XGBoost no auxílio de diagnóstico de câncer de bexiga

Filipe Ferreira de Oliveira
Graduação em Engenharia da Computação
Universidade Federal do Espírito Santo
Vitória, Brasil
filipe.oliveira.51@edu.ufes.br

***Abstract*—O câncer de bexiga é uma doença prevalente e recorrente no trato urinário, apresentando grandes desafios no diagnóstico e no monitoramento contínuo. A detecção precoce é essencial para melhorar os resultados dos pacientes, mas os métodos tradicionais frequentemente são invasivos e não são suficientemente sensíveis ou específicos. Este estudo investiga a aplicação do aprendizado de máquina, com ênfase no algoritmo XGBoost, para aprimorar a precisão diagnóstica do câncer de bexiga através da espectroscopia Raman de amostras de urina. Utiliza-se um conjunto de dados com 553 amostras com diferentes condições urológicas, que foram pré-processadas com Correção de Dispersão Multiplicativa (MSC) e Análise de Componentes Principais (PCA). É feita a classificação com e sem o uso de PCA para avaliar seu impacto no desempenho. Os resultados mostraram que o XGBoost alcançou uma precisão balanceada de 63,75% sem PCA e 70,50% com PCA, demonstrando uma melhoria significativa. Esses resultados destacam o potencial de integrar técnicas avançadas de aprendizado de máquina com análise espectroscópica para desenvolver ferramentas diagnósticas não invasivas e mais precisas para o câncer de bexiga. Futuras pesquisas buscarão expandir o conjunto de dados e incorporar novos métodos de pré-processamento e classificação para aumentar ainda mais a precisão e robustez das previsões.**

I. Introdução

O câncer é um problema de saúde pública mundial. Na última década, houve um aumento de 20% na incidência e espera-se que, para 2030, ocorram mais de 25 milhões de casos novos [1]. Assim o diagnóstico precoce tem um papel fundamental.

Podemos destacar o uso progressivo de aprendizado de máquina no auxílio do diagnóstico médico [2], por apresentar eficácia no reconhecimento de padrões presentes em diferentes tipos de exames clínicos e também por se mostrar como um novo instrumento para os profissionais da saúde como sendo uma nova fonte de informação a partir de dados já obtidos [3].

Um dos algoritmos de aprendizado de máquina mais comumente utilizados está o XGBoost, que se destaca pela sua alta performance e capacidade de lidar com grandes volumes de dados, tornando-se uma ferramenta valiosa na detecção precoce e precisa de doenças complexas como o câncer.

Neste estudo, é proposto a utilização do XGBoost na classificação de amostras de espectro Raman em casos de câncer de bexiga, presentes na base de dados apresentada por [5].

II. Motivação

Diante da relevância do câncer na área da saúde, é imperativo a intervenção científica de diversas áreas para combater este problema tão complexo [6]. Uma das contribuições significativas da inteligência artificial, por meio do aprendizado de máquina, é a utilização do poder computacional para tarefas de classificação e predição, proporcionando maior precisão e eficiência na análise de grandes volumes de dados [7].

Outra questão que requer atenção é a disponibilidade de dados clínicos, que representa um grande desafio para o avanço de pesquisa na área da saúde. A escassez de dados pode ser atribuída a vários fatores, incluindo questões de privacidade, variação na qualidade dos registros médicos e a fragmentação dos dados entre diferentes instituições de saúde [9]. A coleta e a integração de dados clínicos são processos complicados devido às rigorosas regulamentações de proteção de dados e às dificuldades técnicas em consolidar informações de múltiplas fontes. Além disso, a heterogeneidade dos dados, que inclui imagens médicas, informações genéticas e registros clínicos, dificulta a criação de conjuntos de dados robustos e representativos [10]. Diante dessas complexidades, ressalta-se a importância do aprofundamento em bases de dados que são disponibilizadas abertamente, como a que será discutida neste estudo.

O câncer de bexiga é uma das neoplasias mais comuns do trato urinário. Sua alta taxa de recorrência e à necessidade de monitoramento contínuo são aspectos que dificultam seu tratamento pelos sistemas de saúde. Este tipo de câncer é frequentemente diagnosticado em estágios avançados, o que complica o tratamento e reduz as chances de sobrevivência a longo prazo [8]. A precisão no diagnóstico inicial e o acompanhamento rigoroso são essenciais para melhorar os resultados clínicos dos pacientes.

Uma forma acessível e não invasiva de diagnóstico de certos tipo de câncer pode ser feita pela análise de material biológico através da espectroscopia [11]. É possível destacar a espectroscopia Raman que se mostra muito útil para diagnóstico de diversos tipos de câncer [12]. Diante disso foi conveniente adotar a base da dados disponibilizada publicamente pelo estudo em [5], no qual faz o uso da espectroscopia Raman para diagnóstico do câncer de bexiga através de amostras de urina coletadas.

No presente estudo, é proposto a utilização do algoritmo XGBoost devido sua grande aceitação no meio científico. A tarefa de classificação dos espectros será afim de detectar casos de câncer de bexiga.

III. Revisão de literatura

A espectroscopia Raman é uma técnica analítica que fornece informações detalhadas sobre a composição molecular das amostras. Ela fornece uma ”impressão digital” molecular, permitindo a identificação de mudanças bioquímicas associadas a diferentes estados de saúde. É particularmente útil na análise de biofluidos, como a urina, por ser não invasiva e altamente sensível. Espectros Raman podem ser utilizados para diferenciar entre amostras saudáveis e aquelas afetadas por câncer, incluindo câncer de bexiga [5].

É possível notar o potencial do aprendizado de máquina na área de classificação para diferentes doenças. Para câncer de bexiga, é possível observar a utilização de Máquina de vetores de suporte(SVM) juntamente com PCA para classificação de amostras cancerígenas, com valor de predição preditiva de 92.9% [13]. Outro trabalho também pode ser observado no qual é feita a aplicação SVM e algoritmos de floresta aleatória(*Random Forest*) em espectros Raman para a detecção de câncer de pulmão, alcançando sensitividade e especifidade de 100% [14].

XGBoost-*Extreme Gradient Boosting* é um dos algoritmos de aprendizado de máquina mais poderosos e eficientes para tarefas de classificação e regressão. Proposto em [4], combina o princípio do *boosting* com árvores de decisão, corrigindo iterativamente os erros das iterações anteriores. Este método é conhecido por sua robustez, eficiência computacional e capacidade de lidar com dados desbalanceados.

IV. Metodologia

Os testes conduzidos neste estudo são inspirados no trabalho de [5], tomando como base o primeiro estudo feito no mesmo, intitulado de *”Study 1”* como mostrado na Fig. 1. O artigo apresenta um algoritmo proprietário executado em MATLAB, porém neste presente estudo foram feitos testes em Python.

Study	Datasets	Classifications
Study 1	Urology Clinic Patients + Surine™	BCA-Positive, BCA-Negative, Surine™
Study 2	Urology Clinic Patients + Healthy Volunteers + Surine™	BCA-Positive, BCA-Negative, Surine™
Study 3	Urology Clinic Patients + Healthy Volunteers + Nephrology Clinic Patients + Surine™	BCA-Positive, BCA-Negative, Surine™
Study 4	Urology Clinic Patients + Healthy Volunteers + Nephrology Clinic Patients + Surine™	Urology Clinic Patients, Nephrology Clinic Patients, Healthy Volunteers, Surine™
Study 5	Urology Clinic Patients + 9 Healthy Volunteers + Surine™	GU Cancer, Other GU Disease, Healthy, Surine™

https://doi.org/10.1371/journal.pone.0237070.t002

Fig. 1. Características das populações estudadas e categorias de patologia do trato geniturinário estudadas. https://doi.org/10.1371/journal.pone.0237070.t001

Os dados utilizados são compostos de 553 amostras de pacientes de uma clínica urológica, com diversas classificações quanto ao tipo de doença. Para cada indivíduo, há em torno de 10 amostras. As subclasses presentes são:

- **Câncer de bexiga** (*Bladder cancer*) - 170 amostras
- Doenças geniturinárias (*GU Disease*) - 156 amostras
- Câncer de bexiga inativo (*BCA Inactive*) - 79 amostras
- Câncer geniturinário (*GU Cancer*) - 78
- Saudável (*Healthy*) - 70 amostras

Inicialmente, foi realizada a truncagem do espectro Raman para 250-1950 cm^{-1} que originalmente variava de 0-4000 cm^{-1}.

Para o pré-processamento dos dados, o artigo base faz o uso de Normalização pelo Desvio Padrão (SNV), *Baseline correction* e aplicação de PCA. É feito no presente estudo, correção de espectro com a utilização de MSC, normalização, *Baseline correction* e posteriormente, outro teste é realizado fazendo uso de PCA.

Dois testes foram conduzidos a fim de observar o impacto do uso de PCA na classificação. O primeiro teste consiste na classificação dos dados pré-processados somente, e no segundo teste é feita a aplicação de PCA após o pré-processamento.

A escolha dos hiper-parâmetros do classificador foram realizadas utilizando *GridSearch* por validação cruzada em 5 *folds*.

V. Resultados

A divisão de treino e teste foi realizada levando em consideração a existência de várias amostras de um mesmo paciente. Foi feita a separação de modo que as amostras do mesmo paciente não estejam presentes simultaneamente nas partições de treino e teste. A proporção utilizada foi de 70% dos dados para o treino de maneira estratificada em relação as subclasses.

O pipeline de pré-processamento utilizado é composto das seguintes técnicas, em ordem de aplicação: MSC(*Multiplicative Scatter Correction*), Escalonador padrão(*Standard scaler*) e Correção de linha de base(*baseline correction*). O efeito do processamento pode ser visto na Fig. 2 e Fig. 3.

O resultado da classificação dos dois testes é mostrado na Tabela I com as seguintes medidas: acurácia(ACC), acurácia balanceada(BAC), recall(REC), precision(PREC) e f1-score(F1). Para o primeiro teste, sem PCA, foi atingido uma acurácia balanceada de 63,75%. Utilizando-se o PCA os resultados apresentaram uma pequena melhora, mostrando uma acurácia balanceada de 70,50%.

A distribuição das classificações dos dois testes é mostrada nas matrizes de confusão ilustradas nas Figuras 4 e 5.

VI. Discussão

A disponibilidade de casos positivos é um fator crítico e frequentemente limitante nas tarefas de diagnóstico de doenças. O conhecimento das características presentes nesses casos é essencial para qualquer classificador, seja de fonte humana ou automatizada. Contudo, a obtenção de um número substancial de amostras positivas é desafiadora, especialmente para doenças raras ou em estágios iniciais. No presente estudo, a distribuição de casos positivos de câncer de bexiga no conjunto de dados foi de 30%. Essa proporção de amostras positivas

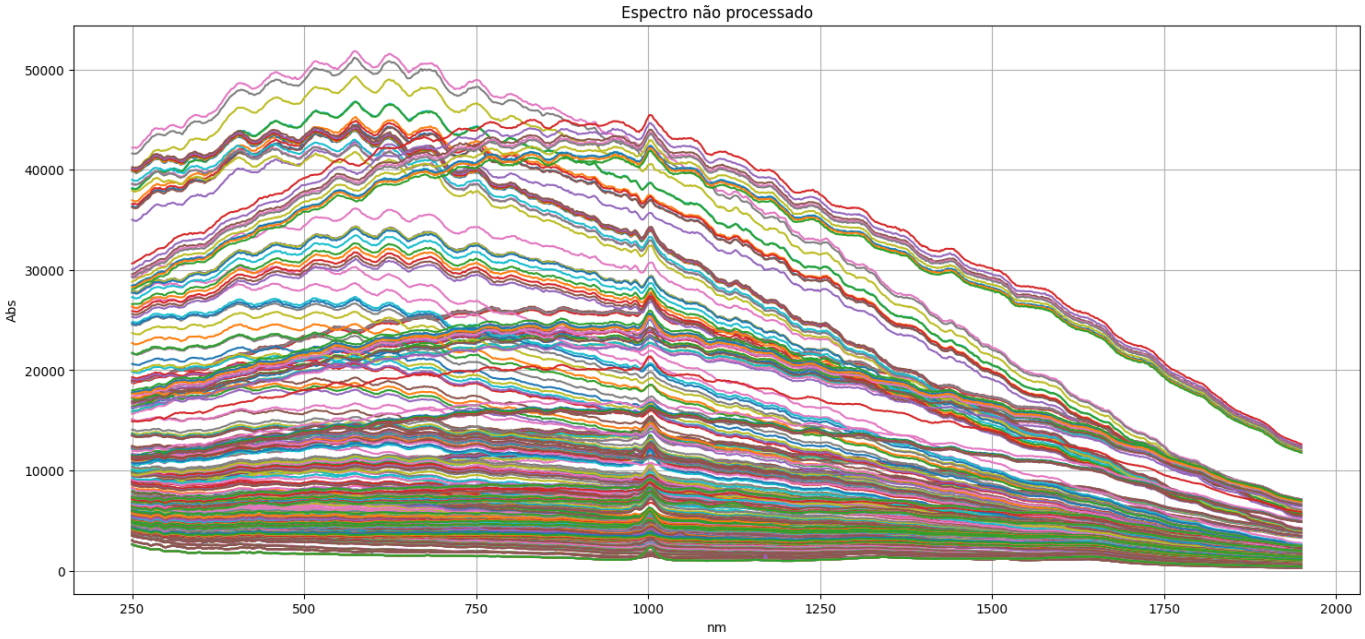

Fig. 2. Espectro antes do processamento

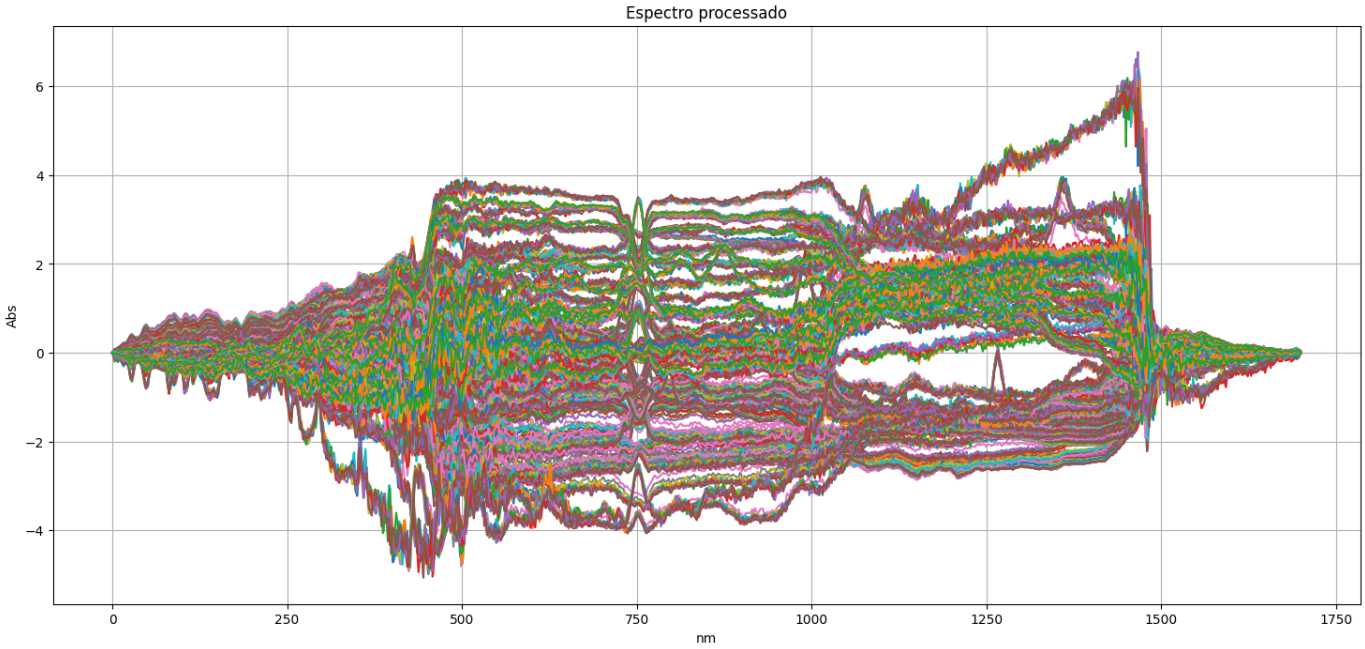

Fig. 3. Espectro depois do processamento

TABLE I
TABELA DE RESULTADOS

Teste	**ACC**	**BAC**	**REC**	**PREC**	**F1**
XGBoost + Pré-processamento	63,30%	63,75%	63,30%	68,24%	64,50%
XGBoost + Pré-processamento + PCA	71,28%	70,50%	71,28%	73,75%	72,01%

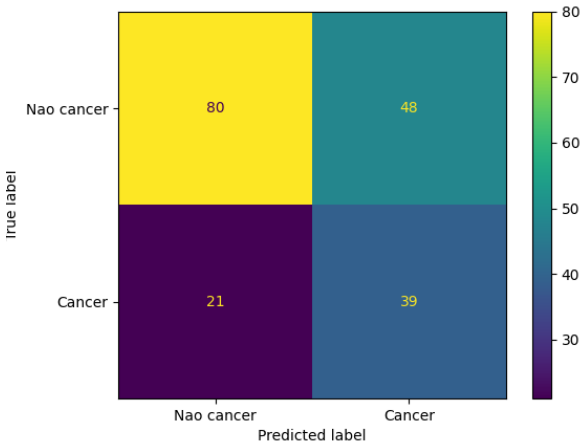

Fig. 4. Matriz de confusão para o teste sem PCA

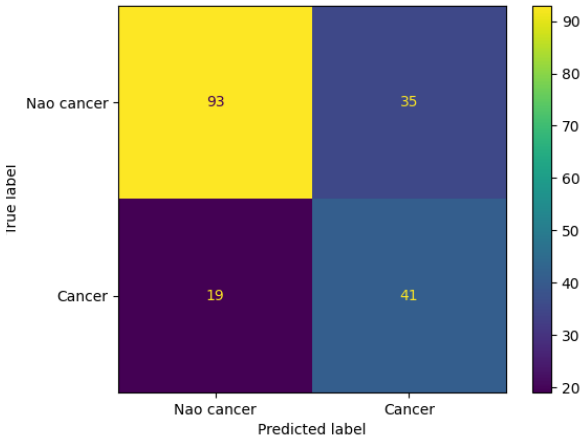

Fig. 5. Matriz de confusão para o teste com PCA

pode influenciar significativamente os resultados obtidos nos testes.

Existem inúmeras formas de pre processamento e combinações possíveis quando se trata de algoritmos de aprendizado de máquina. A combinação dos mesmos fornecem diferentes entradas para o modelo a ser estudado, proporcionando minimização de ruído e uma melhoria na comparabilidade dos dados. Diferentes combinações e aplicações devem ser estudadas a fim de se encontrar a que seja mais adequada para o problema específico.

É possível notar que o presente estudo reafirma a necessidade de grandes conjuntos de dados rotulados para o treinamento eficaz de um modelo de classificação. É visto que a quantidade de dados heterogêneos é uma limitação significativa se tratando de desempenho mediante a acurácia de predição de um algoritmo. Além disso, a generalização dos resultados para populações mais diversas requer validação adicional com dados de diferentes origens e condições clínicas.

Futuras pesquisas tem o objetivo de expansão do conjunto de dados e na validação clínica em larga escala para confirmar a aplicabilidade do modelo em contextos reais. Além disso, a integração de outras técnicas de pré-processamento e diferentes modelos de classificação pode oferecer melhorias adicionais na precisão e robustez na tarefa preditiva.

REFERENCES

[1] Santos, M. d. O., Lima, F. C. d. S. d., Martins, L. F. L., Oliveira, J. F. P., Almeida, L. M. d., and Cancela, M. d. C. (2023). Estimativa de Incidência de Câncer no Brasil, 2023-2025. Revista Brasileira de Cancerologia, 69(1):e–213700.

[2] Ramirez, C. A. M., Greenop, M., Almoshawah, Y. A., Hirsch, P. L. M., and Rehman, I. U. (2023). Advancing cervical cancer diagnosis and screening with spectroscopy and machine learning. Expert Review of Molecular Diagnostics, 23(5):375–390.

[3] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23(1):89–109.

[4] Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785.

[5] Huttanus, H. M., Vu, T., Guruli, G., Tracey, A., Carswell, W., Said, N., Du, P., Parkinson, B. G., Orlando, G., Robertson, J. L., and Senger, R. S. (2020). Raman chemometric urinalysis (Rametrix) as a screen for bladder cancer. PLOS ONE, 15(8):1–21.

[6] Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1), 7-30. https://doi.org/10.3322/caac.21590.

[7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056.

[8] Antoni, S., Ferlay, J., Soerjomataram, I., Znaor, A., Jemal, A., and Bray, F. (2017). Bladder cancer incidence and mortality: A global overview and recent trends. European Urology, 71(1), 96-108. https://doi.org/10.1016/j.eururo.2016.06.010.

[9] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7.

[10] Shilo, S., Rossman, H., and Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. Nature Medicine, 26(1), 29-38. https://doi.org/10.1038/s41591-019-0727-5.

[11] Zhu, J., Zhang, S., Wang, R., Fang, R., Lei, L., Zheng, J., and Chen, Z. (2023). Urine based near-infrared spec-troscopy analysis reveals a noninvasive and convenient diagnosis method for cancers: a pilot study. PeerJ, 11:e15895.

[12] Ralbovsky, N. M. and Lednev, I. K. (2019). Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 219:463–487.

[13] Wang L, Fan JH, Guan ZF, Liu Y, Zeng J, He DL, Huang LQ, Wang XY, Gong HL. [Study on bladder cancer tissues with Raman spectroscopy]. Guang Pu Xue Yu Guang Pu Fen Xi. 2012 Jan;32(1):123-6. Chinese. PMID: 22497142.

[14] Qian, Kun and Wang, Yan and Hua, Lin and Chen, Anyu and Zhang, Yi. (2018). New method of lung cancer detection by saliva test using surface-enhanced Raman spectroscopy: New detection method for lung cancer. Thoracic Cancer. 9. 10.1111/1759-7714.12837.