

Elaboração de Modelo Preditivo de Imagens 2D de Estruturas Químicas de Fármacos Anti-Hipertensivos

Bruna Rodrigues Cardoso^{1,2}, Rodrigo Bonacin¹, Ana Carolina Borges Monteiro¹

bruna.cardoso@cti.gov.br, rodrigo.bonacin@cti.gov.br,
amonteiro@cti.gov.br

¹Divisão de Metodologias da Computação – DIMEC
CTI/MCTI Renato Archer – Campinas/SP

²Pontifícia Universidade Católica de Campinas – Campinas/SP

Abstract. *This work evaluates the efficiency of Deep Convolutional Networks (CNNs) combined with different Machine Learning algorithms for classifying antihypertensive and non-antihypertensive drugs based on the analysis of two-dimensional (2D) images of their chemical structures. Traditional models such as Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), and Support Vector Machines (SVM) were used, as well as pre-trained CNN architectures such as MobileNet50 and ResNet50, applied as visual feature extractors. The experiments showed that the combination of MobileNetV2 and the Random Forest classifier achieved the best results in terms of accuracy 83.33% and F1-score of 80.62%, outperforming the other approaches. The results demonstrate the potential of using 2D visual representations combined with Machine Learning techniques to support the discovery and analysis of antihypertensive drugs.*

Resumo. *Este trabalho aborda a avaliação da eficiência das redes Convolucionais Profundas (CNN) em conjunto com diferentes algoritmos Aprendizado de Máquina para a classificação de fármacos anti-hipertensivos e não anti-hipertensivos, a partir da análise de imagens bidimensionais (2D) das estruturas químicas. Para isso, foram utilizados modelos tradicionais como Random-Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR) e Support Vector Machines (SVM), além de arquiteturas de CNN pré-treinadas, como a MobileNetV2 e a ResNet50, aplicadas como extratoras de características visuais. Os experimentos realizados evidenciaram que a combinação MobileNetV2 com o classificador Random Forest obteve os melhores resultados em termo de acurácia com 83,33% e F1-score 80,62%, superando as demais abordagens. Os resultados demonstraram o potencial do uso de representações visuais em 2D associadas a técnicas de Aprendizado de Máquina para apoiar processos de descoberta e análise de fármacos-anti-hipertensivos.*

1. Introdução

Hipertensão arterial é uma doença crônica caracterizada pelos níveis elevados da pressão sanguínea nas artérias, com valores iguais ou superiores a 140/90 mmHg [1]. Estima-se que 1,28 bilhão de adultos com idade entre 30 e 70 anos, sofram com a condição em todo o mundo, sendo que dois terços dessa população vivem em países de baixa e média renda [2]. Além disso, pode ser considerado fator de risco para desenvolvimento de doenças cardiovasculares, acidente vascular cerebral (AVC) e insuficiência renal. O tratamento adequado da hipertensão é essencial para reduzir essas complicações, e o uso de medicação anti-hipertensiva é o principal método para controle eficaz da pressão arterial e a melhora da qualidade de vida dos pacientes.

A pesquisa para desenvolvimento e comercialização de novos fármacos anti-hipertensivos é de elevado custo para a indústria farmacêutica, pois exige gastos com pesquisas, testes clínicos e processos regulatórios. Visando esse fato, técnicas baseadas em Inteligência Artificial (IA) têm emergido como ferramentas para auxiliar e otimizar a descoberta de medicamentos, podendo diminuir tempo e recursos necessários para a triagem de compostos bioativos.

Neste trabalho, são exploradas diferentes abordagens de Aprendizado de Máquina aplicadas à classificação de compostos químicos representados por imagens bidimensionais (2D) das estruturas químicas de fármacos anti-hipertensivos e não anti-hipertensivos. Esses modelos foram desenvolvidos e avaliados a partir de estruturas químicas extraídas da base de dados DrugBank, a qual reúne informações amplamente utilizadas em pesquisas na área de farmacologia. O objetivo é avaliar a eficácia de técnicas baseadas em redes convolucionais profundas (CNN), como MobileNetV2 e ResNet50, para que seja possível identificar fármacos com ação anti-hipertensiva. Foram utilizados diferentes modelos de aprendizado de máquina, como *Random-Forest* (RF), *Multilayer Perceptron* (MLP), *Logistic Regression* (LR) e *Support Vector Machines* (SVM) sobre as *features* extraídas para encontrar combinações eficazes na distinção entre fármacos anti-hipertensivos e não anti-hipertensivos. Os resultados obtidos oferecem evidências para o uso da IA como ferramenta estratégica na indústria farmacêutica, ao possibilitar a redução de custos e o aumento da eficiência em etapas do processo de pesquisa, desenvolvimento e descoberta de novos medicamentos.

O restante deste documento está estruturado da seguinte maneira: a seção 2 apresenta a metodologia utilizada, a seção 3 apresenta os resultados obtidos, a seção 4 faz considerações e conclusões com base nos resultados obtidos e, por fim, a seção 5 apresenta os próximos passos desta pesquisa.

2. Metodologia

Para o desenvolvimento deste trabalho, foi utilizado um banco de dados composto por imagens bidimensionais (2D) de estruturas químicas, classificadas em duas categorias: fármacos anti-hipertensivos com 128 imagens e não anti-hipertensivos composto por 50 imagens. As imagens foram coletadas da base de conhecimento DrugBank [3], a qual reúne informações detalhadas sobre medicamentos.

Inicialmente, as imagens das estruturas químicas dos fármacos foram submetidas a uma etapa de pré-processamento com o objetivo de garantir uniformidade no conjunto

de dados. As técnicas aplicadas incluíram o redimensionamento das imagens para dimensões fixas e assim permitir a padronização da entrada para os modelos de aprendizado de máquina. A normalização dos valores dos pixels, foram convertidos para o intervalo entre 0 e 1 para melhorar a estabilidade e o desempenho durante o treinamento, e quando necessário a conversão das imagens para escala de cinza, com o objetivo de reduzir a complexidade computacional e evitar perda de informações. Essas etapas prepararam de forma adequada os dados para os classificadores e foram implementadas utilizando a biblioteca OpenCV (Open Source Computer Vision Library), amplamente utilizada em visão computacional e processamento de imagens [4] escolhida devido à eficiência em operações como redimensionamento, transformação e manipulação de matrizes de pixels. A Pillow (PIL Fork) é uma biblioteca em Python destinada ao processamento de imagens [5] aplicada neste trabalho por sua simplicidade na leitura, manipulação e conversão de formatos de imagem. Já o TensorFlow/Keras é um framework de aprendizado de máquina e aprendizado profundo [6], adotado por fornecer ferramentas nativas para o pré-processamento de imagens, o que facilitou a integração direta com os modelos de redes neurais convolucionais.

A partir do conjunto das imagens pré-processadas, foram utilizados diferentes modelos de aprendizado de máquina para realizar a classificação binária e desta maneira distinguir fármacos anti-hipertensivos dos não anti-hipertensivos. Para isso, aplicaram-se técnicas baseadas em Redes Neurais Convolucionais (CNN) [7] com transferência de aprendizado, aproveitado o conhecimento adquirido por modelos previamente treinados em grandes bases de dados como o ImageNet [8]. Foram aplicadas duas arquiteturas a MobileNetV2, uma rede neural leve e eficiente [9] e a ResNet50, que utiliza conexões residuais para facilitar o treinamento de redes profundas e melhorar o desempenho [10]. Essas redes foram empregadas como extratoras de características visuais, permitindo a captura de padrões relevantes de imagens. As representações obtidas foram utilizadas como entrada para quatro tipos de técnicas de aprendizado de máquina (ML): *Random-Forest* (RF), *Multilayer Perceptron* (MLP), *Logistic Regression* (LR) e *Support Vector Machines* (SVM). O processo seguiu o modelo de treinamento 80/20, onde 80% do conjunto de dados é utilizado para treinar e 20% para testes.

Para avaliar o desempenho dos classificadores as métricas adotadas foram acurácia, precisão e F1-score.

3. Resultados

A partir dos modelos de Redes Neurais Convolucionais utilizados como extratores de características (MobileNetV2 e ResNet50), em conjunto com diferentes algoritmos de aprendizado de máquina, foi possível avaliar a capacidade de classificação binária entre fármacos anti-hipertensivos e não anti-hipertensivos. As métricas escolhidas para a avaliação foram: acurácia e o F1-score, pois juntas oferecem uma visão criteriosa do desempenho. A acurácia quantifica a proporção de acertos totais, enquanto o F1-score, que combina precisão e recall, mede o equilíbrio entre a capacidade de identificar corretamente os exemplos positivos e a cobertura das classes. Devido a isso, neste estudo, o F1-score foi adotado como principal métrica para determinação da classificação final.

Na Figura 1 é possível observar os resultados da combinação do algoritmo da Rede Neural Convolutacional MobileNetV2 aplicada junto aos quatro tipos de aprendizado de máquina: Random Forest, MLP, Logistic Regression e SVM. O modelo Random Forest apresentou melhor desempenho ao identificar estruturas químicas de fármacos pertencentes à classe de anti-hipertensivos. Conforme apresentado, o resultado da acurácia é de 83,33%, precisão de 86,45%, Recall de 83,33% e F1-Score de 80,62%. Esses valores indicam elevada capacidade preditiva e mostra distribuição proporcional de acertos entre as classes. Além disso, o F1-score por integrar precisão e sensibilidade em uma única métrica reforça a confiabilidade do desempenho. Em seguida, os modelos de aprendizado de máquina MLP, Logistic Regression e SVM apresentaram respectivamente acurácia de 72,22%, precisão de 74,04% e F1-score de 72,92% (MLP e Logistic Regression, com desempenho iguais), e acuraria de 69,44%com F1-score de 70% (SVM), embora apresentem resultados satisfatórios, não superam a performance alcançada pelo Random Forest na atribuição da classificação proposta.

Extractor	Modelo	Accuracy	Precision	Recall	F1-score
MobileNetV2	Random Forest	0.833333	0.864583	0.833333	0.806240
MobileNetV2	MLP	0.722222	0.740741	0.722222	0.729293
MobileNetV2	Logistic Regression	0.722222	0.740741	0.722222	0.729293
MobileNetV2	SVM	0.694444	0.724823	0.694444	0.705018

Figura 1. Resultados das Rede Neural MobileNetV2 de acordo com o modelo de aprendizado de máquina

Na Figura 2 é apresentada a Matriz de Confusão do melhor resultado apresentado: MobileNetV2 e Random Forest. Dentre as 36 imagens das estruturas químicas do fármacos anti-hipertensivos - 20% das imagens do banco de dados foram utilizadas no conjunto de teste – 26 foram classificados como verdadeiros positivos, nenhuma estrutura foi classificada erroneamente como não anti-hipertensivo, 6 amostras não anti-hipertensivas foram classificadas erroneamente como anti-hipertensivos e 4 amostras foram classificadas de forma correta como não anti-hipertensivos.

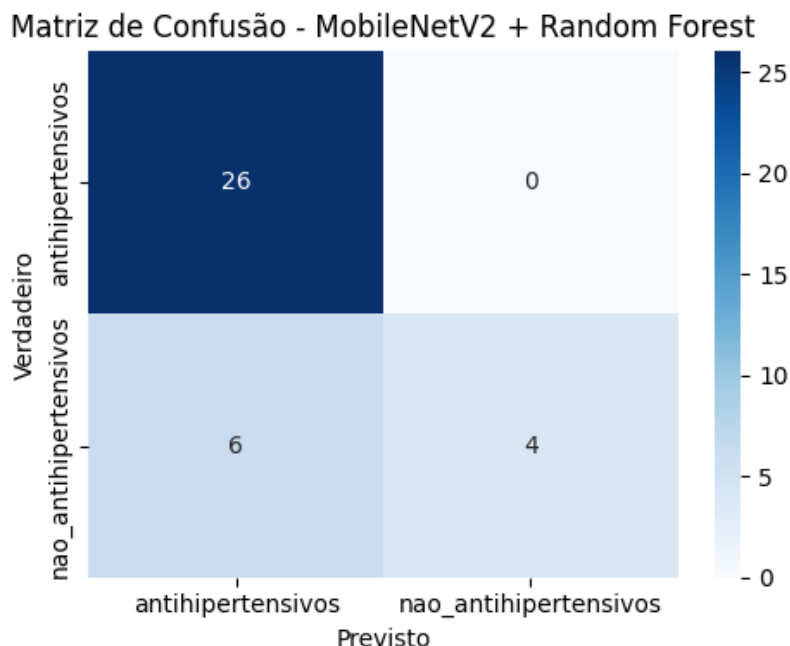


Figura 2. Matriz de Confusão – Resultado entre MobileNetV2 e Random Forest

A figura 3 apresenta os resultados da combinação do algoritmo da Rede Neural Convolucional ResNet50 aplicada aos quatro classificadores de aprendizado de máquina: Random Forest, MLP, Logistic Regression e SVM. Entre esses, o modelo SVM apresentou melhor desempenho, com valores de acurácia, precisão, recall e F1-score de 66,66%. Esse resultado, apesar de moderado indica que o SVM conseguiu dispor de forma mais eficiente as características extraídas pela ResNet50. Em seguida, os modelos Random Forest e MLP apresentaram desempenho de acurácia de 72,22% e 61,11%, respectivamente e de F1-score de 64,81 e 59,67%, respectivamente. O Logistic Regression obteve menor desempenho alcançando acurácia de 58,83% e F1-score de 58,91%. Esses resultados de forma geral sugerem que o ResNet50 apresentou menor capacidade de extração de características no conjunto de imagens químicas analisada quando comparada à MobileNetV2.

Extractor	Modelo	Accuracy	Precision	Recall	F1-score
ResNet50	SVM	0.666667	0.666667	0.666667	0.666667
ResNet50	Random Forest	0.722222	0.669935	0.722222	0.648148
ResNet50	MLP	0.611111	0.585317	0.611111	0.596708
ResNet50	Logistic Regression	0.583333	0.595758	0.583333	0.589169

Figura 3. Resultados das Rede Neural ResNet50 de acordo com o modelo de aprendizado de máquina

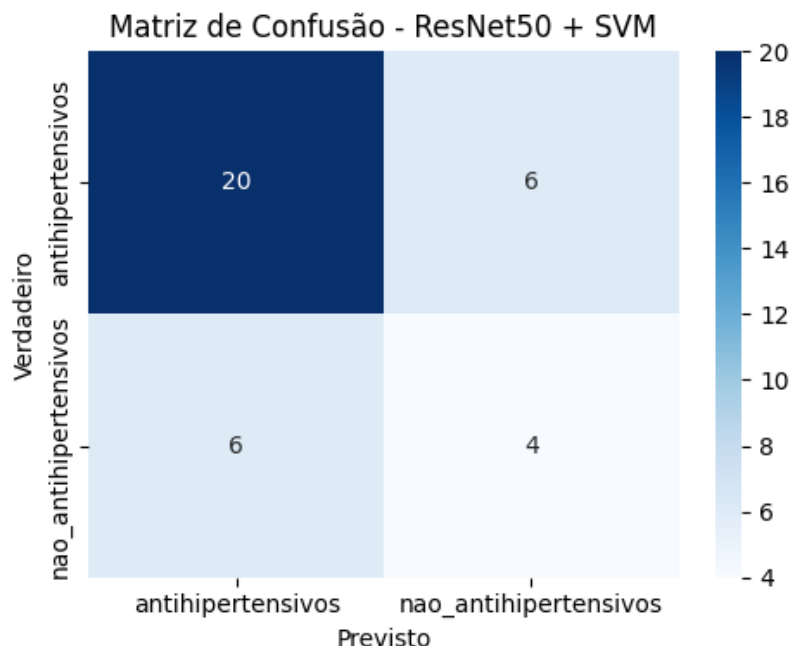


Figura 4. Matriz de Confusão – Resultado entre ResNet50 e SVM

Na Figura 4 é apresentada a Matriz de Confusão do melhor resultado apresentado com ResNetV2 e SVM. Dentre as 36 imagens das estruturas químicas do fármacos anti-hipertensivos - 20% das imagens do banco de dados foram utilizadas no conjunto de teste – 20 foram classificados como verdadeiros positivos, 6 estruturas foram classificadas erroneamente como não anti-hipertensivo, 6 amostras não anti-hipertensivas foram classificadas erroneamente como anti-hipertensivos e 4 amostras foram classificadas de forma correta como não anti-hipertensivos. Esses resultados indicam desempenho inferior na detecção da classe de interesse, principalmente ao analisar o falso-negativo. Sendo assim, observa-se que o modelo apresentou limitação na extração das principais características discriminantes das estruturas químicas para classifica-las como anti-hipertensivo ou não anti-hipertensivo.

4. Considerações e Conclusões

Este estudo analisou a eficiência dos modelos de Redes Neurais Convolucionais utilizados como extratores de características associados a modelos de aprendizagem de máquina na predição de estruturas de fármacos anti-hipertensivos, com base nas estruturas moleculares em 2D.

Os resultados obtidos indicam que a escolha da arquitetura da Rede Neural Convolucional utilizada como extratora de características influencia diretamente o desempenho dos modelos de classificação. Entre os métodos testados, a MobileNetV2 em conjunto com o algoritmo Random Forest, apresentou desempenho superior ao alcançar os maiores valores de acurácia e F1-score. Esses resultados são promissores, pois indicam o potencial da abordagem em identificar fármacos com ação anti-hipertensiva de forma confiável. Além disso, a utilização de arquiteturas de CNN como a MobileNetV2 demonstra a relevância do aprendizado profundo na análise de imagens

química, pois permite a extração de padrões complexos que dificilmente seriam capturados por técnicas tradicionais de representação molecular. Dessa forma, a integração de CNNs pré-treinadas com algoritmos de aprendizado supervisionado pode representar um caminho eficiente para aplicações em farmacologia computacional, contribuindo para a triagem virtual de compostos e a otimização de processos de descoberta de novos medicamentos.

O modelo ResNet50 em conjunto com Logistic Regression apresentou menor desempenho, o que indica que foi menos eficaz na identificação de fármacos anti-hipertensivos. Isso indica que apesar de ser um modelo eficaz em outros contextos, não foi o mais adequado para este tipo específico de análise, o que pode ser devido a configuração dos dados utilizados desta pesquisa.

5. Perspectivas e Trabalho Futuro

Os resultados obtidos neste estudo evidenciam o potencial uso de arquitetura de Redes Neurais Convolucionais como extratoras de características em conjunto com classificadores tradicionais de aprendizado de máquina para a triagem de compostos anti-hipertensivos.

Como trabalhos futuros pretende-se ampliar a base de dados para a utilização de representações tridimensionais (3D) das estruturas moleculares, para explorar técnicas de aprendizado profundo capazes de capturar informações espaciais e mais detalhadas do que as imagens bidimensionais. Essa linha de pesquisa poderá aproximar ainda mais os modelos computacionais da realidade química e biológica dos compostos, favorecendo aplicações práticas de descoberta de fármacos e consolidando a inteligência artificial como recurso estratégico no desenvolvimento de novos fármacos.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) sob o número 124618 pelo fomento ao desenvolvimento dessa pesquisa.

6. Referências

- [1] MINISTÉRIO DA SAÚDE. Hipertensão (pressão alta). Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/h/hipertensao>. Acesso em: 16 abr. 2025.
- [2] ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). Hypertension. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/hypertension>. Acesso em: 18 abr. 2025.
- [3] DRUGBANK. DrugBank Online. Disponível em: <https://go.drugbank.com/>. Acesso em: 18 fev. 2025.
- [4] OPENCV. OpenCV-Python. Disponível em: <https://pypi.org/project/opencv-python/>. Acesso em: 23 jan. 2025.

- [5] PYTHON IMAGING LIBRARY. Pillow. Disponível em: <https://pypi.org/project/pillow/>. Acesso em: 23 jan. 2025.
- [6] TENSORFLOW. TensorFlow – uma plataforma open source de aprendizado de máquina. Disponível em: <https://www.tensorflow.org/?hl=pt-br>. Acesso em: 23 jan. 2025.
- [7] YAMASHITA, R., NISHIO, M., DO, R.K.G. *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 9, 611–629 (2018). <https://doi.org/10.1007/s13244-018-0639-9>
- [8] DENG, J. *et al.* ImageNet: A Large-Scale Hierarchical Image Database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, Miami, FL. Proceedings [...]. Piscataway: IEEE, 2009. p. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [9] K. DONG, C. ZHOU, Y. RUAN and Y. Li, "MobileNetV2 Model for Image Classification," 2020 2nd International Conference on Information Technology and Computer Application (ITCA), Guangzhou, China, 2020, pp. 476-480, doi: 10.1109/ITCA52113.2020.00106.
- [10] K. HE, X. ZHANG, S. REN and J. SUN, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.