

Modelos de Aprendizado de Máquina na Detecção de Câncer de Mama Utilizando Características GLCM e de Primeira Ordem[☆]

Machine Learning Models in Breast Cancer Detection Using GLCM and First Order Features

Matheus Santos Silva, Dany Sanchez Dominguez[†], Hélder Conceição de Almeida, Paulo Eduardo Ambrosio, Susana Marrero Iglesias

¹Universidade Estadual de Santa Cruz, Ilhéus, Brasil

[†]**Autor correspondente:** dsdominguez@uesc.br

Resumo

Entre as mulheres, o câncer de mama é um dos tipos de câncer com maior incidência e letalidade no mundo. Apesar da alta taxa de letalidade, o câncer de mama tem alta porcentagem de cura e diagnósticos favoráveis quando diagnosticado em estágios iniciais. A mamografia é considerada o melhor método de detecção, contudo, suas imagens são complexas, o que torna a análise diagnóstica suscetível a erros. Uma das formas de reduzir os erros é o uso de métodos computadorizados para auxílio ao diagnóstico. Com o objetivo de contribuir para o diagnóstico preciso desta doença, neste trabalho, comparamos três modelos de aprendizado de máquina na detecção de câncer de mama usando o banco de imagens de mamografia MIAS, a partir de características extraídas da matriz de co-ocorrência de níveis de cinza (GLCM) e de primeira ordem. Os modelos avaliados são o *K-Nearest neighbor* (KNN), *random forest* e XGBoost. Os resultados mostram que os modelos testados não obtiveram resultados com alto grau de precisão. Entre os modelos avaliados, o XGBoost obteve o melhor resultado de detecção.

Palavras-chave

Características de primeira ordem • Matriz de co-ocorrência de níveis de cinza • Aprendizado de máquina • Detecção de câncer de mama

Abstract

Among women, breast cancer is one of the types of cancer with the highest incidence and lethality in the world. Despite the high lethality rate, breast cancer has a high percentage of cure and favorable diagnoses when diagnosed in early stages. Mammography is considered the best method of detection, however, its images are complex, which makes the analysis susceptible to errors. One of the ways to reduce diagnostic errors is the use of computerized methods to aid diagnosis. To contribute to the accurate diagnosis of this disease, in this work, we compared three machine learning models for breast cancer detection using the MIAS mammography image database, based on features extracted from the gray level co-occurrence matrix and first order features. The models evaluated are K-Nearest neighbor (KNN), random forest and XGBoost. The result show that the tested models did not obtain result with high degree of accuracy. Among the models evaluated, XGBoost obtained the best result.

[☆] Este artigo é uma versão estendida do trabalho apresentado no XXVII ENMC Encontro Nacional de Modelagem Computacional e XV ECTM Encontro de Ciência e Tecnologia de Materiais, ocorridos em Ilhéus – BA, de 1 a 4 de outubro de 2024.

Keywords

First order features • Gray level co-occurrence matrix • Machine Learning • Breast cancer detection

1 Introdução

Entre as mulheres, o câncer de mama é um dos tipos de câncer com maior incidência e letalidade no mundo. Ele é a segunda principal causa de mortes por câncer entre as mulheres em escala global e a principal causa nos países em desenvolvimento. Estima-se que a cada ano aproximadamente 22% dos novos casos de câncer correspondem ao câncer de mama, sendo responsável por um alto número de mortes de mulheres adultas [1]. Contudo, as medidas de prevenção, diagnóstico e controle da doença não tem acompanhado o mesmo ritmo de crescimento dos casos [2].

Apesar da alta taxa de letalidade, o câncer de mama tem alta porcentagem de cura e prognósticos favoráveis quando diagnosticados em estágios iniciais [1]. Entre os meios de detecção precoce de câncer de mama, os mais eficazes são o exame clínico de mama (ECM) e a mamografia [2]. O ECM pode detectar tumores superficiais de até 1,0 cm e exige apenas um médico ou enfermeira treinados. Já a mamografia, é realizada utilizando um equipamento de mamografia e é necessário que um especialista avalie as imagens geradas pelo exame. A mamografia é considerada o melhor método de detecção precoce, pois consegue detectar tumores menores em quaisquer localização [3].

Um exame de mamografia geralmente é composto por 4 imagens, duas para cada mama [3]. Essas imagens podem ser extremamente complexas, o que torna desafiador sua análise, até mesmo para especialista experientes, sendo altamente suscetível a erros [4]. Muitos fatores podem influenciar na complexidade dessas imagens, desde lesões extremamente pequenas a fatores relacionados a aquisição das imagens, que demanda manutenção e calibração dos aparelhos e técnicos qualificados para a realização do exame [3]. Para reduzir os erros de diagnósticos, existem algumas soluções, como o uso de uma segunda opinião e o uso de sistemas computadorizados de auxílio ao diagnóstico (CAD, *Computed Aided Diagnosis*). O uso de uma segunda opinião pode ser inviável na maioria dos hospitais devido à falta de profissionais treinados [3]. Já os sistemas CAD vêm ganhando grande notoriedade, principalmente devido ao advento da inteligência artificial e da *big data*, que tem reduzido as limitações enfrentadas pelos métodos [4].

O uso de características extraídas da matriz de co-ocorrência de níveis de cinza (GLCM, *Gray Levels Co-occurrence Matrix*) tem sido aplicado na classificação de imagens e detecção em diversas áreas, como na classificação de *Lasem Batik* [5], na classificação de imagens histopatológicas [6] e na detecção de tumores no cérebro [7]. Estudos recentes, como [8], [9] e [10] também demonstraram, com sucesso, a aplicação de características extraídas de matriz GLCM em conjunto com métodos de aprendizado de máquina na detecção de câncer de mama em imagens de mamografias.

Neste trabalho implementamos três modelos de aprendizado de máquina para detecção de câncer de mama, utilizando característica extraídas da GLCM e de primeira ordem, e comparamos seus resultados no banco público de imagens de mamografia MIAS. O objetivo deste trabalho é realizar uma exploração preliminar do problema visando avaliar a consistência do banco e o uso de técnicas simples de aprendizado de máquina em sua abordagem. Os modelos implementados são o *K-Nearest neighbor* (KNN), *random forest* e XGBoost.

2 Conjunto de Dados e Modelos

Para realização desta pesquisa, utilizamos o banco de imagem de mamografia *Mammographic Image Analysis Society* (MIAS). O MIAS é um banco de imagens público desenvolvido por um grupo de pesquisa do Reino Unido, sendo um dos bancos mais utilizados em pesquisas na área de análise de imagens médicas. Ele é composto por 330 imagens no formato *Portable Gray Map* (PGM), com todas as imagens apresentando dimensões de 1024x1024 pixels. Dentre as 330 imagens do banco, 207 tem patologias classificadas como normais, 69 como benignas e 54 como malignas.

Para extração das características das imagens, utilizamos a matriz GLCM, um dos métodos mais tradicionais para extração de características de imagens [11], em conjunto com características de primeira ordem. Essas características são amplamente utilizadas em problemas de detecção [12]. A matriz GLCM considera uma direção e um ângulo para comparar a ocorrência de níveis de cinza de dois pixels vizinhos [13]. Por outro lado, as características de primeira ordem são extraídas com base na distribuição de intensidade dos pixels.

O desenvolvimento desta pesquisa foi realizado em duas etapas, sendo elas: (1) Pré-processamento dos dados e extração das características; (2) Treinamento e avaliação dos modelos de detecção.

Na etapa de pré-processamento dos dados e extração das características, iniciamos realizando um recorte retangular nas imagens em torno da região de interesse (ROI) especificada nos metadados das imagens do banco. O recorte realizado nas imagens tem largura e altura iguais a 2 vezes o raio da região de interesse. Em seguida, seguimos com o aprimoramento das imagens do banco de dados. Para isso, utilizamos, respectivamente, o filtro de mediana para reduzir o ruído e *Contrast Limited Adaptive Histogram Equalization* (CLAHE) para melhorar o contraste das imagens. A Figura 1 a seguir mostra um exemplo de imagem após o procedimento de recorte e aprimoramento de imagem.

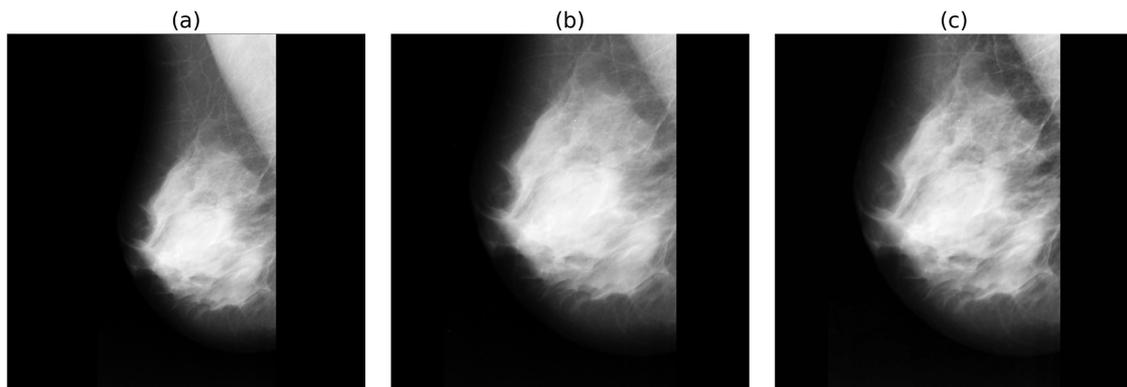


Figura 1: (a) Imagem original; (b) Imagem após recorte; (c) Imagem recortada após filtro de mediana e CLAHE.

Como próximo passo, realizamos a extração das características GLCM e de primeira ordem das imagens. As características GLCM extraídas foram dissimilaridade, correlação, homogeneidade, contraste, energia e momento angular de segunda ordem (ASM). Os ângulos utilizados foram 0° , 45° , 90° e 135° graus, com uma distância de 1. As características de primeira ordem utilizadas foram a média, desvio padrão, suavidade, terceiro momento, uniformidade e entropia. Para melhor entendimento das características GLCM e de primeira ordem extraídas, mais informações podem ser encontradas em [13] e [9], respectivamente. As características, valores angulares e distâncias foram selecionados com base na literatura, e em testes de avaliação de modelos.

Após extrair as características, extraímos dos metadados do banco MIAS os diagnósticos da patologia e os atribuímos para suas respectivas imagens. As imagens com patologia classificadas como normais foram descartadas, resultando em apenas 123 imagens. Para balancear as classes, utilizamos a estratégia de *oversampling*, que gera novas amostras da classe minoritária. Optamos por utiliza essa abordagem devido à baixa quantidade de imagens do banco.

Como últimos passos da etapa de pré-processamento e extração de características, realizamos a padronização dos dados, tratamos os rótulos categóricos e dividimos os dados em conjuntos de treino e testes. Para padronizar, utilizamos o método *StandardScaler*, que transforma os dados em novos valores com média 0 e desvio padrão constante. Para divisão dos dados, utilizamos 80% dos dados para treinamento, os quais serão utilizados para treinar e realizar ajustes de parâmetros dos modelos por meio da validação cruzada, e os outros 20% foram utilizados para avaliar o modelo. A Figura 2 a seguir apresenta o fluxograma das etapas de pré-processamento dos dados e extração das características.

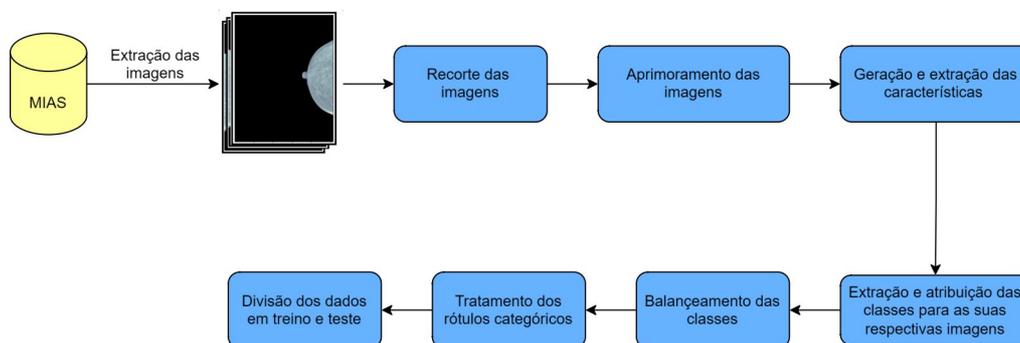


Figura 2: Fluxograma das etapas de pré-processamento dos dados e extração das características.

Como passo inicial da fase de treinamento e avaliação dos modelos de detecção, iniciamos com a seleção dos modelos. Os modelos escolhidos foram o KNN, *random forest* e XGBoost. Optamos por esses modelos devido ao fato de já terem sido amplamente utilizados em outras pesquisas de detecção que empregam características extraídas da matriz GLCM e/ou de primeira ordem.

Em seguida, realizamos a seleção dos melhores parâmetros dos modelos e apuramos quais características extraídas são mais relevantes por meio da validação cruzada no conjunto de treino utilizando K igual a 5. Ao analisar, observamos que os modelos obtiveram resultados melhores com a seguinte combinação das características: média, desvio padrão, suavidade, uniformidade, dissimilaridade (distância 1, ângulo 135°), correlação (distância 1, ângulos 0°, 45° e 90°), homogeneidade (distância 1, ângulo 0°), contraste (distância 1, ângulo 0°), energia (distância 1, ângulo 0°) e ASM (distância 1, ângulos 0°, 45°, 90° e 135°). Após encerrado o treinamento, o último passo, envolve avaliar os resultados dos modelos no conjunto de teste. A métrica utilizada para avaliação dos modelos foi a acurácia.

3 Resultados e Discussões

Nesta seção serão apresentados os resultados dos modelos implementados neste trabalho na detecção de câncer de mama utilizando características extraídas da matriz GLCM e de primeira ordem. Na Figura 3 a seguir, ilustramos as saídas dos modelos para cada cenário possível na detecção de câncer, ou seja, câncer detectado, câncer não detectado e falhas. Essas imagens, apresentadas na Fig. 3, podem mostrar como geralmente são as imagens para cada uma das alternativas, dessa forma, auxiliando na compreensão e na identificação de fatores que possam contribuir para o aprimoramento desses modelos.

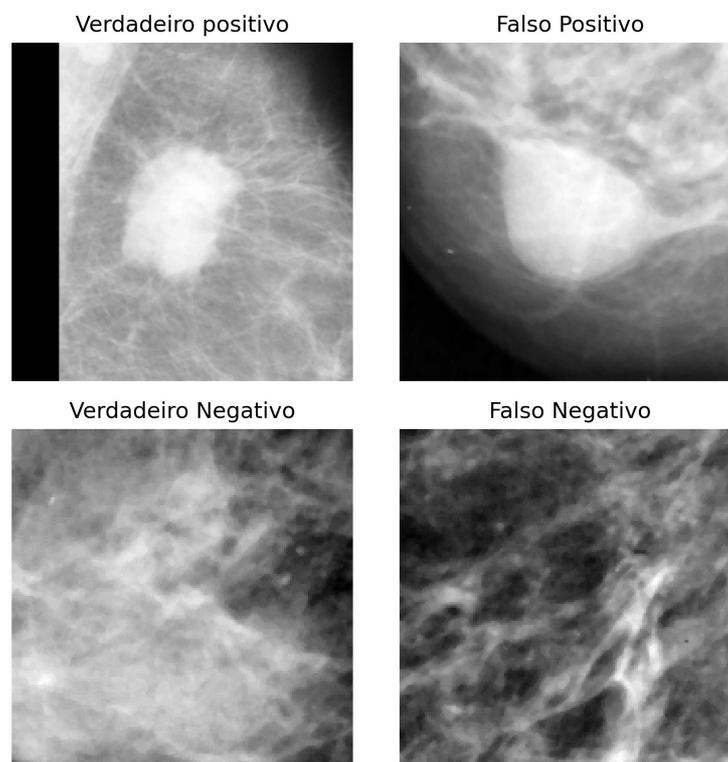


Figura 3: Saídas de câncer detectado, não detectado e falhas geradas pelos modelos.

Na Tabela 1 a seguir, apresentamos os resultados de detecção dos modelos no conjunto de teste do banco de imagens MIAS. Ao analisar os resultados, notamos que os modelos não obtiveram resultados com alto grau de precisão. Em geral, os modelos apresentaram resultados de detecção próximos uns dos outros, com valores entre 60% e 70% de acurácia. Entre os resultados, o modelo XGBoost foi o que obteve melhor resultado, com 69,15% de acurácia, enquanto o *Random Forest* teve pior resultado, com 63,33% de acurácia, sendo ligeiramente inferior ao resultado obtido pelo KNN.

Tabela 1: Resultados de detecção dos modelos no conjunto de teste.

Modelo	Acurácia (%)
<i>K-Nearest Neighbor</i>	64,96
<i>Random Forest</i>	63,33
<i>XGBoost</i>	69,15

Na Tabela 2, apresentamos os resultados da matriz de confusão do XGBoost, aquele modelo que apresentou o melhor resultado de detecção considerando sua acurácia. O XGBoost obteve 17 acertos de um total de 24 amostras do conjunto de teste, apresentando resultados semelhantes entre os acertos como verdadeiro positivo e verdadeiro negativo. Já entre as detecções falhas, a matriz de confusão mostra um marcado desbalanço, com o modelo apresentando maior quantidade de erros como falso negativo. Este comportamento, pode ser considerada uma característica negativa uma vez que o falso negativo pode trazer graves prejuízos ao paciente.

Tabela 2: Matriz de confusão do resultado de detecção do XGBoost.

	Detectado Positivo	Detectado Negativo
Real Positivo	9	5
Real Negativo	2	8

Na literatura encontramos diversos trabalhos [9, 10, 14, 15] que oferecem resultados com alta acurácia para o problema abordado nesta pesquisa. Ao comparar os resultados obtidos pelos modelos propostos com os encontrados na literatura, vimos que nossos resultados apresentaram uma grande discrepância em relação aos reportados, onde valores de acurácia superiores a 90% foram alcançados. Essa grande discrepância nos resultados pode ser atribuída ao uso de técnicas mais avançadas para delimitação da região de interesse, a métodos de aprimoramento de imagens mais complexos e ao uso de conjuntos de dados mais robustos. Uma das dificuldades encontradas nessa pesquisa para obter melhores resultados foi a ocorrência de *overfitting*, devido ao baixo número de imagens disponíveis no banco de dados MIAS. Diversas abordagens para mitigar o *overfitting* foram utilizadas sem trazer melhoria significativa na acurácia da detecção.

4 Conclusões e Trabalhos Futuros

Neste trabalho, empregamos três modelos de aprendizado de máquina na detecção de câncer de mama utilizando características extraídas da matriz GLCM e de primeira ordem. Para isso, realizamos procedimentos de recorte da região de interesse e de aprimoramento da imagem, usando os filtros de mediana e CLAHE. O banco de imagens médicas de mamografia utilizado foi o MIAS, e os modelos considerados foram o KNN, *random forest* e XGBoost.

Os resultados de detecção apresentados mostraram que os modelos obtiveram resultados de previsão semelhantes, com a acurácia variando entre 60% e 70%. Entre os modelos, o XGBoost foi o que obteve melhor resultado de acurácia alcançando um valor próximo de 70%. Além disso, foi visto que o XGBoost apresentou maior dificuldade nos casos de patologia positiva, com a quantidade de erros de detecção associada ao tipo falso negativo. Os resultados encontrados na literatura são superiores aos reportados neste trabalho, o que sugere que outros conjuntos de dados devem ser avaliados e novas etapas de pré-processamento devem ser incorporadas aos modelos.

O diagnóstico precoce do câncer de mama tem papel fundamental no aumento da taxa de sobrevivência a doença e em prognósticos com resultados favoráveis. Com isso, é importante que diagnósticos sejam precisos e confiáveis. Os métodos computacionais de auxílio ao diagnóstico podem exercer um papel fundamental nisso, assim podendo

fornecer informações importantes que ajudem os especialistas a terem maior confiança no diagnóstico ou até mesmo identificar tumores não vistos pelo especialista. Este trabalho tem como desdobramentos futuros explorar novas abordagens e métodos para aprimoramento de imagens de mamografia, e avaliar novos métodos de aprendizado de máquina que possam aprimorar os resultados de detecção em futuras pesquisas.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) e a Universidade Estadual de Santa Cruz pelo financiamento (parcial) oferecido para execução desta pesquisa. Adicionalmente, agradecemos ao Centro de Computação Avançada e Multidisciplinar (CCAM) da UESC pelo suporte computacional oferecido.

Referências

- [1] I. C. B. Ohl, R. I. B. Ohl, S. R. R. Chavaglia, e R. E. Goldman, “Public actions for control of breast cancer in Brazil: integrative review,” *Revista Brasileira de Enfermagem*, vol. 69, no. 4, pp. 793–803, 2016. Disponível em: <https://doi.org/10.1590/0034-7167.2016690424i>
- [2] P. A. da Silva e S. da S. Riul, “Breast cancer: risk factors and early detection,” *Revista Brasileira de Enfermagem*, vol. 64, no. 6, pp. 1016–1021, 2011. Disponível em: <https://doi.org/10.1590/S0034-71672011000600005>
- [3] B. R. N. Matheus, “BancoWeb: base de imagens mamográficas para auxílio em avaliações de esquemas CAD,” Dissertação de Mestrado, Mestrado em Processamento de Sinais e Instrumentação, Universidade de São Paulo, São Carlos, 2010.
- [4] M. K. Santos, J. R. Ferreira Júnior, D. T. Wada, A. P. M. Tenório, M. H. N. Barbosa, e P. M. A. Marques, “Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: Advances in imaging towards to precision medicine,” *Radiologia Brasileira*, vol. 52, no. 6, pp. 387–396, 2019. Disponível em: <https://doi.org/10.1590/0100-3984.2019.0049>
- [5] C. Irawan, E. N. Ardyastiti, D. R. I. M. Setiadi, E. H. Rachmawanto, e C. A. Sari, “A survey: Effect of the number of GLCM features on classification accuracy of lasem batik images using K-nearest neighbor,” em *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia: IEEE, 2018, pp. 33–38. Disponível em: <https://doi.org/10.1109/ISRITI.2018.8864443>
- [6] Ş. Öztürk e B. Akdemir, “Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM and SFTA,” *Procedia Computer Science*, vol. 132, pp. 40–46, 2018. Disponível em: <https://doi.org/10.1016/j.procs.2018.05.057>
- [7] A. K. Aggarwal, “Learning Texture Features from GLCM for Classification of Brain Tumor MRI Images using Random Forest Classifier,” *WSEAS Transactions on Signal Processing*, vol. 18, pp. 60–63, 2022. Disponível em: <https://doi.org/10.37394/232014.2022.18.8>
- [8] V. P. Singh, A. Srivastava, D. Kulshreshtha, A. Chaudhary, e R. Srivastava, “Mammogram Classification Using Selected GLCM Features and Random Forest Classifier,” *International Journal of Computer Science and Information Security*, vol. 14, no. 6, pp. 82–87, 2016. Disponível em: <https://sites.google.com/site/ijcsis/>
- [9] R. A. N. Diaz, N. N. T. Swandewi, e K. D. P. Novianti, “Malignancy Determination Breast Cancer Based on Mammogram Image with K-Nearest Neighbor,” em *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, Denpasar, Indonesia: IEEE, 2019, pp. 233–237. Disponível em: <https://doi.org/10.1109/ICORIS.2019.8874873>
- [10] T. T. Htay e S. S. Maung, “Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image,” em *18th International Symposium on Communications and Information Technologies*, Bangkok, Thailand: IEEE, pp. 171–175, 2018. Disponível em: <https://doi.org/10.1109/ISCIT.2018.8587920>

- [11] B. S. V, “Grey Level Co-Occurrence Matrices: Generalisation and Some New Features,” *International Journal of Computer Science, Engineering and Information Technology*, vol. 2, no. 2, pp. 151–157, 2012. Disponível em: <https://doi.org/10.5121/ijcseit.2012.2213>
- [12] P. K. Mall, P. K. Singh, e D. Yadav, “GLCM based feature extraction and medical X-RAY image classification using machine learning techniques,” em *2019 IEEE Conference on Information and Communication Technology*, Allahabad, India: IEEE, 2019, pp. 1–6. Disponível em: <https://doi.org/10.1109/CICT48419.2019.9066263>
- [13] Z. Abbas, M. U. Rehman, S. Najam, e S. M. Danish Rizvi, “An Efficient Gray-Level Co-Occurrence Matrix (GLCM) based Approach Towards Classification of Skin Lesion,” *2019 Amity International Conference on Artificial Intelligence*, Dubai, United Arab Emirates: IEEE, 2019, pp. 317–320. Disponível em: <https://doi.org/10.1109/AICAI.2019.8701374>
- [14] L. K. Kumari e B. N. Jagadesh, “A Robust Feature Extraction Technique for Breast Cancer Detection using Digital Mammograms based on Advanced GLCM Approach,” *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 8, no. 30, p. e3, 2022. Disponível em: <https://doi.org/10.4108/eai.11-1-2022.172813>
- [15] K. C. Tatikonda, C. M. Bhuma, e S. K. Samayamantula, “The Analysis of Digital Mammograms Using HOG and GLCM Features,” *2018 9th International Conference on Computing, Communication and Networking Technologies*, Bengaluru, India: IEEE, 2018, pp. 1–7. Disponível em: <https://doi.org/10.1109/ICCCNT.2018.8493809>
- [16] M. Pratiwi, Alexander, J. Harefa, e S. Nanda, “Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network,” *Procedia Computer Science*, vol. 59, p. 83–91, 2015. Disponível em: <https://doi.org/10.1016/j.procs.2015.07.340>