



Classificação do Câncer de Mama Usando Aprendizado de Máquina

Orientanda: Isabela Medeiros Belo Lopes

Orientador: Prof. Robson Cavalcanti Lins

Projeto: Desenvolvimento de Metodologia e Soluções Computacionais
para Promover Cidades Inteligentes.

Sumário

Introdução	Pág 3-4
Visão computacional	Pág 5
Objetivos	Pág 6
Inteligência artificial na saúde	Pág 7
Base de dados	Pág 8
Resultados e discussão	Pág 9-27
Conclusão	Pág 28

Notícias recentes



The screenshot shows a news article from HCP.org.br. At the top, there's a navigation bar with links for 'Atendimento', 'Ensino e Pesquisa', 'Tudo sobre câncer', 'Conteúdos', 'Doe', and a search bar with the phone number '81 3217.8090'. The main headline reads 'Outubro Rosa: campanha reforça a importância da detecção precoce do câncer de mama'. Below it, a sub-headline says 'Cuidar, um legado que atravessa gerações' and 'HCP, há 80 anos cuidando de milhares de vidas com amor.' There's also a small image of two women smiling.

<https://www.hcp.org.br/2025/10/01/outubro-rosa-campanha-reforca-a-importancia-da-deteccao-precoce-do-cancer-de-mama/>

DATA: 01/10/2025



The screenshot shows a news article from the Brazilian government website (gov.br). The header includes the 'gov.br' logo, 'Governo Federal', and links for 'Órgãos do Governo', 'Acesso à Informação', 'Legislação', 'Acessibilidade', and 'Entrar com gov.br'. The main headline is 'Ministério da Saúde e INCA apresentam publicação com dados atualizados sobre câncer de mama no Brasil'. Below it, a sub-headline says 'OUTUBRO ROSA 2025'. The text discusses the importance of early detection and mentions 73,600 new cases in 2025. It also notes actions to expand access to modern exams and treatments through the SUS. The article was published on 03/10/2025 at 17h13 and updated on the same day at 17h17.

<https://www.gov.br/saude/pt-br/assuntos/noticias/2025/outubro/ministerio-da-saude-e-inca-apresentam-publicacao-com-dados-atualizados-sobre-cancer-de-mama-no-brasil>

DATA: 03/10/2025

Notícias recentes

Diagnóstico precoce pode garantir até 95% de cura do câncer de mama, diz Unimed

Estudos apontam que até o fim de 2025 mais de 108 mil mulheres sejam afetadas pela doença.

Por Unimed Cuiabá
10/10/2025 15h44 · Atualizado há 4 dias

<https://g1.globo.com/mt/mato-grosso/especial-publicitario/unimed-cuiaba/noticia/2025/10/10/diagnostico-precoce-pode-garantir-ate-95percent-de-cura-do-cancer-de-mama-diz-unimed.ghtml>

DATA: 10/10/2025

NOTÍCIAS

IA ajuda médicos a detectar mais casos de câncer de mama

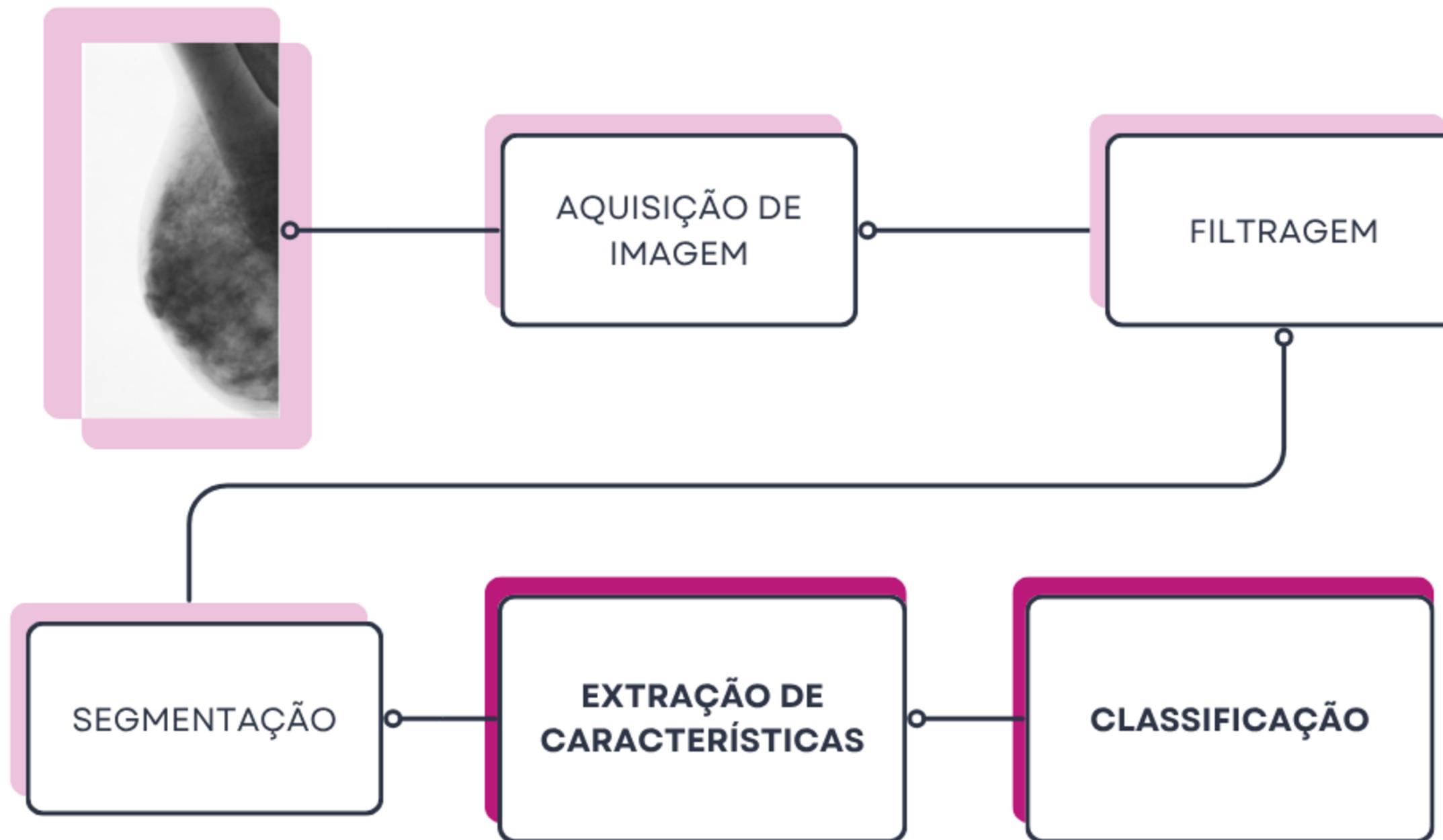
Resultado reforça potencial da tecnologia para aumentar a velocidade com que os radiologistas analisam exames de mamografias, diminuindo sua carga de trabalho e acelerando o diagnóstico da doença

Publicado em 23/01/2025 - Última modificação em 23/01/2025 às 22h05

<https://www.ipea.gov.br/cts/pt/central-de-conteudo/noticias/noticias/464-ia-ajuda-medicos-a-detectar-mais-casos-de-cancer-de-mama>

DATA: 23/01/2025

Visão Computacional



Objetivos

Este trabalho teve por objetivo realizar uma avaliação comparativa de modelos de aprendizagem de máquina para auxiliar no diagnóstico de câncer de mama.

- Estudar os fundamentos de aprendizado de máquina;
- Pesquisar bases de dados de imagem voltadas ao diagnóstico do câncer de mama;
- Estudar modelos de aprendizagem de máquina;
- Estudar funcionalidades das bibliotecas Tensorflow e Sklearn;
- Avaliar modelos de aprendizado de máquina para apoio no diagnóstico do câncer de mama;
- Analisar os resultados.

Inteligência Artificial e Machine Learning na saúde

A Inteligência Artificial (IA) busca criar sistemas capazes de simular o raciocínio humano. Dentro dela, o Machine Learning (ML) permite que máquinas aprendam padrões automaticamente a partir de dados.

- Redes Neurais Convolucionais (RNCs)
- Random Forest (RF)
- K-nearest neighbors (KNN)
- Support Vector Machine (SVM)

Base de Dados

- Base de Dados: MIAS – Mini Mammographic Database.
- Contém 320 imagens mamográficas
- Cada imagem é rotulada e apresenta 2 classes:
 - Normais (0)
 - Anormais (1)



Resultados e Discussão

Ambiente de Execução

- Google Colab
- Linguagem: Python
- Pré-processamento: segmentação e redimensionamento para 150×150 px
- Download e extração: automáticos (gdown e ZipFile)
- Normalização: pixels escalonados de $[0, 255] \rightarrow [0, 1]$
- Divisão:
 - Treino: 286 imagens (187 normais / 99 anormais)
 - Validação: 44 imagens (22 normais / 22 anormais)
 - Proporção: 87% treino / 13% validação

Ambiente de Execução

Data Augmentation

- Aumenta a diversidade do conjunto de treino e reduz overfitting
- Aplica transformações nas imagens: rotação, deslocamento, zoom, espelhamento
- Torna o modelo mais robusto a variações visuais

Balanceamento de Classes

- Treino levemente desbalanceado: 187 normais × 99 anormais
- Aplicação de pesos inversos de classe na função de perda
- Equilibra o impacto das classes e evita viés para a classe majoritária

Bibliotecas Principais

- scikit-learn (sklearn): Criação e treino do modelos Random Forest, K-Nearest Neighbors e Support Vector Machine, cálculo de métricas e validação cruzada.
- TensorFlow / Keras: Extração de características das imagens usando a rede pré-treinada ResNet152V2.
- NumPy / Matplotlib: Manipulação de arrays e exibição dos resultados e imagens.



Keras

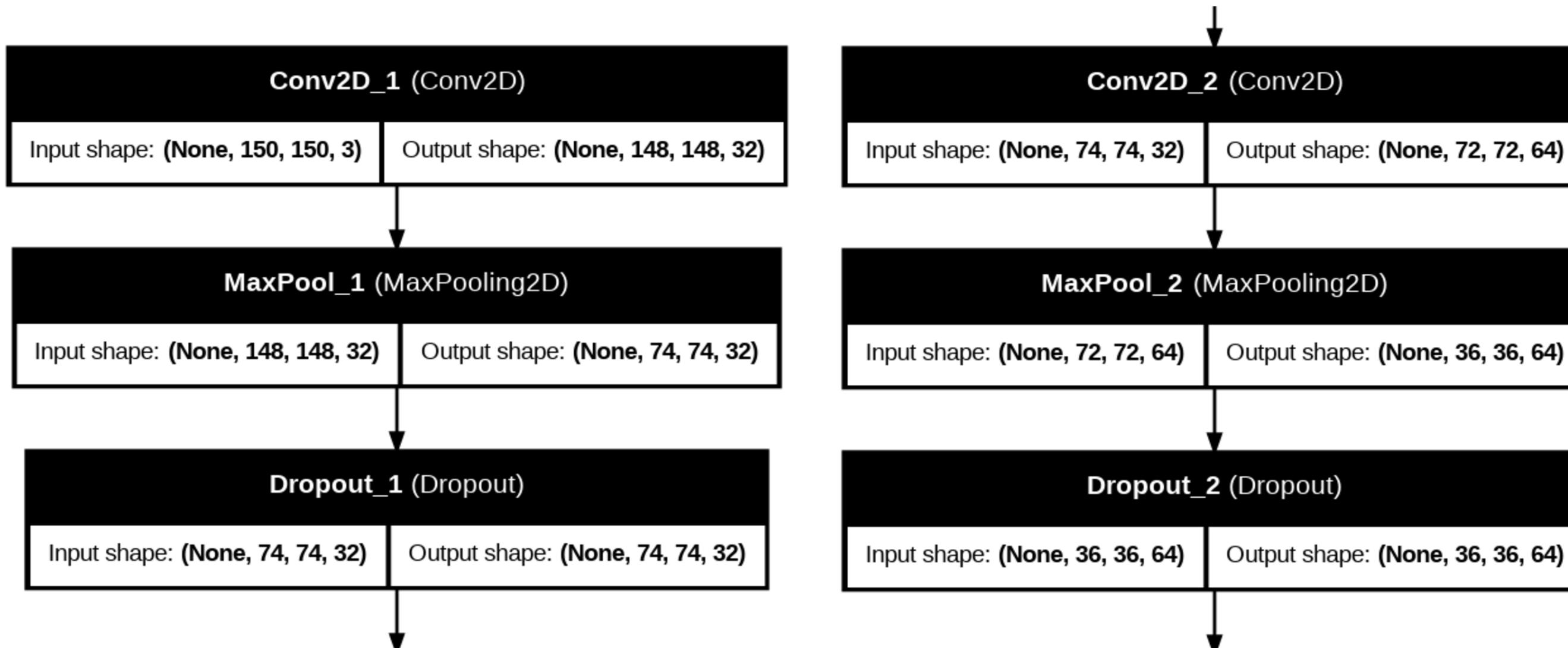


NumPy

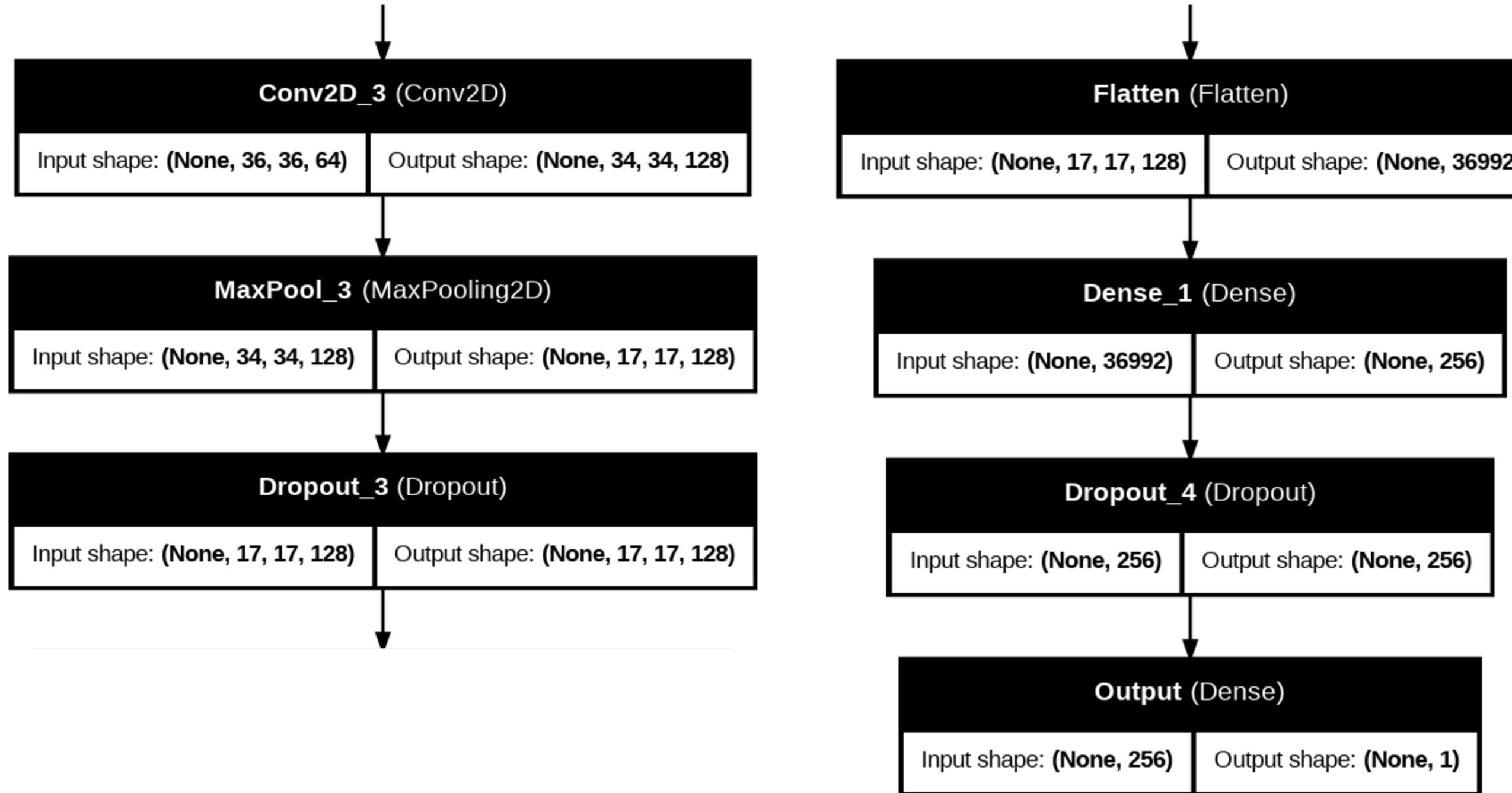


Redes Neurais Convolucionais

Inspirada no cérebro humano, é formada por neurônios artificiais organizados em camadas. Cada neurônio recebe entradas, realiza cálculos e gera saídas. Durante o treino, a rede ajusta seus pesos para reduzir o erro e aprender padrões nos dados.



Redes Neurais Convolucionais



- EarlyStopping (patience=10): monitora acurácia, restaura melhores pesos e evita overfitting

Transfer Learning

Reutilização do conhecimento de um modelo já treinado para acelerar e melhorar o treinamento em uma nova tarefa relacionada.

- Modelo Base: ResNet152V2 pré-treinada no ImageNet
- Congelamento: Atua como extrator de características
- Processamento: Imagem é convertida em um vetor numérico representando características aprendidas
- Hibridização:
 - Deep Learning → extrai características complexas e robustas
 - ML clássico (Random Forest, KNN, SVM) → classificar características
 - Vantagens: menos tempo de treino e interpretabilidade via ML

Random Forest

É um modelo de aprendizado supervisionado que combina várias árvores de decisão para aumentar a precisão e garantir maior robustez nas previsões. Seu principal objetivo é reduzir o erro e evitar o overfitting.

Elementos e Funcionamento:

- Árvore de Decisão: Modelo base que cria regras para classificar ou prever valores.
- Bootstrap: Amostragem aleatória com reposição que gera subconjuntos diferentes para treinar cada árvore.
- Votação / Média: Combina as previsões para produzir o resultado final mais confiável.

K-Nearest Neighbors

É um algoritmo de aprendizado supervisionado que classifica novos dados pela votação dos K vizinhos mais próximos. Não realiza treinamento explícito, apenas armazena os dados e utiliza métricas de distância para identificar a classe mais frequente ao redor da nova amostra.

- Baseado em instâncias
- Classifica novos pontos usando distâncias entre amostras
- Tomada de decisão
 - Classificação: a nova amostra recebe a classe mais frequente entre os K vizinhos
 - Regressão: a nova amostra recebe a média dos valores dos K vizinhos

Support Vector Machine

Modelo de aprendizado supervisionado que busca a melhor fronteira de separação entre as classes. Ele projeta os dados em um espaço de alta dimensão e define um hiperplano que maximiza a margem entre as classes, equilibrando complexidade e erro de classificação.

Elementos e Funcionamento:

- Hiperplano: Superfície que separa as classes maximizando a distância entre os pontos mais próximos.
- Vetores de Suporte: Amostras que definem a posição e orientação do hiperplano.
- Parâmetro C: Controla o equilíbrio entre margem ampla e menor erro de classificação.

GridSearchCV

Técnica que busca automaticamente a melhor combinação de hiperparâmetros para maximizar o desempenho de um modelo.

- Busca Exaustiva
- Validação Cruzada
- Escolhe os parâmetros com melhor pontuação média

Principais Parâmetros:

- estimator: Modelo a ser ajustado.
- param_grid: Conjunto de valores a testar.
- cv: 10 (validação cruzada 10-fold)
- scoring: 'accuracy'
- n_jobs: -1 (usa todos os núcleos)

Resultados do Grid Search

Random Forest

- n_estimators: 50
- max_depth: 10
- min_samples_split: 10
- min_samples_leaf: 2
- max_features: 'sqrt'

K-Nearest Neighbors

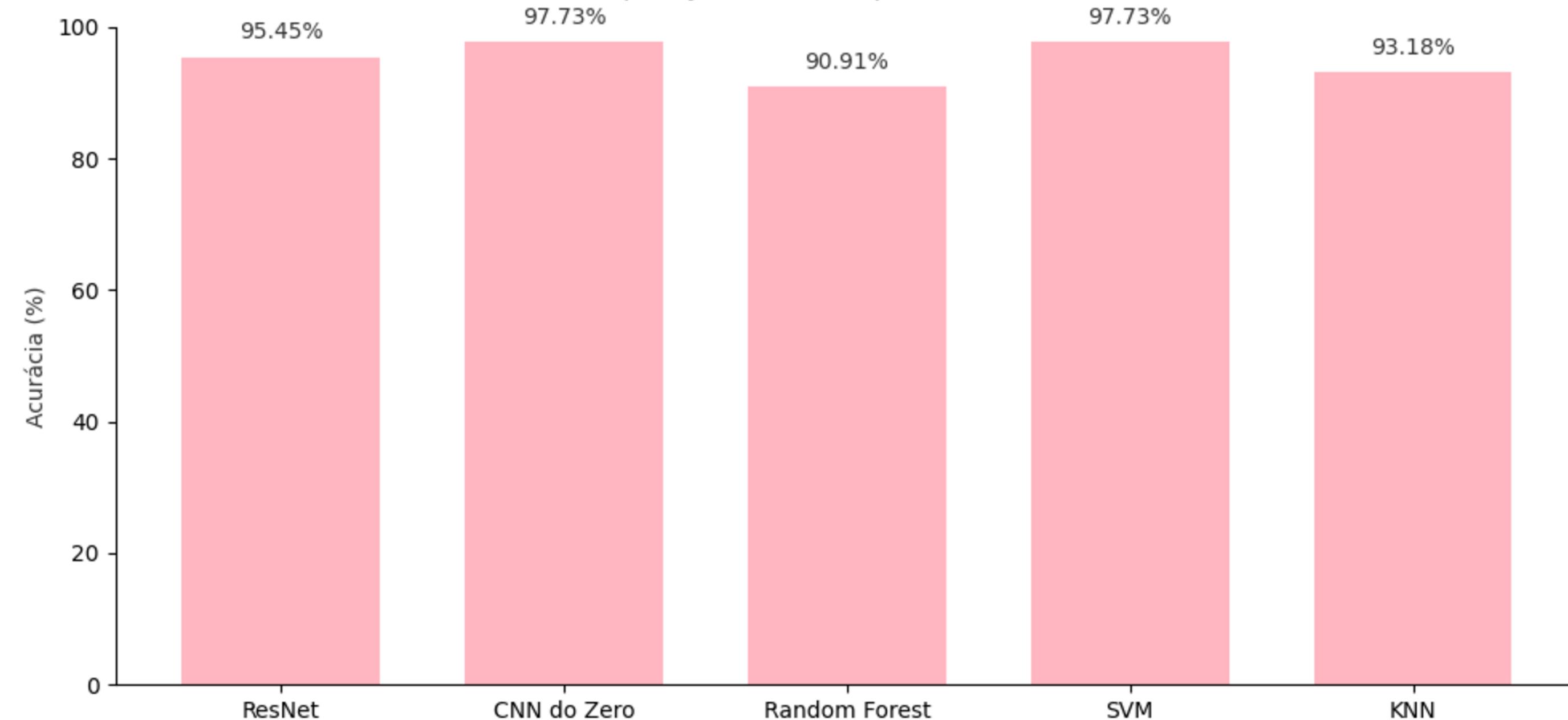
- n_neighbors: 7
- weights: 'uniform'
- metric: 'euclidean'

Support Vector Machine

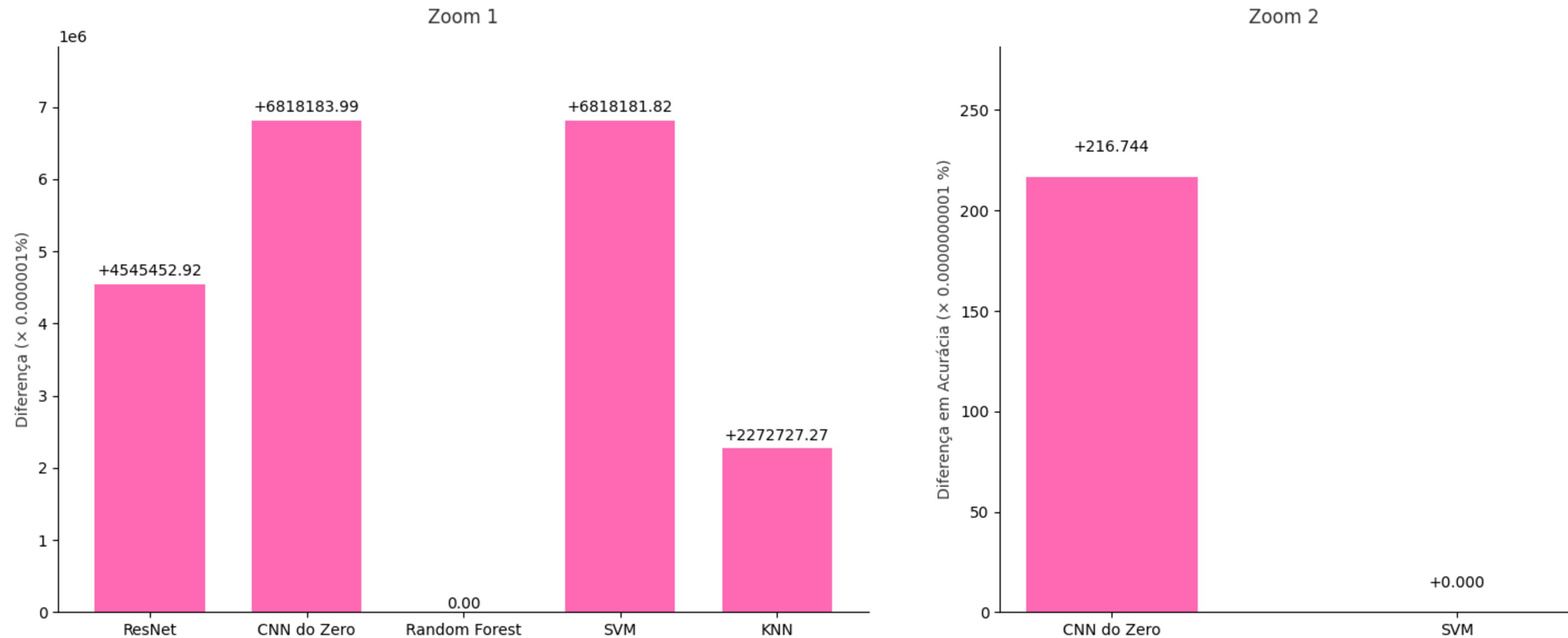
- C: 0.1
- kernel: 'linear'

Comparativo de Desempenho

Visão Geral



Comparativo de Desempenho Ampliado

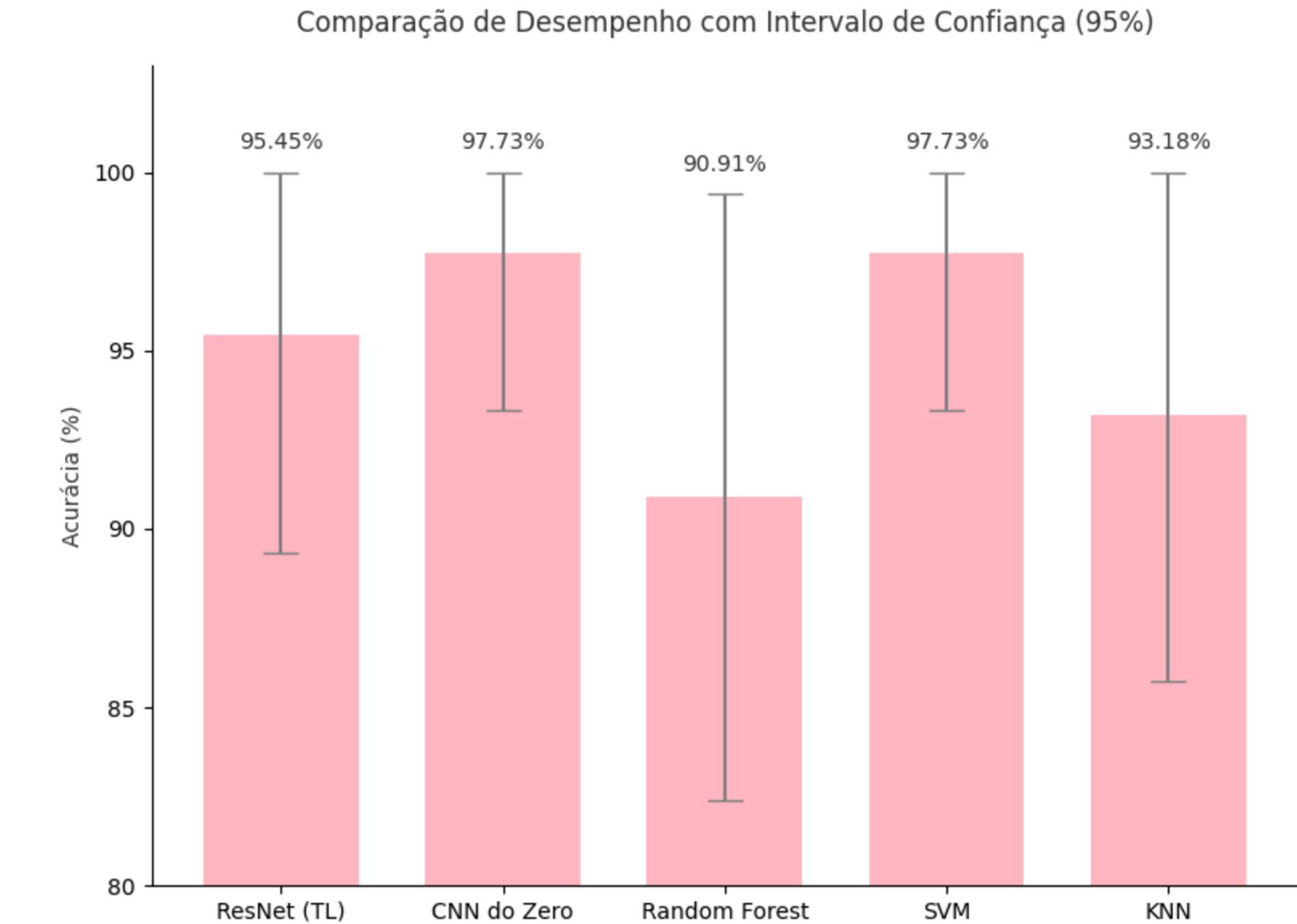


Comparativo de Desempenho

Intervalo de Confiança

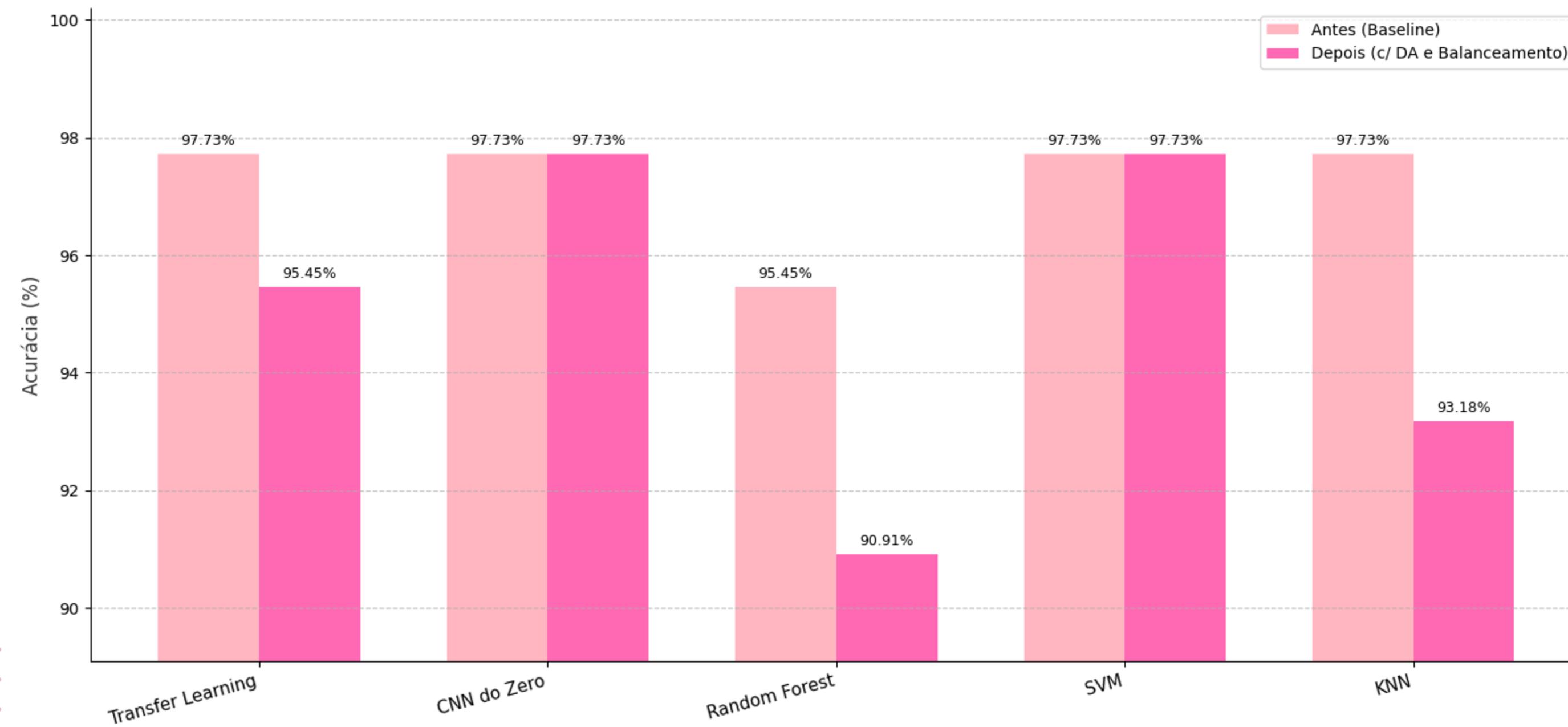
- IC definido em 95%
- Mede a incerteza da acurácia observada do modelo
- Indica a faixa de valores em que a verdadeira acurácia provavelmente se encontra.

	Acurácia (%)	IC (95%)
CNN do Zero	97.73	[93.32 - 100.00]
SVC	97.73	[93.32 - 100.00]
ResNet (TL)	95.45	[89.30 - 100.00]
KNN	93.18	[85.73 - 100.00]
Random Forest	90.91	[82.41 - 99.40]



Comparativo de Desempenho

Impacto do Data augmentation e Balanceamento de classes



Conclusão

- A aplicação de Aprendizado Profundo (DL), Transfer Learning e Visão Computacional demonstrou ser altamente eficaz na classificação de mamografias, apresentando o melhor desempenho de 97,7%
- Os modelos apresentaram alta robustez, servindo como uma excelente validação para o diagnóstico assistido.
- Esta pesquisa é a base para o aprimoramento contínuo, com foco em:
 - Expansão do dataset
 - Otimização de arquiteturas e escolha de hiperparâmetros
 - Validação externa

Referências

- BAKATOR, Mihalj; RADOSAV, Dragica. Deep learning and medical diagnosis: A review of literature. *Multimodal Technologies and Interaction*, v. 2, n. 3, p. 47, 2018.
- CHING, Travers et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, v. 15, n. 141, p. 20170387, 2018.
- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. Deep learning. Cambridge: MIT Press, 2016.
- INCA. Estimativa 2023: Incidência de Câncer no Brasil. Rio de Janeiro, 2024. Disponível em: gov.br/inca.
- LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*, v. 521, n. 7553, p. 436-444, 2015.
- MAMA. Câncer de Mama. São Paulo: A.C. Camargo, 2023. Disponível em: accamargo.org.br.
- MUNSHI, Raafat M. et al. A novel approach for breast cancer detection using optimized ensemble learning framework and XAI. *Image and Vision Computing*, v. 142, 2024.
- NAQA, Issam; MURPHY, Martin J. What is machine learning? In: Machine learning in radiation oncology. Cham: Springer, 2015. p. 3-11.
- OMS. Breast cancer now most common form of cancer: WHO taking action. Genebra, 2021. Disponível em: who.int.
- OMS. Breast cancer. Genebra, 2024. Disponível em: who.int.
- PICCIALLI, Francesco et al. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, v. 66, p. 111-137, 2021.
- RODRIGUES, Iago et al. Classifying COVID-19 positive X-ray using deep learning models. *IEEE Latin America Transactions*, v. 19, n. 6, p. 884-892, 2021.
- RUSSELL, Stuart; NORVIG, Peter. Artificial intelligence: a modern approach. 3. ed. New Jersey: Pearson, 2016.
- SANTANA, Maíra Araújo de et al. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, v. 34, p. 45-53, 2018.
- YADAV, Rahul Kumar; SINGH, Pardeep; KASHTRIYA, Poonam. Diagnosis of breast cancer using machine learning techniques - A Survey. *Procedia Computer Science*, v. 218, p. 1434-1443, 2023.