IA explicável aplicada para identificar genes influentes na classificação do câncer por meio de dados de expressão gênica de RNA-Seq

Karolayne S. Azevedo*,Luísa C. de Souza*, 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: karolayneazevsantos@gmail.com, luisa.souza.103@ufrn.edu.br matheusdalmolinrs@gmail.com,mfernandes@dca.ufrn.br

***Abstract*—Este artigo faz uso de três técnicas de aprendizagem de máquina (*Machine Learnig* - ML) para classificar os cinco tipos de câncer mais recorrentes em mulheres, a partir de dados de expressão gênica de RNA-Seq. Os desafios incluem: alta dimensionalidade do conjunto de dados e a falta de transparência dos modelos de ML. Para mitigar esses problemas, foi utilizado a técnica SHAP (*SHapley Additive exPlanations*) que é uma técnica de inteligência artificial explicável (*Explainable artificial intelligence* - XAI) utilizada para compreender como esses modelos tomam decisões podendo ser usada como uma estratégia para a seleção de recursos. Como entrada, foram utilizadas 2.105 amostras, sendo 421 amostras referentes a cada tumor, processadas pelos modelos Árvore de Decisão (*Decision Tree*- DT), Floresta Aleatória (*Random Forest*-RF) e Aumento de Gradiente (*ExtremoeXtreme Gradient Boosting*-XGB) treinadas e validadas por meio da técnica de validação cruzada. Os modelos RF, DT e XGB alcançaram precisões de $99,40\%$, $97,60\%$ e $99,34\%$. Posteriormente, a técnica SHAP foi utilizada para obter uma lista de recursos visando compreender quais características influenciaram nas tomadas de decisões dos modelos e consequentemente, nos resultados de predição dos cinco tumores. 122, 90 e 11 genes foram obtidos nos modelos RF, XGB e DT, totalizando 223 resultando em 194 genes únicos.**

***Index Terms*—Explainable AI, machine learning, feature selection, RNA-Seq, cancer, SHAP, gene expression.**

I. Introduction

O câncer é uma das causas mais comuns de morte entre as mulheres, sendo o câncer de mama (*Breast Cancer* - BRCA), um dos mais incidentes, ocupando o segundo lugar entre as causas de morte relacionadas ao câncer em mulheres [1]. Na América, 30% dos casos detectados correspondem ao câncer de mama cuja taxa de mortalidade é de 190 a cada 100.000 casos [2]. O câncer de pulmão é o câncer de maior incidência entre as mulheres (*Lung adenocarcinoma* - LUAD) [3], [4], o câncer de ovário, é considerado um dos mais fatais devido à sua grande dificuldade de detecção, sendo um desafio diagnosticá-lo precocemente (*Ovarian* - OV) [5]. O câncer de colorretal (*Colon Adenocarcinomas* - COAD) é o terceiro câncer mais comum em todo mundo, afetando aproximadamente um milhão de pacientes todos os anos [6], [7]. Globalmente falando, o câncer de Tireóide (*Thyroid cancer* - THCA) é três vezes mais incidentes em mulheres, sendo um dos cânceres mais comumente diagnosticado antes dos 30 anos [8], [9]. Por se tratar um problema de saúde global, ações governamentais e cientificas são urgentes e necessárias. Os países de baixa e média renda, geralmente enfrentam cargas mais altas de câncer, tendo acesso limitado a medidas de prevenção e tratamento do câncer, corroborando para taxas de sobrevivência mais baixas. A detecção precoce desempenha um papel vital na melhoria dos resultados do câncer podendo identificar estágios iniciais da doença [10]. Técnicas de Inteligência Artificial (IA) estão sendo cada vez mais usadas em vários segmentos no que tange as pesquisas cientificas a respeito do câncer e de atendimento ao paciente, auxiliando a detecção, prognóstico, monitoramento e análise de vários tipos de câncer [11], [12].

Trabalhos na literatura abordam os desafios e preocupações associados a implantação efetiva desses modelos na oncologia. Em seu trabalho, [13] menciona que o câncer inclui condições distintas, com padrões únicos e complexos de tratamento. Outra problemática citada por [14] é a falta de transparência de alguns modelos, tendo em vista que muitos, costumam ser considerados caixas pretas que operam por meio de algoritmos complexos e difíceis de interpretar. Essa falta de transparência limita a confiança que pacientes e médicos têm nas previsões dos modelos.

Com intuito de mitigar esse problema, técnicas de inteligência artificial explicável (*Explainable artificial intelligence* - XAI) vêm sendo utilizadas na tentativa de compreender como esses modelos tomam suas decisões e quais recursos ou entradas têm mais influência em suas previsões [15], [16]. A técnica SHAP (*SHapley Additive exPlanations*) faz parte dos recursos XAI, e baseia-se nos valores SHAP para explicar a saída de alguns modelos de aprendizagem de máquina (*Machine Learnig* - ML) e aprendizagem profunda (*Deep Learning*- DL) [17].

Neste contexto, este trabalho, propõe o uso de uma técnica XAI, baseada em valores SHAP, para identificar características mais relevantes em um problema de classificação multiclasse entre cinco tipos de câncer mais recorrentes em mulheres a partir de dados de expressão de genes RNA-seq, extraídas do

Atlas do Genoma do Câncer (The Cancer Genome Atlas - TCGA). Os dados foram aplicados em modelos tradicionais de ML baseados em árvores. Assim, este trabalho traz as seguintes contribuições específicas:

- Uso de um método XAI para explicar o comportamento dos classificadores baseado na biblioteca SHAP da linguagem *python* de programação.
- Possibilidade de análise dos principais genes identificados pela técnica SHAP.
- Obtenção de modelos com bons desempenho baseados em modelos de ML a partir de valores de expressões gênicas de RNA-seq.
- Possibilidade de utilizar a técnica SHAP como método de redução de dimensionalidade para recursos de entradas.

II. Trabalhos Relacionados

Em seu trabalho, [18] fez uso da máquina de vetores de suporte (*Support Vector Machine* -SVM), Regressão Logística (*Logistic Regression*- RL), K-vizinhos mais próximos (*K-nearest neighbors*- KNN), Árvore de Decisão (*Decision tree* - DT), *Naive Bayes* e floresta aleatória (*Random Forest*- RF) para classificar câncer de mama benigo e maligno a partir da base de dados do (*Wisconsin Breast Cancer Dataset*-WBCD) alcançando valores de acurácia entre 94% e 97%.

A classificação de subgrupos de câncer de mama a partir do uso de perfis de mutação genética para foi proposta por [19] utilizando modelos de ML amplamente conhecidos. Contudo, o maior valor de precisão obtido com este estudo foi de 70% por meio do modelo RF.

A classificação e seleção de recursos dos cânceres de cólon, próstata e leucemia foi realizada a partir de dados de expressão gênica. Os valores de precisão obtidos através do algoritmo RF, para o câncer de cólon, próstata e leucemia foram de $85,45\%$, $66,66\%$ e 100% respectivamente [20].

Um método de seleção de recurso, juntamente com o modelo SVM, foi aplicado em dados de expressões gênicas com intuito de classificar amostras de dois subtipos de câncer pulmonar. Para tanto, os autores obtiveram valores de precisão entre $91,00\%$ a $96,70\%$ que variaram de acordo com os seletores de características utilizados previamente [21]. Cinco modelos de ML foram utilizados para diagnóstico de câncer em tecidos ovarianos a partir de um painel de mRNA selecionado. Os maiores valores de especificidade foi obtido por meio do modelo RF [22].

A maior quantidade de trabalhos que utilizam expressão gênica na predição de câncer estão no campo da *deep learning* [23]. As técnicas comumente utilizadas são as redes neurais convolucionais (*convolutional neural networks* - CNN), redes neurais totalmente conectadas (*fully connected neural networks* - FCNNs) e redes neurais recorrentes (*recurrent neural networks* - RNN).

Uma classificação multiclasse foi realizada a partir de cinco tipos de câncer. Os autores converteram dados de RNA-seq em imagens $2D$ e aplicaram numa CNN de múltiplas camadas. A abordagem proposta alcançou uma precisão geral de teste de $96,90\%$ [24]. Seguindo a mesma vertente, [25] fez uso de sequências de RNA-seq em imagens $2D$ para vários tipos de câncer e aplicaram numa CNN obtendo acurácia de $95,65\%$.

Como mencionado anteriormente, para que os modelos sejam aceitos e inseridos na oncologia, é interessante inserir técnicas que possibilite visualizar e compreender as tomadas de decisões dos modelos.

Em seu trabalho, [17] fez uso de uma classificação multiclasse para detectar subtipos de câncer de mama. Para tanto, os autores utilizaram os modelos SVM, RF, Árvores Extremamente Aleatórias(*Extremely Randomized Trees*-ERT) e Aumento de Gradiente (*ExtremoeXtreme Gradient Boosting*-XGB) para obter os resultados de predição, e em seguida aplicaram a técnica SHAP para obter o conjunto de características que influenciaram nos modelos. Os valores de acurácia obtidos encontra-se no intervalo entre 61% e 77%.

Modelos de ML foram aplicados em imagens de ultrassom e ressonância magnética para detectar câncer de próstata. Os valores de precisão alcançados pelo método proposto foram entre 80% e 97% [26]. Por fim, os resultados dos modelos foram submetidos a técnica SHAP.

Amostras de dados de RNA-seq de expressão genótipo-tecido de 47 tecidos diferentes foram aplicados numa CNN de múltiplas camadas na qual obteve resultados de precisão entre 70% e 100%. A técnica SHAP foi utilizada para obter os recursos mais relevantes e compreender os processos biológicos envolvidos na diferenciação e função dos tecidos [27].

A maioria dos trabalhos que apresentaram valores de acurácia e precisão maiores que 90% fizeram uso de técnicas de DL em seus modelos de classificação. Esses modelos, em sua maioria, envolve um alto custo computacional se comparado com as técnicas convencionais de ML. Além disso, alguns trabalhos utilizam outras técnicas para reduzir e para explicar os modelos. O trabalho proposto resulta numa lista única de genes bem reduzida se comparado com outros trabalhos mencionados anteriormente [17], [27].

III. Metodologia

Para a obtenção das características mais importantes na classificação dos tipos de câncer mais frequentes em mulheres, este trabalho fez uso de dados de expressão gênica, que são medidas quantitativas dos RNA mensageiros presentes em uma determinada amostra, relativas a uma condição fisiológica específica.

Esses dados foram inicialmente treinados em três técnicas tradicionais da ML, sendo os modelos RF, DT e XGB os escolhidos, e posteriormente submetidos à biblioteca SHAP para obtenção das principais características que influência as tomadas de decisões dos modelos que resultaram na classificação dos cinco tumores é obtida. Em seguida, foram selecionados recursos cuja importância na previsão de saída em cada modelo fossem maior ou igual à $0{,}1\%$, pois valores abaixo desse corte não colaboraram significativamente com o aumento da performance das técnicas usadas. O fluxograma de atividades é exibido na Figura 1 abaixo. Em seguida, uma lista combinada resultante com os genes únicos de cada matriz de SHAP Values é obtida e processada, novamente, pelos modelos

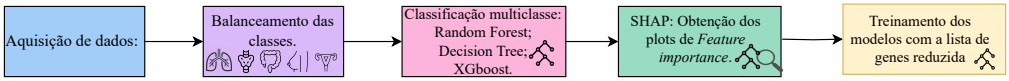


Fig. 1: Fluxograma de atividades para obtenção das características mais importantes na classificação.

(DT, RF, XGB) para verificar se é possível obtermos bons resultados de predição a partir de uma lista de genes reduzida.

A. Base de Dados

Os dados de RNA-seq foram obtidos do Atlas do Genoma do Câncer *(The Cancer Genome Atlas* - TCGA) que contém informações de tecidos cancerigenos, referentes ao câncer de mama (BRCA), câncer de pulmão (LUAD), (THCA), câncer de ovário (OV) e câncer de cólon (COAD) totalizando 3.057 amostras. Como observado na Figura 2, o BRCA representa $36,34\%$ dos tecidos cancerígenos contidos na base de dados, seguido do LUAD $(17,63\%)$ e THCA $(16,52\%)$. Os câncer de cólon e ovário encontram-se em menor número, representando apenas $15,73\%$ e $14,75\%$.

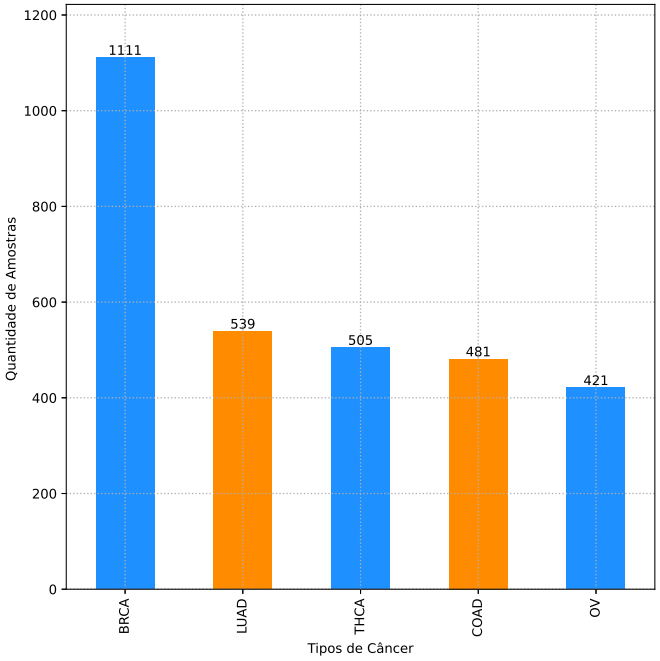


Fig. 2: Quantidade de amostras existentes na base de dados antes de aplicar a técninca debalanceamento *Under Sampling*.

Neste contexto, faz-se necessário realizar o balanceamento dos dados não somente para melhorar o desempenho da rede, mas também evitar problemas como *Overfitting* em virtude da desproporção das quantidades de amostras em relação aos demais tipos de câncer, além de evitar resultados tendenciosos e até mesmo, dificultar o aprendizado e a generalização dos modelos de ML.

Existem várias técnicas que podem ser usadas para resolver o desequilíbrio de dados, neste trabalho, foi utilizado a técnica de reamostragem do conjunto de dados, que consistem em subamostrar (*Under Sampling*) a quantidade de amostras da classe majoritária para equilibrar o conjunto de dados a partir da classe minoritária (COAD). Desse modo, 421 amostras foram extraídas aleatoriamente de cada classe presente da base de dados. Logo, o conjunto de dados utilizado na etapa de treinamento dos modelos, passa a conter $2,105$ amostras pertencentes aos cinco tipos de tecidos cancerígenos, representados por rótulos, que vão de 0 à 4, sendo cada rótulo associado a uma classe. Parte das amostras restantes, foram utilizadas para testar à performance da rede.

B. Random Forest

No geral, o algoritmo RF é eficaz nas tarefas de regressão e classificação multiclasse, sendo inicialmente implementada por [28], baseando-se no conceito de *ensemble learning*, na qual utiliza um conjunto de árvores de decisão aleatórias no processo de aprendizagem [28], [29]. O treinamento é inicializado selecionando através de *bootstraping*, para cada árvore não podada, um diferente subconjunto aleatório de tamanho *N* dos dados de entrada

$$\mathbf{D} = [(x_1, y_1), \ldots, (x_N, y_N)] \tag{1}$$

onde cada vetor de característica $x_i = (x_{i,1}, ..., x_{i,M})^T$ denotam os M preditores, enquanto y_i está associada a resposta esperada.

Em seguida, a separação dos nós de cada árvore individual é determinada encontrando as melhores divisões de *m* associadas a cada nó, onde *m* é um *subset* de preditores selecionados aleatoriamente do conjunto total de preditores disponíveis de quantidade *M*, onde *m*<<*M*. Consequentemente, as árvores de decisão do modelo apresentarão diferentes condições para seus nós, resultando em estruturas diferentes [28], [30], [31]. Finalmente, a previsão é realizada a partir das médias dos resultados individuais das várias árvores de decisão [32].

C. eXtreme Gradient Boosting- XGBoost

As técnicas *Boosting* combinam várias outras técnicas de aprendizado simples com intuito de criar um modelo mais robusto de modo que, cada classificador fraco tenta melhorar a classificação de amostras que foram classificadas erroneamente pelo classificador fraco anterior a fim de melhorar a precisão preditiva do modelo em comparação com um único modelo de aprendizado. Vários algoritmos de aprendizagem de máquina baseiam-se nesta técnica, dentre eles estão os algoritmos Reforço Adaptável (*Adaptive Boosting*-Adaboost), Aumento de Gradiente (*Gradient Boosting*) e XGB, tendo a técnica de árvores de decisão incluídas em muitas dessas estruturas.

O algoritmo XGB incorpora a técnica do aumento de gradiente além de outros recursos voltados para melhoria do próprio algoritmo, como recursos computacionais relacionados

ao hardware que está rodando o modelo, e o conceito de regularização. O XGB procura otimizar a função de custo, minimizando seu gradiente a cada iteração, de modo a obter a melhor árvore, cujo o erro seja o menor possível, conforme a Equação 2.

$$L = \sum_{i=1}^{N} L(y_i, y)) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

onde $\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w^2$ corresponde o termo de regularização da função objetivo L responsável por medir a complexidade do modelo, sendo T o número de folhas das árvores e w as pontuações de saída das folhas que controla o ganho mínimo de redução de perda necessário para dividir um nó interno [33], [34].

D. Explainable Artificial Intelligence

A maior parte dos modelos de IA, incluindo aprendizagem profunda, podem ser entendidos como caixas pretas devido a dificuldade de entendimento do seu processo de tomada de decisão. Essa falta de transparência pode ser um problema em muitos contextos em que é importante entender como os modelos de IA tomam decisões.

Neste sentido, a explicabilidade de um modelo permitiria aumentar a confiabilidade, transparência e a interpretabilidade dos resultados dos modelos de IA e até mesmo viabilizar a redução de custos computacionais impostos por muitas dessas técnicas ao serem aplicadas em conjuntos de dados grandes [34]. As técnicas de XAI são um campo emergente da IA capaz de apontar quais recursos foram mais relevantes para tomadas de decisões de seus algoritmos, bem como os recursos sem prejudicar seu desempenho.

Baseado na teoria dos jogos, o *SHApley Additive exPlanations* é um método capaz de interpretar os recursos mais relevantes nos resultados de predição dos modelos de ML baseados nos valores SHAP [17]. Os valores SHAP podem ser usados para explicar a saída de qualquer modelo de aprendizado de máquina, incluindo redes neurais, árvores de decisão e modelos lineares a partir de um cálculo de importância relativa de cada recurso na previsão do modelo [35], [36]. Isso nos permite entender como o modelo está fazendo sua previsão e quais recursos de entrada estão tendo o maior impacto na saída, matematicamente descrito por:

$$f(x) = g(X^{'}) = \phi_0 + \sum_{i=1}^{M} \phi_i x_i^{'} \tag{3}$$

onde M representa o número de recursos de entrada e ϕ_0 representa uma constante quando todas as entradas estão ausentes, $x_i^{'}$ representa o recurso i observado. O valor SHAP para cada recurso ϕ_i foi proposto e detalhado por [37].

E. Treinamento dos Modelos

Para realizar o treinamento dos modelos, os dados foram divididos e embaralhados aleatoriamente em conjunto treinamento e validação na proporção 80% e 20%. A validação cruzada foi utilizada para validar e verificar o desempenho dos modelos. O valor de *k=10* foi escolhido a partir de uma varredura que permitiu encontrar o valor do *fold* a partir do melhor desempenho dos modelos. Os hiperparâmetros desempenham um importante papel no desempenho, generalização e interpretabilidade nos modelos de IA. Um ajuste em grande foi aplicado aos modelos com intuito de determinar os melhores parâmetros de treinamento, entretanto, ao aplicar o conjunto de teste nos modelos, em muitos casos observou-se a presença de *overfitting*.

Dessa forma, um ajuste manual, a partir das curvas de treinamento do modelo, foi realizado nos principais parâmetros dos modelos. Para o algoritmo DT a profundidade da árvore adotada foi de 3 com intuito de evitar árvores profundas. O número de folhas máximo foi 5 e o critério adotado para a escolha dos nós foi a entropia. Um dos principais parâmetros para ser ajustado no modelo RF corresponde a quantidade de árvores individuais e número mínimo de amostras necessárias para dividir um nó interno. Dessa forma, os valores escolhidos foram 100 e 2, a profundidade máxima foi 3 e o critério utilizado para escolha dos nós foi a perda logarítmica [38]. No XGB utilizou-se a *softmax* para otimizar as probabilidades de classe, por se tratar de um problema de multiclasse, a profundidade máxima das árvores de seus modelos bases foi de 3 [39].

As curvas de aprendizagem de cada modelo foi gerada com intuito de observar seus comportamentos durante o processo de treinamento à medida que ocorre o aumento de amostras conforme a Figura 3.

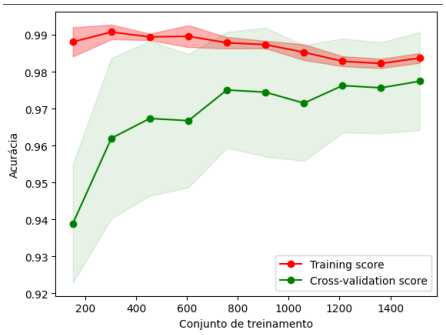
A pontuação de treinamento permanece alta em todos os conjuntos de treinamento de todos os modelos. Contudo, ocorre um crescimento da pontuação para os dados de validação à medida que têm-se um aumento dos dados treinamento. Nas curva do XGB (Figura 3c) e DT (Figura 3a) vemos uma diferença significante entre a pontuação do treinamento quando se tem menos de 600 amostras, acima de 1.000 amostras a diferença entre a precisão do treinamento e da validação é bem menor.

Todas as curvas, indicam que existe uma boa compensação entre o viés e variância e apontam que as quantidades de amostras utilizadas nos treinamentos dos modelos, seriam suficientes, tendo em vista que os modelos se estabilizam com um número considerável de amostras. A diferença reduzida entre as curvas de treinamento e validação consolida a ausência de *overfitting* indicando que o modelo pode generalizar bem e corroboram com os resultados que serão apresentados a seguir (ver Tabela I).

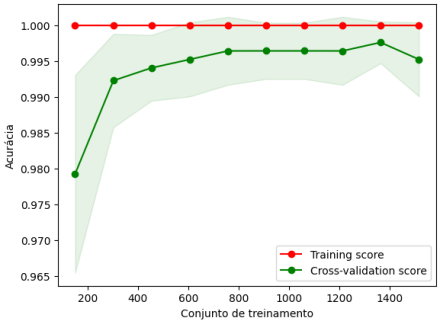
IV. Resultados e Discussões

As métricas de acurácia, precisão, sensibilidade e F1-*score* foram utilizadas para avaliar os desempenhos dos modelos aplicados ao problema de classificação dos cinco tipos de tumores mais recorrentes em mulheres. Dessa forma, esses valores, correspondem à média entre todos os valores obtidos em cada *fold* e podem ser visualizados na Tabela I.

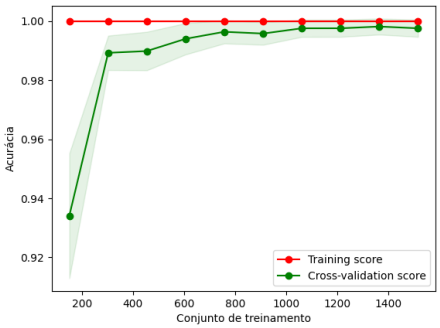
O RF obteve a melhor performance, atingindo o valor de acurácia de $99,82\%$ e precisão $99,83\%$. Em segundo lugar,



(a) Curva de treinamento referente ao modelo RF.



(b) Curva de treinamento referente ao modelo DT.



(c) Curva de treinamento referente ao modelo XGB.

Fig. 3: Curva de aprendizado representando pontuações de acurácia para dados de treinamento e validação versus tamanho dos dados de treinamento.

TABLE I: Comparação das métricas de desempenho para a base de dados original.

Classificadores	Acurácia	Precisão	Sensibilidade	F1-Score
Random Forest	99,40%	99,43%	99,40%	99,40%
Decision Tree	97,60%	97,74%	97,63%	97,64%
XGBoost	99,34%	99,36%	99,34%	99,34%

temos o XGB que também apresenta valores de acurácia, precisão, sensibilidade e F1-*score* acima de $99,00\%$. O algoritmo DT também apresentou um bom desempenho, alcançando valores acima de 97% para todas as métricas adotadas. Destacamos que os grupos analisados pertencem a tipos tumorais altamente heterogêneos, o que pode explicar o alto desempenho obtido pelos modelos empregados, incluindo o DT. O desvio padrão pode indicar a variância dos modelos durante seus treinamentos. O desvio padrão referente ao treinamento do modelo DT em termos de acurácia foi $0,0089$, para o RT $0,003709$ e o XGB $0,00356$ respectivamente. Dessa forma, quanto menor o desvio padrão, mais baixa seriam as variâncias.

Posteriormente, o método SHAP foi aplicado nos três modelos com intuito de observar a contribuição dos recursos que mais tiveram influencia em suas tomadas de decisões. A matriz global dos valores SHAP nos fornece os recursos de maior impacto no processo de classificação dos cinco tipos de câncer. Calculou-se os valores médios de SHAP associados a cada gene em todas as amostras do conjunto de treinamento.

Ao final da seleção 223 genes foram selecionados, sendo 122 genes extraídos do modelo RF, 11 do DT e 90 genes provinientes do XGB, sendo 194 genes únicos. A quantidade final de genes extraídos representam menos de 1% (21.481) do total do conjunto de genes original. Uma discussão a respeito dos principais genes encontrados pela técnica será realizada em IV-A. A etapa final de nossa análise consistiu em retreinar

TABLE II: Comparação entre as métricas de desempenho com o conjunto de genes reduzidos obtidos por meio da técnica SHAP Values.

Classificadores	Acurácia	Precisão	Sensibilidade	F1-*score*
Decision Tree	98,69%	98,74%	98,70%	98,70%
Random Forest	99,76%	99,77%	99,87%	99,86%
XGBoost	99,64%	99,66%	99,79%	99,80%

os modelos utilizando exclusivamente os genes selecionados pela técnica SHAP (223 genes), utilizando todos as condições iniciais de treinamento de acordo com seu respectivo modelo. Os valores referentes as métricas de desempenho, para todos os modelos, desse experimento podem ser visualizados na Tabela II. Mesmo diminuindo a quantidade de genes, os modelos tiveram um aumento de aproximadamente 1% em seus valores de acurácia, precisão, sensibilidade e F1-*score*. Esses resultados, sinalizam que os genes selecionados por essa técnica podem ser suficientes para obter bons resultados a um

custo computacional menor dento em vista a redução massiva da base de dados.

A. *Características Selecionadas pela técninca SHAP*

A matriz global dos valores SHAP nos fornece os recursos de maior impacto nas cinco classes bem como sua contribuição. As Figuras 4, 5 e 6 representam os valores globais da técnica SHAP obtidos para os algoritmos RF, DT e XGB. O padrão de contribuição dos valores obtidos pela técnica SHAP, difere entre as três classes. A técnica RF e a DT apresentam genes com contribuições para todas as classes, mesmo que de forma desequilibrada, o XGB resulta em genes com valores concentrados em apenas uma das classes.

Os resultados obtidos pela técnica SHAP para classificação de tecidos utilizando dados de RNA-seq, demonstraram que os genes com maiores valores de SHAP refletiam os processos biológicos relevantes de acordo correspondente com a classe predita [27]. Os genes CDX1, PAX8, SFTA3, TBX4 e TG foram classificados entre os 20 primeiros das técnicas de ML aplicadas. CDX1 (*Caudal Type Homeobox 1*) foi o único gene identificado com alto valor de SHAP em todos os modelos. No modelo RF, o CDX1 contribuiu para a classificação de todas as classes tumorais, e no modelo DT, ele contribuiu principalmente para LUAD, BRCA e COAD. Já no modelo XGB, a contribuição foi exclusiva para a classe COAD, o que é coerente com o papel que este gene desempenha no desenvolvimento intestinal e no câncer de cólon [40]–[42].

O gene TBX4 (T-Box Transcription Factor 4) contribuiu principalmente para a classe LUAD no modelo RF e exclusivamente para esta classe no modelo XGB. Este gene desempenha diversos papéis na embriogênese, incluindo a expressão no mesênquima pulmonar e múltiplos papéis no desenvolvimento pulmonar [43]–[45]. O gene TG *Thyroglobulin*) codifica a proteína precursora dos hormônios tireoidianos, que são essenciais para o crescimento, desenvolvimento e controle do metabolismo [46]–[48]. No modelo XGB, este gene se relacionou exclusivamente com THCA e no modelo DT, além de THCA, também se relacionou com a classe OV. O gene PAX8 (*Paired Box 8*) foi identificado nos modelos RF e DT. Nos dois modelos, os valores de SHAP estavam associados a todas as classes de tumor, sendo os maiores valores referentes ao OV, BRCA e THCA. O PAX8 é um fator de transcrição que atua no desenvolvimento embrionário da tireoide, rins e tratos genitais masculinos e femininos [49]. Recentemente, diversos trabalhos estão mostrando o papel de PAX8 no câncer de ovário [50], [51] e no câncer de mama [52]. Por último, o gene SFTA3 (*Surfactant Associated 3*) desempenha um papel especialmente no pulmão como surfactante e possui possíveis propriedades imunológicas [53]. De fato, SFTA3 contribuiu mais para a classe LUAD, seguido de BRCA nos modelos RF e DT.

B. *Explicabilidade vs Interpretabilidade*

As contribuições dos genes mais explicativos obtidos por meio da técnica XAI dos modelos RF e DT para cada classe tumoral nem sempre estão diretamente relacionadas a uma contribuição biológica clara. Embora o modelo possa identificar que um gene contribui para todas as classes, é possível que essa contribuição ocorra de maneira negativa para algumas classes e positiva para outras.

No modelo XGB cada gene contribuiu para uma única classe e se relaciona positivamente com a biologia. É importante reconhecer que alguns genes podem estar relacionados ao desenvolvimento do tecido normal onde o tumor se origina ou outro aspecto, que não seja ligado à doença. Para estabelecer uma relação direta entre os genes identificados pelo método XAI utilizado e o fenótipo. Dessa forma, são necessárias investigações adicionais que superem os desafios apontados.

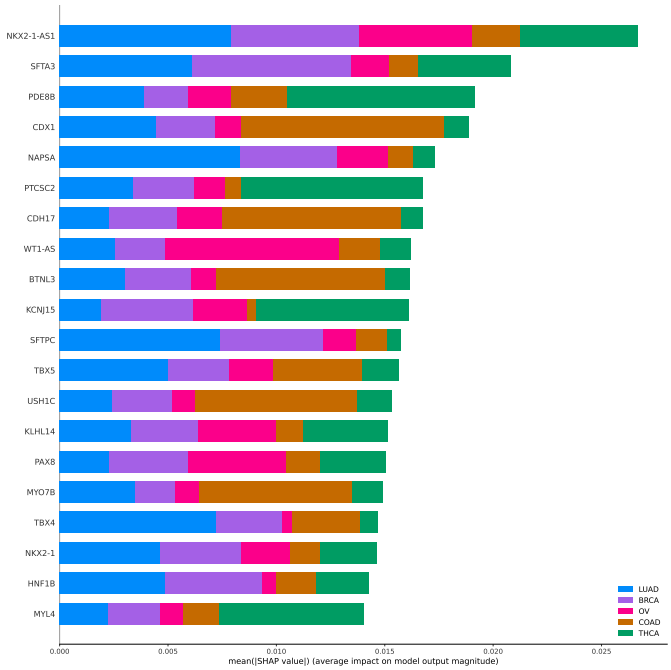


Fig. 4: Random Forest.

V. Conclusão

Este trabalho, propõe o uso de uma técnica XAI, baseada em valores SHAP, para identificar características relevantes de um problema de classificação multiclasse entre os cinco tipos de cânceres mais recorrentes em mulheres a partir de dados de expressão de genes RNA-seq. Os resultados obtidos demonstraram um desempenho excelente, com métricas de desempenho iguais ou superiores aos estudos discutidos na Seção II, que também empregaram modelos tradicionais de ML com dados de expressão gênica.

A utilização de uma técnica de XAI permitiu a seleção de atributos que melhor explicavam as predições dos modelos, aumentando a transparência e confiança nos resultados iniciais. Além disso, os resultados preliminares mostram a possibilidade da utilização da SHAP como método de redução de dimensionalidade para recursos de entradas. É importante mencionar que muitos dos genes obtidos pela técnica XAI estão realmente associados aos tipos tumorais utilizados. Os

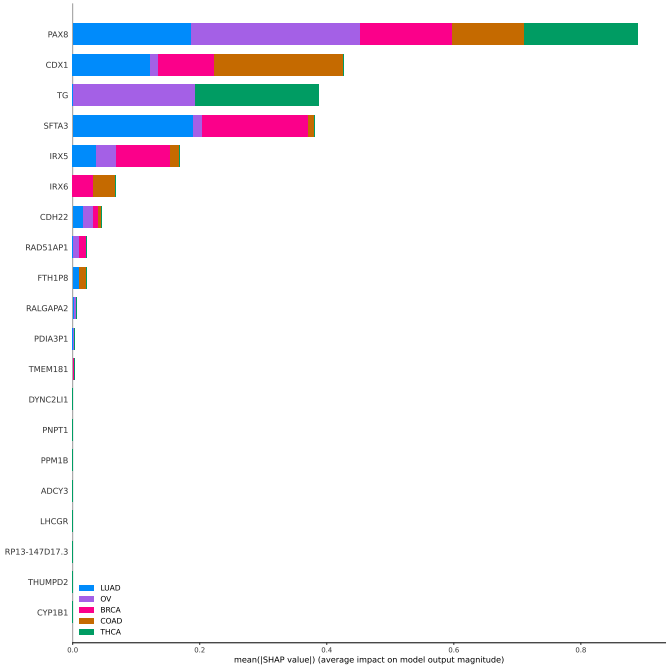


Fig. 5: Decision Tree.

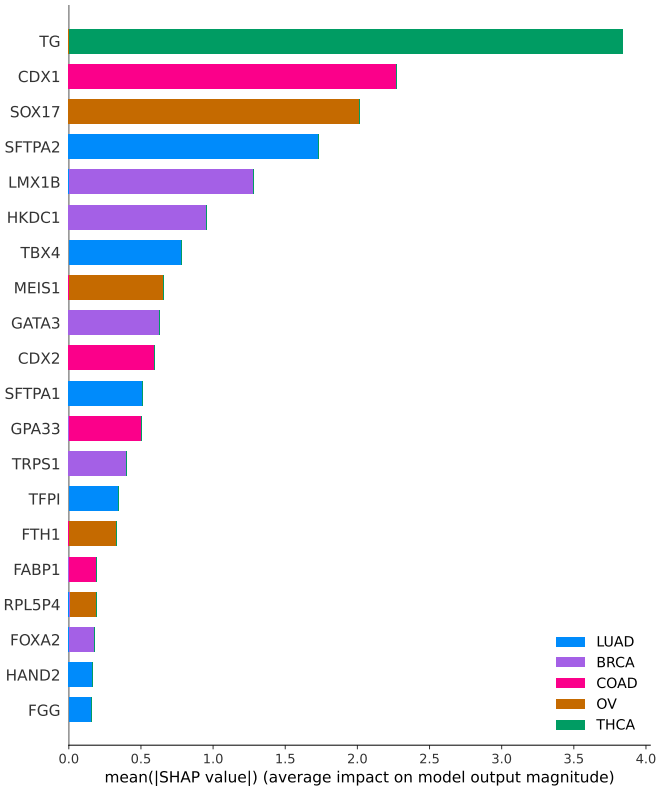


Fig. 6: XGboost.

desafios relacionados à seleção de atributos são significativos, e métodos de XAI, juntamente com outras ferramentas, podem ser empregados na busca de novos potenciais biomarcadores relevantes para doenças complexas, como o câncer.

Acknowledgment

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.

References

[1] M. Fahad Ullah, "Breast cancer: current perspectives on the disease status," *Breast Cancer Metastasis and Drug Resistance: Challenges and Progress*, pp. 51–64, 2019.

[2] X. Wang, I. Ahmad, D. Javeed, S. A. Zaidi, F. M. Alotaibi, M. E. Ghoneim, Y. I. Daradkeh, J. Asghar, and E. T. Eldin, "Intelligent hybrid deep learning model for breast cancer detection," *Electronics*, vol. 11, no. 17, p. 2767, 2022.

[3] M. M. Fidler-Benaoudia, L. A. Torre, F. Bray, J. Ferlay, and A. Jemal, "Lung cancer incidence in young women vs. young men: a systematic analysis in 40 countries," *International journal of cancer*, vol. 147, no. 3, pp. 811–819, 2020.

[4] L. L. Tsai, N.-Q. Chu, W. A. Blessing, P. Moonsamy, and Y. L. Colson, "Lung cancer in women," *The Annals of Thoracic Surgery*, vol. 114, no. 5, pp. 1965–1973, 2022.

[5] C. Stewart, C. Ralyea, and S. Lockwood, "Ovarian cancer: an integrated review," in *Seminars in oncology nursing*, vol. 35, no. 2. Elsevier, 2019, pp. 151–156.

[6] Y. Cai, N. J. Rattray, Q. Zhang, V. Mironova, A. Santos-Neto, K.-S. Hsu, Z. Rattray, J. R. Cross, Y. Zhang, P. B. Paty *et al.*, "Sex differences in colon cancer metabolism reveal a novel subphenotype," *Scientific reports*, vol. 10, no. 1, p. 4905, 2020.

[7] H. Wen, F. Li, I. Bukhari, Y. Mi, C. Guo, B. Liu, P. Zheng, and S. Liu, "Comprehensive analysis of colorectal cancer immunity and identification of immune-related prognostic targets," *Disease Markers*, vol. 2022, 2022.

[8] E. F. van Velsen, A. M. Leung, and T. I. Korevaar, "Diagnostic and treatment considerations for thyroid cancer in women of reproductive age and the perinatal period," *Endocrinology and Metabolism Clinics*, vol. 51, no. 2, pp. 403–416, 2022.

[9] Z. Tang, J. Zhang, Q. Zhou, S. Xu, Z. Cai, and G. Jiang, "Thyroid cancer "epidemic": a socio-environmental health problem needs collaborative efforts," 2020.

[10] C. Mattiuzzi and G. Lippi, "Current cancer epidemiology," *Journal of epidemiology and global health*, vol. 9, no. 4, p. 217, 2019.

[11] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," *Cancer letters*, vol. 471, pp. 61–71, 2020.

[12] O. Elemento, C. Leslie, J. Lundin, and G. Tourassi, "Artificial intelligence in cancer research, diagnosis and therapy," *Nature Reviews Cancer*, vol. 21, no. 12, pp. 747–752, 2021.

[13] I. S. Chua, M. Gaziel-Yablowitz, Z. T. Korach, K. L. Kehl, N. A. Levitan, Y. E. Arriaga, G. P. Jackson, D. W. Bates, and M. Hassett, "Artificial intelligence in oncology: Path to implementation," *Cancer Medicine*, vol. 10, no. 12, pp. 4138–4149, 2021.

[14] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival," *Scientific reports*, vol. 11, no. 1, p. 6968, 2021.

[15] K. Hauser, A. Kurz, S. Haggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon *et al.*, "Explainable artificial intelligence in skin cancer recognition: A systematic review," *European Journal of Cancer*, vol. 167, pp. 54–69, 2022.

[16] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, 2022.

[17] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable multi-class classification based on integrative feature selection for breast cancer subtyping," *Mathematics*, vol. 10, no. 22, p. 4271, 2022.

[18] S. Ara, A. Das, and A. Dey, "Malignant and benign breast cancer

classification using machine learning algorithms," in *2021 International Conference on Artificial Intelligence (ICAI)*. IEEE, 2021, pp. 97–101.

[19] S. Vural, X. Wang, and C. Guda, "Classification of breast cancer patients using somatic mutation profiles and machine learning approaches," *BMC systems biology*, vol. 10, no. 3, pp. 263–276, 2016.

[20] M. Ram, A. Najafi, and M. T. Shakeri, "Classification and biomarker genes selection for cancer gene expression data using random forest," *Iranian journal of pathology*, vol. 12, no. 4, p. 339, 2017.

[21] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1866, no. 8, p. 165822, 2020.

[22] P. N. Yeganeh and M. T. Mostafavi, "Use of machine learning for diagnosis of cancer in ovarian tissues with a selected mrna panel," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2429–2434.

[23] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[24] N. E. M. Khalifa, M. H. N. Taha, D. Ezzat Ali, A. Slowik, and A. E. Hassanien, "Artificial intelligence technique for gene expression by tumor rna-seq data: A novel optimized deep learning approach," *IEEE Access*, vol. 8, pp. 22 874–22 883, 2020.

[25] J. M. de Guia, M. Devaraj, and C. K. Leung, "Deepgx: Deep learning using gene expression for cancer classification," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2019, pp. 913–920.

[26] M. R. Hassan, M. F. Islam, M. Z. Uddin, G. Ghoshal, M. M. Hassan, S. Huda, and G. Fortino, "Prostate cancer classification from ultrasound and mri images using deep learning based explainable artificial intelligence," *Future Generation Computer Systems*, vol. 127, pp. 462–472, 2022.

[27] 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.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[29] A. Cutler, D. Cutler, and J. Stevens, "Random forests. ensemble machine learning," in *Ensemble Machine Learning*, 2012, pp. 157–175.

[30] I. Reis, D. Baron, and S. Shahaf, "Probabilistic random forest: A machine learning algorithm for noisy data sets," *The Astronomical Journal*, vol. 157, no. 1, p. 16, 2018.

[31] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.

[32] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[33] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[34] Y. Meng, N. Yang, Z. Qian, and G. Zhang, "What makes an online review more helpful: an interpretation framework using xgboost and shap values," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 3, pp. 466–490, 2020.

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

[36] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, "Failure mode and effects analysis of rc members based on machine-learning-based shapley additive explanations (shap) approach," *Engineering Structures*, vol. 219, p. 110927, 2020.

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

[38] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6673–6690, 2017.

[39] B. Quinto, *Next-generation machine learning with spark: Covers XGBoost, LightGBM, Spark NLP, distributed deep learning with keras, and more*. Apress, 2020.

[40] R.-J. Guo, E. Huang, T. Ezaki, N. Patel, K. Sinclair, J. Wu, P. Klein, E.-R. Suh, and J. P. Lynch, "Cdx1 inhibits human colon cancer cell proliferation by reducing β-catenin/t-cell factor transcriptional activity," *Journal of Biological Chemistry*, vol. 279, no. 35, pp. 36 865–36 875, 2004.

[41] E. Pilozzi, M. R. Onelli, V. Ziparo, P. Mercantini, and L. Ruco, "Cdx1 expression is reduced in colorectal carcinoma and is associated with promoter hypermethylation," *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, vol. 204, no. 3, pp. 289–295, 2004.

[42] A. Hryniuk, S. Grainger, J. G. Savory, and D. Lohnes, "Cdx1 and cdx2 function as tumor suppressors," *Journal of Biological Chemistry*, vol. 289, no. 48, pp. 33 343–33 354, 2014.

[43] M. G. Haarman, W. S. Kerstjens-Frederikse, and R. M. Berger, "Tbx4 variants and pulmonary diseases: getting out of the 'box'," *Current opinion in pulmonary medicine*, vol. 26, no. 3, p. 277, 2020.

[44] L. Naiche, R. Arora, A. Kania, M. Lewandoski, and V. E. Papaioannou, "Identity and fate of tbx4-expressing cells reveal developmental cell fate decisions in the allantois, limb, and external genitalia," *Developmental dynamics*, vol. 240, no. 10, pp. 2290–2300, 2011.

[45] R. Arora, R. J. Metzger, and V. E. Papaioannou, "Multiple roles and interactions of tbx4 and tbx5 in development of the respiratory system," *PLoS genetics*, vol. 8, no. 8, p. e1002866, 2012.

[46] F. Coscia, A. Taler-Verčič, V. T. Chang, L. Sinn, F. J. O'Reilly, T. Izoré, M. Renko, I. Berger, J. Rappsilber, D. Turk *et al.*, "The structure of human thyroglobulin," *Nature*, vol. 578, no. 7796, pp. 627–630, 2020.

[47] C. E. Citterio, H. M. Targovnik, and P. Arvan, "The role of thyroglobulin in thyroid hormonogenesis," *Nature Reviews Endocrinology*, vol. 15, no. 6, pp. 323–338, 2019.

[48] B. Di Jeso and P. Arvan, "Thyroglobulin from molecular and cellular biology to clinical endocrinology," *Endocrine reviews*, vol. 37, no. 1, pp. 2–36, 2016.

[49] R. R. Kakun, Z. Melamed, and R. Perets, "Pax8 in the junction between development and tumorigenesis," *International Journal of Molecular Sciences*, vol. 23, no. 13, p. 7410, 2022.

[50] P. Gokulnath, A. A. Soriano, T. de Cristofaro, T. Di Palma, and M. Zannini, "Pax8, an emerging player in ovarian cancer," *Ovarian Cancer: Molecular & Diagnostic Imaging and Treatment Strategies*, pp. 95–112, 2021.

[51] D. Chaves-Moreira, M. A. Mitchell, C. Arruza, P. Rawat, S. Sidoli, R. Nameki, J. Reddy, R. I. Corona, L. K. Afeyan, I. A. Klein *et al.*, "The transcription factor pax8 promotes angiogenesis in ovarian cancer through interaction with sox17," *Science signaling*, vol. 15, no. 728, p. eabm2496, 2022.

[52] S. Lu, E. Yakirevich, J. Hart, L. Wang, and Y. Wang, "Pax8 expression in breast cancer," *Applied Immunohistochemistry & Molecular Morphology*, vol. 29, no. 4, pp. 293–298, 2021.

[53] M. Schicht, F. Rausch, S. Finotto, M. Mathews, A. Mattil, M. Schubert, B. Koch, M. Traxdorf, C. Bohr, D. Worlitzsch *et al.*, "Sfta3, a novel protein of the lung: three-dimensional structure, characterisation and immune activation," *European Respiratory Journal*, vol. 44, no. 2, pp. 447–456, 2014.