed and Accuracy**. 2024. Disponível em: https://medium.com/@nikhil-rao-20/yolov11-explained-next-level-object-detection-with-enhanced-speed-and-accuracy-2dbe2d376f71. Acesso em: 4 jul. 2025.

CUBUK, Ekin D.; ZOPH, Barret; SHLENS, Jonathon; LE, Quoc V. **RandAugment: Practical automated data augmentation with a reduced search space**. arXiv:1909.13719, 2019. Disponível em: https://arxiv.org/pdf/1909.13719. Acesso em: 9 jul. 2025.