Combined Classifiers for Detection of Breast Cancer Metastasis in Histopathological Images

Luis Jhonne Carvalhal de Melo[1], **Omar Andres Carmona Cortes**[2], **Antônio Fernando Lavareda Jacob Júnior**[1]

[1]Programa de Pós-Graduação em Engenharia de Computação e Sistemas - Centro de Ciências Tecnológicas - Universidade Estadual do Maranhão - São Luís - MA - Brasil

[2]Departamento de Computação - IFMA

1uis.jhonne@gmail.com, omar@ifma.edu.br, antoniojunior@professor.uema.br

Abstract. *Breast cancer is a disease responsible for the majority of deaths in Brazil and primarily affects women. Among the consequences of its occurrence are genetic predisposition, sedentary lifestyle, and late menopause. It is a disease that requires diagnosis as quickly as possible so that the patient can begin treatment. Furthermore, studies suggest that pathologists can achieve an accuracy of around 72% when analyzing an exam consisting of thousands of histopathological images of lymph node sections. In this context, this work describes a classifier combining ResNet50, Random Forest (RF), and Support Vector Machine (SVM) trained by the PatchCamelyon dataset. Results indicate that the proposed method achieved an accuracy and F1-score of 81%.*

Resumo. *O câncer de mama é uma doença responsável pela maioria das mortes no Brasil e afeta principalmente mulheres. Entre as consequências de sua ocorrência estão a predisposição genética, o sedentarismo e a menopausa tardia. É uma doença que requer diagnóstico o mais rápido possível para que o paciente possa iniciar o tratamento. Além disso, estudos sugerem que patologistas podem atingir uma precisão de cerca de 72% ao analisar um exame composto por milhares de imagens histopatológicas de cortes de linfonodos. Nesse contexto, este trabalho descreve um classificador que combina ResNet50, Random Forest (RF) e Support Vector Machine (SVM) treinados pelo conjunto de dados PatchCamelyon. Os resultados indicam que o método proposto atingiu uma precisão e F1-score de 81%.*

1. Introdução

O câncer é um problema de saúde pública global, com estimativas apontando para mais de 25 milhões de novos casos até 2030 [PREVENTION 2023a]. Entre os tipos de câncer, o de mama é o que mais causa óbitos entre as mulheres no Brasil. Em 2019, foram registradas 18.295 mortes (18.068 mulheres e 227 homens), e a estimativa para o ano de 2025 é de 74.000 novos casos [PREVENTION 2023b]. O diagnóstico desse tipo de câncer é particularmente desafiador, pois requer a avaliação de diversas características clínicas e histopatológicas. Entre as causas estão predisposição genética, fatores hereditários e hábitos de vida não saudáveis. No Brasil, a prevenção e o controle do câncer de mama incluem estratégias de rastreamento e diagnóstico precoce, sendo a mamografia a principal ferramenta, capaz de antecipar o diagnóstico em até três anos antes do surgimento dos sintomas [Younis et al. 2022].

Além disso, o diagnóstico histopatológico, realizado por patologistas, é crucial para a detecção precisa do câncer. No entanto, fatores como fadiga, pressão por resultados rápidos e a complexidade das análises podem impactar negativamente o desempenho dos profissionais. O estudo de [Camelyon 2017] demonstrou que, sob restrições de tempo, a taxa de acerto na detecção de micro-metástases foi de apenas 38%. Diante desses desafios, algoritmos de aprendizado de máquina, especialmente redes neurais convolucionais (CNNs), têm se mostrado ferramentas essenciais para apoiar a decisão diagnóstica, aumentando a precisão e a eficiência, conforme [Litjens 2022].

Este trabalho propõe uma análise comparativa entre uma CNN, algoritmos tradicionais de aprendizado de máquina e dois modelos combinados (*ensemble*), com o objetivo de detectar a presença ou ausência de metástase em imagens histopatológicas. Para isso, foi utilizado o *dataset* PatchCamelyon [Veeling et al. 2018], que consiste em imagens de linfonodos com e sem metástase. Foram avaliados cinco modelos de classificação: a CNN ResNet50 [He et al. 2016], Random Forest (RF) [Breiman 2001], Support Vector Machine (SVM) [Cortes and Vapnik 1995], um meta-classificador proposto que combina ResNet50 e RF, e outro que integra ResNet50, RF e SVM. Os resultados foram submetidos a testes estatísticos rigorosos, incluindo Shapiro-Wilk (para normalidade), Levene (para homogeneidade de variâncias), Kruskal-Wallis, ANOVA e Tukey, a fim de validar as diferenças de desempenho entre os modelos.

Para cumprir seu objetivo este trabalho está dividido da seguinte forma: a Seção 2 ilustra alguns trabalhos correlatos, destacando produções recentes com PatchCamelyon; a Seção 3 apresenta uma revisão breve da literatura, os métodos propostos nesta investigação e as métricas utilizadas; a Seção 4 mostra os experimentos computacionais com sua configuração, pré-processamento e resultados; finalmente, a Seção 5 apresenta as considerações finais e trabalhos futuros.

2. Trabalhos Relacionados

As CNNs desempenham um papel fundamental em segmentação e classificação de imagens e são um marco no aprendizado de máquina baseado em imagens [Silva and Cortes 2020]. Pesquisas que relacionam aprendizagem de máquina e aplicações médicas e biomédicas, incluem os estudos de [Gayathri et al. 2020] e [Su et al. 2021], [Wei et al. 2023] e [Aftab et al. 2025].

As redes ResNet-18, ResNet-152 e GoogLeNet foram avaliadas por [Silva and Cortes 2020] para a detecção do câncer de mama em imagens histopatológicas. [Ismail and Sovuthy 2019] compararam ResNet-50 e VGG16 em mamografias. Além disso, [Singh et al. 2020] abordaram o problema de desbalanceamento de dados em *datasets* usando VGG19.

No contexto do dataset PatchCamelyon, [Camelyon 2017] propuseram a ConcatNet, comparando seus resultados com GoogLeNet, ResNet e VGG16, alcançando um AUC de 0.924. Posteriormente, [Silva and Cortes 2023] apresentaram o modelo UnetVGG19, um *ensemble* de Rede Neural Convolucional com Aprendizado de Transferência, que atingiu um AUC de 0.9565 e uma perda de 0.2869. No entanto, esses trabalhos utilizaram *datasets* limitados e com imagens de alta qualidade. Diferentemente, o classificador proposto neste estudo enfrentou o desafio de treinar com o *dataset* completo, incluindo imagens de qualidade variável, tornando o processo mais desafiante para

o modelo e os resultados ainda mais relevantes para aplicações no mundo real.

3. Materiais e Métodos

3.1. Fundamentação Teórica

Inspiradas pela organização do córtex visual humano, as CNNs utilizam convoluções para extrair automaticamente características hierárquicas das imagens. Dessa forma, capturam padrões como bordas, texturas e formas complexas em camadas sucessivas [LeCun et al. 2015]. A ResNet50, introduzida por [He et al. 2016], é uma rede neural convolucional residual e amplamente reconhecida por sua capacidade de treinar redes profundas sem problemas significativos de desaparecimento gradiente.

Por sua vez, RF são algoritmos de aprendizado de máquina baseados em *ensembles* os quais combinam múltiplas árvores de decisão para melhorar a robustez e a precisão do modelo. Cada árvore é treinada em uma amostra aleatória do conjunto de dados com substituição (*bagging*) e as decisões finais são obtidas por meio de um processo de votação ou média [Breiman 2001]. O RF é altamente eficiente para lidar com dados estruturados e são amplamente utilizados em classificação, regressão e análise de variáveis. Sua capacidade de operar com dados de alta dimensionalidade e de lidar com valores ausentes o torna uma escolha versátil e confiável em muitos cenários.

O SVM é um método de aprendizado supervisionado amplamente utilizado em tarefas de classificação e regressão. Ele é particularmente eficaz em problemas com fronteiras de decisão complexas. O SVM busca encontrar um hiperplano ótimo que **maximize** a margem entre as classes no espaço de características. Para isso, utiliza a técnica conhecida como *kernel trick*, a fim de lidar com dados não linearmente separáveis [He et al. 2016]. Entre suas vantagens estão a robustez em espaços de alta dimensionalidade e a capacidade de generalizar bem em cenários com conjuntos de dados limitados. No entanto, a escolha adequada de parâmetros, como o tipo de *kernel* e o valor de regularização, é crucial para o desempenho do modelo, especialmente em problemas com dados ruidosos [Hastie et al. 2009].

A combinação de ResNet50, Random Forest (RF) e SVM em um *ensemble stacking* une as forças do aprendizado profundo, das árvores de decisão e das margens de separação para melhorar a precisão e a generalização do modelo. Nesse contexto, a ResNet50 pode ser utilizada para extrair características profundas e discriminativas de imagens. Já o Random Forest e o SVM atuam como classificadores complementares, cada um adotando abordagens distintas para modelar os dados [He et al. 2016]. O meta-modelo no *ensemble* aprende a combinar as saídas desses modelos para otimizar a decisão final, beneficiando-se da capacidade do SVM de lidar com fronteiras de decisão complexas e da robustez do Random Forest frente a dados ruidosos.

Estudos demonstram a alta precisão e robustez de *ensembles* em tarefas de classificação de imagens médicas e em detecção de anomalias, como em [Wang et al. 2021]. Esse tipo de combinação é particularmente útil em cenários com alta complexidade e variabilidade nos dados, exigindo modelos capazes de capturar diferentes aspectos das informações. Além disso, o uso de meta-modelos robustos para aprender as interações entre as previsões dos diversos modelos contribui para aumentar a capacidade do sistema de evitar erros de *overfitting* e melhorar a generalização para novos dados.

Este trabalho propôs dois modelos *ensembles* para detecção de câncer de mama: a primeira proposta combina a ResNet50 com o RF; a segunda proposta combina a ResNet50, RF e o SVM. Além disso, comparou-se o desempenho de cada classificador (CNN, RF e SVM) separadamente. A Tabela 1 resume os modelos que foram comparados. Utilizou-se as métricas de acurácia, perda baseada em *binary_crossentropy*, além da precisão, *recall* e F1-score para avaliação dos resultados. O método de Shapiro-Wilk foi utilizado para verificar a normalidade das métricas obtidas. A homogeneidade de variâncias foi testada com o método de Levene. Outrossim compara-se esses resultados utilizando os métodos estatísticos Kruskal-Wallis (ou ANOVA) e Tukey.

Tabela 1. Modelos e Algoritmos a serem analisados

Nome	Descrição
ResNet50	Rede Neural convolucional
RF	Modelo tipo *bagging* que utiliza árvores de decisão
SVM	Com kernel RBF, Linear e Polinomial.
Ensemble 1	Combinação utilizando ResNet50 + RF
Ensemble 2	Combinação entre ResNet50 + RF + SVM

3.2. Solução proposta 1

A Figura 1 apresenta o diagrama da arquitetura do primeiro modelo *ensemble* proposto, na qual estão disposto os níveis do classificador e o metamodelo combinador. O nível 0 é composto pela ResNet50, uma CNN de extração de características, e o algoritmo do Random Forest (ou floresta aleatória), um algoritmo do tipo *bagging*. Todos os modelos de nível zero foram treinados sobre os *pixels* brutos das imagens. Por este motivo, esta proposta também pode ser considerada um *ensemble* de *ensemble*. Utiliza-se Regressão Logística no nível 1 para aprender a melhor forma de combinar os resultados.

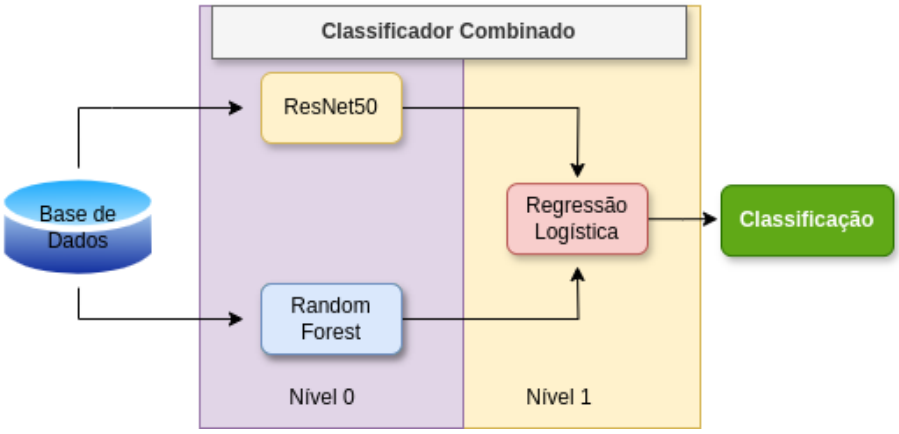

Figura 1. Diagrama da arquitetura proposta do Ensemble 1

3.3. Solução proposta 2

A Figura 2 apresenta o diagrama da arquitetura para o segundo modelo *ensemble* proposto, na qual estão disposto os níveis do classificador e o metamodelo combinador. O

nível 0 é composto por três modelos (ou algoritmos) de classificação e também foram treinados com os *pixels* brutos das imagens. O primeiro modelo é o ResNet50, uma CNN de extração de características, a RF é um algoritmo do tipo *bagging* e o SVM representa um algoritmo tradicional bem conhecido na literatura de aprendizagem de máquina. Por este motivo, a presente proposta consiste em um *stacked* de *classifier*. Utiliza-se Regressão Logística no nível 1 para aprender a melhor forma de combinar os resultados.

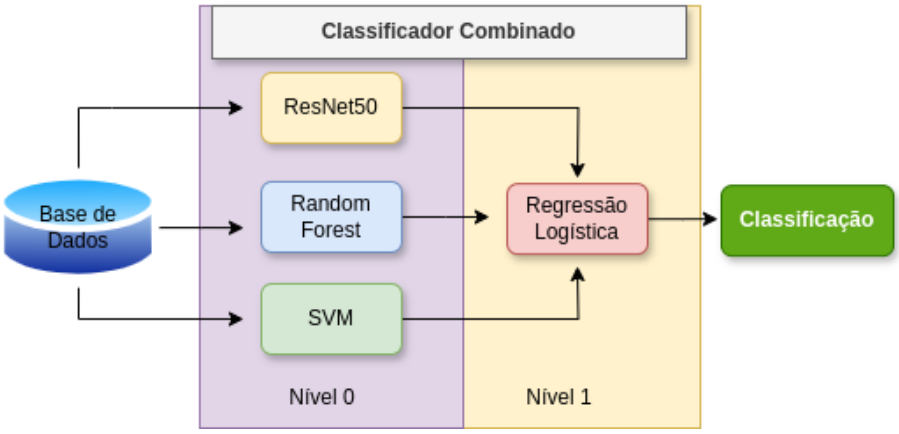

Figura 2. Diagrama da arquitetura proposta do Ensemble 2

3.4. Métricas de Avaliação e Comparação

Foram utilizadas as métricas de Acurácia, Precisão, F1-score e *Recall* para todos os modelos. Além dessas, para avaliação da ResNet50 foram plotados os gráficos para análise da Acurácia, Perda e AUC. Para comparação dos desempenhos dos diferentes modelos serão utilizados os métodos:

- Método Shapiro-Wilk para decidir se os dados seguem uma distribuição normal.
- Teste de Levene para verificar se as variâncias entre os grupos são homogêneas.
- Método Kruskal-Wallis para verificar se há diferenças estatisticamente significativas entre os resultados obtidos, em caso de não normalidade dos dados.
- Método ANOVA para verificar se há diferenças estatisticamente significativas entre os resultados obtidos, caso os dados sigam uma distribuição normal.
- Método Tukey: teste *post-hoc* após encontrar diferenças significativas, o teste de Tukey identifica quais pares diferem significativamente.

4. Experimentos Computacionais

4.1. Configuração

O ambiente computacional utilizado para implementação e execução dos testes foi a máquina local com 32GB de RAM, processador Intel Core i5 2.40GHz, sem GPU disponível. Foi utilizado a linguagem Python e as bibliotecas Scikit-Learn, TensorFlow e Keras para criação dos modelos, executando no Kubuntu Linux 24.04.

Cada modelo e os *ensembles* foram treinados, validados e testados com o mesmo *dataset*. Para treinamento, utilizou-se validação cruzada, com K-Fold $= 5$. Cada modelo utilizado - CNN e classificadores - necessita de parâmetros e configuração própria, conforme descrito a seguir.

A metodologia adotada no teste foi realizar a predição de 10 *subdatasets* diferentes e randômicos. Para cada *subdataset* gerado foram calculados as métricas dos cinco modelos apresentados.

A ResNet50 foi o modelo base, com camada de entrada do tamanho das imagens (96, 96, 3), pesos inicias da Imagenet e camadas congeladas. Foram adicionados 3 callbacks: ReduceLROnPlateau para diminuir a taxa de aprendizagem em caso de platô na métrica, ModelCheckpoint para salvar os melhores pesos e EarlyStopping para parada antecipada em caso de não melhora no aprendizado. Todos os *callbacks* utilizaram a perda na validação (*val_loss*) como métrica.

Já os classificadores e meta-modelo utilizados neste trabalho foram criados conforme a descrição na Tabela 2. O classificador SVM foi treinado com três *kernels* diferentes (RBF, Linear e Polinomial), a fim de encontrar o melhor para ser utilizado no *ensemble*.

Tabela 2. Parametrização dos modelos

Classificador	**Classe do Scikit-Learn**	**Parametrização**
Random Forest (RF)	RandomForestClassifier	n_estimators $= 100$ random_state $= 42$
Support Vector Machine (SVM)	SVC	kernel ='rbf'
Support Vector Machine (SVM)	SVC	kernel ='linear'
Support Vector Machine (SVM)	SVC	kernel ='poly'
Meta-modelo (Regressão Logística)	LogisticRegression	N/A

4.2. Dataset

O conjunto de dados PatchCamelyon [Veeling et al. 2018] é uma referência para testar algoritmos de detecção de imagens histopatológicas em tecido metastático. Este banco de dados contém 327.800 imagens microscópicas de tecidos, classificados binariamente, indicando presença ou ausência de metástase. A proporção base é $50/50$ para as duas classes (com ou sem metástase). A resolução da imagem é 96×96, com cores de 3 canais e formato de arquivo TIF.

4.3. Pré-Processamento

Para o treinamento da CNN, realizou-se o método de *data augmentation* com parâmetros conforme a tabela 3, além da validação cruzada com k $= 5$. O treinamento aconteceu em duas etapas com 10 épocas cada, totalizando 20 épocas por *fold*. A primeira etapa foi realizada com as camadas congeladas, com o objetivo de aprender características gerais. Após esse período, as 50 camadas foram descongeladas e a taxa de aprendizagem (*learning rate*) foi diminuída, a fim de obter um ajuste fino das características do *dataset*.

A Figura 3 exibe a evolução do treinamento (Perda, Acurácia e AUC) da ResNet50 ao longo das épocas (10 épocas com as camadas congeladas e 10 épocas com as camadas descongeladas). Os outros modelos de aprendizagem também foram treinados com *k-fold* $= 5$, sem a necessidade de épocas.

Tabela 3. Parametrização para data augmentation

Parâmetros	Parametrização
rotation_range	20
width_shift_range	0.2
height_shift_range	0.2
shear_range	0.15
zoom_range	0.15
horizontal_flip	True
fill_mode	nearest
brightness_range	0.8-1.2

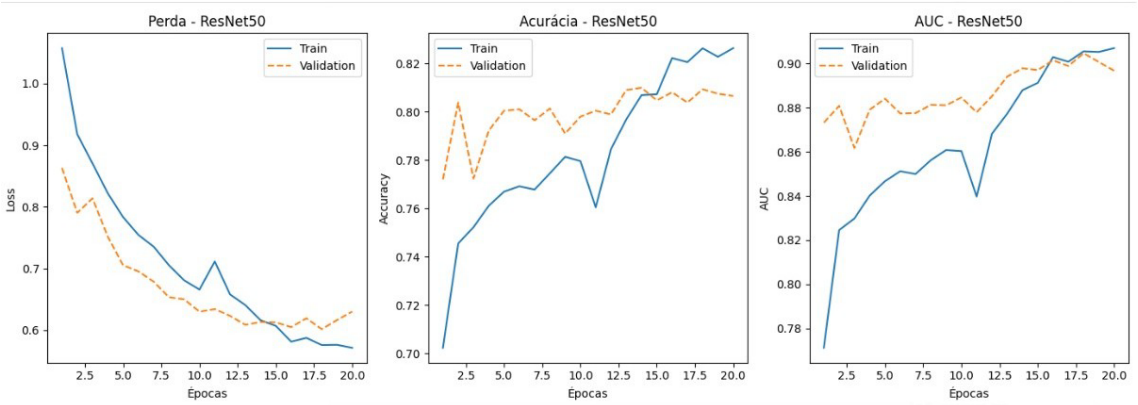

Figura 3. Evolução da perda, Acurácia e AUC por épocas

4.4. Resultados

Nesta seção apresentar-se-á os resultados obtidos para a classificação do *dataset* utilizando os modelos citados na Tabela 1. Primeiramente serão apresentadas as métricas de treinamento e validação dos modelos, em seguida o resultado dos testes e por fim as comparações estatísticas.

A média e desvio padrão de cada métrica foram tabuladas e são mostradas na Tabela 4. O *kernel* RBF foi o que apresentou melhor desempenho no SVM e foi o escolhido para compor o Ensemble 2. Conforme pode ser observado, o Ensemble 1 (ResNet50 + RF) apresentou vantagem em todas as métricas de validação.

Após o final do treinamento por *folds*, utilizou-se todo o *dataset* de validação para gerar o relatório de classificação e matriz de confusão para cada modelo. O desempenho do modelo combinado 1 (ResNet50 + RF) pode ser visto na matriz de confusão obtida na predição do *ensemble 1* no *dataset* de validação, conforme Figura 4. O desempenho do modelo combinado 2 (ResNet50 + RF + SVM) na predição sobre o *dataset* de validação pode ser visto na matriz de confusão da Figura 5.

Os cinco modelos com melhor desempenho no treinamento (ResNet50, RF, SVM - Rbf, Ensemble 1 e Ensemble 2) foram salvos em arquivo para serem testados e os resultados serão mostrados a seguir. Foram calculadas as métricas de acurácia, precisão, *recall* e *f1-score* para serem comparadas. A Tabela 5 abaixo fornece a média dos resultados

Tabela 4. Resultados das métricas de validação do modelo

Modelo	Acurácia	Precisão	Recall	F1-Score
ResNet50	0.8000 ± 0.0097	0.8094 ± 0.0059	0.8000 ± 0.0097	0.7982 ± 0.0106
RF	0.7318 ± 0.0024	0.7340 ± 0.0024	0.7318 ± 0.0024	0.7309 ± 0.0025
SVM (rbf)	0.7496 ± 0.0044	0.7505 ± 0.0044	0.7496 ± 0.0044	0.7493 ± 0.0044
SVM (linear)	0.5353 ± 0.0090	0.5354 ± 0.0090	0.5353 ± 0.0090	0.5352 ± 0.0091
SVM (poly)	0.6342 ± 0.0063	0.6343 ± 0.0063	0.6342 ± 0.0063	0.6340 ± 0.0063
Ensemble 1	**0.8100 ± 0.0017**	**0.8138 ± 0.0015**	**0.8100 ± 0.0017**	**0.8092 ± 0.0017**
Ensemble 2	0.7761 ± 0.0153	0.7806 ± 0.0179	0.7761 ± 0.0153	0.7750 ± 0.0149

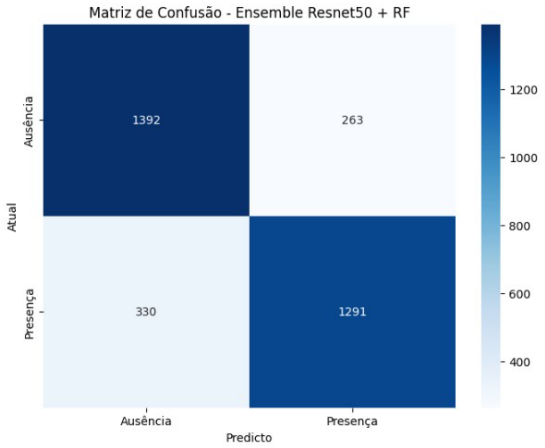

Figura 4. Matriz de confusão Ensemble 1

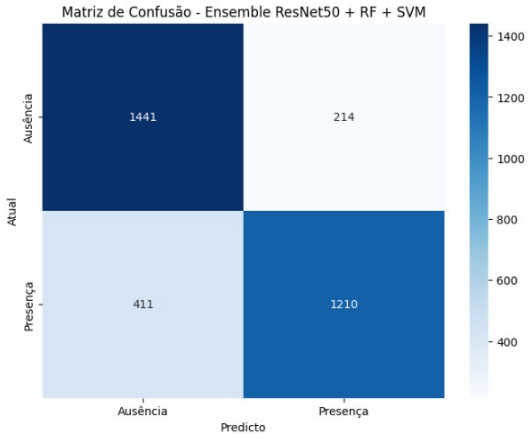

Figura 5. Matriz de confusão Ensemble 2

obtidos.

Tabela 5. Média das métricas obtidas nos testes

Modelo	Acurácia	Precisão	Recall	F1-Score
ResNet50	0.8036 ± 0.0083	0.8108 ± 0.0081	0.8036 ± 0.0083	0.8023 ± 0.0085
RF	0.7370 ± 0.0098	0.7400 ± 0.0100	0.7370 ± 0.0098	0.7360 ± 0.0098
SVM	0.7530 ± 0.0069	0.7543 ± 0.0071	0.7530 ± 0.0069	0.7526 ± 0.0069
Ensemble 1	**0.8180 ± 0.0103**	**0.8178 ± 0.0103**	**0.8189 ± 0.0104**	**0.8180 ± 0.0103**
Ensemble 2	0.8085 ± 0.0084	0.8136 ± 0.0083	0.8085 ± 0.0084	0.8076 ± 0.0085

Os testes comparativos estatísticos foram executados paras as métricas de precisão, *recall* e *F1-score*. Os resultados obtidos para Shapiro-Wilk podem ser vistos na Tabela 6 indicam que a suposição de normalidade foi atendida, com todos os p-valores acima de 0.05. O Teste de Homogeneidade de Levene (Tabela 7) com p-valores > 0.05 confirma que as variâncias entre os grupos são homogêneas e permite aplicar o teste de ANOVA para encontrar diferenças significativas.

O teste ANOVA foi aplicado e obteve p-valores muitos pequenos (< 0.05) e indicou que há diferenças significativas entre as métricas do modelo. O resultado completo

Tabela 6. Teste de Normalidade Shapiro-Wilk - Testes

Modelo	Precisão	Recall	F1-Score
ResNet50	0.5015	0.3853	0.3762
RF	0.2871	0.4929	0.4618
SVM	0.6869	0.9690	0.9743
Ensemble 1	0.1745	0.2307	0.2238
Ensemble 2	0.2771	0.5032	0.4609

Tabela 7. Resultados Teste de Levene

Teste	Precisão	Recall	F1-Score
Teste de Levene	0.8698	0.9108	0.9148

do teste de ANOVA está na Tabela 8.

Tabela 8. Resultados Teste de ANOVA

Métrica	H-statistic	p-value
Precisão	142.2196	4.4132e-23
Recall	136.8266	9.0573e-23
F1-Score	135.1483	1.1390e-22

Após a análise dos resultados obtidos acima e executando o teste *post-hoc* de Tukey para as métricas é possível identificar quais os grupos que possuem diferenças. O resultado pode ser visto na Tabela 9.

Tabela 9. Resultados Teste de Tukey (Post hoc)

Modelo1	Modelo2	Precisão	Recall	F1-Score
Ensemble1	Ensemble 2	False	False	False
Ensemble1	RF	True	True	True
Ensemble1	ResNet	False	True	True
Ensemble1	SVM	True	True	True
Ensemble2	RF	True	True	True
Ensemble2	ResNet	False	False	False
Ensemble2	SVM	True	True	True
RF	ResNet	True	True	True
RF	SVM	True	True	True
ResNet	SVM	True	True	True

Considerando a precisão, não há diferenças significativas entre Ensemble 1 *versus* Ensemble 2, Ensemble 1 *versus* ResNet50 e Ensemble 2 *versus* ResNet50, mas há diferenças entre:

- Ensemble 1 *versus* RF e SVM: Ensemble significativamente superior em média;

- Ensemble 1 *versus* RF e SVM: Ensemble significativamente superior em média;
- RF *versus* ResNet50 e SVM: RF superior;
- ResNet50 *versus* SVM: SVM tem média significativamente menor que ResNet50.

. O teste de Tukey para o *recall* fornece a interpretação detalhada entre os pares. As diferenças significativas (reject = True) encontradas foram: Ensemble 1 tem *recall* significativamente maior que RF, ResNet50 e SVM. Ensemble 2 tem *recall* superior ao RF e SVM. Já RF é superior ao ResNet50 e SVM. O *recall* do ResNet50 é superior ao SVM. Já nos pares abaixo não foram encontradas diferenças significativas (reject = False): Ensemble 1 *versus* Ensemble 2 e Ensemble 2 *versus* ResNet50.

Realizando a análise detalhada do Teste de Tukey para F1-Score, percebe-se que não houve diferenças significativas (reject = False) para Ensemble 1 *versus* Ensemble 2 e Ensemble 2 *versus* ResNet50. As diferenças significativas foram:

- Ensemble 1 é significativamente maior que RF, ResNet50 e SVM;
- Ensemble 2 versus RF e Ensemble 2 versus SVM, sendo Ensemble 2 superior;
- Ensemble 2 *versus* SVM: SVM tem F1 significativamente menor que Ensemble 2;
- RF tem F1 significativamente maior que ResNet50;
- RF tem F1 significativamente maior que SVM;
- SVM tem F1 significativamente menor que ResNet50.

A partir dos resultados obtidos nos testes e nos comparativos estatísticos, algumas considerações podem ser feitas: Ensemble 1 e Ensemble 2 apresentam as melhores médias de precisão, *recall* e *f1-score*, com desempenho superior ao RF, SVM e ResNet50. Não há diferença significativa entre Ensemble 1 e Ensemble 2. Embora o RF tenha média significativamente maior que SVM e ResNet50, seu desempenho ainda é inferior ao dos modelos Ensemble. ResNet50 tem recall e F1 significativamente inferior ao Ensemble 1, mas não é significativamente diferente do Ensemble 2; Supera o SVM, mas é superado pelo RF. SVM tem o pior desempenho em *recall* e F1, sendo significativamente inferior a todos os outros modelos. O *kernel* RBF apresentou o melhor desempenho no SVM.

Os *stackings* Ensemble 1 e Ensemble 2 são as melhores opções em todas as métricas. RF e ResNet50 são alternativas intermediárias, enquanto o SVM apresentou o pior desempenho. Adicionar o SVM ao Ensemble 1, ou seja, o Ensemble 2 não melhorou as métricas.

5. Conclusões

Este trabalho avaliou diferentes modelos para classificação de imagens: ResNet50, RF, SVM (com *kernel* linear, polinomial e rbf), Ensemble 1 e Ensemble 2. Os modelos treinados apresentaram um resultado satisfatório, como Ensemble 1 alcançando 81% de acurácia, mas ainda possível de ser melhorado. Como sugestão para trabalhos futuros pode-se citar: a) utilizar métodos para explorar parametrizações que resultem em modelos de melhores desempenhos e b) melhorar o desempenho da CNN utilizada, incrementando a quantidade de épocas.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001

Referências

Aftab, M., Mehmood, F., Zhang, C., and et al. (2025). Ai in oncology: Transforming cancer detection through machine learning and deep learning applications. *arXiv preprint arXiv:2501.15489*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Camelyon, C. Bejnordi, B. E. V. M. v. D. P. J. v. G. B. K. N. L. G. v. d. L. J. A. W. M. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Gayathri, S., Gopi, V. P., and Palanisami, P. (2020). A lightweight cnn for diabetic retinopathy classification from fundus images. *Biomedical Signal Processing and Control*, 62:102115.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Ismail, N. S. and Sovuthy, C. (2019). Breast cancer detection based on deep learning technique. In *2019 International UNIMAS STEM 12th Engineering Conference (EnCon)*, pages 89–92.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–444.

Litjens, G. Bandi, P. (2022). He-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, pages 2–3.

PREVENTION (2023a). Estimate — 2023. In of Health, B. M., editor, *Cancer Incidence in Brazil*, pages 39–40. Brazilian National Institute of Cancer Prevention.

PREVENTION (2023b). Estimate — 2023. page 31.

Silva, D. and Cortes, O. (2020). On convolutional neural networks and transfer learning for classifying breast cancer on histopathological images using gpu. In *XXVII Brazilian Congress on Biomedical Engineering*.

Silva, D. and Cortes, O. (2023). Metastasis detection of breast cancer using ensemble deep learning. In *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 104–114, Porto Alegre, RS, Brasil. SBC.

Singh, R., Ahmed, T., Kumar, A., Singh, A. K., Pandey, A. K., and Singh, S. K. (2020). Imbalanced breast cancer classification using transfer learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 83–93.

Su, Y., Li, D., and Chen, X. (2021). Lung nodule detection based on faster r-cnn framework. *Computer Methods and Programs in Biomedicine*, 200:105866.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. (2018). Rotation equivariant cnns for digital pathology. In Springer, editor, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241.

Wang, Z., Liu, Y., and Wang, Y. (2021). Medical image analysis with convolutional neural networks: A review. *Journal of Biomedical Informatics*, 77:102–116.

Wei, L., Niraula, D., Gates, E. D. H., Fu, J., Luo, Y., Nyflot, M. J., Bowen, S. R., El Naqa, I. M., and Cui, S. (2023). Artificial intelligence (ai) and machine learning (ml) in precision oncology: a review on enhancing discoverability through multiomics integration. *British Journal of Radiology*, 96(1150):20230211. Open Access.

Younis, Y. S., Ali, A. H., Alhafidh, O. K. S., Yahia, W. B., Alazzam, M. B., Hamad, A. A., and Meraf, Z. (2022). Early diagnosis of breast cancer using image processing techniques. In Velmurugan, P., editor, *Applications of Nanomaterials and Nanotechnology in Engineering, Environment and Life Sciences*, pages 2–5. Hindawi.