Aplicação de aprendizado de máquina para diagnóstico do câncer de mama: uma abordagem baseada em Random Forest

Silva JVR[1], Araújo BS[1], Amorim JR[1], Nunes HFO[1], Viana NI[2]

[1]Acadêmicos do curso de Medicina da Universidade do Estado de Minas Gerais;

[2]Doutora em Ciências e Docente da Universidade do Estado de Minas Gerais

joao.241216338@discente.uemg.br

Introdução: O câncer de mama é a forma mais comum de câncer entre as mulheres, representando uma das principais causas de morbidade e mortalidade mundial. A detecção precoce é fundamental para o sucesso do tratamento e melhora na sobrevida. O avanço da tecnologia, tem possibilitado o uso de ferramentas computacionais para auxiliar no diagnóstico, sendo o aprendizado de máquina uma abordagem promissora. O modelo Random Forest, uma técnica de aprendizado supervisionado, é eficaz para identificar padrões complexos em grandes volumes de dados, tornando-o uma ferramenta relevante no diagnóstico precoce de tumores de mama. Objetivo: O objetivo deste estudo foi aplicar o algoritmo Random Forest para classificar tumores de mama como benignos ou malignos, utilizando um conjunto de dados clínicos e físicos extraídos de amostras de tecido tumoral. Métodos: O conjunto de dados utilizado foi o Conjunto de Dados de Câncer de Mama de Wisconsin, que contém 569 amostras de tumores com 30 características extraídas de imagens digitalizadas. As variáveis independentes foram as características físicas dos tumores, e a variável dependente foi o diagnóstico (benigno ou maligno). Os dados foram divididos em conjuntos de treino (80%) e teste (20%). Após a limpeza e preparação dos dados, o modelo Random Forest foi treinado utilizando 100 árvores de decisão. A normalização dos dados foi realizada para garantir consistência nas escalas das variáveis. A avaliação da performance do modelo foi feita com base em métricas de acurácia, precisão, recall e F1-score. Resultados: O modelo Random Forest obteve uma acurácia de 96,5%, demonstrando sua eficácia em classificar corretamente os tumores de mama. A precisão foi de 0.98, o recall foi de 0.93 e o F1-score foi de 0.95 para tumores malignos. As características mais influentes para a decisão do modelo foram o raio, perímetro e textura, que estão diretamente associados ao crescimento e comportamento dos tumores. Conclusão: Este estudo demonstrou que o modelo Random Forest é uma ferramenta eficaz para a classificação de tumores de mama, com resultados altamente precisos. A aplicação de aprendizado de máquina, especialmente com o uso de Random Forest, tem o potencial de melhorar o diagnóstico precoce, otimizar os tratamentos e, consequentemente, aumentar as chances de sucesso terapêutico. No entanto, mais estudos com dados ampliados e a utilização de outros algoritmos podem aumentar a robustez e a precisão do modelo.

Palavras-chave: Aprendizado de máquina, Câncer de mama, Random Forest, Diagnóstico precoce, Oncologia molecular

Referências bibliográficas:

1. Dallagassa MR, Oliveira AJC. Uso de machine learning no diagnóstico de câncer de mama através de ultrassonografia. J Health Inform. 2024;16(Esp):36-40. Disponível em: https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/1289. Acesso em: 24 mar. 2025.
2. Rakhlin A, Shvets A, Iglovikov V, Kalinin AA. Deep convolutional neural networks for breast cancer histology image analysis. In: Campilho A, Karray F, ter Haar Romeny B, eds. Image Analysis and Recognition. Cham: Springer; 2018. p. 737-744. (Lecture Notes in Computer Science; vol 10882). doi:10.1007/978-3-319-93000-8_83.
3. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride RB, Sieh W. Deep learning to improve breast cancer early detection on screening mammography. Sci Rep. 2019;9(1):12495. doi:10.1038/s41598-019-48995-4.
4. Khamparia A, Pandey B, Gupta D, Khanna A, Bhattacharya S, Tiwari P. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. Comput Electr Eng. 2020;85:106653. doi:10.1016/j.compeleceng.2020.10665