Obtenção de novos potenciais biomarcadores para o câncer de mama com o auxílio da inteligência artificial explicável

Luísa C. de Souza*,Karolayne S. Azevedo*, Matheus G. S. Dalmolin*† e Marcelo A. C. Fernandes*†‡
*InovAI Lab, nPITI/IMD, UFRN, Natal, RN, Brasil.
*Centro Multiusuário de Bioinformática (BioME) - IMD, UFRN, Natal, RN, Basil.
‡Departamento de Engenharia da Computação e Automação (DCA), UFRN, Natal, RN, Basil.
Email: luisa.souza.103@ufrn.edu.br,karolayneazevsantos@gmail.com, matheusdalmolinrs@gmail.com,mfernandes@dca.ufrn.br

***Resumo*—O câncer de mama é a neoplasia maligna mais comum em ambos os sexos, representando um quarto dos diagnósticos de câncer em mulheres. Para melhorar a identificação e o desenvolvimento de terapias mais eficazes, é fundamental encontrar potenciais biomarcadores de prognóstico. Neste estudo, uma abordagem foi desenvolvida utilizando técnicas de aprendizado de máquina e inteligência artificial explicável para identificar genes relevantes na distinção entre câncer de mama e tecido normal. Os resultados deste estudo revelaram um total de 74 genes com potencial importância na classificação. Algumas características desses genes mostraram evidências de influenciar a expressão do câncer de mama, tornando-os possíveis biomarcadores prognósticos. Essa abordagem inovadora, que combina aprendizado de máquina e inteligência artificial explicável, pode fornecer insights valiosos sobre os mecanismos moleculares subjacentes ao câncer de mama. A identificação de genes com relevância na classificação pode contribuir para o desenvolvimento de terapias mais direcionadas e personalizadas, melhorando assim o prognóstico e o tratamento dos pacientes com câncer de mama.**

***Index Terms*—Câncer de mama, Biomarcadores, Aprendizagem de máquina, Inteligência Artificial Explicável, SHAP Values, Interpretabilidade.**

I. Introdução

Câncer de mama (*Breast Cancer Gene* - BRCA) é atualmente o carcinoma com maior número de diagnósticos mundial, dados da GLOBOCAN 2020 mostram que BRCA constituiu 2.261.419 novos casos em 2020 em ambos os sexos, número mais alarmante quando considerado apenas o sexo feminino, onde representa $24,5\%$ dos diagnósticos, sendo o quinto mais fatal, correspondendo a $6,9\%$ das mortes causadas por cânceres [1], [2].

O carcinoma da mama é uma patologia considerada complexa devido a sua diversidade de alterações moleculares, composições celulares e de comportamentos clínicos, que podem ser relacionados à mudanças nas expressões genéticas associadas [3], [4]. Isto porque o desenvolvimento e crescimento de neoplasias malignas no tecido da mama é influenciado por fatores genéticos que incluem disparidades de expressão e mutações em genes de predisposição ao câncer de alto e moderado risco [5]. Entre os fatores não genéticos estão os não modificáveis como histórico familiar, idade, gênero e raça, e os fatores modificáveis, como exposição à químicos, alimentação, uso de drogas, álcool ou nicotina [6]. O número de fatores de risco significante, e suas diferentes influências no desenvolvimento dos tumores, contribuem para a heterogeneidade da patologia.

A heterogeneidade tumoral dificulta o prognóstico e a tomada de decisão do tratamento específico eficaz. Dessa forma, novos estudos vêm buscando analisar expressões gênicas para classificação molecular de tecidos da mama [3]. Posto que há evidências de que a análise dos perfis de expressão gênica e mutações pode ser utilizada para identificar diferentes subtipos de câncer de mama, esses dados podem desempenhar um papel crucial no monitoramento da progressão da doença e na otimização do tratamento, sendo adotados como biomarcadores prognósticos [7].

Técnicas de inteligência artificial (IA), mais precisamente de aprendizagem de máquina, como classificadores baseados em árvores de decisão e *ensemble learning*, vêm sendo amplamente aplicadas na predição e descrição de cânceres de mama, cólon, ovário, pulmão, leucemia entre outros [8]. Entre estas técnicas, os de *Gradient boosting decision tree* (GBDT) criam modelos robustos ao sequenciar classificadores mais simples iterativamente, de forma que cada uma de suas árvores de decisão otimiza os pesos da função de custo com base nos erros do classificador anterior [9]. O trabalho apresentado em [4] utilizou os algoritmos de gradient boosting machine (GBM), extreme gradient boosting (XGBoost) e light gradient boosting (LightGBM) para classificar dados do *Breast Cancer Wisconsin Dataset* [https://archive.ics.uci.edu/datasets] que conta com 10 características para o BRCA. Os resultados obtidos mostraram que o LightGBM exibiu melhor performance com acurácia de $95,3\%$. Na tentativa de prôpor biomarcadores para o câncer de mama triplo negativo (*Triple-negative breast cancer* -TNBC), o trabalho [10] usou cinco técnicas de aprendizagem de máquina empregados em um dataset contendo as principais expressões gênicas que diferenciam o TNBC de outros subtipos de BRCA obtidas pelo método *Recursive Feature Eliminarion*, constatou-se que o XGBoost apresentou

o melhor desempenho para um subconjunto de 25 e 20 genes. Dos 45 genes desses dois conjuntos de dados, 34 genes foram encontrados como regulados de forma diferencial, dos quais dois são novos e podem ser potenciais genes de prognóstico.

No entanto, algoritmos de aprendizagem de máquina realizam classificação ou estratificações baseando-se em descorbertas de associações e correlações de padrões nos dados de treinamento, sem oferecer explicações ou a racionalização do processo além do sentido estatístico oferecido pelos preditores [11]. Dessa forma, com a crescente aplicação das técnicas de inteligência artificial para tomadas de decisão com impacto na vida da sociedade, ampliou de mesmo modo a demanda por a transparência, a interpretabilidade e a responsabilidade dos resultados obtidos pelos modelos [12]. Dessa demanda surgiu o conceito de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* - XAI), um conjunto de métodos capazes de oferecer explicação sobre as características com mais impacto para a saída dos modelos de predição, gerando a capacidade de interpretabilidade para humanos às decisões feitas pela máquina [13].

O emprego de técnicas de XAI é adequado principalmente em aplicações da área de medicina e saúde, onde informações sobre os padrões aprendidos pela aprendizagem de máquina podem ser mais importantes que as acurácias dos modelos [14]. Na pesquisa proposta em [15], foi apresentada uma abordagem de XAI para traçar perfis de pacientes oncológicos com o objetivo de designar terapias personalizadas. Para isso, os autores basearam-se em uma redução dimensional adaptativa com o desejo de destacar as características clínicas mais importantes para agrupar pacientes com neoplasia maligna de mama em *clusters* com características semelhantes. Já em [16], os autores tinham como objetivo a investigação de eventos de doença invasiva do câncer de mama, como recorrência, cânceres contralaterais e secundários, em duas janelas de tempo, 5 e 10 anos após o primeiro diagnóstico. Utilizando quatro algoritmos de aprendizagem de máquina e uma técnica de XAI, observaram que para a janela de 5 anos as características com mais impacto foram idade, diâmetro do tumor, tipo de cirurgia e multiplicidade enquanto características relacionadas à terapia, incluindo hormônios, esquemas de quimioterapia e invasão linfovascular, dominam a predição da janela de 10 anos.

Perante o exposto, o presente trabalho realizou a aplicação de técnicas de aprendizagem de máquina que realizaram a classificação de dados de expressões gênicas de tecidos saudáveis e de tecidos acometidos pelo carcinoma de mama, e que foram subsequentemente associados ao método *Shapley Additive Explanations*(SHAP), uma ferramenta de inteligência artificial explicável, obtendo uma lista com genes de maior importância na classificação que apresentam potencial para se tornarem novos biomarcadores para o prognóstico do câncer de mama.

II. Materiais e Métodos

A Figura 1 exibe o fluxograma de atividades desenvolvidas no presente trabalho para a obtenção e análise de expressões gênicas que influenciam na classificação do câncer de mama com algoritmos de aprendizagem de máquina associadas à métodos de obtenção de *feature importance*, tal qual a técnica SHAP [17]. Cada atividade foi detalhada nas próximas seções.

A. Dataset

Para este estudo, foram utilizados dados de expressão gênica tecidos com a presença do câncer de mama e tecido normal. Esses dados de expressão gênica se referem à um processo complexo de regulação no qual informações contidas em genes são utilizadas na produção de moléculas funcionais, geralmente associadas a síntese de proteínas [18]. A regulação destes elementos exercem um papel fundamental no desenvolvimento e funcionalidade dos organismos humanos, de forma que, qualquer distúrbio nesse processo pode levar ao aparecimento de afecções, como câncer [19].

Para obtenção dos dados, seguimos segundo proposto em [20], que oferece protocolos com base no ambiente de software RStudio [21], para análises de genes diferencialmente expressos em diversos tecidos com cânceres humanos e saudáveis, oferecendo os parâmetros para aquisição dos dados clínicos, e de RNA-Seq, de onde as informações de expressão gênica são obtidos. As amostras de BRCA que pertencem ao The Cancer Genome Atlas Program (TCGA) [22] somaram um total de 1092 casos. Os dados do tecido normal foram obtidos do Genotype-Tissue Expression (GTEx) [23], que disponibiliza dados de expressão genética específicos para diversos tecidos saudáveis, contando com 178 amostras para tecido mamário. Os dados foram unidos realinhando uniformemente as leituras com o genoma hg38 como referência. O conjunto resultante contém 1270 casos entre BRCA e tecido normal (*normal tissue*), com 9412 expressões gênicas, sendo adicionada à última coluna da tabela os rótulos das duas classes.

B. Algoritmos de Classificação

Baseados na técnica GBDT, os algoritmos XGBoost e LightGBM, detalhados a seguir, foram selecionados para o presente estudo pois têm sido amplamente utilizados na análise de expressões gênicas devido sua fácil implementação, flexibilidade na construção da arquitetura e suas performances em grandes datasets com rápida velocidade de treinamento, mantendo altas acurácias [24]. São capazes de auxiliar na identificação de genes diferencialmente expressos, análise de importância de características e modelagem de padrões complexos entre expressões gênicas e fenótipos. Suas aplicações tem contribuído significativamente para a compreensão dos mecanismos moleculares envolvidos em diversas doenças e obtenção de biomarcadores [25].

1) Extreme Gradient Boosting (XGBoost): Proposto originalmente em [26] e posteriormente implementado por [27], se trata de uma técnica escalável, veloz e precisa para classificação e regressão de dados baseado em *ensemble* de árvores de decisão e boosting [28]. XGBoost utiliza o somatório dos valores previstos em cada árvore para obter o

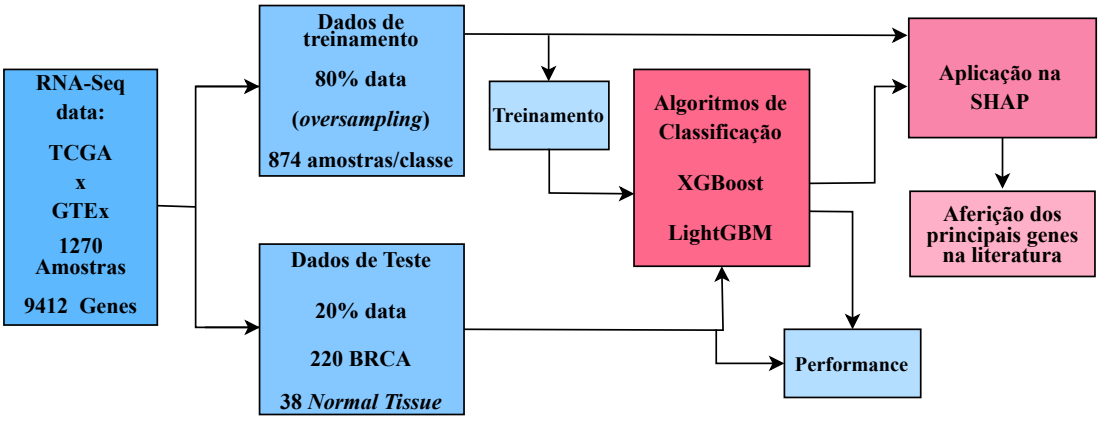


Figura 1: Fluxograma de atividades para obtenção de expressões gênicas influentes na classificação do câncer de mama.

valor de saída ou previsão $\hat{y}_i$ de cada amostra x_i, sendo $\hat{y}_i$ expresso como

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

onde K é o número total de árvores, F é o modelo dos classificadores, e f_k a k-ésima árvore de decisão [27]. Sendo l a função de custo, que representa o erro entre o valor real y_i e o valor predito $\hat{y}_i$, e $\Omega(f_k)$ a função de regularização que previne *overfitting* e define a complexidade do modelo, que apresenta-se como

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2 \tag{2}$$

onde T é o número de folhas da k-ésima árvore com γ sendo a penalização de regularização de cada folha, ou limiar de divisão dos nós de cada folha quando o valor da redução da função de perda for superior a γ. E w sendo o peso da saída das folhas de cada k-ésima árvore com λ sendo o termo de penalização de regularização de cada peso. Uma função objetivo é definida inicialmente para descrever o modelo, consistindo em uma parte com a função de erro de treinamento e a parte com o termo de regularização [29]. A função objetivo é dada por

$$L = \sum_{i=1}^{N} l(y_i, \hat{y}_i)) + \sum_{k=1}^{K} \Omega(f_k). \tag{3}$$

O algoritmo constrói uma árvore de decisão por iteração, com a otimização da função objetivo utilizando a expansão de Taylor de segunda ordem e adicionando um algoritmo de descida de gradiente para minimizar a perda. Os parâmetros de regularização tornam o XGBoost um algoritmo mais generalizável para aplicações práticas. [24], [25], [30], [31].

2) Light Gradient Boosting Machine (LightGBM): Desenvolvido pela Microsoft em 2017, o classificador LightGBM [9] utiliza o algoritmo Gradient-based One-Side Sampling (GOSS) que ordena as amostras de acordo com seus gradientes e utiliza apenas aqueles que têm maior influência na informação de ganho [24]. Além disso, emprega algoritmos baseados em histograma, que discretizam os autovalores contínuos em k valores inteiros, construindo um histograma com esses k valores. Esse histograma é analisado após o algoritmo percorrer todos os dados para realizar uma escolha eficiênte de pontos de divisão, resultando em uma diminuição no tempo de treinamento e na memória requerida [25], [29], [31], [32]. Um dos pontos no qual esse classificador difere dos demais algoritmos baseados em GBDT é a forma da construção das árvores de decisão. O LightGBM usa a estratégia *leaf-wise* com limitação de profundidade para dividir as árvores focando nas folhas que reduzem a função de erro, ao invés das técnicas tradicionais que fazem uso de algoritmos *level-wise*, o que previne *overlearning* dado que o crescimento do modelo é horizontal e as árvores de decisão são menos profundas [4].

C. Shapley Additive Explanations (SHAP)

Com o objetivo de promover interpretabilidade dos resultados dos algoritmos de aprendizagem de máquina, os trabalhos [33], [34] propuseram a Shapley additive explanation (SHAP) como uma medida unificada para representar as *features importance*, baseando-se na técnica da teoria dos jogos cooperativos conhecida como *Shapey values* [35]. O princípio básico de funcionamento do algoritmo é baseado no uso dos valores de SHAP para o cálculo da contribuição de cada característica, que neste trabalho são os genes, para a saída do algoritmo de classificação ou regressão [14], [36]. Matematicamente, o valor de SHAP ϕ de uma característica i para uma predição p é dado por

$$\phi_i(p) = \sum_{S \subseteq F \setminus i} \frac{|S|!(F-|S|-1)!}{F!}(p(S \cup i) - p(S)) \tag{4}$$

onde F é o conjunto das características, e a expressão $S \subseteq F$ demonstra que o modelo necessita ser treinado com todo o subconjunto de F [37]. Sendo ainda $\frac{|S|!(F-|S|-1)!}{F!}$ a parte da equação que indica os pesos ponderados das contribuições marginais e $(p(S \cup i) - p(S))$ exprime a diferença entre a

predição do modelo com a presença da característica i e a predição sem a presença dela [38]. Resultando na explicação da saída do algoritmo de aprendizagem de máquina para um gene específico.

Para a aplicação da SHAP em algoritmos baseados em árvores de decisão, como o caso do XGBoost e LightGBM, é indicada a utilização da *TreeExplainer*, pois a otimização do explicador faz com que ele combine os efeitos das contribuições locais das características para cada previsão, compreendendo a partir disto a estrutura geral do classificador, ocasionando em cálculos mais rápidos dos valores de SHAP [14], [38].

III. Resultados e Discussões

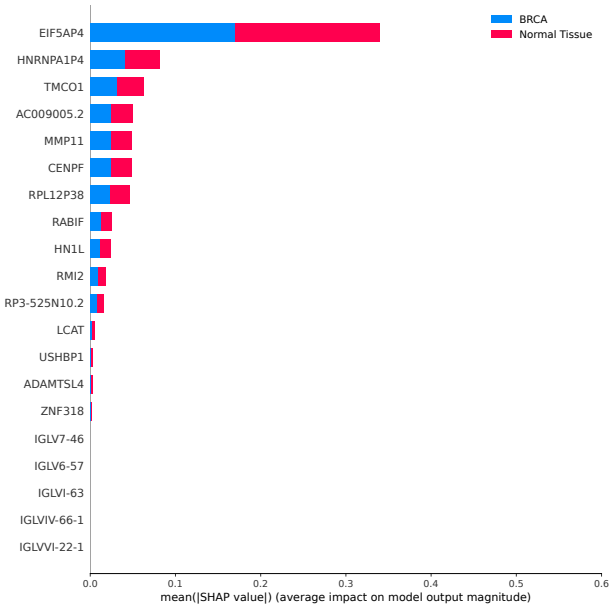
Para o treinamento dos algoritmos de aprendizagem de máquina, os dados foram separados em 80% para treinamento e 20% para teste, porém, dado que o dataset não apresentava quantidades equílibradas de amostras, foi necessário a realização de balanceamento dos dados de treinamento, através de *oversampling*, ou seja, repetição dos casos com menor quantidade. O total de amostras para treinamento foi de 874 para cada classe, e os dados de teste somaram 258, sendo 220 da classe BRCA e 38 de *Normal Tissue*. Os classificadores foram treinados com validação cruzada com k-fold=10 e os resultados de acurácia (ACC), sensibilidade (SEN), especificidade (ESP) e f1-*score* obtidos foram como exibidos na Tabela I.

Tabela I: Comparação de performance da classificação com LGBM e XGBoost.

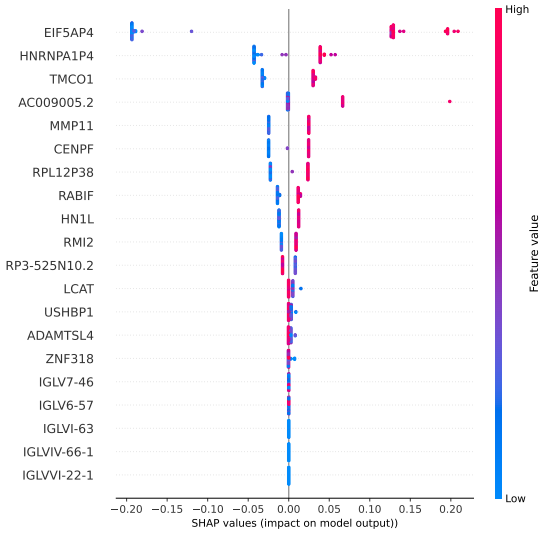
Classificador	Métricas de Performance			
	ACC	SEN	PRE	F1-Score
XGBoost	99,885%	99,886%	99,886%	99,885%
LightGBM	99,771%	99,770%	99,775%	99,771%

Após o treinamento dos classificadores, a técnica SHAP, implementada em Pyhton foi aplicada aos dados de treinamento com o *TreeExplainer* [33] para o cálculo dos valores de Shapey. Associado ao classificador XGBoost, o SHAP obteve um total de 16 genes como os que apresentam maior *feature importance* para a saída do modelo, enquanto que aplicado ao algoritmo LightGBM, selecionou um total de 58 genes. O resultado da SHAP é sumarizado nos plots de *feature importance* e *summary plot* expostos nas Figuras 2 e 3.

As Figuras 2a e 3a exibem os genes de maior importância em ordem decrescente para cada classificador, calculados pela média dos valores absolutos de Shapley por característica nos dados, e é possível observar que os genes HNRNPA1P4 e EIF5AP4 apareceram nas primeiras colocações para o XGBoost e LightGBM, o que pode indicar a necessidade de análise destes genes especificamente associados a tumores de mama. Porém como este tipo de plot não oferece informações além da importância de cada gene, foi plotado também o *summary plot* que oferece informações sobre a importância e o impacto de cada característica na predição [38]. A partir deles vemos que valores baixos de expressão dos genes HNRNPA1P4 e



(a)

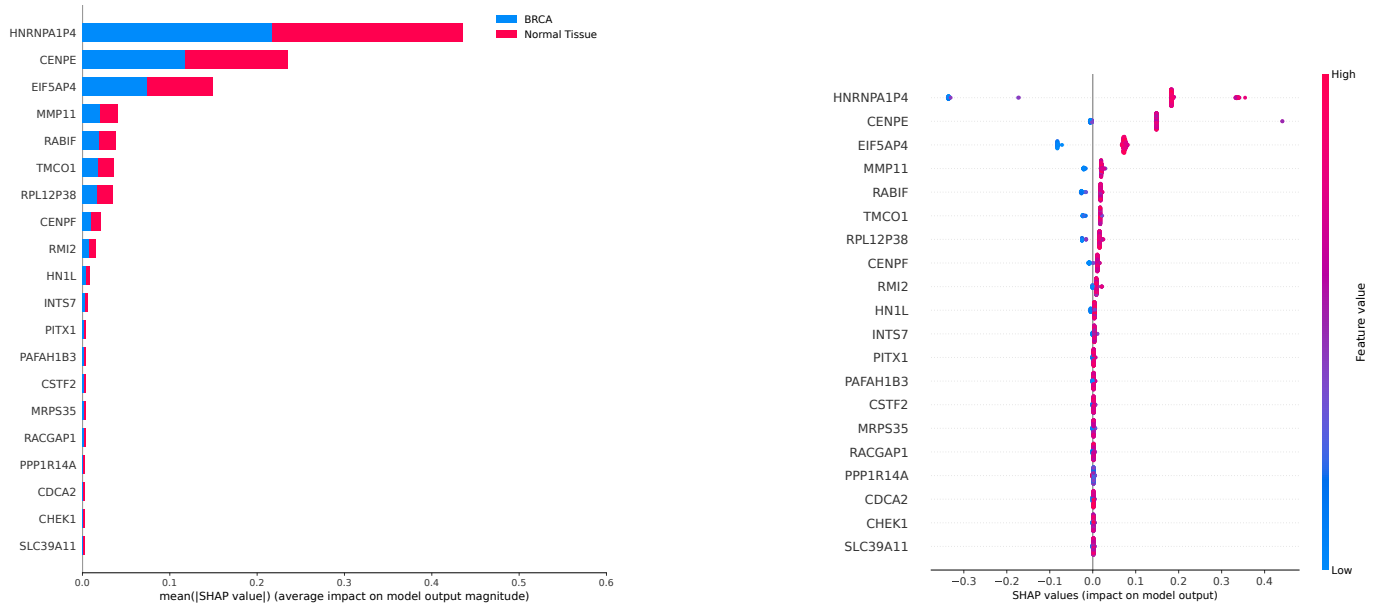


(b)

Figura 2: Plots do classificador XGBoost. (a)SHAP *Feature Importance* mostra em forma hierárquica as contribuições de cada um dos 16 genes obtidos na classificação do XGBoost; (b) SHAP *Summary Plot* permite observar que, com excessão do gene RP3-525N10.2, altos valores da expressão gênica aumentam o risco de câncer de mama.

EIF5AP4 estão associados à baixa probabilidade de câncer de mama, enquanto o contrário também é verdadeiro, e vemos que altos valores estão relacionados a presença do câncer de mama. A Tabela II foi esquematizada com o objetivo de detalhar os genes observados pela técnica SHAP, com as principais importâncias para ambos os algoritmos de classificação que podem ser potenciais novos biomarcadores para o prognóstico de câncer de mama.

Os pseudogenes são gerados quando mutações que ocor-



(a) (b)

Figura 3: Plots do classificador LightGBM.(a)SHAP *Feature Importance* exibe os 20 genes mais importantes dos 58 obtidos pelo SHAP na classificação do LightGBM; (b) SHAP *Summary Plot* mostra que os pontos com altos valores de expressão gênica (cor rosa), estão mais presentes no eixo positivo dos SHAP *Values* o que significa classificá-los como câncer de mama.

Tabela II: Principais expressões gênicas selecionadas pela SHAP para ambos os algoritmos de classificação.

Gene	**Descrição**	**Principal Função**	**Patologia Associada**	**Probabilidade SHAP (XGBoost)**	**Probabilidade SHAP (LightGBM)**	**Ref**
HNRNPA1P4	Pseudogene.	Regulação do splicing alternativo.	Esclerose Lateral Amiotrófica.	$\sim 9\%$	$\sim 42\%$	[39], [40].
EIF5AP4	Pseudogene.	Fator de iniciação de tradução.	Potencial de associação com pneumonia severa.	$\sim 34\%$	$\sim 16\%$	[41].
MMP11	Codificador de Proteína.	Regulador na remodelação da matriz.	Adenocarcinoma, Carcinoma Basocelular, Dermatofibroma.	$\sim 6\%$	$\sim 5\%$	[42], [43].
TMCO1	Codificador de Proteína.	Regulação de íons de cálcio no retículo endoplasmático.	Dismorfismo Craniofacial, Anomalias esqueléticas e Síndrome do Desenvolvimento Intelectual Prejudicado 1.	$\sim 8\%$	$\sim 4.5\%$	[44].
RABIF	Codificador de Proteína.	Fator de liberação de Guanina.	-	$\sim 3\%$	$\sim 5\%$	[45].
RPL12P38	Pseudogene.	-	-	$\sim 6\%$	$\sim 4\%$	[46].
CENPF	Codificador de Proteína.	Associado a função do cinetócoro e segregação cromossômica na mitose.	Síndrome de Stromme, Associada à vários. tipos de tumor.	$\sim 5\%$	$\sim 2.5\%$	[47], [48].
RMI2	Codificador de Proteínas.	Associado ao processamento de intermediários de recombinação homóloga.	Síndrome de Bloom, Adenocarcinoma pulmonar, Microcefalia e Restrição do Crescimento.	$\sim 2\%$	$\sim 2\%$	[49], [50].
HN1L	Codificador de Proteínas.	Necessária para liberação de cálcio intracelular evocado por NAADP.	Associado a proliferação celular maligna.	$\sim 3\%$	$\sim 2\%$	[51].

reram durante a evolução em um gene codificante resultam na perda da sua habilidade de codificar aminoácidos, sendo considerados como unidades não funcionais. Porém novos estudos apontaram evidências que os pseudogenes podem exercer papéis importantes em diversos processos fisiológicos e patológicos, tal como o câncer [52]. Reconhecendo esta tendência, o estudo [53] destacou o potencial que a análise das expressões de pseudogenes como biomarcadores para um prognóstico de tumor, podem auxiliar na compreensão dos mecanismos do câncer. o trabalho [54] mostra que o gene HNRNPA1P4 é expresso de forma diferente em tecidos com carcinoma hepatocelular de como é expresso em tecido

normal.

Em [46], os autores buscaram analisar o impacto dos macrófagos associados a tumores (TAM) no câncer de mama, com o intuito de identificar os genes regulados por TAMs e desenvolveu uma assinatura genética de prognóstico com base neles, entre os 45 genes encontrados no trabalho, o RPL12P38 era um deles. Porém, para mais análises, não foi encontrada documentação descritiva do gene.

Centromere protein F (CENPF) é uma proteína associada com a progessão de vários tipos de tumores, incluindo carcinoma hepatocelular, adenocarcinoma pulmonar e câncer de mama [47]. Em [48], os autores investigaram o papel da CENPF na resistência do câncer de mama triplo negativo à quimioterapia, pois observaram que se tratava de um gene altamente expresso nesse tipo de câncer. Encontraram evidências de que CENPF pode ser associado a um prognóstico negativo em pacientes recebendo quimioterapia.

Em [55] foi discutido como o gene HN1L é superexpresso em alguns tecidos cancerosos, incluindo o de mama. Além disso, foi observado que o silenciamento do gene HN1L inibiu significativamente a invasão e metástase de células cancerígenas de mama in vitro.

Além disso, na Figura 3 é possível ver que integrado com o LightGBM, o SHAP também selecionou o CHEK1, conhecido pela sua influência no câncer de mama triplo negativo, é considerado como um biomarcador de prognóstico, dado que inibidores químicos do CHEK1 podem obter uma diminuição da sobrevida de células do TNBC [56].A presença de genes conhecidos pela sua influência carcinoma de mama na análise realizada mostram a efetividade do SHAP em obter genes relevantes com potencial para serem biomarcadores.

IV. Conclusão

Neste trabalho foi proposto uma metodologia para seleção de novos possíveis biomarcadores para o prognóstico de câncer de mama, com o emprego de algoritmos de aprendizagem de máquina associadas a SHAP, uma técnica de inteligência artificial explicável que se propõe a interpretar e explicar os resultados de saída dos classificadores. Por se tratarem de dois algoritmos com princípios de funcionamento semelhantes, dado que ambos são baseados na técnica GBDT, as características captadas pela técnica SHAP foram similares para o XGBoost e LightGBM. As análises subsequentes dos genes obtidos mostraram que alguns deles já eram conhecidos pela sua influência e ou superexpressão em tumores no tecido mamário, contribuindo para suas relevâncias como novos biomarcadores do câncer de mama. Porém, é importante mencionar que os genes e suas contribuições, negativas e positivas, são diretamente relacionadas ao modelo de classificação, podendo ser, como visto na literatura, ou não associadas a características biológicas. O que convida novas análises funcionais com tais genes para observar seus papéis no prognóstico do câncer de mama. Dessa forma, um dos possíveis trabalhos futuros a ser desenvolvido é a observação do comportamento de novos algoritmos de classificação com apenas os 74 genes escolhidos pela SHAP para o XGBoost e LightGBM.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte e financiamento.

Referências

[1] H. Sung, J. Ferlay, R. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. ca cancer clin 2021; 71: 209-49," *This report provides the latest global cancer statistics of incidence and mortality worldwide*, 2022.

[2] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, 2022.

[3] F. K. Al-Thoubaity, "Molecular classification of breast cancer: A retrospective cohort study," *Annals of medicine and surgery*, vol. 49, pp. 44–48, 2020.

[4] S. Akbulut, I. B. Cicek, and C. Colak, "Classification of breast cancer on the strength of potential risk factors with boosting models: A public health informatics application." *Medical Bulletin of Haseki/Haseki Tip Bulteni*, vol. 60, no. 3, 2022.

[5] K. L. Britt, J. Cuzick, and K.-A. Phillips, "Key steps for effective breast cancer prevention," *Nature Reviews Cancer*, vol. 20, no. 8, pp. 417–436, 2020.

[6] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz, and A. Stanisławek, "Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review," *Cancers*, vol. 13, no. 17, p. 4287, 2021.

[7] S. Cocco, M. Piezzo, A. Calabrese, D. Cianniello, R. Caputo, V. Di Lauro, G. Fusco, G. Di Gioia, M. Licenziato, and M. De Laurentiis, "Biomarkers in triple-negative breast cancer: state-of-the-art and future perspectives," *International journal of molecular sciences*, vol. 21, no. 13, p. 4579, 2020.

[8] D. Che, Q. Liu, K. Rasheed, and X. Tao, "Decision tree and ensemble learning algorithms with their applications in bioinformatics," *Software tools and algorithms for biological systems*, pp. 191–199, 2011.

[9] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[10] A. Thalor, H. K. Joon, G. Singh, S. Roy, and D. Gupta, "Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer," *Computational and structural biotechnology journal*, vol. 20, pp. 1618–1631, 2022.

[11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[12] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[13] M. Yap, R. L. Johnston, H. Foley, S. MacDonald, O. Kondrashova, K. A. Tran, K. Nones, L. T. Koufariotis, C. Bean, J. V. Pearson *et al.*, "Verifying explainability of a deep learning tissue classifier trained on rna-seq data," *Scientific reports*, vol. 11, no. 1, p. 2641, 2021.

[14] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[15] N. Amoroso, D. Pomarico, A. Fanizzi, V. Didonna, F. Giotta, D. La Forgia, A. Latorre, A. Monaco, E. Pantaleo, N. Petruzzellis *et al.*, "A roadmap towards breast cancer therapies supported by explainable artificial intelligence," *Applied Sciences*, vol. 11, no. 11, p. 4881, 2021.

[16] R. Massafra, A. Fanizzi, N. Amoroso, S. Bove, M. C. Comes, D. Pomarico, V. Didonna, S. Diotaiuti, L. Galati, F. Giotta *et al.*, "Analysing breast cancer invasive disease event classification through explainable artificial intelligence," *Frontiers in Medicine*, vol. 10, p. 95, 2023.

[17] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," *IEEE Access*, vol. 9, pp. 144 352–144 360, 2021.

[18] C. Buccitelli and M. Selbach, "mrnas, proteins and the emerging principles of gene expression control," *Nature Reviews Genetics*, vol. 21, no. 10, pp. 630–644, 2020.

[19] P. Fafournoux, A. Bruhat, and C. Jousse, "Amino acid regulation of gene expression," *Biochemical Journal*, vol. 351, no. 1, pp. 1–12, 2000.

[20] H.-M. Chen and J. A. MacDonald, "Network analysis of tcga and gtex gene expression datasets for identification of trait-associated biomarkers in human cancer," *STAR protocols*, vol. 3, no. 1, p. 101168, 2022.

[21] J. Allaire, "Rstudio: integrated development environment for r," *Boston, MA*, vol. 770, no. 394, pp. 165–171, 2012.

[22] C. Hutter and J. C. Zenklusen, "The cancer genome atlas: creating lasting value beyond its data," *Cell*, vol. 173, no. 2, pp. 283–285, 2018.

[23] L. J. Carithers and H. M. Moore, "The genotype-tissue expression (gtex) project," pp. 307–308, 2015.

[24] D. Zhang and Y. Gong, "The comparison of lightgbm and xgboost coupling factor analysis and prediagnosis of acute liver failure," *IEEE Access*, vol. 8, pp. 220 990–221 003, 2020.

[25] D. Wang, Y. Zhang, and Y. Zhao, "Lightgbm: an effective mirna classification method in breast cancer patients," in *Proceedings of the 2017 international conference on computational biology and bioinformatics*, 2017, pp. 7–11.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[28] O. Sagi and L. Rokach, "Approximating xgboost with an interpretable decision tree," *Information sciences*, vol. 572, pp. 522–542, 2021.

[29] S. Wu, Q. Yuan, Z. Yan, and Q. Xu, "Analyzing accident injury severity via an extreme gradient boosting (xgboost) model," *Journal of advanced transportation*, vol. 2021, pp. 1–11, 2021.

[30] S. S. Dhaliwal, A.-A. Nahid, and R. Abbas, "Effective intrusion detection system using xgboost," *Information*, vol. 9, no. 7, p. 149, 2018.

[31] A. Ogunleye and Q.-G. Wang, "Xgboost model for chronic kidney disease diagnosis," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 17, no. 6, pp. 2131–2140, 2019.

[32] Z. Zhong, Y. Li, Z. Han, and Z. Yang, "Ship target detection based on lightgbm algorithm," in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*. IEEE, 2020, pp. 425–429.

[33] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[34] S. Lundberg and S.-I. Lee, "An unexpected unity among methods for interpreting model predictions," *arXiv preprint arXiv:1611.07478*, 2016.

[35] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using improved shapley additive explanation," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 546–546.

[36] L. E. Tideman, L. G. Migas, K. V. Djambazova, N. H. Patterson, R. M. Caprioli, J. M. Spraggins, and R. Van de Plas, "Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized shapley additive explanations," *Analytica Chimica Acta*, vol. 1177, p. 338522, 2021.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[38] R. Rynazal, K. Fujisawa, H. Shiroma, F. Salim, S. Mizutani, S. Shiba, S. Yachida, and T. Yamada, "Leveraging explainable ai for gut microbiome-based colorectal cancer classification," *Genome Biology*, vol. 24, no. 1, pp. 1–13, 2023.

[39] Q. Liu, S. Shu, R. R. Wang, F. Liu, B. Cui, X. N. Guo, C. X. Lu, X. G. Li, M. S. Liu, B. Peng *et al.*, "Whole-exome sequencing identifies a missense mutation in hnrnpa1 in a family with flail arm als," *Neurology*, vol. 87, no. 17, pp. 1763–1769, 2016.

[40] C. J. David, M. Chen, M. Assanah, P. Canoll, and J. L. Manley, "Hnrnp proteins controlled by c-myc deregulate pyruvate kinase mrna splicing in cancer," *Nature*, vol. 463, no. 7279, pp. 364–368, 2010.

[41] C. Feng, H. Huang, S. Huang, Y.-Z. Zhai, J. Dong, L. Chen, Z. Huang, X. Zhou, B. Li, L.-L. Wang *et al.*, "Identification of potential key genes associated with severe pneumonia using mrna-seq," *Experimental and Therapeutic Medicine*, vol. 16, no. 2, pp. 758–766, 2018.

[42] R. Hourihan, G. O'sullivan, and J. Morgan, "Transcriptional gene expression profiles of oesophageal adenocarcinoma and normal oesophageal tissues." *Anticancer research*, vol. 23, no. 1A, pp. 161–165, 2003.

[43] A. B. Undén, B. Sandstedt, K. Bruce, M.-A. Hedblad, and M. Ståhle-Bäckdahl, "Stromelysin-3 mrna associated with myofibroblasts is overexpressed in aggressive basal cell carcinoma and in dermatofibroma but not in dermatofibrosarcoma," *Journal of investigative dermatology*, vol. 107, no. 2, pp. 147–153, 1996.

[44] D. Cilliers, Y. Alanay, K. Boduroglu, E. Utine, E. Tunçbilek, and J. Clayton-Smith, "Cerebro-facio-thoracic dysplasia: expanding the phenotype," *Clinical dysmorphology*, vol. 16, no. 2, pp. 121–125, 2007.

[45] F. Müller-Pillasch, F. Zimmerhackl, U. Lacher, N. Schultz, H. Hameister, G. Varga, H. Friess, M. Büchler, G. Adler, and T. Gress, "Cloning of novel transcripts of the human guanine-nucleotide-exchange factor mss4: In situchromosomal mapping and expression in pancreatic cancer," *Genomics*, vol. 46, no. 3, pp. 389–396, 1997.

[46] M. Long, J. Wang, and M. Yang, "Transcriptomic profiling of breast cancer cells induced by tumor-associated macrophages generates a robust prognostic gene signature," *Cancers*, vol. 14, no. 21, p. 5364, 2022.

[47] M. Li, J. Zhao, R. Yang, R. Cai, X. Liu, J. Xie, B. Shu, and S. Qi, "Cenpf as an independent prognostic and metastasis biomarker corresponding to cd4+ memory t cells in cutaneous melanoma," *Cancer Science*, vol. 113, no. 4, p. 1220, 2022.

[48] D. Wang, W. Xu, M. Huang, W. Ma, Y. Liu, X. Zhou, Q. Yang, and K. Mu, "Cenpf knockdown inhibits adriamycin chemoresistance in triple-negative breast cancer via the rb-e2f1 axis," *Scientific Reports*, vol. 13, no. 1, p. 1803, 2023.

[49] W. Zhan, Y. Liu, Y. Gao, R. Gong, W. Wang, R. Zhang, Y. Wu, T. Kang, and D. Wei, "Rmi2 plays crucial roles in growth and metastasis of lung cancer," *Signal Transduction and Targeted Therapy*, vol. 5, no. 1, p. 188, 2020.

[50] D. F. Hudson, D. J. Amor, A. Boys, K. Butler, L. Williams, T. Zhang, and P. Kalitsis, "Loss of rmi2 increases genome instability and causes a bloom-like syndrome," *PLoS genetics*, vol. 12, no. 12, p. e1006483, 2016.

[51] L. Li, T.-T. Zeng, B.-Z. Zhang, Y. Li, Y.-H. Zhu, and X.-Y. Guan, "Overexpression of hn1l promotes cell malignant proliferation in non-small cell lung cancer," *Cancer Biology & Therapy*, vol. 18, no. 11, pp. 904–915, 2017.

[52] L. Xiao-Jie, G. Ai-Mei, J. Li-Juan, and X. Jiang, "Pseudogene in cancer: real functions and promising signature," *Journal of medical genetics*, vol. 52, no. 1, pp. 17–24, 2015.

[53] L. Han, Y. Yuan, S. Zheng, Y. Yang, J. Li, M. E. Edgerton, L. Diao, Y. Xu, R. G. Verhaak, and H. Liang, "The pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes," *Nature communications*, vol. 5, no. 1, p. 3963, 2014.

[54] J. Wang, X. Wang, A. Bhat, Y. Chen, K. Xu, Y.-y. Mo, S. S. Yi, and Y. Zhou, "Comprehensive network analysis reveals alternative splicing-related lncrnas in hepatocellular carcinoma," *Frontiers in Genetics*, vol. 11, p. 659, 2020.

[55] D. Jiao, J. Zhang, P. Chen, X. Guo, J. Qiao, J. Zhu, L. Wang, Z. Lu, and Z. Liu, "Hn1l promotes migration and invasion of breast cancer by up-regulating the expression of hmgb1," *Journal of cellular and molecular medicine*, vol. 25, no. 1, pp. 397–410, 2021.

[56] L. Albiges, A. Goubar, V. Scott, C. Vicier, C. Lefèbvre, S. Alsafadi, F. Commo, M. Saghatchian, V. Lazar, P. Dessen *et al.*, "Chk1 as a new therapeutic target in triple-negative breast cancer," *The Breast*, vol. 23, no. 3, pp. 250–258, 2014.