

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Cauê Grassi Ribeiro da Silva

**Classificação de lesões cutâneas em imagens
infravermelhas com redes neurais convolucionais**

Uberlândia, Brasil

2026

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Cauê Grassi Ribeiro da Silva

**Classificação de lesões cutâneas em imagens
infravermelhas com redes neurais convolucionais**

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção do título de
Bacharel em Ciência da Computação.

Orientador: Prof. Henrique Coelho Fernandes

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Ciência da Computação

Uberlândia, Brasil

2026

Cauê Grassi Ribeiro da Silva

Classificação de lesões cutâneas em imagens infravermelhas com redes neurais convolucionais

Trabalho de conclusão de curso apresentado
à Faculdade de Computação da Universidade
Federal de Uberlândia, como parte dos requi-
sitos exigidos para a obtenção do título de
Bacharel em Ciência da Computação.

Trabalho aprovado em _____ de _____ de 2026.

Prof. Henrique Coelho Fernandes
Orientador

Gustavo Pedreira de Freitas
Convidado(a) 1

Prof. Jefferson Rodrigo de Souza
Convidado(a) 2

Uberlândia, Brasil
2026

*À minha avó, Neyde (in memoriam),
que não viu o fim desta caminhada,
mas esteve presente em cada passo dela.*

Agradecimentos

Agradeço, primeiramente e de forma especial, à minha avó, Neyde (*in memoriam*), que não pôde acompanhar a conclusão desta etapa, mas cujo amor, exemplo e presença permanecem comigo e sustentaram cada passo desta caminhada.

Aos meus pais, agradeço por todo o amor, pelo apoio incondicional e por nunca deixarem de acreditar em mim, mesmo diante das dificuldades; a eles devo a base de tudo o que conquistei. Ao meu irmão, pela companhia, pela parceria e por estar sempre presente ao longo desta jornada.

Ao meu orientador, Professor Henrique Coelho Fernandes, agradeço pela orientação, pela disponibilidade e pela confiança depositada neste trabalho, fundamentais para a sua realização.

Por fim, agradeço à professora Marcia Aparecida Fernandes, cuja rigidez e exigência, longe de me afastarem, despertaram em mim um apreço ainda maior pelo curso e a certeza de ter feito a escolha certa pela Ciência da Computação.

Resumo

Este trabalho investiga o uso de redes neurais convolucionais na classificação de lesões cutâneas benignas e malignas a partir de imagens de termografia infravermelha dinâmica, com foco no apoio ao diagnóstico de câncer de pele. Partindo da hipótese de que o sinal discriminativo está associado à dinâmica de reaquecimento da lesão, e não a um quadro estático isolado, cada sequência térmica foi tratada como a unidade principal de avaliação. Caracterizado como pesquisa científica aplicada de natureza experimental, o estudo empregou rótulos confirmados por biópsia, validação cruzada de cinco dobras agrupada por paciente e, como métrica principal, a AUC agrupada com intervalo de confiança de 95% por reamostragem *bootstrap* estratificada (também verificada com *bootstrap* agrupado por paciente). Foram comparados uma rede convolucional própria treinada do zero, transferência de aprendizado (ResNet18) e modelos clássicos (SVM e *Random Forest*) sobre representações extraídas da ResNet18 ajustada e atributos de dinâmica térmica. Os melhores modelos atingiram uma AUC de aproximadamente 0,66, sem diferença detectável em relação ao *Random Forest* no *bootstrap* pareado, enquanto a rede treinada do zero apresentou desempenho compatível com o nível de acaso. Nenhuma das intervenções testadas (entrada temporal multicanal, enriquecimento de atributos, agregação mais expressiva e combinação de modelos) superou o teto observado. Os resultados indicam o tamanho amostral como uma das principais restrições nesta base, embora limitações de aquisição, representação radiométrica e modelagem temporal também possam ter contribuído para o teto observado. O trabalho contribui com um pipeline reprodutível e com uma avaliação criteriosa da viabilidade da termografia dinâmica combinada a aprendizado de máquina como ferramenta complementar ao diagnóstico dermatológico.

Palavras-chave: câncer de pele. termografia infravermelha. redes neurais convolucionais. aprendizado de máquina. diagnóstico assistido por computador.

Abstract

This work investigates the use of convolutional neural networks for the classification of benign and malignant skin lesions from dynamic infrared thermography images, aiming to support skin cancer diagnosis. Based on the hypothesis that the discriminative signal is associated with the lesion's reheating dynamics rather than with an isolated static frame, each thermal sequence was treated as the primary unit of evaluation. Characterized as applied experimental research, the study used biopsy-confirmed labels, five-fold patient-grouped cross-validation and, as the main metric, the pooled AUC with a 95% confidence interval obtained by stratified bootstrap resampling (also verified with a patient-grouped bootstrap). A custom convolutional network trained from scratch, transfer learning (ResNet18) and classical models (SVM and Random Forest) built on representations extracted from the fine-tuned ResNet18 together with thermal-dynamics features were compared. The best models reached an AUC of approximately 0.66, with no detectable difference from the Random Forest under paired bootstrap analysis, whereas the network trained from scratch showed performance compatible with chance level. None of the tested interventions (multichannel temporal input, feature enrichment, more expressive pooling and model ensembling) surpassed this ceiling. The results suggest that sample size is one of the main constraints on this dataset, although acquisition, radiometric representation and temporal modeling limitations may also have contributed to the observed performance ceiling. The work contributes a reproducible pipeline and a careful assessment of the feasibility of dynamic thermography combined with machine learning as a complementary tool for dermatological diagnosis.

Keywords: skin cancer. infrared thermography. convolutional neural networks. machine learning. computer-aided diagnosis.

Lista de ilustrações

Figura 1	– Fluxo metodológico geral, da base de imagens à avaliação. As duas trilhas de modelagem — redes convolucionais e modelos clássicos sobre <i>embeddings</i> e atributos de dinâmica térmica — compartilham a mesma divisão por paciente e o mesmo protocolo de avaliação.	25
Figura 2	– Trajetória de intensidade média ao longo da sequência térmica para um exemplo de lesão benigna e um de lesão maligna (ambos do mesmo paciente), ilustrando a natureza dinâmica do sinal registrado pela DTI. A figura é apenas ilustrativa: o eixo horizontal é o índice do quadro (não o tempo físico), as duas sequências não estão temporalmente alinhadas, e a diferença de nível médio entre as curvas não representa uma tendência estatística da base nem implica que as classes possam ser separadas pelo nível de intensidade.	26
Figura 3	– Amostras representativas da base sob rótulo de biópsia: lesões benignas (linha superior) e malignas (linha inferior), de pacientes distintos. As três primeiras colunas mostram os termogramas em tons de cinza (canal único); a última reapresenta um exemplo em falsa cor (<i>colormap inferno</i>), apenas para visualização.	27
Figura 4	– Distribuição das 168 sequências utilizáveis por classe, sob rótulo confirmado por biópsia.	28
Figura 5	– Etapas do pré-processamento aplicado a cada quadro térmico, do quadro bruto ao PNG final. A equalização de histograma foi intencionalmente omitida para preservar as relações de intensidade térmica entre regiões e entre quadros.	29
Figura 6	– Efeito do pré-processamento sobre quadros térmicos reais da base (uma lesão por linha): quadro original, após remoção de ruído, após a conversão de escala para $[0, 1]$ e após redimensionamento, todos em tons de cinza; a última coluna reapresenta o quadro em falsa cor (<i>colormap inferno</i>), apenas para visualização. Como a conversão de escala é linear, as três primeiras etapas parecem visualmente semelhantes.	29
Figura 7	– Arquitetura da ThermalCNN treinada do zero. Três blocos convolucionais reduzem progressivamente a resolução espacial e ampliam a profundidade de canais; a camada densa de 256 dimensões alimenta exclusivamente o classificador próprio da rede. Os modelos clássicos (SVM/RF) não utilizam essa representação — seus <i>embeddings</i> são extraídos separadamente da ResNet18 (512 dimensões), conforme a Seção 4.4.	30

Figura 8 – Matrizes de confusão agrupadas das 168 sequências (execução de referência), para o <i>Random Forest</i> , a ResNet18 e a ThermalCNN. Percentuais normalizados por classe real: a diagonal da linha “Benigno” corresponde à especificidade e a da linha “Câncer”, à sensibilidade (classe maligna).	38
Figura 9 – Curvas ROC agrupadas dos modelos na execução de referência (168 sequências).	38
Figura 10 – AUC agrupada e intervalo de confiança de 95% (<i>bootstrap</i> estratificado) por modelo. A linha tracejada indica o nível de acaso ($AUC = 0,5$).	39
Figura 11 – Curvas de treinamento (dobra 0) da ResNet18 e da ThermalCNN: perda de treino e validação (acima) e AUC de validação por época (abaixo). Para a ResNet18, mostra-se a fase de ajuste fino; as 10 épocas iniciais com o <i>backbone</i> congelado foram omitidas.	45

Lista de tabelas

Tabela 1	– Síntese dos trabalhos relacionados discutidos.	23
Tabela 2	– Hiperparâmetros utilizados no treinamento e na avaliação.	35
Tabela 3	– Desempenho em nível de sequência na execução de referência (168 sequências). AUC agrupada com IC 95% por <i>bootstrap</i> estratificado; sensibilidade, especificidade, precisão, F1-score e acurácia no limiar padrão. O <i>ensemble</i> combina RF, SVM e ResNet18; a ThermalCNN foi treinada do zero.	37
Tabela 4	– Diferença de AUC agrupada entre o <i>Random Forest</i> (melhor modelo) e os demais, com intervalo de confiança de 95% por <i>bootstrap</i> pareado (estratificado por classe). Um intervalo que contém o zero indica ausência de diferença detectável.	37
Tabela 5	– AUC agrupada (IC 95%) por configuração experimental e modelo. As estimativas são próximas e os intervalos amplamente sobrepostos. As diferenças entre configurações são apresentadas de forma descritiva; a comparação pareada formal (Tabela 4) refere-se aos modelos da execução de referência, para os quais não se detecta diferença distinguível de zero.	40

Lista de abreviaturas e siglas

ABCDE	Assimetria, Borda, Cor, Diâmetro e Evolução
AUC	Area Under the Curve (área sob a curva ROC)
BN	Batch Normalization
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
DL	Deep Learning
DTI	Dynamic Thermal Imaging
GAP	Global Average Pooling
HOG	Histogram of Oriented Gradients
IC	Intervalo de Confiança
IRT	Infrared Thermography
LBP	Local Binary Patterns
ML	Machine Learning
ReLU	Rectified Linear Unit
RF	Random Forest
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SVM	Support Vector Machine
TCC	Trabalho de Conclusão de Curso

Sumário

1	INTRODUÇÃO	13
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	Melanoma e câncer de pele	15
2.2	Diagnóstico assistido por computador em dermatologia	15
2.3	Termografia infravermelha aplicada a lesões cutâneas	16
2.4	Tratamento de imagens térmicas e extração de informação	16
2.5	Aprendizado de máquina e redes neurais convolucionais	16
2.6	Métricas e avaliação experimental em saúde	17
3	TRABALHOS RELACIONADOS	19
3.1	Quantificação do sinal térmico dinâmico	19
3.2	Termografia infravermelha com aprendizado de máquina	20
3.3	Revisões e diretrizes de aquisição	20
3.4	Aprendizado profundo aplicado à dermatologia	21
3.5	Síntese e posicionamento	22
4	MATERIAIS E MÉTODOS	24
4.1	Formulação da tarefa e unidade de análise	24
4.2	Base de imagens e definição dos rótulos	25
4.3	Pré-processamento	27
4.4	Arquiteturas avaliadas	28
4.5	Codificação do sinal dinâmico	30
4.6	Protocolo de validação	31
4.7	Métricas e reporte estatístico	32
4.8	Desenho experimental e ablações	34
4.9	Hiperparâmetros	34
5	RESULTADOS	36
5.1	Resultado central: teto de desempenho observado	36
5.2	Ablações e resultados negativos	37
5.2.1	Entrada temporal por empilhamento de canais	38
5.2.2	Rótulo por biópsia versus por pasta	39
5.2.3	Enriquecimento de atributos, <i>pooling</i> e <i>ensemble</i>	40
6	DISCUSSÃO	41
6.1	Evidências de limitação pelo tamanho amostral	41

6.2	Limitações metodológicas independentes do tamanho amostral . . .	42
6.3	Comparação com a literatura	43
6.4	O desempenho de acaso da rede treinada do zero	44
6.5	Implicação clínica: sensibilidade e falsos negativos	44
6.6	A hipótese da dinâmica térmica e o que faltou capturar	45
6.7	Ameaças à validade	46
7	CONSIDERAÇÕES FINAIS	47
	REFERÊNCIAS	49

1 Introdução

O câncer de pele corresponde a um conjunto de neoplasias cutâneas que surgem a partir do crescimento anormal de células da pele. Entre essas neoplasias, o melanoma merece atenção especial por seu maior potencial de invasão e metástase, sobretudo quando o diagnóstico ocorre de forma tardia ([NATIONAL CANCER INSTITUTE, 2025](#)). Na prática clínica, a avaliação inicial de lesões suspeitas costuma envolver exame visual, aplicação de critérios como a regra ABCDE e, em muitos casos, dermatoscopia, que amplia a visualização de estruturas não perceptíveis a olho nu ([DINNES et al., 2018](#)). Embora esses recursos tenham elevado a qualidade diagnóstica, a interpretação ainda depende fortemente da experiência do profissional e da disponibilidade de atendimento especializado.

Nos últimos anos, sistemas de diagnóstico assistido por computador (CAD, do inglês *computer-aided diagnosis*) passaram a ocupar papel relevante no apoio à avaliação dermatológica. Em especial, o avanço do aprendizado profundo sobre bases de imagens médicas ampliou a possibilidade de construir modelos capazes de aprender padrões discriminativos úteis para tarefas de classificação de lesões ([ESTEVA et al., 2017](#)). Paralelamente, a termografia infravermelha vem sendo explorada como modalidade complementar de aquisição, pois permite observar distribuições térmicas associadas a processos fisiológicos, inflamatórios e vasculares que podem diferenciar tecidos saudáveis e alterados ([VERSTOCKT et al., 2022](#)). Dessa forma, a combinação entre imagens térmicas e técnicas de aprendizado de máquina configura uma linha de pesquisa promissora para apoio ao diagnóstico de câncer de pele.

A relevância do tema decorre tanto de seu impacto clínico quanto de seus desafios computacionais. Do ponto de vista médico, soluções de triagem podem contribuir para priorização de casos suspeitos e para ampliação do suporte diagnóstico em contextos com acesso limitado a especialistas. Do ponto de vista científico, o problema envolve questões como qualidade da aquisição térmica, pré-processamento de imagens, definição de regiões de interesse, escolha de arquiteturas de classificação e análise criteriosa de métricas, especialmente em cenários nos quais falsos negativos têm implicações graves. Assim, investigar estratégias computacionais para classificação de lesões em imagens infravermelhas representa uma contribuição potencial tanto para a literatura em visão computacional aplicada à saúde quanto para o desenvolvimento de ferramentas complementares à prática clínica.

Diante desse contexto, o problema investigado neste trabalho pode ser formulado da seguinte maneira: redes neurais convolucionais conseguem aprender padrões térmicos

relevantes em imagens infravermelhas de lesões cutâneas a ponto de apoiar, de maneira útil, a distinção entre classes de interesse em câncer de pele? A hipótese adotada é que, sob um pipeline adequado de preparação dos dados, treinamento e avaliação, modelos convolucionais podem capturar assinaturas térmicas discriminativas e produzir desempenho potencialmente relevante como apoio ao diagnóstico.

O objetivo geral deste TCC é investigar a aplicação de redes neurais convolucionais na classificação de lesões cutâneas em imagens infravermelhas, avaliando a termografia como ferramenta complementar ao diagnóstico de câncer de pele. Como objetivos específicos, pretende-se:

1. realizar levantamento bibliográfico sobre câncer de pele, diagnóstico assistido por computador, termografia e aprendizado profundo aplicado a imagens médicas;
2. organizar e analisar a base de imagens infravermelhas disponibilizada para o estudo;
3. definir e aplicar etapas de pré-processamento adequadas aos termogramas;
4. implementar e comparar arquiteturas convolucionais, transferência de aprendizado e classificadores clássicos sobre representações aprendidas e atributos de dinâmica térmica, avaliando também diferentes estratégias de codificação da dinâmica térmica e de agregação em nível de sequência;
5. avaliar os modelos por meio de métricas de classificação sob um protocolo de validação controlado e realizar análise comparativa com a literatura.

O trabalho caracteriza-se como uma pesquisa científica aplicada, conforme a classificação de Wazlawick ([WAZLAWICK, 2020](#)), pois não se limita à construção de um artefato computacional, mas busca produzir conhecimento sobre o comportamento de uma abordagem baseada em aprendizado profundo para um problema específico. Ainda que scripts, modelos treinados e notebooks sejam produzidos como artefatos, o foco principal recai sobre a formulação do problema, a definição do método experimental, a avaliação quantitativa e a discussão crítica dos resultados.

Quanto à organização do texto, esta monografia está estruturada da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica; o Capítulo 3 discute os trabalhos relacionados; o Capítulo 4 detalha os materiais e métodos efetivamente empregados no experimento; o Capítulo 5 reporta os resultados obtidos; o Capítulo 6 discute esses resultados à luz da literatura e das limitações do estudo; e, por fim, o Capítulo 7 apresenta as considerações finais.

2 Fundamentação Teórica

Este capítulo reúne os conceitos essenciais que sustentam a proposta do trabalho, abordando os aspectos médicos e computacionais necessários para compreender a classificação de lesões cutâneas em imagens infravermelhas.

2.1 Melanoma e câncer de pele

O termo câncer de pele engloba diferentes neoplasias cutâneas, geralmente divididas em cânceres de pele não melanoma e melanoma. Embora o melanoma seja menos frequente do que outros tipos, ele tende a ser mais agressivo e apresenta maior risco de metástase, o que torna a detecção precoce especialmente importante ([NATIONAL CANCER INSTITUTE, 2025](#)). Na rotina clínica, a avaliação inicial costuma apoiar-se em critérios visuais e na experiência do especialista, enquanto a confirmação diagnóstica pode exigir exames complementares, como biópsia e análise histopatológica.

Historicamente, o exame visual foi por muito tempo a principal forma de avaliar lesões suspeitas. A incorporação da dermatoscopia representou um avanço importante ao permitir a observação de estruturas não visíveis a olho nu, aumentando a precisão em diversos cenários clínicos ([DINNES et al., 2018](#)). Mesmo assim, a interpretação continua sujeita à variabilidade entre avaliadores, o que reforça o interesse por ferramentas computacionais de apoio à decisão.

2.2 Diagnóstico assistido por computador em dermatologia

Os sistemas de diagnóstico assistido por computador buscam apoiar especialistas por meio da análise automatizada de imagens e outros dados clínicos. Em dermatologia, esses sistemas podem contribuir para triagem, padronização da análise e apoio à decisão clínica. O crescimento de bases de imagens e de métodos de aprendizado profundo impulsionou essa área, permitindo o treinamento de modelos com desempenho expressivo em tarefas de classificação de lesões cutâneas ([ESTEVA et al., 2017](#)).

Ainda assim, o uso de inteligência artificial em saúde exige cuidados metodológicos. Questões como transparência, qualidade da base de dados, reprodutibilidade experimental e adequação das métricas são frequentemente apontadas como fatores centrais para a validação de modelos clínicos ([LIU et al., 2020](#); [RAJKOMAR](#); [DEAN](#); [KOHANE, 2019](#)). Em aplicações médicas, não basta obter boa acurácia média: é necessário interpretar o impacto dos diferentes tipos de erro e explicitar as limitações do experimento.

2.3 Termografia infravermelha aplicada a lesões cutâneas

A termografia infravermelha é uma técnica de aquisição não invasiva que registra padrões de temperatura na superfície corporal. No contexto dermatológico, ela pode revelar alterações térmicas associadas a vascularização, metabolismo e inflamação, o que a torna potencialmente útil para a análise de lesões cutâneas (VERSTOCKT et al., 2022). Dependendo do protocolo adotado, a termografia pode ser passiva, quando registra o estado térmico natural da superfície, ou ativa/dinâmica, quando observa a resposta térmica após algum estímulo controlado.

Estudos na área indicam que o sucesso dessa abordagem depende fortemente de aspectos de aquisição e padronização, como condições ambientais, posicionamento da câmera, distância ao paciente, emissividade da pele e consistência do protocolo experimental (VERSTOCKT et al., 2022; GODOY et al., 2015). Isso mostra que a modalidade oferece oportunidades relevantes, mas também exige cuidado com a reprodutibilidade e com a comparabilidade entre experimentos.

2.4 Tratamento de imagens térmicas e extração de informação

Antes do treinamento de modelos, imagens térmicas geralmente passam por etapas de preparação, que podem incluir normalização, redimensionamento, filtragem, organização de rótulos e delimitação de regiões de interesse (VERSTOCKT et al., 2022; SOTO; GODOY, 2024). Em pipelines tradicionais, também podem ser extraídas características estatísticas, espaciais e térmicas da lesão ou de sua vizinhança (MAGALHÃES et al., 2021). Já em abordagens baseadas em redes neurais convolucionais, parte dessas representações é aprendida automaticamente a partir dos dados.

Mesmo em modelos profundos, o pré-processamento permanece importante, pois ajuda a reduzir variabilidades indesejadas e torna a entrada mais consistente para o treinamento. Em aplicações médicas, essa etapa também influencia a interpretabilidade dos resultados e a robustez do modelo em diferentes condições de aquisição.

2.5 Aprendizado de máquina e redes neurais convolucionais

O aprendizado de máquina supervisionado consiste em treinar modelos a partir de exemplos rotulados, buscando aprender relações entre as entradas e suas respectivas classes. Em problemas de imagens, redes neurais convolucionais se destacam por sua capacidade de extrair automaticamente padrões locais e hierárquicos, o que as tornou amplamente utilizadas em visão computacional e análise de imagens médicas (ESTEVA et al., 2017; MAGALHÃES et al., 2021).

No contexto deste trabalho, a adoção de CNNs se justifica pela possibilidade de aprender características discriminativas diretamente dos termogramas, sem depender exclusivamente de atributos definidos manualmente. Entretanto, o desempenho dessas redes depende de fatores como o tamanho da base, a distribuição entre classes, o pré-processamento adotado, a arquitetura escolhida e o controle de sobreajuste durante o treinamento.

Arquiteturas convolucionais profundas enfrentam a dificuldade de treinar muitas camadas; a ResNet (HE et al., 2016) contornou esse problema com *conexões residuais* (atalhos que somam a entrada de um bloco à sua saída), permitindo redes mais profundas e estáveis. Duas técnicas tornaram-se padrão nessas redes: a normalização em lote (IOFFE; SZEGEDY, 2015), que padroniza as ativações e acelera a convergência, e o *dropout* (SRI-VASTAVA et al., 2014), que desativa neurônios aleatoriamente durante o treino para reduzir o sobreajuste. Quando os dados de treino são escassos — situação comum em imagens médicas —, a *transferência de aprendizado* é uma estratégia eficaz: parte-se de uma rede pré-treinada em uma base ampla e genérica, como a ImageNet (RUSSAKOVSKY et al., 2015), e ajustam-se seus pesos à tarefa de interesse, reaproveitando representações visuais já aprendidas.

Além das redes profundas, classificadores clássicos permanecem relevantes, sobretudo com poucos dados. A máquina de vetores de suporte (SVM) (CORTES; VAPNIK, 1995) busca o hiperplano de maior margem entre as classes, empregando núcleos (*kernels*) para separar dados não linearmente separáveis. A *Random Forest* (BREIMAN, 2001) combina múltiplas árvores de decisão treinadas sobre subamostras aleatórias, reduzindo variância e sobreajuste. Uma estratégia híbrida — adotada neste trabalho — usa a penúltima camada de uma CNN como extratora de atributos (*embeddings*) e alimenta esses vetores a SVM e *Random Forest*, combinando a capacidade de representação da rede com a robustez dos classificadores clássicos em amostras pequenas.

2.6 Métricas e avaliação experimental em saúde

A avaliação de classificadores em saúde deve considerar não apenas o número de acertos globais, mas também a natureza dos erros cometidos (LIU et al., 2020; RAJKOMAR; DEAN; KOHANE, 2019). Métricas como precisão, sensibilidade, especificidade e F1-score são especialmente relevantes porque ajudam a interpretar o comportamento do modelo em diferentes classes. Em aplicações relacionadas ao câncer de pele, a sensibilidade merece destaque, pois reflete a capacidade do sistema de identificar corretamente casos positivos (malignos).

Além das métricas numéricas, matrizes de confusão e comparações com trabalhos da literatura auxiliam na interpretação dos resultados. Contudo, tais comparações precisam

ser feitas com cautela, pois diferenças na base de dados, na modalidade de aquisição e nos protocolos experimentais podem afetar significativamente o desempenho reportado. Cabe ainda observar que, quando a amostra de teste é pequena, estimativas pontuais de desempenho são ruidosas; o *bootstrap* (EFRON; TIBSHIRANI, 1993) — reamostragem com reposição do conjunto de teste — permite quantificar essa incerteza por meio de intervalos de confiança, sendo preferível a reportar um valor isolado.

3 Trabalhos Relacionados

Este capítulo revisa trabalhos que dialogam diretamente com a proposta deste TCC, organizados em quatro eixos: a evidência de que o sinal térmico dinâmico carrega informação discriminativa; o uso de aprendizado de máquina sobre termogramas; as revisões e diretrizes de aquisição; e o aprendizado profundo aplicado à dermatologia. Ao final, uma síntese posiciona a contribuição deste trabalho frente às lacunas identificadas na literatura.

3.1 Quantificação do sinal térmico dinâmico

Pirtini Çetingül e Herman, no Laboratório de Transferência de Calor da Universidade Johns Hopkins, estão entre os primeiros a quantificar a assinatura térmica de lesões melanocíticas por imagem infravermelha transiente (dinâmica). O protocolo resfria a pele da lesão e do tecido vizinho — com bolsa de gel ou fluxo de ar de um tubo de vórtice — e registra a reação durante o reaquecimento, partindo da hipótese de que a maior atividade metabólica e a vascularização do melanoma produzem uma resposta térmica distinta da pele sadia. Comparando o comportamento de tecido sadio e maligno com o resultado de biópsia, os autores propuseram um modelo quantitativo capaz de evidenciar o contraste entre as classes ([ÇETINGÜL; HERMAN, 2011](#)). Esse trabalho é uma das bases conceituais deste TCC: sugere que o sinal discriminativo está associado à *dinâmica de reaquecimento*, e não a um quadro estático isolado — exatamente a premissa que motiva o tratamento das sequências térmicas como séries temporais neste estudo.

Bonmarin e Le Gal exploram uma variante ativa distinta, a termografia *lock-in*: em vez de um único resfriamento seguido de recuperação, a pele é estimulada por um fluxo de ar com temperatura modulada periodicamente, e o processamento *lock-in* isola as regiões cujas propriedades termofísicas diferem do tecido sadio. Em estudo de viabilidade voltado à detecção precoce de melanoma cutâneo, a técnica mostrou-se promissora para revelar lesões com contraste térmico modificado ([BONMARIN; LE GAL, 2014](#)). O trabalho reforça o valor da estimulação térmica ativa e ilustra que há mais de uma forma de excitar e ler o sinal dinâmico — a modulação periódica difere do protocolo de resfriamento e recuperação adotado na base utilizada neste TCC.

Godoy et al. avaliam a *dynamic thermal imaging* (DTI) como abordagem não invasiva de triagem, com um protocolo padronizado de resfriamento seguido da análise da recuperação térmica das lesões. Os resultados evidenciam que lesões malignas reaquecem de maneira diferente do tecido benigno, sustentando a utilidade clínica do sinal térmico dinâmico e a necessidade de protocolos de aquisição consistentes ([GODOY et al., 2015](#)). Esse paradigma de aquisição — resfriamento e recuperação — é justamente o da base

TR_heatSkin empregada neste trabalho, o que torna o estudo uma referência direta para o protocolo experimental.

3.2 Termografia infravermelha com aprendizado de máquina

Magalhães, Tavares, Mendes e Vardasca investigam sistematicamente o potencial da termografia infravermelha *passiva* como ferramenta auxiliar ao diagnóstico de câncer de pele, comparando duas estratégias de aprendizado de máquina — *ensemble learning* e aprendizado profundo — sobre parâmetros térmicos extraídos dos termogramas. A abordagem profunda superou o *ensemble*, atingindo precisão de aproximadamente 0,97 e sensibilidade de 0,94 na distinção entre melanoma e nevos, e precisão em torno de 0,93 entre melanoma e lesões não melanoma (MAGALHÃES et al., 2021). É a principal referência metodológica para a estratégia de *features* adotada aqui — combinar representações aprendidas por uma CNN com atributos térmicos derivados das sequências, reduzindo a dependência exclusiva de descritores visuais artesanais como HOG e LBP. Vale notar, contudo, que esses números elevados foram obtidos em uma tarefa mais separável (melanoma *versus* nevo) e sobre uma amostra específica, o que ajuda a contextualizar o teto mais modesto obtido neste TCC na tarefa binária benigno *versus* maligno em sentido amplo.

Soto e Godoy propõem o primeiro algoritmo automático de detecção de câncer de pele por termografia *ativa*, estruturado em três etapas: segmentação da lesão por uma rede U-Net (RONNEBERGER; FISCHER; BROX, 2015), robusta mesmo em imagens de baixa qualidade adquiridas em ambiente clínico, extração de um conjunto inédito de atributos espaço-temporais e classificação por máquina de vetores de suporte (SVM). Fixando a sensibilidade de referência em 85%, o sistema atingiu especificidade em torno de 88% com segmentação feita por especialista e de cerca de 69% com a segmentação automática por U-Net (SOTO; GODOY, 2024). É o trabalho metodologicamente mais próximo deste TCC, também combinando termografia ativa, atributos espaço-temporais e SVM. As principais diferenças são o uso, aqui, de uma aproximação não supervisionada da região de interesse (limiar de Otsu) em vez de segmentação dedicada, e a adoção de um protocolo de avaliação mais conservador (validação cruzada por paciente e intervalos de confiança por *bootstrap*).

3.3 Revisões e diretrizes de aquisição

Em uma revisão da literatura, Magalhães, Vardasca e Mendes levantaram o uso recente da termografia infravermelha médica em neoplasias cutâneas, consultando as bases Scopus, PubMed e ISI Web of Science e selecionando, após critérios de exclusão, catorze publicações a partir de um conjunto inicial de cinquenta e cinco (MAGALHÃES; VARDASCA; MENDES, 2018). A revisão confirma que a IRT tem sido empregada para

caracterizar e distinguir neoplasias malignas e benignas e para acompanhar o tratamento, e aponta a termografia dinâmica como direção promissora — ao mesmo tempo em que evidencia a heterogeneidade metodológica e o tamanho reduzido das amostras como limitações recorrentes do campo.

Verstockt et al. apresentam uma revisão focada no *setup* de medição, no procedimento e nos equipamentos para detecção de câncer de pele por termografia, abrangendo tanto a modalidade passiva quanto a ativa. A contribuição é particularmente útil por explicitar os fatores que afetam a aquisição dos termogramas e a reprodutibilidade — condições ambientais, posicionamento e distância da câmera, emissividade da pele e consistência do protocolo (VERSTOCKT et al., 2022). Essas diretrizes fundamentam decisões metodológicas deste TCC, como a preservação do gradiente de intensidade térmica (sem equalização de histograma) e a atenção à padronização da aquisição.

Mais recentemente, Kesztyüs et al. revisaram o uso da termografia infravermelha não invasiva para triagem, diagnóstico e monitoramento do câncer de pele, confirmando que o interesse pela modalidade permanece ativo e que a padronização da aquisição e a validação clínica seguem entre as questões em aberto (KESZTYÜS et al., 2025). A atualidade dessa revisão reforça que o problema investigado neste TCC continua relevante e ainda não resolvido na literatura.

3.4 Aprendizado profundo aplicado à dermatologia

Esteve et al. marcaram a área ao demonstrar classificação de câncer de pele em nível de dermatologista com aprendizado profundo. Uma rede Inception v3, pré-treinada e ajustada por transferência de aprendizado sobre 129.450 imagens clínicas abrangendo 2.032 doenças, igualou o desempenho de 21 dermatologistas certificados em duas tarefas binárias (carcinoma queratinocítico *versus* ceratose seborreica benigna; melanoma maligno *versus* nevo benigno) (ESTEVA et al., 2017). Embora empregue imagens clínicas e dermatoscópicas — e não infravermelhas —, é a referência conceitual para o uso de CNNs e transferência de aprendizado no domínio, estratégia central deste TCC. O contraste de escala é instrutivo: o desempenho expressivo apoiou-se em mais de cem mil imagens, ordens de grandeza acima das 168 sequências disponíveis aqui.

Haenssle et al. compararam uma CNN (arquitetura Inception v4) a 58 dermatologistas na classificação dermatoscópica de melanoma, sobre um conjunto de teste de 100 imagens avaliado em dois níveis de informação clínica. A rede superou a maioria dos especialistas, e os autores enfatizam que a ferramenta deve servir como *apoio* à decisão de biopsiar — não como substituta do dermatologista (HAENSSLE et al., 2018). Esse enquadramento converge com a leitura adotada neste TCC, que posiciona o classificador como eventual ferramenta complementar de triagem, e não como sistema autônomo de

diagnóstico.

3.5 Síntese e posicionamento

Em conjunto, esses trabalhos delineiam três constatações convergentes: o sinal térmico *dinâmico* carrega informação discriminativa entre lesões benignas e malignas (ÇETINGÜL; HERMAN, 2011; BONMARIN; LE GAL, 2014; GODOY et al., 2015); o aprendizado de máquina e profundo sobre termogramas é viável e competitivo (MAGALHÃES et al., 2021; SOTO; GODOY, 2024); e as CNNs alcançam desempenho de destaque em dermatologia, ainda que sobre grandes bases de imagens RGB/dermatoscópicas (ESTEVA et al., 2017; HAENSSLE et al., 2018). As revisões da área (MAGALHÃES; VARDASCA; MENDES, 2018; VERSTOCKT et al., 2022; KESZTYÜS et al., 2025), por sua vez, apontam lacunas persistentes: a falta de padronização na aquisição, a escassez de bases termográficas adequadas e públicas, os tamanhos amostrais reduzidos e a necessidade de avaliações experimentais rigorosamente controladas. É nesse último ponto que este TCC se posiciona: sobre uma base de termografia dinâmica, combina rótulos confirmados por biópsia, validação cruzada agrupada por paciente e reporte por intervalos de confiança, oferecendo uma avaliação criteriosa do que a modalidade permite alcançar sob restrição de dados. A Tabela 1 sintetiza os trabalhos discutidos.

Tabela 1 – Síntese dos trabalhos relacionados discutidos.

Trabalho	Modalidade	Método / pipeline	Contribuição principal
Çetingül & Her- man (2011)	DTI (ativa)	Resfriamento + imagem transiente; quantificação da assinatura térmica	Evidência de que a dinâmica de reaque- cimento distingue o melanoma
Bonmarin & Le Gal (2014)	IRT ativa (<i>lock- in</i>)	Estimulação térmica pe- riódica + processamento <i>lock-in</i>	Deteção precoce por contraste ter- mofísico
Godoy et al. (2015)	DTI (ativa)	Resfriamento + recupera- ção térmica padronizada	Utilidade clínica do sinal térmico dinâ- mico
Magalhães et al. (2021)	IRT passiva	Parâmetros térmicos + <i>en- semble</i> /aprendizado pro- fundo	Comparação de es- tratégias de ML; DL superior
Soto & Godoy (2024)	IRT ativa	U-Net (ROI) + atributos espaço-temporais + SVM	Pipeline automá- tico robusto em ambiente clínico
Magalhães et al. (2018)	Revisão	Levantamento sistemático (14 estudos)	Panorama e lacu- nas da IRT em neo- plasias
Verstockt et al. (2022)	Revisão	<i>Setup</i> , procedimento e equipamentos	Diretrizes e fato- res de reprodutibi- lidade
Keszytyüs et al. (2025)	Revisão	Triagem, diagnóstico e mo- nitoramento	Atualização; padro- nização em aberto
Esteva et al. (2017)	RGB e dermatos- copia	CNN Inception v3 (129.450 imagens)	Classificação em ní- vel de dermatolo- gista
Haenssle et al. (2018)	Dermatoscopia	CNN Inception v4 <i>vs.</i> 58 dermatologistas	CNN supera espe- cialistas; uso como apoio

Fonte: Elaborada pelo autor (2026).

4 Materiais e Métodos

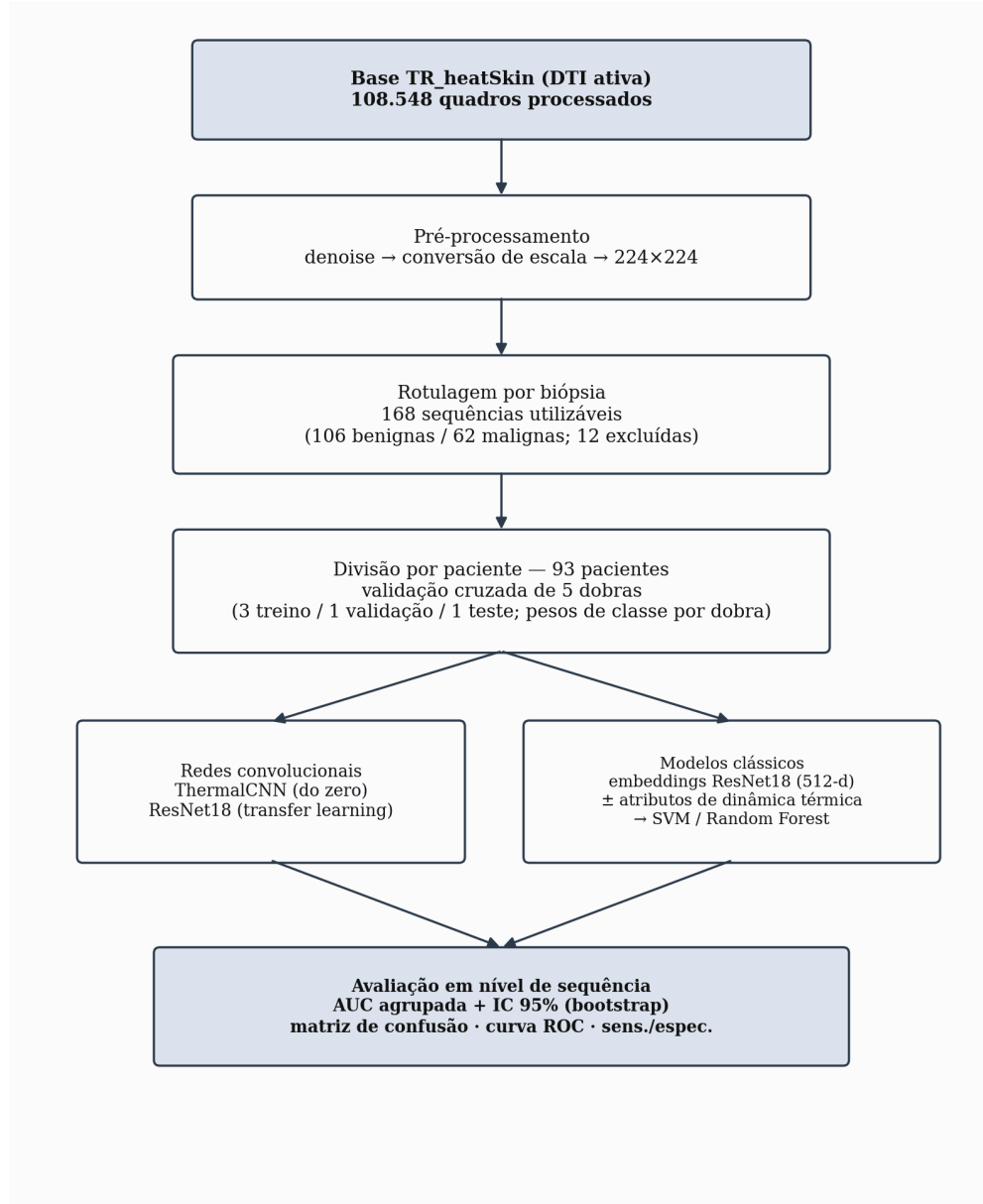
Este trabalho adota uma abordagem experimental em Computação, com natureza de pesquisa científica aplicada, conforme a classificação de Wazlawick (WAZLAWICK, 2020). O foco não recai apenas sobre a implementação de um sistema, mas sobre a construção de um pipeline reprodutível e a produção de evidências quantitativas, obtidas sob validação controlada, acerca da viabilidade de classificar lesões cutâneas benignas e malignas a partir de termografia infravermelha dinâmica. Todo o código foi implementado em Python 3.11, com PyTorch (PASZKE et al., 2019) para as redes neurais, OpenCV e NumPy para o processamento de imagens e *scikit-learn* (PEDREGOSA et al., 2011) para os modelos clássicos; o versionamento foi feito com Git. Para favorecer a reprodutibilidade, fixou-se a semente aleatória (*seed* = 42) nas etapas estocásticas e cada execução do experimento foi persistida em um diretório próprio, contendo os hiperparâmetros exatos, o registro de treinamento e as previsões por sequência, de modo que nenhuma execução sobrescreve outra e todas permanecem independentemente auditáveis. Reconhece-se que a fixação da semente não garante, isoladamente, determinismo bit a bit (que depende também de versões de biblioteca e de configurações de *hardware*); a rastreabilidade é assegurada, sobretudo, pelo registro completo de configuração, logs e previsões de cada execução. A Figura 1 resume o fluxo metodológico completo, do dado bruto à avaliação; as seções seguintes detalham cada etapa.

4.1 Formulação da tarefa e unidade de análise

A tarefa investigada é a classificação binária entre lesões histopatologicamente benignas e malignas, segundo o laudo de biópsia. Vale destacar que ambas as classes contêm uma lesão: não se trata de distinguir pele sadia de lesão, mas de diferenciar dois tipos de lesão a partir de sua assinatura térmica. Frequentemente um mesmo paciente contribui com lesões de ambas as classes, o que torna o agrupamento por paciente uma exigência metodológica, e não uma escolha opcional.

Uma constatação empírica orientou todo o desenho experimental: um único quadro (*frame*) térmico estático de uma lesão benigna é quase indistinguível de um quadro de uma lesão maligna, com a AUC em nível de *frame* isolado situando-se em torno de 0,45–0,49 nas primeiras execuções, ou seja, próxima do acaso. O sinal discriminativo não está, portanto, na aparência de um quadro isolado, mas, presumivelmente, na dinâmica de reaquecimento da lesão ao longo da sequência térmica capturada pela DTI. A Figura 2 ilustra essa trajetória temporal para um exemplo de cada classe. Por essa razão, a unidade principal de avaliação é a sequência térmica completa de uma lesão, e é sobre essa unidade

Figura 1 – Fluxo metodológico geral, da base de imagens à avaliação. As duas trilhas de modelagem — redes convolucionais e modelos clássicos sobre *embeddings* e atributos de dinâmica térmica — compartilham a mesma divisão por paciente e o mesmo protocolo de avaliação.



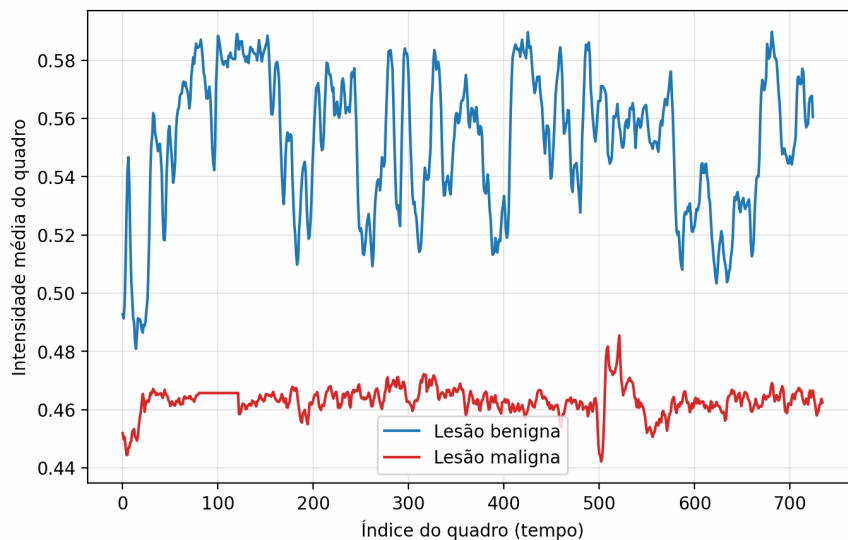
Fonte: Elaborada pelo autor (2026).

que as métricas finais foram calculadas — embora, como se detalha adiante, as redes convolucionais ainda sejam treinadas com quadros individuais.

4.2 Base de imagens e definição dos rótulos

Utilizou-se a base privada *TR_heatSkin*, disponibilizada pelo orientador, adquirida por termografia infravermelha dinâmica ativa (protocolo de aclimação, resfriamento controlado e registro da recuperação térmica). As imagens são organizadas na estrutura

Figura 2 – Trajetória de intensidade média ao longo da sequência térmica para um exemplo de lesão benigna e um de lesão maligna (ambos do mesmo paciente), ilustrando a natureza dinâmica do sinal registrado pela DTI. A figura é apenas ilustrativa: o eixo horizontal é o índice do quadro (não o tempo físico), as duas sequências não estão temporalmente alinhadas, e a diferença de nível médio entre as curvas não representa uma tendência estatística da base nem implica que as classes possam ser separadas pelo nível de intensidade.

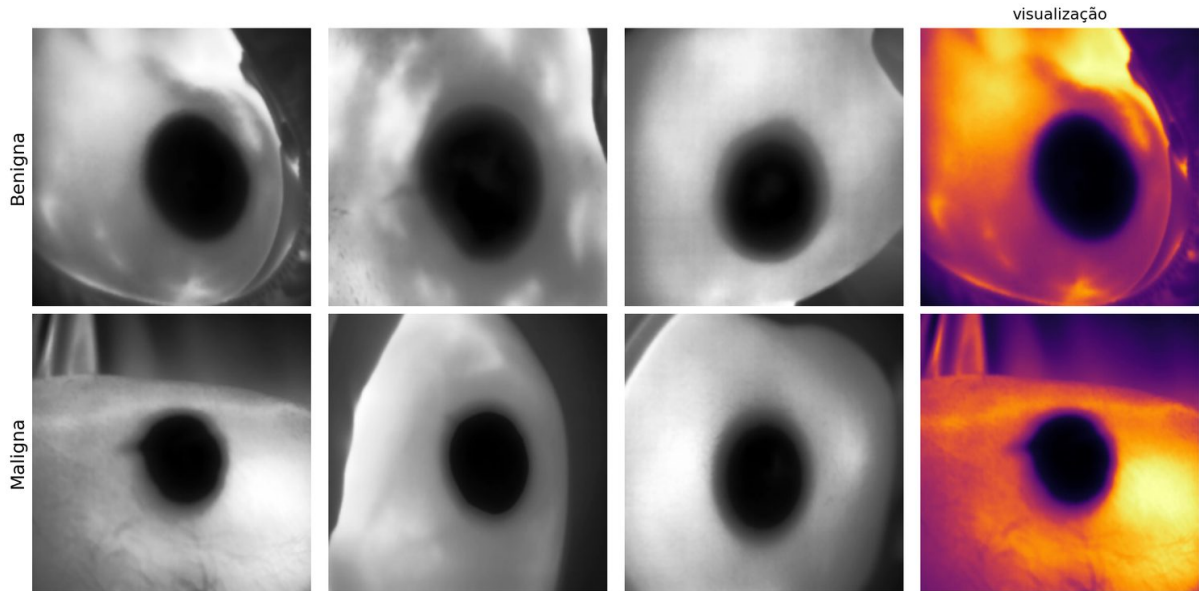


Fonte: Elaborada pelo autor (2026).

`Classe/Paciente/frames_sequenciais.png` e, embora descritas no protocolo de coleta como RGB, a inspeção dos arquivos mostrou que as imagens utilizadas possuem um único canal, armazenado como matriz bidimensional *uint8*, com dimensões originais de 512×480 px. Após o pré-processamento, obtiveram-se 108.548 quadros processados. A Figura 3 apresenta amostras representativas das duas classes.

Os diretórios **Healthy/Sick** presentes no disco constituem apenas uma heurística de nomenclatura. Para produzir uma tarefa binária limpa, os rótulos foram definidos a partir da planilha de biópsias (`Labels_Skin_Cancer.xlsx`), que contém o diagnóstico histopatológico confirmado de cada lesão. O procedimento de rotulagem liga cada pasta de imagem ao seu diagnóstico e não infere um diagnóstico ausente ou ambíguo: sequências cujo rótulo não pôde ser resolvido de forma inequívoca foram excluídas e registradas em relatório para adjudicação pelo orientador. Foram excluídos, por exemplo, pacientes com múltiplas lesões de tipos biopsiados distintos que não podem ser associadas por nome ao vídeo correspondente, registros em que o tipo de câncer contradiz o texto do laudo, e pastas cujo paciente está ausente da planilha; casos de pré-câncer também foram excluídos da tarefa binária. O resultado é um conjunto de 168 sequências utilizáveis (106 benignas e 62 malignas), de 93 pacientes — ou seja, em média cerca de 1,8 sequência por paciente, já que um mesmo paciente pode contribuir com mais de uma lesão biopsiada de forma independente (por exemplo, uma lesão benigna e uma lesão suspeita, investigadas

Figura 3 – Amostras representativas da base sob rótulo de biópsia: lesões benignas (linha superior) e malignas (linha inferior), de pacientes distintos. As três primeiras colunas mostram os termogramas em tons de cinza (canal único); a última reapresenta um exemplo em falsa cor (*colormap inferno*), apenas para visualização.



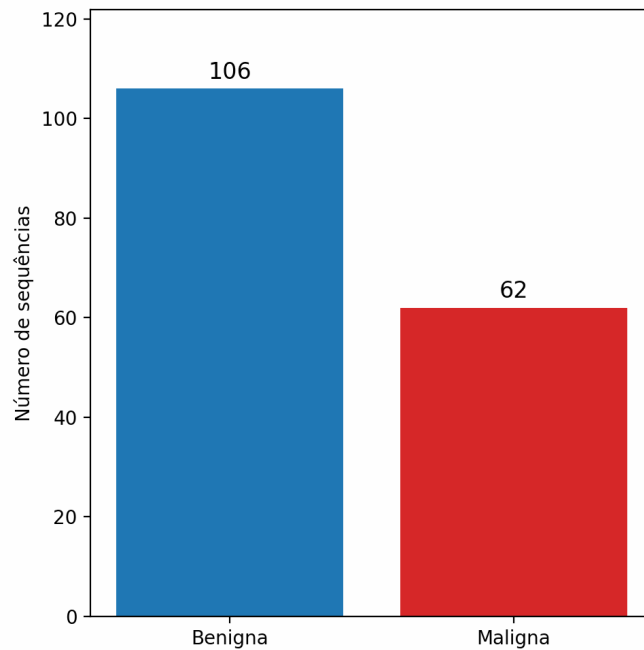
Fonte: Elaborada pelo autor (2026).

separadamente), cada lesão constituindo uma sequência térmica própria e sendo tratada como uma unidade de análise distinta —, com 12 sequências excluídas por ambiguidade (7 inconsistentes, 2 multilesão, 2 ausentes da planilha e 1 indeterminada). A Figura 4 mostra a distribuição resultante entre as classes. Essa possibilidade de múltiplas sequências por paciente é, aliás, a razão pela qual o agrupamento por paciente (e não apenas por sequência) na validação cruzada é uma exigência metodológica, detalhada na Seção 4.6: sem esse agrupamento, sequências do mesmo paciente poderiam aparecer simultaneamente em treino e teste, vazando identidade individual entre os conjuntos.

4.3 Pré-processamento

Cada quadro foi submetido ao seguinte pipeline: carregamento em tons de cinza, remoção de ruído por filtro gaussiano 5×5 , conversão dos valores de *uint8* para ponto flutuante em $[0, 1]$ (divisão por 255, um reescalonamento constante idêntico para todos os quadros, que preserva as diferenças relativas de intensidade) e redimensionamento para 224×224 px por interpolação de área (*INTER_AREA*); ao gravar, os valores são reescalonados para $[0, 255]$ e salvos em PNG *uint8*. Cabe notar que os pixels representam intensidades de 8 bits, e não temperaturas radiométricas calibradas. Uma decisão metodológica relevante foi não aplicar equalização de histograma: como as relações de intensidade entre regiões e quadros podem conter informação relevante para a classificação, a equalização poderia

Figura 4 – Distribuição das 168 sequências utilizáveis por classe, sob rótulo confirmado por biópsia.



Fonte: Elaborada pelo autor (2026).

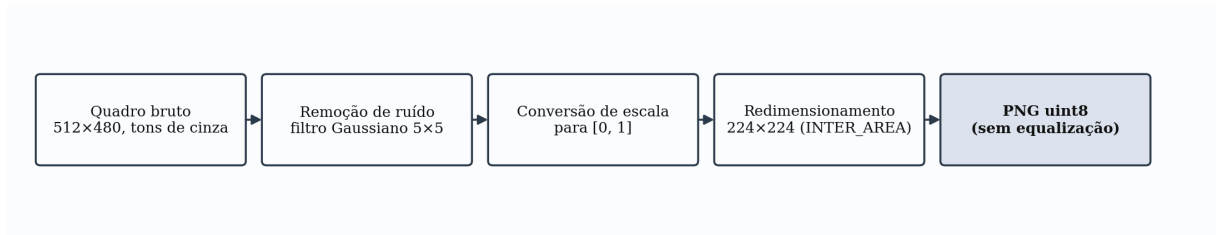
destruir parte dessa estrutura. A padronização final por média e desvio-padrão foi calculada, em cada iteração da validação cruzada, exclusivamente sobre as três dobras de treinamento daquela iteração e aplicada a seus conjuntos de treino, validação e teste (os valores média = 0,4811 e desvio = 0,2242 são de uma dessas dobras, a título de exemplo). Como o mesmo par é aplicado a todos os quadros de uma iteração, preserva-se a comparabilidade de intensidade entre quadros de uma mesma sequência; a padronização é feita em tempo de execução, no *DataLoader*, e não gravada em disco. No treinamento, aplicou-se aumento de dados geométrico leve (espelhamento horizontal e rotação de $\pm 10^\circ$); os conjuntos de validação e teste receberam apenas as transformações determinísticas (conversão de escala, redimensionamento e padronização por média e desvio-padrão), sem aumento de dados. A Figura 5 ilustra a sequência de etapas aplicada a cada quadro, e a Figura 6 mostra seu efeito sobre quadros reais da base.

4.4 Arquiteturas avaliadas

Foram avaliados quatro modelos principais, todos adaptados à entrada de canal único, além de um *ensemble* formado pela combinação de três deles (Capítulo 5).

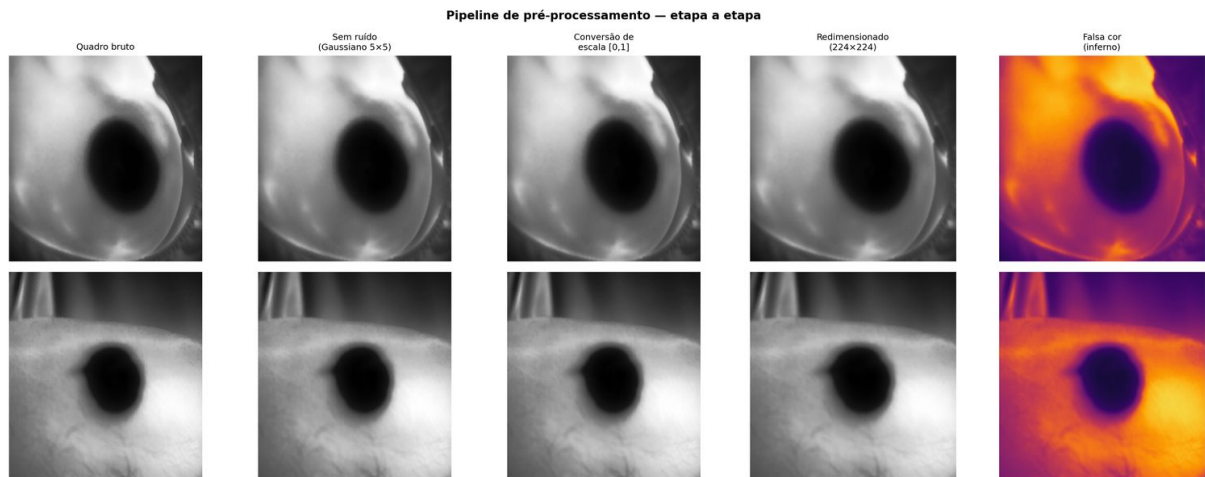
A ThermalCNN é uma rede convolucional própria, composta por três blocos convolucionais (convolução, normalização em lote (IOFFE; SZEGEDY, 2015), ReLU e *max pooling*; $1 \rightarrow 32 \rightarrow 64 \rightarrow 128$ canais) seguidos de um classificador (*pooling* adaptativo, camada densa de 256 unidades, *dropout* (SRIVASTAVA et al., 2014) de 0,5 e camada de

Figura 5 – Etapas do pré-processamento aplicado a cada quadro térmico, do quadro bruto ao PNG final. A equalização de histograma foi intencionalmente omitida para preservar as relações de intensidade térmica entre regiões e entre quadros.



Fonte: Elaborada pelo autor (2026).

Figura 6 – Efeito do pré-processamento sobre quadros térmicos reais da base (uma lesão por linha): quadro original, após remoção de ruído, após a conversão de escala para $[0, 1]$ e após redimensionamento, todos em tons de cinza; a última coluna reapresenta o quadro em falsa cor (*colormap inferno*), apenas para visualização. Como a conversão de escala é linear, as três primeiras etapas parecem visualmente semelhantes.



Fonte: Elaborada pelo autor (2026).

saída), totalizando 126.658 parâmetros treináveis; sua arquitetura é detalhada na Figura 7. A ThermalCNN foi avaliada como classificador independente, treinado do zero.

Como estratégia de transferência de aprendizado, adaptaram-se arquiteturas pré-treinadas na base ImageNet (RUSSAKOVSKY et al., 2015) para aceitar um único canal, calculando a média dos pesos da primeira convolução ao longo da dimensão de cor, o que preserva as representações pré-treinadas sem replicar canais. O *backbone* adotado nos experimentos reportados foi a ResNet18 (HE et al., 2016) (a EfficientNet-B0 (TAN; LE, 2019) também foi implementada, mas não integra os resultados finais). A ResNet18 foi treinada em duas fases: uma primeira com o *backbone* congelado (10 épocas) e uma segunda de ajuste fino das últimas camadas com taxa de aprendizado reduzida.

Por fim, os modelos clássicos, máquina de vetores de suporte (SVM) (CORTES; VAPNIK, 1995) e *Random Forest* (RF) (BREIMAN, 2001), foram treinados não sobre

atributos artesanais (HOG/LBP), mas sobre *embeddings* extraídos de uma CNN. Os *embeddings* foram extraídos da saída do *backbone* da ResNet18 — a arquitetura de transferência de aprendizado adotada no desenho experimental —, após o *pooling* global e antes da camada classificadora: substituindo essa camada por uma identidade, expõe-se a representação de 512 dimensões da rede. A ResNet18 foi ajustada separadamente em cada dobra, utilizando-se apenas o *checkpoint* selecionado na validação daquela dobra. Essas representações por quadro foram então agregadas em nível de sequência (Seção 4.6) e, conforme a configuração experimental, concatenadas aos atributos escalares de dinâmica térmica. A ThermalCNN, por sua vez, não teve suas representações utilizadas pelos modelos clássicos. Essa estratégia de usar *embeddings* de CNN em vez de atributos manuais alinha-se à de Magalhães et al. (2021).

Figura 7 – Arquitetura da ThermalCNN treinada do zero. Três blocos convolucionais reduzem progressivamente a resolução espacial e ampliam a profundidade de canais; a camada densa de 256 dimensões alimenta exclusivamente o classificador próprio da rede. Os modelos clássicos (SVM/RF) não utilizam essa representação — seus *embeddings* são extraídos separadamente da ResNet18 (512 dimensões), conforme a Seção 4.4.



Fonte: Elaborada pelo autor (2026).

4.5 Codificação do sinal dinâmico

Como o sinal discriminativo, por hipótese, está associado à dinâmica térmica, duas estratégias foram implementadas para capturá-lo. A primeira, utilizada nos modelos clássicos, baseia-se em atributos escalares de dinâmica térmica: para cada sequência, resume-se a trajetória de intensidade média dos quadros (inclinação de aquecimento/resfriamento, diferença entre o primeiro e o último quadro e variabilidade), formando um vetor concatenado ao *embedding* antes dos classificadores clássicos. Variantes desses atributos foram exploradas, restringindo o cálculo aos pixels de maior intensidade de cada quadro (segmentados por limiar de Otsu (OTSU, 1979), como aproximação não supervisionada da região térmica de interesse, predominantemente peri-lesional) e ampliando o conjunto para descritores de forma da curva de reaquecimento e de heterogeneidade espacial.¹

¹ Cabe uma ressalva: como a lesão aparece nos termogramas como uma região de *menor* intensidade (mais fria) que a pele ao redor, a máscara de maior intensidade (Otsu) tende a capturar sobretudo o

A segunda estratégia foi uma entrada temporal multicanal para as próprias CNNs: em vez de um quadro, empilhavam-se $C = 5$ quadros ao longo da sequência (o quadro-âncora e quatro quadros amostrados em posições igualmente espaçadas, ou suas diferenças em relação ao âncora, formando um mapa explícito de reaquecimento), adaptando a primeira convolução para os cinco canais. Essa codificação foi implementada e testada, mas, como reportado no Capítulo 5, mostrou-se prejudicial, de modo que a configuração padrão adotada foi a de quadro único.

4.6 Protocolo de validação

A validação foi conduzida por validação cruzada de 5 dobras agrupada por paciente, sobre os 93 pacientes da base. Todos os quadros de um mesmo paciente pertencem sempre à mesma dobra (nunca se divide por *frame*), o que evita o vazamento de dados (*data leakage*) que ocorreria caso quadros quase idênticos do mesmo paciente aparecessem simultaneamente em treino e teste. Os pacientes são distribuídos entre as dobras em ordem estratificada pela razão de classes, de forma que cada paciente é usado exatamente uma vez como teste e uma vez como validação. Em cada uma das cinco iterações, uma dobra é reservada para teste, a dobra seguinte para validação (usada no *early stopping* e na seleção do melhor modelo) e as três dobras restantes para treino — ou seja, aproximadamente 100 sequências de treino, 34 de validação e 34 de teste por iteração.

Duas salvaguardas complementares foram adotadas. Primeiro, a ponderação de classe por dobra: a razão global entre classes é de aproximadamente 63/37 (106 benignas e 62 malignas), e a distribuição de cada dobra pode desviar-se desse valor. Como salvaguarda, calcularam-se pesos por frequência inversa a partir do conjunto de treino de cada dobra, alimentados à função de perda; sem essa ponderação, uma dobra desfavorável pode fazer o modelo colapsar na classe majoritária. Segundo, o ajuste de hiperparâmetros de SVM e RF foi feito com busca aleatória sob *GroupKFold* agrupado por paciente, evitando que a validação interna da busca sofresse do mesmo vazamento. Essa busca interna utilizou exclusivamente os pacientes das três dobras externas de treinamento; as dobras externas de validação (usada no *early stopping* das redes) e de teste não participaram da seleção dos hiperparâmetros dos modelos clássicos.

Além disso, os modelos clássicos operam em nível de sequência: os *embeddings* por quadro de cada sequência são agregados (por média, ou por média, desvio e máximo) em um único vetor, de modo que SVM e RF treinam e predizem uma linha por sequência (cerca de 100 sequências de treino por dobra, extraídas das mesmas três dobras de treino usadas pelas redes). Isso alinha a unidade de treinamento dos modelos clássicos à unidade

tecido peri-lesional mais quente; trata-se, portanto, de uma aproximação grosseira e não supervisionada de uma região térmica peri-lesional de interesse, e não de uma segmentação da lesão propriamente dita.

principal de avaliação e elimina a inflação causada pelo tratamento de quadros quase duplicados como amostras independentes.

4.7 Métricas e reporte estatístico

A avaliação parte da matriz de confusão construída em nível de sequência, tratando a classe maligna (câncer) como positiva e a benigna como negativa. Em relação ao rótulo confirmado por biópsia, cada sequência classificada enquadra-se em uma de quatro situações: verdadeiro positivo (VP), câncer corretamente identificado; falso negativo (FN), câncer classificado como benigno, o erro de maior gravidade clínica; falso positivo (FP), lesão benigna classificada como câncer (alarme falso); e verdadeiro negativo (VN), benigna corretamente identificada.

As redes convolucionais são treinadas com quadros individuais, subamostrados a cada dez ($frame_stride = 10$), o que reduz os 108.548 quadros para cerca de 10 mil no total; em cada dobra, o treino utiliza aproximadamente 6 mil desses quadros (três das cinco dobras). Não há balanceamento do número de quadros por sequência, de modo que sequências mais longas contribuem com mais exemplos ao treinamento (ameaça retomada na Seção 6.7). Na predição, cada quadro recebe uma probabilidade e a predição da sequência é obtida pela média dessas probabilidades, seguida do *argmax*. Dessa forma, tanto as redes quanto os modelos clássicos produzem exatamente uma predição por sequência, e é sobre elas que a matriz de confusão e as métricas são calculadas. O “limiar padrão” reportado corresponde a esse *argmax* (equivalente a 0,5 sobre a probabilidade da classe positiva); o ajuste do limiar por validação (índice de Youden) foi avaliado, mas mantido desativado por instabilidade com apenas ~ 34 sequências de validação.

A partir dessas contagens definiram-se as métricas das Equações 4.1–4.5:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (4.1)$$

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (4.2)$$

$$\text{Sensibilidade (recall)} = \frac{VP}{VP + FN} \quad (4.3)$$

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (4.4)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \quad (4.5)$$

No contexto de câncer, a sensibilidade (Equação 4.3) recebe atenção especial, pois quantifica a fração de lesões malignas efetivamente detectadas: um falso negativo corresponde a uma lesão maligna não sinalizada, de consequências mais graves do que um alarme falso.

O treinamento das redes minimizou a entropia cruzada ponderada por classe. Para N amostras, com rótulo verdadeiro y_i , probabilidade p_{i,y_i} atribuída pelo modelo à classe correta e peso de classe w_{y_i} (frequência inversa calculada por dobra), a função de perda é dada pela Equação 4.6:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N w_{y_i} \log(p_{i,y_i}) \quad (4.6)$$

O peso w_{y_i} compensa o desbalanceamento de classes por dobra, impedindo que uma partição desfavorável leve o modelo a colapsar na classe majoritária.

Além das métricas de limiar fixo, a curva ROC descreve o compromisso entre a sensibilidade (taxa de verdadeiros positivos) e a fração $1 - \text{especificidade}$ (taxa de falsos positivos) à medida que o limiar de decisão varia. A AUC (área sob essa curva) resume o compromisso em um único valor, interpretável como a probabilidade de o modelo atribuir um escore maior a uma sequência maligna do que a uma benigna escolhidas ao acaso; um classificador aleatório obtém $\text{AUC} = 0,5$ e um perfeito, $\text{AUC} = 1,0$. Por ser insensível ao limiar, a AUC é conveniente para comparar modelos; ainda assim, ela é reportada em conjunto com a sensibilidade e a especificidade, pois, em um contexto de rastreamento, o desempenho em um limiar específico também é relevante.

A métrica principal de comparação é a AUC agrupada (*pooled*). A justificativa é estatística: com apenas 168 sequências, cada dobra de teste contém cerca de 34 sequências, e a média de cinco valores de AUC (um por dobra) é um estimador ruidoso. Em vez disso, reuniram-se as predições de todas as dobras, cada sequência testada exatamente uma vez, e calculou-se uma única AUC sobre esse conjunto agrupado, acompanhada de um intervalo de confiança de 95% por reamostragem *bootstrap* (EFRON; TIBSHIRANI, 1993) estratificada por classe. Como algumas sequências pertencem ao mesmo paciente, verificou-se também uma variante do *bootstrap* agrupada por paciente (reamostrando pacientes, e não sequências), que produziu intervalos praticamente idênticos. Para comparar dois modelos, contudo, não se recorre à simples sobreposição dos intervalos individuais — que não constitui um teste formal —, mas ao *bootstrap* pareado da diferença de AUC (a mesma reamostragem aplicada aos dois modelos), cujo intervalo de confiança indica se a diferença é distinguível de zero. Esses testes são reportados no Capítulo 5.

4.8 Desenho experimental e ablações

A parte experimental não se limitou a treinar os modelos uma vez: conduziu-se uma sequência de ablações controladas, cada uma isolando o efeito de uma decisão de projeto, todas registradas como execuções independentes. As principais foram: (i) um *baseline* funcional com quadro único; (ii) a introdução da entrada temporal multicanal, seguida de (iii) uma execução de isolamento controlado que mantém rede, rótulos e dobras idênticos e apenas liga ou desliga a entrada temporal; (iv) a comparação entre rótulos por biópsia e por pasta sob metodologia idêntica; e (v) uma execução de enriquecimento (*features* restritas à ROI de maior intensidade, conjunto rico de atributos, *pooling* `meanstdmax` e *ensemble* por média de probabilidades de ResNet18, SVM e RF). Os resultados de cada ablação são reportados a seguir.

4.9 Hiperparâmetros

A Tabela 2 reúne os hiperparâmetros efetivamente utilizados, para reprodutibilidade. As redes foram otimizadas com Adam; SVM e RF tiveram seus hiperparâmetros ajustados por busca aleatória com validação cruzada interna agrupada por paciente (*GroupKFold*). Os valores são também persistidos, em formato legível por máquina, no arquivo `train_config.json` de cada execução.

Tabela 2 – Hiperparâmetros utilizados no treinamento e na avaliação.

Componente	Valor
<i>Redes convolucionais (ThermalCNN e ResNet18)</i>	
Otimizador	Adam
Taxa de aprendizado	10^{-3} (ResNet18, fase de ajuste fino: 10^{-4})
<i>Weight decay</i>	5×10^{-4}
Tamanho de lote	32
Épocas (máximo)	50
<i>Early stopping</i>	paciência 7 (sobre a AUC de validação)
ResNet18 — fase 1 (<i>backbone</i> congelado)	10 épocas, paciência 4
Escalonador de taxa	<i>ReduceLROnPlateau</i> (fator 0,5; paciência 3)
Função de perda	entropia cruzada ponderada por classe
<i>Modelos clássicos (busca aleatória; GroupKFold interno, 5 dobras, 12 iterações)</i>	
SVM — <i>kernel</i>	RBF
SVM — <i>C</i>	{0,1; 1; 10; 50}
SVM — γ	{ <i>scale</i> , <i>auto</i> , 0,01; 0,001}
RF — número de árvores	{100, 200, 300, 500}
RF — profundidade máxima	{sem limite, 10, 20, 30}
RF — <i>min. samples leaf</i>	{1, 2, 4}
Pesos de classe (SVM/RF)	<i>balanced</i>
<i>Avaliação</i>	
Reamostragens <i>bootstrap</i>	2000
Semente aleatória	42

Fonte: Elaborada pelo autor (2026).

5 Resultados

Este capítulo reporta os resultados da parte experimental. A configuração de referência, por ser a mais recente e a mais completa, combina validação cruzada de 5 dobras agrupada por paciente, rótulos por biópsia, modelos clássicos em nível de sequência, atributos térmicos restritos à ROI de maior intensidade, conjunto rico de atributos, *pooling* *meanstdmax* e *ensemble* por média de probabilidades de ResNet18, SVM e RF, avaliada sobre as 168 sequências (106 benignas / 62 malignas)¹. Os valores de AUC agrupada e seus intervalos de confiança foram recalculados diretamente das predições por sequência salvas nessa execução.

5.1 Resultado central: teto de desempenho observado

A Tabela 3 apresenta o resultado principal. *Random Forest*, SVM e ResNet18 (com ajuste fino) atingem uma AUC agrupada em torno de 0,64–0,66, com intervalos de confiança de 95% que se estendem até cerca de 0,74 e se sobrepõem amplamente. Um *bootstrap* pareado da diferença de AUC confirma que essa diferença não é distinguível de zero: entre o melhor modelo (RF) e a ResNet18, a diferença é de apenas +0,012, com IC 95% de $[-0,024; +0,048]$ — praticamente inalterado sob reamostragem agrupada por paciente ($[-0,027; +0,048]$). O modelo clássico (RF) é, portanto, tão competitivo quanto a rede profunda pré-treinada. A ThermalCNN treinada do zero, em contraste, permanece em nível de acaso (AUC $\approx 0,48$): sua diferença para o RF é de +0,177, com IC 95% de $[+0,062; +0,294]$, que exclui o zero. O *ensemble* não supera o melhor modelo individual. A Tabela 4 reporta o *bootstrap* pareado da diferença de AUC do RF (melhor modelo) frente aos demais: todas as comparações entre os modelos de melhor desempenho incluem o zero, ao passo que a comparação com a ThermalCNN o exclui.

As matrizes de confusão agrupadas (Figura 8) tornam esse comportamento visualmente explícito.

Dois pontos merecem destaque já nesta tabela. Primeiro, a sensibilidade dos melhores modelos é modesta ($\approx 0,44$): eles acertam a maioria das lesões benignas (especificidade $\approx 0,80$), mas deixam passar mais da metade dos casos de câncer no limiar padrão, como se vê na linha “Câncer” da Figura 8 — limitação discutida no Capítulo 6. Segundo, o desempenho de acaso da ThermalCNN treinada do zero é reportado explicitamente como um resultado negativo: com apenas ~ 100 sequências de treino por dobra, a rede não conseguiu aprender representações discriminativas úteis a partir do zero, resultado compatível

¹ No repositório, essa configuração corresponde à execução `roi_rich_ensemble`; os demais experimentos citados neste capítulo (184810, 112444) seguem a mesma convenção de nomenclatura.

Tabela 3 – Desempenho em nível de sequência na execução de referência (168 sequências). AUC agrupada com IC 95% por *bootstrap* estratificado; sensibilidade, especificidade, precisão, F1-*score* e acurácia no limiar padrão. O *ensemble* combina RF, SVM e ResNet18; a ThermalCNN foi treinada do zero.

Modelo	AUC	IC 95%	Sens.	Espec.	Prec.	F1	Acur.
Random Forest	0,660	[0,573; 0,746]	0,44	0,83	0,60	0,50	0,69
SVM	0,640	[0,547; 0,727]	0,47	0,82	0,60	0,53	0,69
ResNet18	0,648	[0,559; 0,734]	0,44	0,80	0,56	0,49	0,67
<i>Ensemble</i>	0,652	[0,564; 0,736]	0,42	0,82	0,58	0,49	0,67
ThermalCNN	0,483	[0,391; 0,575]	0,53	0,41	0,34	0,42	0,45

Nota: sob *bootstrap* agrupado por paciente, os intervalos são praticamente idênticos — p.ex., RF [0,576; 0,750], ResNet18 [0,561; 0,739] e ThermalCNN [0,403; 0,565].

Fonte: Elaborada pelo autor (2026).

Tabela 4 – Diferença de AUC agrupada entre o *Random Forest* (melhor modelo) e os demais, com intervalo de confiança de 95% por *bootstrap* pareado (estratificado por classe). Um intervalo que contém o zero indica ausência de diferença detectável.

Comparação	Δ AUC	IC 95% da diferença
RF – SVM	+0,020	[−0,018; +0,060]
RF – ResNet18	+0,012	[−0,024; +0,048]
RF – <i>Ensemble</i>	+0,008	[−0,013; +0,030]
RF – ThermalCNN	+0,177	[+0,062; +0,294]

Fonte: Elaborada pelo autor (2026).

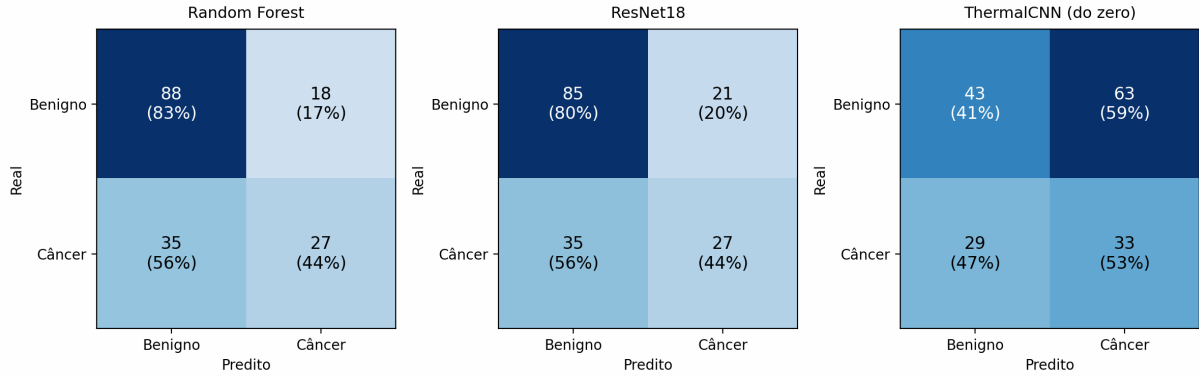
com a limitação amostral e com a alta correlação entre os quadros; é a transferência de aprendizado (ResNet18 pré-treinada) que viabiliza o sinal.

As curvas ROC (Figura 9) confirmam esse quadro: *Random Forest*, SVM e ResNet18 acompanham-se de perto acima da diagonal do acaso, enquanto a ThermalCNN praticamente coincide com ela. Já a Figura 10 sintetiza o achado central do trabalho: os intervalos de confiança dos três modelos individuais de melhor desempenho e do *ensemble* se sobrepõem quase por completo, ao passo que a ThermalCNN apresenta estimativa pontual inferior e intervalo que inclui o nível de acaso. Em conjunto com o *bootstrap* pareado reportado acima, esse quadro sustenta que a diferença entre AUC 0,64 e 0,66 não é distinguível nesta amostra, ponto retomado no Capítulo 6.

5.2 Ablações e resultados negativos

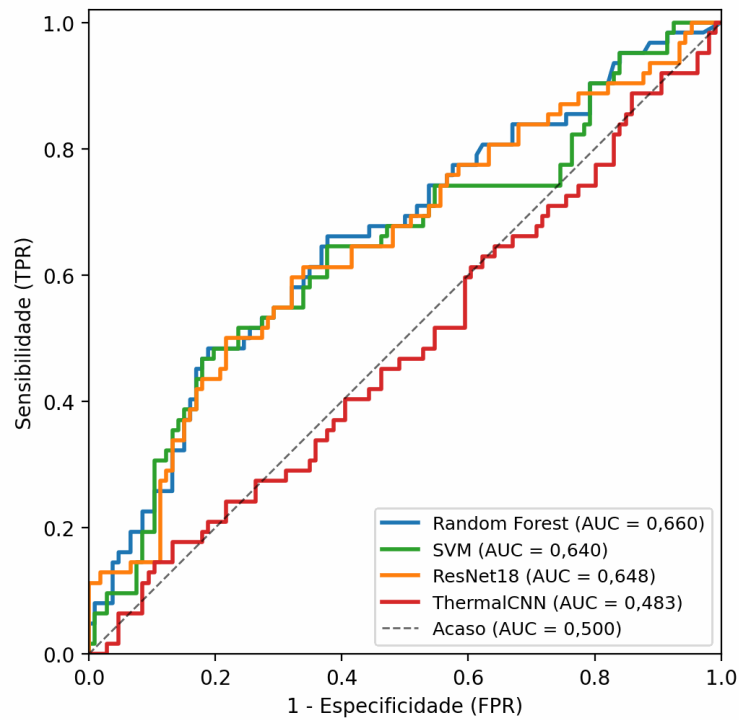
Uma parte central do rigor deste trabalho está em documentar as intervenções que não funcionaram, evitando o viés de reportar apenas o que deu certo.

Figura 8 – Matrizes de confusão agrupadas das 168 seqüências (execução de referência), para o *Random Forest*, a *ResNet18* e a *ThermalCNN*. Percentuais normalizados por classe real: a diagonal da linha “Benigno” corresponde à especificidade e a da linha “Câncer”, à sensibilidade (classe maligna).



Fonte: Elaborada pelo autor (2026).

Figura 9 – Curvas ROC agrupadas dos modelos na execução de referência (168 seqüências).

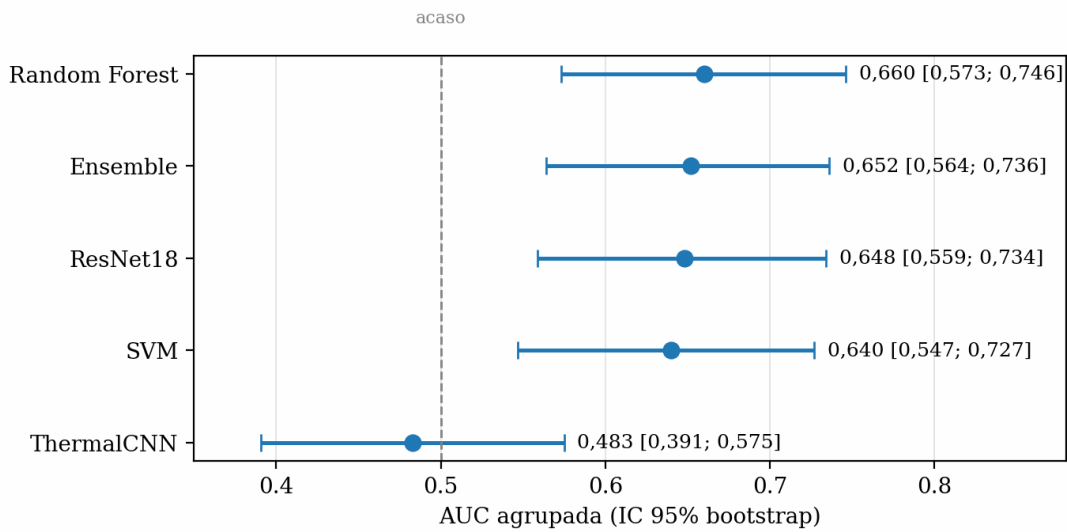


Fonte: Elaborada pelo autor (2026).

5.2.1 Entrada temporal por empilhamento de canais

A hipótese inicial era que fornecer vários quadros empilhados como canais permitiria à CNN capturar a dinâmica de aquecimento. O efeito medido foi o oposto. Em uma execução de isolamento controlado (rede, rótulos e dobras idênticos, variando apenas a entrada temporal), desligar a entrada temporal elevou a AUC dos três modelos: ResNet18 de 0,634 para 0,675, RF de 0,632 para 0,677 e SVM de 0,624 para 0,691 (ganhos de +0,04 a

Figura 10 – AUC agrupada e intervalo de confiança de 95% (*bootstrap* estratificado) por modelo. A linha tracejada indica o nível de acaso (AUC = 0,5).



Fonte: Elaborada pelo autor (2026).

+0,07). Esses valores correspondem à AUC agregada em nível de paciente (média das cinco dobras) e, portanto, não são diretamente comparáveis à AUC agrupada por sequência reportada nas tabelas. A interpretação é que empilhar quadros como canais permite à convolução combinar as posições temporais, mas sem uma modelagem sequencial explícita da ordem, da duração ou das dependências entre instantes, o que adiciona superfície de sobreajuste sem capturar a evolução da sequência. A hipótese do reaquecimento não é refutada, mas essa codificação não a captura; o caminho que efetivamente carrega o sinal dinâmico são os atributos escalares de dinâmica térmica já fornecidos a SVM/RF.

5.2.2 Rótulo por biópsia versus por pasta

Comparando execuções sob metodologia idêntica, a diferença de desempenho entre as duas fontes de rótulo é pequena ($\sim 0,01$ – $0,02$ de AUC agrupada) e, na configuração comparável, ligeiramente favorável à biópsia (Tabela 5: RF com 0,657 sob rótulo por biópsia *versus* 0,643 sob rótulo por pasta). Ou seja, a rerotulagem por biópsia, epistemicamente mais correta, não penaliza o desempenho. A queda mais acentuada observada em uma execução anterior (de $\approx 0,71$ para $\approx 0,63$; execução com entrada temporal, não incluída nas tabelas) devia-se majoritariamente à entrada temporal, e não à rerotulagem por biópsia — o que reforça a adoção do rótulo por biópsia, mais fidedigno por corresponder ao diagnóstico histopatológico.

5.2.3 Enriquecimento de atributos, *pooling* e *ensemble*

O conjunto de melhorias da execução de enriquecimento (atributos restritos à ROI de maior intensidade, conjunto rico de descritores, *pooling* *meanstdmax* e *ensemble*) também não moveu o teto frente ao *baseline* de quadro único: o RF passou de 0,657 para 0,660 (dentro do ruído), o SVM caiu de 0,656 para 0,640 e o *ensemble* (0,652) ficou abaixo do melhor modelo individual, o que indica que combinar modelos correlacionados sobre o mesmo teto de sinal não agrega. A ResNet18 não reaparece entre as linhas de enriquecimento porque suas predições são idênticas às do *baseline*: essas mudanças afetam apenas a entrada dos modelos clássicos, e não o treinamento da rede. A Tabela 5 consolida a comparação entre execuções e mostra o ponto essencial: todos os intervalos de confiança se sobrepõem amplamente.

Tabela 5 – AUC agrupada (IC 95%) por configuração experimental e modelo. As estimativas são próximas e os intervalos amplamente sobrepostos. As diferenças entre configurações são apresentadas de forma descritiva; a comparação pareada formal (Tabela 4) refere-se aos modelos da execução de referência, para os quais não se detecta diferença distinguível de zero.

Configuração	Modelo	AUC	IC 95%
Baseline (biópsia)	ResNet18	0,648	[0,559; 0,734]
Baseline (biópsia)	Random Forest	0,657	[0,569; 0,743]
Baseline (biópsia)	SVM	0,656	[0,571; 0,736]
Rótulo por pasta	Random Forest	0,643	[0,562; 0,724]
Enriquecimento (biópsia)	Random Forest	0,660	[0,573; 0,746]
Enriquecimento (biópsia)	SVM	0,640	[0,547; 0,727]
Enriquecimento (biópsia)	<i>Ensemble</i>	0,652	[0,564; 0,736]

Nota: no repositório, as configurações correspondem às execuções 184810 (*baseline*), 112444 (rótulo por pasta) e *roi_rich_ensemble* (enriquecimento).

Fonte: Elaborada pelo autor (2026).

Em síntese, o melhor resultado obtido é uma AUC agrupada de aproximadamente 0,66 (com IC de 95% chegando a $\sim 0,74$), alcançado por RF, SVM e ResNet18 sem diferença detectável em relação ao RF, e nenhuma das intervenções testadas (entrada temporal por canais, atributos de ROI mais ricos, *pooling* mais expressivo ou *ensemble*) apresentou ganho consistente sobre o *baseline* nas comparações descritivas realizadas. A interpretação desse padrão é o objeto do próximo capítulo.

6 Discussão

Os resultados do Capítulo 5 convergem para um achado central: existe um teto de desempenho observado em torno de 0,66 de AUC agrupada, atingido por modelos de naturezas distintas sem diferenças detectáveis em relação ao *Random Forest* no *bootstrap* pareado, e que nenhuma das intervenções de engenharia testadas conseguiu ultrapassar. Este capítulo argumenta que esse teto reflete a combinação de duas frentes distintas, tratadas em seções separadas a seguir: uma restrição de tamanho amostral (Seção 6.1), evidenciada pelo padrão de convergência das intervenções testadas, e um conjunto de limitações metodológicas que atuam independentemente do número de exemplos — a representação em *uint8*, a aproximação da ROI por limiar de Otsu, a ausência de segmentação dedicada e de uma arquitetura genuinamente temporal, e o uso de uma única base (Seção 6.2). Os experimentos permitem caracterizar bem a primeira frente; a segunda não pôde ser isolada estatisticamente nas configurações avaliadas, mas não deve ser lida como secundária apenas por isso.

6.1 Evidências de limitação pelo tamanho amostral

A principal evidência de que o tamanho amostral constitui uma restrição importante é o padrão de resultados negativos observado nas diferentes intervenções de modelagem. Foram testadas intervenções que atacam frentes independentes do problema: a arquitetura (ResNet18 com ajuste fino *versus* rede própria), a representação da dinâmica (entrada temporal multicanal, atributos ricos restritos à ROI de maior intensidade), a agregação (*pooling meanstdmax*) e a combinação de modelos (*ensemble*). Todas resultaram em valores contidos no intervalo de confiança do *baseline*. Quando aprimoramentos independentes e razoáveis convergem para a mesma faixa, e essa faixa tem largura de aproximadamente $\pm 0,09$ de AUC, a interpretação mais parcimoniosa é que, sob o *pipeline* e as representações testadas, a incerteza amostral domina os ganhos marginais de modelagem — e não que cada intervenção tenha falhado por má execução.

Essa incerteza é uma consequência direta do número de exemplos. Com 168 sequências e 93 pacientes, cada dobra de teste contém cerca de 34 sequências; foi justamente por isso que se adotou a AUC agrupada com *bootstrap* como unidade de reporte, em vez da média por dobra. Os intervalos de confiança largos e mutuamente sobrepostos não são um detalhe de apresentação: são a manifestação quantitativa do fato de que a amostra não permite distinguir um modelo de AUC 0,64 de outro de AUC 0,66. Nesse regime, o retorno de mais engenharia de atributos ou de arquitetura tende a ser decrescente; ganhos futuros provavelmente dependerão da ampliação da base e/ou de representações temporais

e radiométricas mais adequadas, validadas sobre uma amostra maior.

6.2 Limitações metodológicas independentes do tamanho amostral

A convergência descrita na Seção 6.1 mostra que o tamanho amostral é uma restrição importante, mas atribuir a ele a totalidade do teto de 0,66 seria uma simplificação: pelo menos quatro limitações metodológicas atuam de forma independente do número de exemplos, e cada uma delas, isoladamente, já poderia impor um teto de desempenho mesmo numa amostra maior.

Primeiro, a aproximação da região de interesse por limiar de Otsu captura a região peri-lesional mais quente, não a lesão propriamente dita: a inspeção visual das máscaras resultantes mostra que a área realçada contorna, mas não cobre, a região mais fria correspondente ao núcleo da lesão. Essa região adjacente pode carregar sinal fisiologicamente relevante — o aquecimento do tecido ao redor também reflete atividade vascular e metabólica — mas é, ainda assim, um *proxy* impreciso e sistematicamente enviesado, não apenas ruidoso. Uma segmentação supervisionada da lesão (por exemplo, via U-Net treinada com anotação de especialista, como em [Soto e Godoy \(2024\)](#)) tende a concentrar o sinal de forma mais direta, independentemente de quantos exemplos estivessem disponíveis para o restante do *pipeline*.

Segundo, a conversão das imagens para tons de cinza *uint8* descarta resolução radiométrica: diferenças finas de temperatura absoluta, potencialmente relevantes para diferenciar tecidos, são comprimidas em 256 níveis relativos a cada quadro, sem calibração térmica em graus Celsius nem correção por emissividade. Trata-se de um limite da representação dos dados adquiridos, não da quantidade de exemplos disponíveis — mais dados no mesmo formato não resolveriam essa perda de informação.

Terceiro, a codificação temporal avaliada (empilhamento multicanal) não modela explicitamente a ordem, a duração ou as dependências entre instantes da sequência de reaquecimento, ainda que tenha sido treinada sobre toda a amostra disponível; conforme discutido na Seção anterior, seu desempenho inferior ao *baseline* é compatível tanto com falta de dados quanto com uma codificação estruturalmente inadequada para capturar a dinâmica — as duas explicações não são mutuamente exclusivas, e os experimentos realizados não permitem isolar a contribuição de cada uma.

Quarto, os resultados restringem-se a uma única base privada, adquirida sob um protocolo específico, em um único centro e com um único equipamento. Mesmo que essa base fosse consideravelmente ampliada, a validade externa para outros equipamentos, protocolos de resfriamento e populações permaneceria não demonstrada, pois trata-se de uma limitação de diversidade da fonte de dados, não apenas de volume.

Essas quatro limitações coexistem com a restrição amostral discutida na seção anterior e não competem com ela por importância: a ampliação da base é uma condição necessária para reduzir a incerteza estatística observada, mas não é, isoladamente, suficiente para superar o teto de desempenho, caso a representação, a segmentação e a modelagem temporal permaneçam como estão.

6.3 Comparação com a literatura

Os números obtidos aqui são visivelmente mais modestos do que os reportados por trabalhos de referência, e a comparação exige cautela, pois diferenças de base, modalidade, tarefa e protocolo afetam fortemente o desempenho (VERSTOCKT et al., 2022). Magalhães et al. (2021), com uma abordagem de aprendizado profundo sobre parâmetros térmicos, relatam precisão de $\approx 0,97$ e sensibilidade de $\approx 0,94$ na diferenciação entre melanoma e nevos. Contudo, esse é um par de classes mais separável (melanoma *versus* nevo) do que a tarefa deste trabalho (lesão benigna *versus* maligna em sentido amplo, incluindo carcinomas), e a metodologia de extração de *features* por *embeddings* de CNN aqui adotada é justamente inspirada nesse trabalho. Soto e Godoy (2024), também com termografia ativa e classificação por SVM, reportam especificidades da ordem de 88% (com segmentação por especialista) a $\approx 69\%$ (com segmentação automática por U-Net) fixando a sensibilidade de referência em 85%; seu pipeline emprega segmentação dedicada da lesão e um conjunto de atributos espaço-temporais projetados especificamente, ao passo que este trabalho usa uma aproximação não supervisionada da ROI (limiar de Otsu) e atributos de dinâmica mais simples. Uma referência particularmente adequada para comparação é o estudo de Malheiros et al. (2026), conduzido no Hospital de Amor (Barretos) sobre exatamente o mesmo recorte deste trabalho — benigno *versus* maligno em sentido amplo, com padrão histopatológico —, e não sobre o par mais separável melanoma *versus* nevo. Em 68 lesões (56 malignas e 12 benignas) de 64 indivíduos, a termografia *ativa* (por resfriamento) atingiu sensibilidade de 91,1% e especificidade de 75,0%, enquanto a modalidade *passiva* ficou em apenas 17,9% de sensibilidade, apesar de 100% de especificidade. Esse contraste sustenta dois pontos deste trabalho: que a modalidade *dinâmica* é o caminho mais promissor, e que o teto modesto aqui observado não decorre da modalidade em si, mas do tamanho e das condições da amostra. Fixando a mesma sensibilidade de referência (85%), os modelos deste trabalho atingem especificidade em torno de apenas 25% (RF), muito inferior aos $\approx 69\text{--}88\%$ daqueles autores — um contraste coerente com a AUC mais modesta obtida aqui e com a maior separabilidade da tarefa por eles considerada.

Uma leitura criteriosa dessa comparação é que o presente trabalho não compete em desempenho absoluto com esses estudos, mas oferece algo metodologicamente distinto e complementar: um protocolo de avaliação conservador (validação cruzada agrupada por paciente, rótulos confirmados por biópsia e reporte por AUC agrupada com intervalo

de confiança) que expõe, em vez de mascarar, o efeito do tamanho amostral. Parte da distância entre 0,66 e os números da literatura é atribuível à dificuldade da tarefa e ao rigor do protocolo, e não apenas à qualidade dos modelos.

6.4 O desempenho de acaso da rede treinada do zero

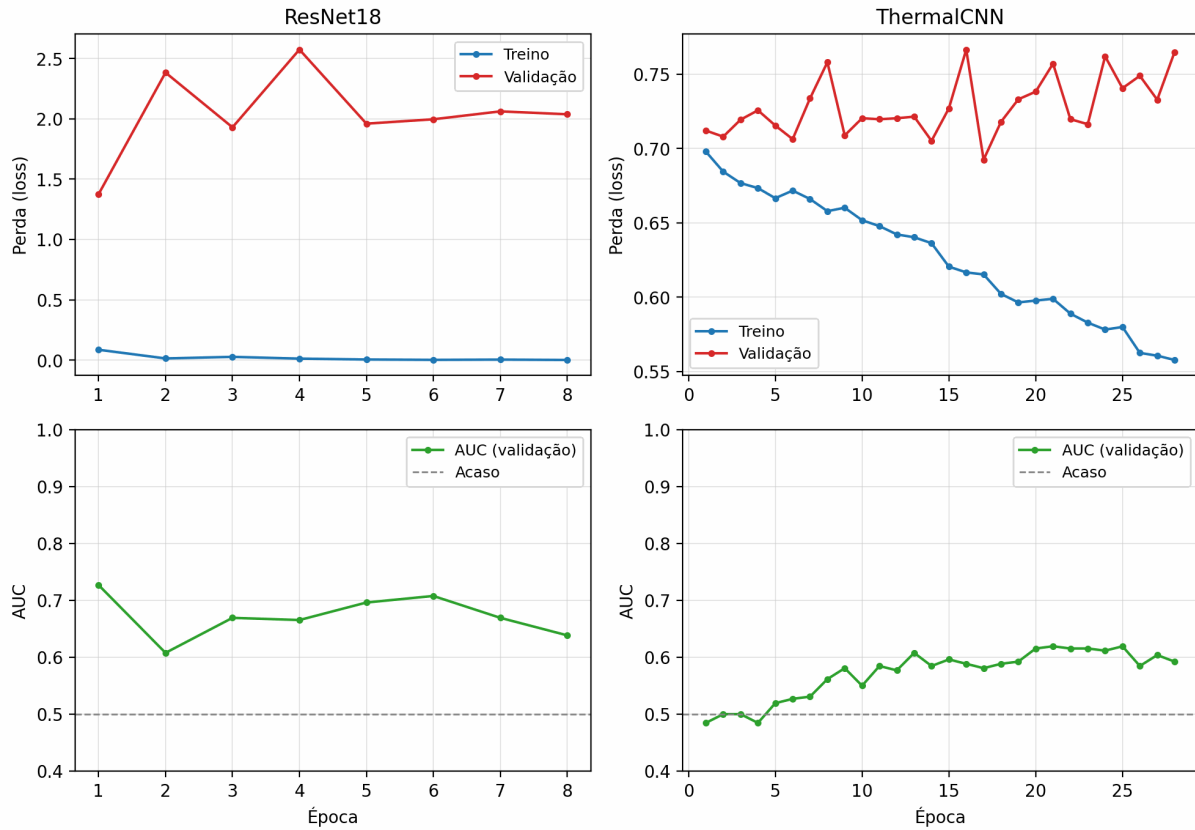
A ThermalCNN treinada do zero permanecer em nível de acaso é coerente com a hipótese do tamanho amostral, e não a contradiz. Redes convolucionais aprendem representações úteis quando dispõem de muitos exemplos; com ~ 100 sequências de treino por dobra, a rede não conseguiu induzir representações que generalizassem para o teste — resultado compatível com a limitação amostral, embora arquitetura, hiperparâmetros e a unidade de treinamento por quadro também possam ter contribuído. O fato de a transferência de aprendizado (ResNet18 pré-treinada em ImageNet, com a primeira convolução adaptada ao canal único) alcançar o teto de $\approx 0,66$ enquanto a rede própria fracassa reforça que, neste regime de dados, o conhecimento pré-treinado é o que viabiliza o sinal. Trata-se de um resultado negativo que se reporta explicitamente, por ser informativo, e não de uma falha a ocultar.

As curvas de treinamento (Figura 11) tornam esse diagnóstico concreto. Na ResNet18, a perda de treino cai rapidamente para próximo de zero enquanto a perda de validação sobe e a AUC de validação oscila em torno de 0,67, assinatura clássica de sobreajuste sob amostra pequena, contida pelo *early stopping* em poucas épocas. Na ThermalCNN, a perda de treino cai lentamente (de $\approx 0,70$ para $\approx 0,56$) enquanto a perda de validação sobe, e a AUC de validação chega a estabilizar-se em torno de 0,60 na dobra 0. Esse sinal aparente, contudo, não se traduz em desempenho de teste: os *checkpoints* assim selecionados produzem, no conjunto agrupado das cinco dobras, AUC de 0,483. Com apenas ~ 34 sequências de validação por dobra, a própria seleção de modelo é ruidosa — mais uma manifestação da restrição amostral.

6.5 Implicação clínica: sensibilidade e falsos negativos

Do ponto de vista clínico, a limitação mais relevante não é a AUC de 0,66 em si, mas a sensibilidade modesta ($\approx 0,44$ no limiar padrão) dos melhores modelos. Em rastreamento de câncer, um falso negativo (uma lesão maligna classificada como benigna) tem consequências potencialmente graves, de modo que a sensibilidade é a métrica crítica (LIU et al., 2020; RAJKOMAR; DEAN; KOHANE, 2019). O limiar de decisão pode ser deslocado para elevar a sensibilidade à custa da especificidade, mas, dada a AUC atual, esse deslocamento produziria um número elevado de falsos positivos. Em outras palavras, o classificador, no estágio atual, não está pronto para uso clínico nem como ferramenta

Figura 11 – Curvas de treinamento (dobra 0) da ResNet18 e da ThermalCNN: perda de treino e validação (acima) e AUC de validação por época (abaixo). Para a ResNet18, mostra-se a fase de ajuste fino; as 10 épocas iniciais com o *backbone* congelado foram omitidas. Para a ThermalCNN, mostra-se a fase de ajuste fino; as 10 épocas iniciais com o *backbone* congelado foram omitidas.



Fonte: Elaborada pelo autor (2026).

de triagem autônoma; seu valor é o de evidência experimental sobre a viabilidade da abordagem, situando de forma realista o que uma amostra desse porte permite alcançar.

6.6 A hipótese da dinâmica térmica e o que faltou capturar

Um ponto conceitualmente importante é que o fracasso da entrada temporal por empilhamento de canais não refuta a hipótese de que o sinal discriminativo reside na dinâmica de aquecimento. O que os experimentos mostram é que aquela codificação específica não captura a dinâmica: empilhar quadros como canais permite à convolução combinar as posições temporais, mas sem uma modelagem sequencial explícita da ordem, da duração ou das dependências entre instantes, o que adiciona parâmetros e superfície de sobreajuste. Entre as codificações temporais avaliadas, os atributos escalares de trajetória térmica mostraram-se mais adequados que o empilhamento multicanal, embora seu ganho isolado sobre os *embeddings* não tenha sido determinado de forma conclusiva. Uma arquitetura genuinamente sequencial (por exemplo, *pooling* temporal ou mecanismos de atenção sobre *embeddings* por quadro) permanece uma hipótese plausível, porém só faz sentido

testá-la sobre uma amostra maior, sob pena de repetir o padrão de sobreajuste observado.

6.7 Ameaças à validade

Quanto à *validade interna*, as principais ameaças (vazamento de dados e rótulos incorretos) foram mitigadas por construção: o agrupamento por paciente na validação cruzada e no ajuste de hiperparâmetros impede que quadros do mesmo paciente cruzem os subconjuntos, e o uso de rótulos confirmados por biópsia, com exclusão explícita de casos ambíguos, reduz o ruído de rotulagem. Uma ameaça remanescente é a unidade de treinamento das redes: embora a avaliação final seja em nível de sequência, as CNNs de quadro único são treinadas com quadros individuais. Em consequência, sequências com mais quadros exercem maior influência no treinamento, e a forte correlação entre quadros vizinhos reduz o número efetivo de observações independentes — o que reforça a interpretação de que a amostra efetiva é menor do que o número de quadros sugere. Adicionalmente, os *embeddings* usados por SVM e RF provêm de uma ResNet18 ajustada nas mesmas dobras de treino: extrator e classificadores compartilham os dados de treino de cada dobra (sem tocar validação e teste), o que acopla as duas etapas, embora não introduza vazamento para os conjuntos de avaliação. Quanto à *validade externa*, os resultados restringem-se a uma única base privada, adquirida sob um protocolo específico de termografia ativa; a generalização para outros equipamentos, protocolos e populações não está demonstrada e depende de padronização da aquisição, tema recorrente na literatura ([VERSTOCKT et al., 2022](#); [GODOY et al., 2015](#)). Quanto à *validade de construção*, a decisão de preservar o gradiente de intensidade térmica (sem equalização de histograma) é adequada à natureza do dado, mas a conversão para tons de cinza *uint8* descarta parte da resolução radiométrica original. Por fim, a *validade estatística* é limitada pelo tamanho amostral: os intervalos de confiança largos são a expressão dessa limitação, e foram reportados de forma transparente em vez de substituídos por estimativas pontuais superconfiantes.

7 Considerações Finais

Este trabalho investigou a classificação de lesões cutâneas benignas e malignas a partir de termografia infravermelha dinâmica, comparando uma rede convolucional própria, transferência de aprendizado (ResNet18) e modelos clássicos (SVM e *Random Forest*) sob um protocolo experimental rigoroso. A fundamentação teórica e os trabalhos relacionados situaram o problema, e os capítulos de materiais e métodos, resultados e discussão consolidaram a parte experimental efetivamente realizada.

Retomando os objetivos específicos formulados na introdução: o levantamento bibliográfico sobre câncer de pele, diagnóstico assistido por computador, termografia e aprendizado profundo foi realizado (Capítulos 2 e 3); a base de imagens foi organizada e caracterizada, com rótulos confirmados por biópsia; as etapas de pré-processamento foram definidas e aplicadas aos termogramas; arquiteturas convolucionais, transferência de aprendizado e classificadores clássicos foram implementados, treinados e comparados; e a avaliação por métricas de classificação, com comparação à literatura, foi conduzida sob um protocolo controlado. Os cinco objetivos específicos foram, portanto, cumpridos.

O principal resultado é um teto de desempenho observado em torno de 0,66 de AUC agrupada (intervalo de confiança de 95% chegando a $\approx 0,74$), atingido por *Random Forest*, SVM e ResNet18, sem diferença detectável entre o *Random Forest* e os demais modelos de melhor desempenho nas comparações pareadas apresentadas. Documentaram-se também os resultados negativos: a entrada temporal por empilhamento de canais, que piorou o desempenho; o enriquecimento de atributos, *pooling* e *ensemble*, que não moveu o teto; e a rede treinada do zero, que permaneceu em nível de acaso. Eles são parte essencial do rigor metodológico e sustentam a conclusão central: o tamanho amostral (168 sequências e 93 pacientes) constitui uma das principais restrições do problema nesta base, uma vez que as intervenções avaliadas permaneceram em faixa semelhante de desempenho e os testes pareados entre os modelos da execução de referência não detectaram diferença. Limitações de aquisição, representação radiométrica, segmentação e modelagem temporal também podem ter contribuído para o teto observado.

Quanto à pergunta de pesquisa — se redes neurais convolucionais conseguem aprender padrões térmicos relevantes a ponto de apoiar, de maneira útil, a distinção entre lesões benignas e malignas —, a resposta é *parcialmente afirmativa*: há sinal detectável acima do acaso, pois o limite inferior do intervalo de confiança do melhor modelo (0,573) é maior que 0,5; esse sinal, contudo, é insuficiente, nesta amostra, para utilidade clínica direta, sobretudo pela sensibilidade modesta. A hipótese inicial confirma-se, assim, apenas em parte: assinaturas térmicas discriminativas existem e são capturáveis, mas o desempenho

alcançável sob as atuais restrições de dados, representação e segmentação fica aquém do necessário para aplicação prática.

Reconhece-se como principal limitação clínica a sensibilidade modesta dos modelos, que impede, no estágio atual, o uso da abordagem como ferramenta autônoma de triagem. Como trabalhos futuros, destacam-se a ampliação da base (incluindo a expansão multi-classe já prevista), a validação em bases e protocolos independentes e a investigação de arquiteturas genuinamente sequenciais (*pooling* temporal ou atenção sobre *embeddings* por quadro) sobre uma amostra maior, caminho pelo qual a hipótese da dinâmica térmica poderá ser reavaliada sem o sobreajuste observado neste estudo. Ainda que modesto em desempenho absoluto, o trabalho contribui com um *pipeline* reprodutível e com uma avaliação criteriosa e estatisticamente cuidadosa do que a termografia dinâmica, combinada a aprendizado de máquina, permite alcançar sob restrição de dados.

Referências

- BONMARIN, M.; LE GAL, F.-A. Lock-in thermal imaging for the early-stage detection of cutaneous melanoma: a feasibility study. **Computers in Biology and Medicine**, v. 47, p. 36–43, 2014. Citado 2 vezes nas páginas 19 e 22.
- BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 17 e 29.
- ÇETİNGÜL, M. P.; HERMAN, C. Quantification of the thermal signature of a melanoma lesion. **International Journal of Thermal Sciences**, v. 50, n. 4, p. 421–431, 2011. Citado 2 vezes nas páginas 19 e 22.
- CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. Citado 2 vezes nas páginas 17 e 29.
- DINNES, J.; DEEKS, J. J.; CHUCHU, N.; FERRANTE DI RUFFANO, L.; MATIN, R. N.; THOMSON, D. R.; WONG, K. Y.; ALDRIDGE, R. B.; ABBOTT, R.; FAWZY, M. et al. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. **Cochrane Database of Systematic Reviews**, John Wiley & Sons, Ltd, v. 2018, n. 12, p. CD011902, 2018. Citado 2 vezes nas páginas 13 e 15.
- EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. New York: Chapman & Hall/CRC, 1993. Citado 2 vezes nas páginas 18 e 33.
- ESTEVA, A.; KUPREL, B.; NOVOA, R. A.; KO, J.; SWETTER, S. M.; BLAU, H. M.; THRUN, S. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115–118, 2017. Citado 5 vezes nas páginas 13, 15, 16, 21 e 22.
- GODOY, S. E.; RAMIREZ, D. A.; MYERS, S. A. et al. Dynamic infrared imaging for skin cancer screening. **Infrared Physics & Technology**, v. 70, p. 147–152, 2015. Citado 4 vezes nas páginas 16, 19, 22 e 46.
- HAENSSLE, H. A.; FINK, C.; SCHNEIDERBAUER, R. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. **Annals of Oncology**, v. 29, n. 8, p. 1836–1842, 2018. Citado 2 vezes nas páginas 21 e 22.
- HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778. Citado 2 vezes nas páginas 17 e 29.
- IOFFE, S.; SZEGEDY, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2015. p. 448–456. Citado 2 vezes nas páginas 17 e 28.
- KESZTYÜS, D.; BAE, H.; WILSON, C.; SCHÖN, M. P.; KESZTYÜS, T. Non-invasive infrared thermography for screening, diagnosis and monitoring of skin cancer. **JDDG: Journal der Deutschen Dermatologischen Gesellschaft**, v. 23, n. 1, p. 7–17, 2025. Citado 2 vezes nas páginas 21 e 22.

- LIU, X.; RIVERA, S. C.; MOHER, D.; CALVERT, M. J.; DENNISTON, A. K. et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. **Nature Medicine**, Nature Publishing Group, v. 26, n. 9, p. 1364–1374, 2020. Citado 3 vezes nas páginas 15, 17 e 44.
- MAGALHÃES, C.; TAVARES, J. M. R. S.; MENDES, J.; VARDASCA, R. Comparison of machine learning strategies for infrared thermography of skin cancer. **Biomedical Signal Processing and Control**, v. 69, p. 102872, 2021. Citado 5 vezes nas páginas 16, 20, 22, 30 e 43.
- MAGALHÃES, C.; VARDASCA, R.; MENDES, J. Recent use of medical infrared thermography in skin neoplasms. **Skin Research and Technology**, v. 24, n. 4, p. 587–591, 2018. Citado 2 vezes nas páginas 20 e 22.
- MALHEIROS, F.; SILVA, E. R.; FAGUNDES, P. N.; FARIA, J. J. R.; LEITE, R. D. V.; GUERRA, I.; TANAKA, V. D.; ALVARES, B. A.; VAZQUEZ, V. d. L. Active versus passive infrared thermography for skin cancer detection: a diagnostic accuracy study. **Cancers**, v. 18, n. 5, p. 829, 2026. Citado na página 43.
- NATIONAL CANCER INSTITUTE. **Melanoma Treatment (PDQ®) – Health Professional Version**. 2025. <<https://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq>>. National Institutes of Health. Acesso em: jul. 2026. Citado 2 vezes nas páginas 13 e 15.
- OTSU, N. A threshold selection method from gray-level histograms. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 9, n. 1, p. 62–66, 1979. Citado na página 30.
- PASZKE, A.; GROSS, S.; MASSA, F. et al. Pytorch: an imperative style, high-performance deep learning library. In: **Advances in Neural Information Processing Systems (NeurIPS)**. [S.l.: s.n.], 2019. v. 32, p. 8024–8035. Citado na página 24.
- PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 24.
- RAJKOMAR, A.; DEAN, J.; KOHANE, I. Machine learning in medicine. **New England Journal of Medicine**, Mass Medical Soc, v. 380, n. 14, p. 1347–1358, 2019. Citado 3 vezes nas páginas 15, 17 e 44.
- RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: convolutional networks for biomedical image segmentation. In: **Medical Image Computing and Computer-Assisted Intervention (MICCAI)**. [S.l.: s.n.], 2015. p. 234–241. Citado na página 20.
- RUSSAKOVSKY, O.; DENG, J.; SU, H. et al. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 2015. Citado 2 vezes nas páginas 17 e 29.
- SOTO, R. F.; GODOY, S. E. An automatic approach to detect skin cancer utilizing active infrared thermography. **Heliyon**, v. 10, n. 23, p. e40608, 2024. Citado 5 vezes nas páginas 16, 20, 22, 42 e 43.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. **Journal of Machine Learning Research**, v. 15, n. 1, p. 1929–1958, 2014. Citado 2 vezes nas páginas 17 e 28.

TAN, M.; LE, Q. V. Efficientnet: rethinking model scaling for convolutional neural networks. In: **International Conference on Machine Learning (ICML)**. [S.l.: s.n.], 2019. p. 6105–6114. Citado na página 29.

VERSTOCKT, J.; VERSPEEK, S.; THIESSEN, F.; TJALMA, W. A.; BROCHEZ, L.; STEENACKERS, G. Skin cancer detection using infrared thermography: Measurement setup, procedure and equipment. **Sensors**, v. 22, n. 9, p. 3327, 2022. Citado 6 vezes nas páginas 13, 16, 21, 22, 43 e 46.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. 3. ed. Rio de Janeiro: GEN LTC, 2020. Citado 2 vezes nas páginas 14 e 24.