CLASSIFICAÇÃO AUTOMÁTICA DE IMAGENS CELULARES EM BASE DE DADOS DESBALANCEADA

Murilo Schreiner Ortiz (PIBIC/CNPq), Andréa Timóteo dos Santos, Ednéia Peres Machado, Daniela Etlinger Colonelli, José Carlos Ferreira da Rocha, Alaine Margarete Guimarães (Orientadora), e-mail: alainemg@uepg.br,

Universidade Estadual de Ponta Grossa/DEINFO

Engenharia de Produção, Pesquisa Operacional.

Palavras-Chave: *aprendizado de máquina, citopatologia, classificação automática, data augmentation, base de dados desbalanceada.*

Resumo

O câncer de colo do útero é uma das principais causas de morte feminina e o exame Papanicolau apresenta alta taxa de falsos negativos. Este estudo, realizado em parceria entre o Instituto Adolfo Lutz e a Universidade Estadual de Ponta Grossa, visa melhorar a detecção precoce de células cancerígenas através abordagens de de Aprendizado de Máquina. Visando enfrentar o desafio de se ter uma base de dados referente a análises celulares desbalanceada, esse trabalho teve o objetivo de utilizar técnicas de *data augmentation* para balanceamento de classes e os modelos ResNet50 e Reptile para classificação. Como resultado, acurácia de até 89,83% foi alcançada.

Introdução

O câncer de colo do útero é uma das principais causas de morte entre mulheres, e o exame preventivo Papanicolau é fundamental para sua detecção precoce. No entanto, devido à sua realização manual, o exame apresenta uma alta taxa de falsos negativos (Rezende et al., 2021). Evidentemente, os exames apresentam maior incidência de resultados negativos do que positivos para a malignidade, caracterizando-se em classes de resultados desbalanceadas.

A integração de métodos de Aprendizado de Máquina (AM) surge como uma solução promissora para melhorar a precisão na detecção de células cancerígenas (Medela et al., 2019). Este estudo explora o uso de técnicas de pré-processamento de dados para reduzir o desbalanceamento entre as classes e a abordagem de Redes Neurais Convolucionais (RNC) para desenvolver um modelo de classificação automático, visando reduzir a taxa de falsos negativos e aumentar a eficácia do diagnóstico.

Materiais e Métodos

Inicialmente foram estabelecidas 4 classes de resultados, sendo elas: Lesão HSIL, Lesão LSIL, Carcinoma e negativo para anormalidade. Devido ao desbalanceamento, foram aplicadas técnicas de *data augmentation* na base de dados desbalanceada (Ortiz et al.; 2023), como rotação. As imagens foram então pré-processadas e utilizadas para treinar e avaliar os modelos, com foco na precisão e robustez na detecção de células cancerígenas.

Foi desenvolvida uma abordagem para a classificação automática de imagens citopatológicas utilizando RNC, especificamente os modelos de Redes Siamesas com ResNet50 e Reptile.

Resultados e Discussão

Os modelos foram avaliados em termos de acurácia e capacidade de detecção de células cancerígenas. O modelo ResNet50 alcançou uma acurácia de até 89,83%, enquanto o Reptile obteve cerca de 85%. Embora os resultados sejam promissores, a análise também destacou a necessidade de uma base de dados maior e mais balanceada uma vez que mesmo com a adequação, classes como a lesão de baixo grau LSIL continuaram com poucas instâncias.

Conclusão

A aplicação de técnicas de AM para a classificação de imagens citopatológicas apresentou resultados indicando acurácia de até 89,83%. No entanto, a limitação imposta pela base de dados desbalanceada sugere a necessidade de ampliação e das amostras para as classes altamente desbalanceadas visando alcançar uma performance ainda melhor.

Agradecimentos

À UEPG, ao CNPq, ao LCAD e ao Instituto Adolfo Lutz.

Referências

Medela, A. et al.. Few Shot Learning in Histopathological Images: Reducing the Need of Labeled Data on Biological Datasets. 2019 **IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)**, p. 1860-1864, 2019.

Ortiz, M., et al. Abordagem One-Few Shot Learning para o Diagnóstico de Câncer de Colo do Útero. In: Encontro Anual de Iniciação Científica UEPG, 32., 2023, Ponta Grossa. **Resumos**. Ponta Grossa: UEPG, 2023.

Rezende, T. et al. Cric searchable image database as a public platform for conventional pap smear cytology data. **Scientific data**. 2021, 8:151.