Construção e Avaliação de Modelos de Aprendizado de Máquina para a Classificação de Mutações Missenses Associadas ao Câncer: Abordagem Utilizando Redes de Interação de Resíduos (RING)

Anderson Leite Camilo Dias

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2023

Anderson Leite Camilo Dias

Construção e Avaliação de Modelos de Aprendizado de Máquina para a Classificação de Mutações Missenses Associadas ao Câncer: Abordagem Utilizando Redes de Interação de Resíduos (RING)

Monografia apresentada ao curso Ciências da Computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em
Ciências da Computação

Orientadora: Prof. Dra. Thais Gaudêncio do Rêgo

Dezembro de 2023

Catalogação na publicação
Seção de Catalogação e Classificação

D541c Dias, Anderson Leite Camilo.
Construção e avaliação de modelos de aprendizado de máquina para a classificação de mutações missenses associadas ao câncer: abordagem utilizando redes de interação de resíduos (RING) / Anderson Leite Camilo Dias. - João Pessoa, 2023.
24 f. : il.

Orientação: Thaís Guadêncio do Rêgo.
TCC (Graduação) - UFPB/CI.

1. Redes de interação de resíduos. 2. Aprendizagem de máquina. 3. Modelo preditivo. 4. Mutações missenses. I. Rêgo, Thaís Guadêncio do. II. Título.

UFPB/CI CDU 004.8

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

Agradecimentos

Gostaria de expressar minha profunda gratidão às professoras Thaís Gaudêncio e Daniela Coelho pelo apoio durante todo o desenvolvimento deste estudo. Agradeço também ao meu colega de estudo, João Alcoforado, cuja colaboração foi inestimável. Não posso deixar de reconhecer o apoio fundamental da minha família, em especial ao meu tio Marcos e Sávia, sem os quais esta jornada acadêmica não seria possível. Um agradecimento especial também vai para Breno, meu companheiro de vida, cujo constante estímulo me motivou a buscar sempre a minha melhor versão. Agradeço, ainda, aos colegas e amigos que compartilharam comigo cada passo dessa trajetória, incluindo Jorge, Arnor e Dandara, entre outros, aos quais serei eternamente grato.

Resumo

O diagnóstico precoce desempenha um papel crucial no tratamento eficaz do câncer, motivando a busca por métodos aprimorados de identificação de mutações genéticas associadas a essa patologia. Este estudo concentra-se na análise das Redes de Interação de Resíduos (RINGs), uma abordagem alternativa de visualização da proteína, como uma fonte promissora de informações complementares para a detecção de mutações *missenses* ligadas a diferentes variantes de câncer. O uso de técnicas de aprendizado de máquina será empregado para avaliar o potencial das RINGs na predição dessas mutações. Observou-se que essas informações desempenham um papel crucial na identificação de mutações genuinamente prejudiciais em modelos de aprendizado de máquina individualizados.

Palavras-chave: Redes de interação de resíduos, Modelo Preditivo, Aprendizagem de Máquina

Abstract

The early diagnosis plays a crucial role in the effective treatment of cancer, driving the quest for enhanced methods of identifying genetic mutations associated with this pathology. This study focuses on the analysis of Residue Interaction Networks (RINGs), an alternative approach for visualizing proteins, as a promising source of supplementary information for the detection of missense mutations linked to various cancer variations. Machine learning techniques will be employed to assess the potential of RINGs in predicting these mutations. It has been observed that such information is vital for identifying genuinely harmful mutations in individualized machine learning models.

Key-words: Residues Interaction Networks, Predictive Model, Machine Learning

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

RING - Redes de Interação de Resíduos

BLCA - *Bladder Urothelial Carcinoma* (Carcinoma Urotelial da Bexiga)

CESC - C*ervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma* (Carcinoma de Células Escamosas Cervical e Adenocarcinoma Endocervical)

STAD - *Stomach Adenocarcinoma* (Adenocarcinoma de Estômago)

BRCA - *Breast Invasive Carcinoma* (Carcinoma Invasivo de Mama)

COAD - *Colon Adenocarcinoma* (Adenocarcinoma de Cólon)

ESCA - *Esophageal Carcinoma* (Carcinoma Esofágico)

HNSC - *Head and Neck Squamous Cell Carcinoma* (Carcinoma de Células Escamosas de Cabeça e Pescoço)

LIHC - *Liver Hepatocellular Carcinoma* (Carcinoma Hepatocelular do Fígado)

PAAD - *Pancreatic Adenocarcinoma* (Adenocarcinoma Pancreático)

READ - *Rectum Adenocarcinoma* (Adenocarcinoma de Reto)

UCEC - *Uterine Corpus Endometrial Carcinoma* (Carcinoma Endometrial do Corpo Uterino)

LUAD - *Lung Adenocarcinoma* (Adenocarcinoma Pulmonar)

OV - *Ovarian Serous Cystadenocarcinoma* (Cistadenocarcinoma Seroso Ovariano)

UCS - *Uterine Carcinosarcoma* (Carcinossarcoma Uterino)

SKCM - *Skin Cutaneous Melanoma* (Melanoma Cutâneo da Pele)

GBM - *Glioblastoma Multiforme* (Glioblastoma Multiforme)

LAML - *Acute Myeloid Leukemia* (Leucemia Mielóide Aguda)

PRAD - *Prostate Adenocarcinoma* (Adenocarcinoma de Próstata)

KIRC - *Kidney Renal Clear Cell Carcinoma* (Carcinoma Renal de Células Claras)

KIRP - *Ovarian Serous Cystadenocarcinoma* (Carcinoma de Células Papilares Renais)

LGG - *Brain Lower Grade Glioma* (Glioma Cerebral de Grau Inferior)

LUSC - *Lung Squamous Cell Carcinoma* (Carcinoma de Células Escamosas do Pulmão)

MESO - *Mesothelioma* (Mesotelioma)

PCPG - *Pheochromocytoma and Paraganglioma* (Feocromocitoma e Paraganglioma)

SARC- *Sarcoma*

TGCT- *Testicular Germ Cell Tumors* (Tumores de Células Germinativas Testiculares)

SUMÁRIO

1. Introdução

Na contemporaneidade, a obtenção de diagnósticos precoces é de extrema importância, particularmente no câncer. A rapidez na identificação de indícios de uma condição médica crítica não apenas agiliza o início de tratamentos, mas também melhora substancialmente a qualidade de vida dos pacientes. Um campo de pesquisa que tem despertado grande interesse é o estudo das mutações *missenses*, uma vez que se acredita que essas variações genéticas estejam intimamente relacionadas à suscetibilidade ao câncer [Malhotra et al. 2019], [Petrosino et al. 2021].

Mutação *missenses* é o termo utilizado, quando há uma alteração de uma das bases do DNA, de tal forma que o tripleto de nucleótidos da qual ela faz parte se altera, passando a codificar um aminoácido incorreto (diferente do que seria esperado na posição correspondente da proteína). Isto pode alterar a função da proteína em maior ou menor grau, dependendo da localização e da importância desse aminoácido em particular [Zhang et al. 2020]. A análise e interpretação de dados genéticos, juntamente com o uso de técnicas de aprendizagem de máquina, tornaram-se ferramentas cruciais para a construção de modelos preditivos, que podem contribuir significativamente para o avanço da ciência e da medicina.

O desafio de desenvolver métodos mais eficazes para identificar mutações *missenses*, associadas ao câncer, motiva grande parte das pesquisas na área. Alguns estudos já exploram estratégias que utilizam informações como frequência alélica [Sobahy et al. 2022] e energia de Gibbs [Nguyen et al. 2021] para refinar a análise da correlação entre mutações e câncer. No entanto, uma vertente de investigação para obtenção mais detalhada de informações a respeito dessas mutações é a análise das Redes de Interação de Resíduos (RINGs), como citado no estudo de [Gyulkhandanyan et al. 2020]. As redes de interação de resíduos (RINGs) são uma representação das estruturas tridimensionais de proteínas, onde os resíduos de aminoácidos são nós e as interações fisico-químicas são arestas em um grafo [Fonseca 2020].

Nesse sentido, o objetivo principal deste estudo é avaliar o comportamento das RINGs e sua aplicação em modelos de aprendizagem de máquina, a fim de determinar se essas redes apresentam relevância e utilidade na predição de mutações *missenses* associadas ao câncer. Serão desenvolvidos modelos de aprendizagem individuais para avaliar seu desempenho ao serem treinados e testados com os conjuntos de dados para o mesmo tipo de câncer. Além disso, serão criados modelos mais generalistas, onde dados de um tipo de câncer será utilizado como conjunto de teste, enquanto os dados dos demais servirão como conjunto de treinamento. Essas análises serão conduzidas tanto com a inclusão, quanto sem a inclusão de dados de RINGs, a fim de determinar o impacto dessas informações na precisão dos modelos.

2. Trabalhos Relacionados

A criação de modelos de aprendizagem de máquina, destinados a prever mutações *missenses* em proteínas associadas ao câncer, tem um potencial significativo para desempenhar um papel crucial no diagnóstico precoce, tratamento específico e, inclusive, na prevenção de doenças decorrentes dessas mutações. Para esta pesquisa, foram escolhidos três estudos correlatos com o propósito de analisar sistematicamente suas fases, resultados e objetivos.

O primeiro estudo, conduzido por [Nguyen et al. 2021], teve como objetivo a previsão quantitativa da energia livre de Gibbs dos efeitos das mutações em pontos únicos e múltiplos. Para essa análise, utilizou-se um conjunto de dados proveniente das fontes: ProNIT 2.0, dbAMERPNI e PrPDH, que consideram interações termodinâmicas entre proteínas e ácidos nucleicos.

No total, o conjunto de dados englobou 856 mutações em pontos únicos e 141 mutações em pontos múltiplos. Embora o estudo não tenha especificado a divisão entre os dados utilizados para teste e treinamento, nem mencionado os atributos utilizados, ele destacou que os modelos foram treinados utilizando validação cruzada com 5, 10 e 20 agrupamentos, um procedimento que aumenta a robustez das análises.

Diversos algoritmos de aprendizado de máquina foram empregados na geração dos modelos, incluindo Gradient Boosting, Extreme Gradient Boosting, Random Forest, Extremely Randomized Trees, AdaBoost, *K*-Nearest Neighbour, Support Vector Regressor e Gaussian Processes.

O critério de avaliação principal baseou-se no coeficiente de correlação de Pearson (PCC) e na raiz do erro quadrático médio (RMSE) no conjunto de treinamento. O modelo que demonstrou o melhor desempenho, com base nesses critérios, foi o de Extremely Randomized Trees, com os valores para Pearson, Spearman, Kendall e RMSE de 0,67, 0,65, 0,47 e 1,06 kcal/mol, respectivamente.

O segundo estudo, conduzido por [Sobahy et al. 2022], tem como objetivo desenvolver uma ferramenta capaz de prever as mutações de nucleotídeos únicos (SNPs), que levam ao surgimento de doenças, utilizando informações de frequências alélicas como parte do processo (AllelePred).

Para o conjunto de dados foi utilizada, inicialmente, a base de dados fornecida no portal SHGP (https://shgp.sa/index.en.html), que contém 168.945 variantes genéticas da população saudita. Para retirar as variantes genéticas criadas artificialmente que havia na base, foi utilizado o banco GnomAD e como resultado dessa etapa sobraram 100.507 amostras. Foram retiradas também as mutações de inserção e deleção, no que resultou em 56.142 variantes. Por fim, foram filtrados os dados com o ClinVar e restaram 6.290 variantes genéticas para o estudo.

Para treinamento do modelo, foram utilizados 70% dos dados selecionados de forma aleatória e para os testes foram os 30% restantes. O algoritmo Random Forest foi utilizado no conjunto de treinamento aplicando validação cruzada com 5 grupos.

Os atributos empregados no treinamento consistiram nos resultados de nove preditores individuais já existentes, nomeadamente: CADD, ExAC_pLI, M-CAP, MetaSVM, MutationTaster, Polyphen2_HDIV, Polyphen2_HVAR, REVEL e SIFT. Além disso, as

frequências alélicas (AFs) também foram incluídas no conjunto de atributos utilizados no processo de treinamento.

Em comparação com outros preditores, que não consideram as frequências alélicas como uma informação relevante, o modelo AllelePred apresentou os resultados: Acurácia (98%), Precisão (96%), Média-F1 (93%), e Cobertura (100%).

O terceiro estudo, conduzido por [Tong et al. 2022], teve como objetivo desenvolver uma ferramenta de previsão de patogenicidade para variantes missense (mvPPT), baseada na técnica de Gradient Boosting.

Como conjunto de dados foram utilizados os seguintes bancos de dados: ClinVar (2020.7), HGMD (versão Pro 2020.3) e UniProt (2020.6), bem como um banco de dados populacional de genomas do Genome Aggregation Database (gnomAD). Não foi explicitada a quantidade total de dados obtidos após os filtros e validações dos bancos de dados utilizados.

Para o treinamento do modelo, foram avaliados os desempenhos de dez algoritmos, incluindo Suport Vector Machine (SVM), Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Extra Forest, Gradient Boosting Machine (GBM), AdaBoost, LightGBM e Bagging. Desses, o de melhor desempenho, de acordo com as métricas: área sob a curva característica de operação do receptor (AUROC) e a área sob a curva de recuperação de precisão (AUPRC), foi o LightGBM, com o maior AUROC (0,970 ± 0,001) e AUPRC (0,952 ± 0,002).

No que diz respeito aos atributos utilizados no mvPPT, eles foram divididos em três categorias distintas:

1. **Avaliação por Ferramentas de Previsão**: Essa categoria incluiu a avaliação por meio de diversas ferramentas de previsão já existentes, tais como Sorting Tolerant From Intolerant (SIFT), MutationAssessor, Protein Variation Effect Analyzer (PROVEAN), GERP++RS, bem como análises PHYlogenéticas específicas do SIte (SiPhy), phyloP e phastCons.
2. **Frequências**: Esta categoria contemplou informações relacionadas às frequências, abrangendo as de alelos, de aminoácidos e genotípicas.
3. **Contexto Genômico**: Contexto genômico da variante, ou seja, informações baseadas em região/gene do Gene Variation Intolerance Rank (GeVIR), VIRLoF, oe_mis_upper (de gnomAD), Previsões de Haploinsuficiência (HIP), Regiões de Codificação Restritas (CCRs), domínio Interpro e sequências de aminoácidos, antes e depois da mutação.

As métricas utilizadas para realizar a comparação da ferramenta do estudo, com outros modelos existentes, foram as seguintes: AUROC, AUPRC, acurácia, precisão, sensibilidade, medida-F1, o log da perda, MCC, especificidade, taxa de falso positivo e Razão de Odds Diagnósticas.

Com base nos indicadores mencionados, foi verificado que o modelo de estudo demonstra uma clara superioridade em comparação aos outros existentes. Por exemplo, alcançou valores de 96%, 71,9%, 92,2%, 32,3%, 95,3% e 47,3% para AUROC, AUPRC, Acurácia, Precisão, sensibilidade e medida-F1, respectivamente.

Nessa perspectiva, é evidente que o avanço no desenvolvimento de preditores, vol-

Autor	Objetivo	Algoritmos	Dataset	Métricas
NGUYEN *et al.* (2021)	Previsão quantitativa (energia livre de Gibbs) dos efeitos das mutações em pontos únicos e múltiplos.	Gradient Boosting, Extreme Gradient Boosting, Random Forest, Extremely Randomized Trees, AdaBoost, K-Nearest Neighbour, Support Vector Regressor e Gaussian Processes	ProNIT 2.0, dbAMERPNI e PrPDH	Coeficiente de correlação de Pearson (PCC) e na raiz do erro quadrático médio (RMSE), Correlações de Spearman e Kendall, precisão, F1 e o coeficiente de correlação de Matthews (MCC).
SOBAHY, MOTWALLI e ALAZMI (2022)	Desenvolver uma ferramenta capaz de prever as variações de nucleotídeos únicos (SNVs) prejudiciais, utilizando informações de frequências alélicas como parte do processo (AllelePred).	Random Forest	SHGP, GnomAD, ClinVar	Precisão, Taxa de Verdadeiros Positivo, Média-F!, Acurácia, Cobertura
TONG *et al.* (2022)	Desenvolver uma ferramenta de previsão de patogenicidade para variantes missense (mvPPT) baseada na técnica de Gradient Boosting	SVM, naive Bayes, logistic regression, decision tree, random forest, extra forest, gradient boosting machine (GBM), AdaBoost, LightGBM e bagging	ClinVar (2020.7), HGMD (versão Pro 2020.3) , UniProt (2020.6), Genome Aggregation Database (gnomAD)	Área Sob a Curva ROC, Área Sob a Curva Precisão-Revocação, Acurácia, Precisão,Taxa de Verdadeiros Positivos (Recall), Média-F1, Pontuação, Perda logarítmica , Coeficiente de Correlação de Matthews , Taxa

Figura 1. Informações sobre algoritmos, métricas, objetivos, e datasets utilizados nos trabalhos correlatos selecionados.

tados para a compreensão das mutações missenses e sua correlação com o surgimento de doenças, representa um campo de pesquisa expansivo e em constante evolução. Há, sem dúvida, espaço para contribuições adicionais que possam aprimorar ainda mais as áreas relacionadas a diagnósticos precoces e medicina de precisão. Com base na análise dessas investigações, foram identificados algoritmos, conjuntos de dados e métricas empregados nos estudos mais recentes (conforme mostrado na Figura 1). Essas informações são de suma importância, pois orientarão as decisões críticas no desenvolvimento deste estudo.

É relevante observar que, embora os estudos mencionados tenham produzido resultados promissores, é fundamental reconhecer as limitações associadas à ausência de informações relacionadas às Redes de Interação de Resíduos (RING). Essas redes podem fornecer mais informações sobre o impacto das mutações nas proteínas e têm o potencial de aprimorar ainda mais a criação de modelos, tornando-os mais abrangentes e informativos. Portanto, a inclusão de dados sobre as RING pode enriquecer significativamente as análises e descobertas, aprimorando a compreensão das complexas interações nas proteínas.

É crucial ressaltar que a estratégia empregada neste estudo se diferencia ao estabelecer um conjunto composto modelos individualistas e generalistas. Dentro desse conjunto, 26 serão treinados de maneira mais generalista, detalhados posteriormente, enquanto os outros quatro serão mais especializados, um para cada tipo de câncer. Adicionalmente, este estudo busca incorporar informações das redes de interação de resíduos (RINGs), com o objetivo de avaliar o potencial dessa integração nos modelos.

3. Metodologia

3.1. Base de Dados

Os dados empregados nesta pesquisa foram extraídos da base de dados usada no estudo intitulado **"Classificação de Mutações Associadas ao Câncer por Meio de Algoritmos de Aprendizagem de Máquina"** [Pereira 2020]. Esta base compreende 334 instâncias relacionadas a 26 tipos de câncer: BLCA (*Bladder Urothelial Carcinoma*), CESC (*Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma*), STAD (*Stomach Adenocarcinoma*), BRCA (*Breast Invasive Carcinoma*), COAD (*Colon Adenocarcinoma*), ESCA (*Esophageal Carcinoma*), HNSC (*Head and Neck Squamous Cell Carcinoma*), LIHC (*Liver Hepatocellular Carcinoma*), PAAD (*Pancreatic Adenocarcinoma*), READ (*Rectum Adenocarcinoma*), UCEC (*Uterine Corpus Endometrial Carcinoma*), LUAD (*Lung Adenocarcinoma*), OV (*Ovarian Serous Cystadenocarcinoma*), UCS (*Uterine Carcinosarcoma*), SKCM (*Skin Cutaneous Melanoma*), GBM (*Glioblastoma Multiforme*), LAML (*Acute Myeloid Leukemia*), PRAD (*Prostate Adenocarcinoma*), KIRC (*Kidney Renal Clear Cell Carcinoma*), KIRP (*Kidney Renal Papillary Cell Carcinoma*), LGG (*Brain Lower Grade Glioma*), LUSC (*Lung Squamous Cell Carcinoma*), MESO (*Mesothelioma*), PCPG (*Pheochromocytoma and Paraganglioma*), SARC (*Sarcoma*), TGCT (*Testicular Germ Cell Tumors*). Importante notar que esta base contém apenas mutações com diagnósticos válidos registrados no ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/).

A base de dados é composta por um total de 31 atributos. Esses atributos englobam informações relacionadas às características de mutações missenses nos tecidos selecionados, bem como características físico-químicas, o nome do tecido (câncer) e informações das RINGs como: Degree_RING, Bfactor_CA_RING, Inter_Lig_tot, Inter_Res_tot, Inter_IAC_Lig_tot, Inter_VDW_Lig_tot, Inter_VDW_Res_tot, Inter_HBOND_Res_tot, Inter_PIPISTACK_Res_tot, Inter_IONIC_Res_tot, Inter_PICATION_Res_tot, triangles_node, clusteringCoef_node e betweennessWeighted_node. Além disso, o rótulo que será avaliado na abordagem de aprendizado de máquina também esta presente na base de dados

Embora a quantidade de dados atualmente disponíveis seja relativamente limitada, optou-se por não utilizar o conceito de *data augmentation*. Isso se deve ao fato de que a aplicação dessa técnica, em contextos de saúde, pode introduzir um viés algorítmico significativo, como discutido por [Dakshit et al. 2023]. Além disso, essa abordagem pode resultar em modelos de aprendizado de máquina potencialmente tendenciosos ou injustos em suas previsões, conforme destacado por [Norori et al. 2021]. Dessa forma, foi escolhida uma abordagem que prioriza a integridade e a ética dos dados na área da saúde.

3.2. Pré-processamento

A fase de pré-processamento foi iniciada com a verificação do rótulo a ser utilizado na abordagem de aprendizado de máquina. O rótulo escolhido foi o atributo "Deleteria", o qual recebe o valor 1, quando a mutação é considerada deletéria e 0, quando a mutação não é deletéria. A Tabela 1 apresenta os atributos após essa fase inicial.

Tabela 1. Descrição dos Atributos

Atributo	**Descrição**
CHROM	Cromossomo no qual a variante genética está localizada.
REF	Alelo de referência, representando a sequência de DNA esperada.
ALT	Alelo alternativo, indicando a variação observada em relação ao alelo de referência.
VariantEffect_EFF	Efeito da variante, descrevendo como a mutação afeta a função ou estrutura da proteína.
Gene_EFF	Gene afetado pela variante genética.
Exon_EFF	Éxon no qual a variante está localizada.
Pos_Point_Mutation_EFF	Posição específica da mutação.
poschangecDNA_EFF	Mudança na sequência de cDNA devido à mutação.
aminBefore	Aminoácido na posição anterior à mutação.
aminAfter	Aminoácido na posição posterior à mutação.
poschangeProt	Mudança na posição da proteína devido à mutação.
Deleteria	Indicação se a mutação é deletéria para a função da proteína.
Blosum62	Pontuação Blosum62, indicando a similaridade entre os aminoácidos afetados pela mutação.
groupChange	Mudança no grupo funcional da proteína devido à mutação.
essencialChange	Indicação se a mutação afeta uma região essencial da proteína.
substitution	Tipo de substituição genética (por exemplo, transição, transversão).
Degree_RING	Grau de conectividade na Rede de Interação de Resíduos.
Bfactor_CA_RING	Fator B na cadeia alfa da proteína na Rede de Interação de Resíduos.
Inter_Lig_tot	Número total de interações com ligantes na Rede de Interação de Resíduos.

Atributo	Descrição
Inter_Res_tot	Número total de interações entre resíduos na Rede de Interação de Resíduos.
Inter_IAC_Lig_tot	Número total de interações átomo-átomo entre ligantes na Rede de Interação de Resíduos.
Inter_VDW_Lig_tot	Número total de interações de Van der Waals entre ligantes na Rede de Interação de Resíduos.
Inter_VDW_Res_tot	Número total de interações de Van der Waals entre resíduos na Rede de Interação de Resíduos.
Inter_HBOND_Res_tot	Número total de interações de ligação de hidrogênio entre resíduos na Rede de Interação de Resíduos.
Inter_PIPISTACK_Res_tot	Número total de interações de empilhamento de anéis entre resíduos na Rede de Interação de Resíduos.
Tecido	Tipo de tecido associado à amostra ou à mutação.

3.3. Separação das bases de dados

Nesta etapa, será realizada a separação das bases, de acordo com o tipo de câncer (tecido), com o objetivo de criar um modelo que seja treinado e testado com o mesmo tipo de câncer. Essa abordagem visa avaliar a precisão da previsão direcionada. Além disso, será realizada uma divisão de dados em X - 1, onde um câncer será utilizado como conjunto de teste, enquanto os demais servirão como conjunto de treinamento. Isso permite avaliar a capacidade do modelo em identificar novos casos de câncer.

É importante destacar que, embora esteja sendo realizada a segmentação das bases de dados para formular hipóteses e obter resultados específicos, todos os registros e atributos serão tratados de forma equitativa e consistente.

3.3.1. Separação das bases de acordo com o tipo de tecido

Primeiramente, foi avaliado o número de tecidos presentes na base e a quantidade de registros associados a cada um deles. Nessa análise, foi identificado um total de 26 tecidos, conforme mencionado na Seção 2.1. No entanto, foi observado um baixo número de instâncias para a grande maioria dos tecidos, conforme apresentado na Tabela 2. Diante disso, foram selecionados 4 tecidos que possuíam um número de registros superior a 20 e 4 bases correspondentes foram criadas para cada um deles, assim como mostrado na Figura 2.

Tabela 2. Quantidade de registros por tipo de tecido

Tipo Tecido	Quantidade
UCEC	94
COAD	40
STAD	31
SKCM	26
CESC	13
KIRC	13
LUAD	12
GBM	11
READ	10
HNSC	9
LUSC	9
PRAD	8
OV	8
ESCA	8
BRCA	7
LIHC	6
PAAD	6
LGG	5
SARC	4
BLCA	4
PCPG	3
UCS	3
MESO	1
KIRP	1
LAML	1
TGCT	1

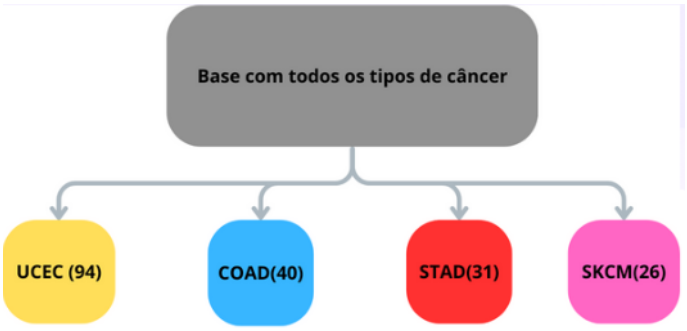

Figura 2. Divisão de bases para modelos individuais

3.3.2. Separação das bases para criação de modelos generalistas

Nesta fase, será realizada a separação dos conjuntos de dados em X - 1 partes, em que o conjunto de treinamento abarcará dados de 25 tipos distintos de câncer e o tipo deixado de fora dessa amostra será utilizado para teste. Nesse caso, foram geradas 26 bases no geral, onde cada uma contém 25 tipos de câncer distintos.

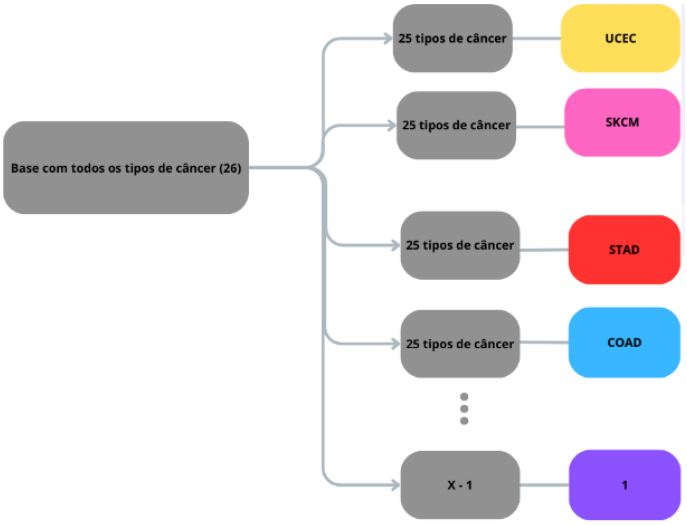

Figura 3. Divisão de bases para modelos generalista

3.3.3. Separação para treinamento e teste

Para os modelos individuais, foi empregada a função `train_test_split` da biblioteca Scikit-Learn, que efetuou a divisão alocando 80% dos dados para o treinamento do modelo e reservando os 20% restantes para a avaliação do desempenho do modelo. Foi realizada essa divisão nas bases individuais citadas na seção 3.3.1. Importante ressaltar que foi utilizada a divisão estratificada para manter a mesma proporção entre as classes deletéria (1) e não deletéria (0).

A separação entre os conjuntos de treinamento e teste para os modelos generalistas

adotou a proporção X-1, conforme descrito na seção 3.3.2 deste estudo. Nesse contexto, o treinamento será conduzido utilizando os 25 tipos de câncer, enquanto o restante será reservado para o conjunto de teste.

Ambos os tipos de modelos foram utilizaram a técnica de validação cruzada estratificada. Essa abordagem envolve a divisão do conjunto de dados em partes (folds), permitindo o treinamento do modelo em algumas partes e a avaliação em outras.

3.4. Tratamento dos atributos não numéricos

Na base de dados, dos 31 atributos disponíveis, 8 deles são classificados como atributos categóricos. No entanto, a maioria dos algoritmos de aprendizado de máquina é projetada para trabalhar com dados numéricos. Portanto, é necessário aplicar uma técnica de codificação para converter esses atributos categóricos em formato numérico.

Para este estudo, foram utilizadas duas técnicas de codificação: o *One-Hot Encoder* e o *Frequency Encoder*. O *One-Hot Encoder* é uma técnica adequada para variáveis categóricas com um número limitado de valores distintos (baixa cardinalidade) [Johnson and Khoshgoftaar 2021]. Ele converte cada categoria única em uma nova coluna binária (0 ou 1), criando assim uma matriz de zeros e uns. Cada coluna representa uma categoria e indica se a observação pertence a essa categoria. Já o *Frequency Encoder* é uma técnica útil quando se lida com variáveis categóricas de alta cardinalidade [Pargent et al. 2022], ou seja, com um grande número de categorias distintas. Em vez de criar uma nova coluna binária para cada categoria, o Frequency Encoder substitui cada categoria pelo número de vezes que ela aparece no conjunto de dados (a frequência).

Nesse contexto, os atributos 'Gene_EFF', 'aminBefore', 'aminAfter' e 'groupChange', devido à sua alta cardinalidade, serão codificados em dados numéricos utilizando a técnica conhecida como *Frequency Encoder*. A execução dessa codificação será efetuada com o auxílio da biblioteca category_encoder, que oferece uma variedade de técnicas de codificação, incluindo a *CountEncoder*, que calcula as frequências das categorias, simplificando assim a transformação de dados categóricos em numéricos. Em contrapartida, os atributos 'REF', 'ALT', 'VariantEffect_EFF' e 'essencialChange' exibem uma baixa cardinalidade e, portanto, serão codificados em numéricos, utilizando a técnica do *One-Hot Encoder*.

Após a conversão dos atributos categóricos em valores numéricos, é fundamental garantir que esses atributos numéricos estejam em escalas comparáveis para garantir o desempenho eficaz dos algoritmos. Para atingir esse objetivo, foi empregada a técnica de normalização Min-Max. Essa etapa foi realizada por meio da função `minmax_scale` da biblioteca Scikit-Learn (sklearn) em Python. A normalização Min-Max ajusta os valores para um intervalo específico, normalmente [0, 1].

3.5. Seleção de atributos mais relevantes

Para minimizar o esforço necessário na aquisição de atributos e reduzir a demanda de processamento, é de suma importância a identificação criteriosa dos atributos mais relevantes para a construção de um modelo [Zebari et al. 2020].

Dessa forma, para a identificação dos atributos mais adequados na criação de um modelo de qualidade, foram utilizados os métodos **MRMR** *(Maximum Relevance Minimum Redundancy)* e **XGBoost** *(Extreme Gradient Boosting)*. O MRMR é uma técnica

Tabela 3. Atributos selecionados por MRMR, com RING

MRMR
Gene_EFF_encoded
REF_G
Inter_VDW_Res_tot
essencialChange_1TO1
triangles_node
Bfactor_CA_RING
Blosum62
Inter_Res_tot
CHROM
Degree_RING
ALT_T
VariantEffect_EFF_NON_SYNONYMOUS_CODING+SPLICE_SITE_REGION
betweennessWeighted_node
REF_T
clusteringCoef_node
REF_C
Inter_PICATION_Res_tot
ALT_G

voltada para a seleção de características com elevada relevância em relação à variável de saída, simultaneamente reduzindo a redundância entre elas. Esse método avalia a relevância e redundância de cada atributo, optando pela seleção daqueles que maximizam a primeira e minimizam a segunda [Bommert et al. 2020]. A biblioteca *mrmr* disponibiliza essa abordagem por meio da função `mrmr_classif`. Por outro lado, o XGBoost é um algoritmo de aprendizado de máquina baseado em árvores de decisão que oferece uma função de importância de características integradas. Essa função atribui pontuações de importância a cada atributo, com base em suas contribuições para a construção das árvores [Zhang et al. 2018]. Para utilizar essa funcionalidade, é necessário importar a classe XGBClassifier do módulo XGBoost.

Para o MRMR foi utilizado a biblioteca “mrmr”na versão 0.2.8. Já o XGBoost, foi utilizado a biblioteca “xgboost”‘na versão 2.0.0.

Tabela 4. Atributos selecionados por XGBoost, com RING

XGBoost
Características
Gene_EFF_encoded
Degree_RING
poschangecDNA_EFF
Bfactor_CA_RING
CHROM
triangles_node
aminAfter_encoded
Pos_Point_Mutation_EFF
Blosum62
betweennessWeighted_node
Inter_VDW_Res_tot
clusteringCoef_node
REF_G
Inter_Res_tot
Exon_EFF
Inter_HBOND_Res_tot
substitution
Inter_VDW_Lig_tot

3.6. Construção dos modelos de aprendizagem

Após o processo de pré-processamento de dados e a divisão da base em conjuntos de treinamento e teste, a fase seguinte compreendeu a geração de modelos preditivos. Nesse contexto, foram desenvolvidos quatro modelos individuais e vinte e seis modelos generalistas, com o propósito de classificar mutações em duas categorias: deletérias (associadas ao câncer) e não deletérias, como descrito nas Seções 3.3.1 e 3.3.2.

Para este estudo, foi empregada uma seleção diversificada de algoritmos de aprendizado supervisionado, incluindo o SVM, Random Forest, AdaBoost, Light GBM, XGBoost e HistGradientBoostingClassifier. A biblioteca e os hiperparâmetros utilizados estão na Tabela 5. Com intuito de aumentar a robustez e a capacidade de generalização dos modelos de aprendizado de máquina, foi utilizada a técnica de validação cruzada estratificada com 5 subconjuntos, ou seja $k = 5$. É válido ressaltar também, que para todos os algoritmos, foi utilizada a classe RandomizedSearchCV da biblioteca sikit-learn para identificação dos melhores hiperparâmetros.

Tabela 5. Hiperparâmetros e bibliotecas utilizadas.

Nome	Biblioteca/Framework	Hiperparâmetro
Support Vector Machines (SVM)	**SVC()** da biblioteca **scikit-learn**	C = uniform(loc=0, scale=10), kernel= ['linear', 'rbf'] e gamma= ['scale', 'auto']
Random Forest	**RandomForestClassifier()** da biblioteca **scikit-learn**	max_depth=9, max_features=None, max_leaf_nodes=9 e n_estimators=50
AdaBoost	**AdaBoostClassifier()** da biblioteca **skit-learn**	base_estimator=DecisionTreeClassifier(max_depth=2), learning_rate=0.5, n_estimators=500 e random_state=42
Light GBM	**LGBMClassifier()** do framework **LightGBM**	bagging_freq=1, colsample_bytree=0.6, max_depth=7, min_data_in_leaf=1, n_estimators=300, num_leaves=128 e subsample=0.05
XGBoost	**XGBClassifier()** da biblioteca **XGBoost**	n_estimators=[100,200], max_depth=[4], learning_rate= [0.1], random_state=[seed], subsample=[1] e colsample_bytree=[1.0]
HistGradientBoostingClassifier	**HistGradientBoostingClassifier()** biblioteca **skit-learn**	early_stopping=True, l2_regularization=2.8379948315360316e-09, max_depth=7, max_leaf_nodes=21, min_samples_leaf=19, n_iter_no_change=2 e validation_fraction=None

É importante destacar que, após a geração inicial dos modelos, aquele algoritmo que demonstrou o desempenho mais eficaz em termos de aprendizado foi estabelecido como o algoritmo de referência. Este algoritmo foi então empregado na obtenção dos resultados desta pesquisa, e a justificativa para a sua seleção será abordada na seção de resultados e discussões deste estudo. Ter um padrão de algoritmo para a geração dos modelos

de aprendizado de máquina, principalmente na área da saúde, é de suma importância, pois permite a avaliação consistente e comparativa dos resultados, garantindo a confiabilidade e reprodutibilidade da pesquisa assim como mencionado por [McDermott et al. 2019].

3.7. Métricas

As métricas utilizadas para este estudo foram:

- **Acurácia (AC)**: A porcentagem de previsões corretas, em relação ao total de previsões. Uma métrica geral de desempenho do modelo.
- **Área sob a Curva de ROC (AUC)**: A Curva ROC (Receiver Operating Characteristic) é uma curva que representa a taxa de verdadeiros positivos (sensibilidade), em relação à taxa de falsos positivos (1 - especificidade) para diferentes limiares de classificação. A Área sob a Curva de ROC (ROC-AUC) é uma métrica que resume a Curva ROC em um único valor entre 0 e 1. Quanto maior o valor do ROC-AUC, melhor o desempenho do modelo. Um modelo perfeito teria um ROC-AUC igual a 1, enquanto um modelo aleatório ou totalmente inadequado teria um ROC-AUC de 0,5, pois se comportaria como uma escolha aleatória entre as classes.
- ***Recall* (ou Taxa de Verdadeiros Positivos - TPR)**: A proporção de exemplos positivos que foram corretamente classificados pelo modelo.
- ***Specificity* (ou Taxa de Verdadeiros Negativos - TNR)**: A proporção de exemplos negativos que foram corretamente classificados pelo modelo.
- **Média-F1 (*F1 Score*)**: Uma métrica que combina precisão e recall em uma única pontuação, útil quando há um desequilíbrio nas classes.
- **Valor Preditivo Negativo (NPV)**: A proporção de exemplos negativos previstos corretamente em relação ao total de previsões negativas.
- **Matriz de confusão**:É uma matriz de dimensão n x n, onde n representa o número de classes a serem previstas. Seu propósito é calcular as métricas associadas à classificação, incluindo a contagem de falsos positivos (FP), falsos negativos (FN), verdadeiros positivos (TP) e verdadeiros negativos (TN).
- **Coeficiente de Correlação de Matthews (MCC)**: Uma métrica que mede a qualidade das previsões de um modelo, levando em consideração todas as células da matriz de confusão. Pode ser usado mesmo em casos de desequilíbrio nas classes.
- **Desvio padrão (DP)**: O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados. Ou seja, o desvio padrão indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de 0 for o desvio padrão, mais homogêneo são os dados [Fonseca 2023].

Essas métricas desempenham papéis diferentes na avaliação de modelos de aprendizado de máquina e ajudam a compreender diferentes aspectos do desempenho do modelo, como a capacidade de distinguir classes, a precisão das previsões e a sensibilidade a falsos positivos e falsos negativos.

4. Resultados e discussões

4.1. Melhor modelo

Nesta seção, será discutido o processo de escolha dos modelos mais apropriados para servirem como referência neste estudo. A avaliação de desempenho baseou-se na média da acurácia (AC) obtida pelos modelos através de validação cruzada estratificada com *k=5*. Os resultados apresentados na Tabela 6 destacam esses valores médios do modelo individual para o câncer UCEC. Já na Tabela 7, são exibidos os resultados que incluem valores médios alcançados na geração do modelo generalista, abrangendo todos os tecidos, exceto UCEC.

Tabela 6. Desempenho do modelo individual para o câncer UCEC.

Algoritmo	AC %	DP
SVM	68,7	0,05
RandomForest	70,6	0,18
AdaBoost	67,6	0,19
XGBoost	75,6	0,13
LightGBM	74,0	0,10
HitsGradientBoosting	70,3	0,20

Abreviações: AC% - Acurácia; DP - Desvio padrão.

Tabela 7. Desempenho do modelo generalista sem o câncer UCEC.

Algoritmo	AC %	DP
SVM	82,1	0,06
RandomForest	84,9	0,04
AdaBoost	79,55	0,12
XGBoost	81,6	0,05
LightGBM	82,1	0,04
HitsGradientBoosting	79.5	0,07

A análise dos resultados apresentados na Tabela 6 para os modelos individuais destaca que o algoritmo de aprendizado de máquina mais eficaz, indicado pela média da acurácia nos cinco grupos da validação cruzada, foi o XGBoost. Este modelo atingiu uma acurácia de 75,6% (AC) ao utilizar os atributos selecionados pelo próprio XGBoost. Na Tabela 7, dedicada aos modelos generalistas, observamos que o algoritmo indicado

seria o Random Forest, alcançando uma acurácia de 84,9% (AC) ao empregar os atributos previamente selecionados pelo XGBoost.

Portanto, para os modelos individuais, tanto com ou sem RING, será adotado o XGBoost com os atributos obtidos pelo próprio XGBoost. Enquanto para os modelos generalistas, independentemente da presença de RING, será utilizado o Random Forest com os atributos selecionados pelo XGBoost. Esta decisão foi respaldada pelos desempenhos superiores observados, especialmente na acurácia média (AC), com foco no câncer UCEC, escolhido devido ao seu maior volume de instâncias.

4.2. Comparação dos modelos gerados

Para os modelos treinados e testatados com o mesmo tipo de câncer, criaram-se quatro modelos distintos, cada um direcionado a um tipo de câncer. Para tal, empregou-se o algoritmo de aprendizado XGBoost, conforme destacado previamente. Uma abordagem análoga foi aplicada aos modelos sem a inclusão de informações de RING.

No que diz respeito aos modelos generalistas, estabeleceu-se um conjunto de 26 modelos, todos construídos com base no algoritmo Random Forest e utilizando atributos selecionados pelo XGBoost. Essa estratégia se aplica tanto aos dados que incorporam informações de RING, quanto aos que as excluem.

Nas tabelas apresentadas em seguida, serão fornecidas duas métricas de acurácia para avaliar o desempenho do modelo proposto.

- **AC^1**: Acurácia média obtida durante o processo de validação cruzada.
- **AC^2**: Acurácia do modelo com o conjunto de teste reservados, conforme explicado na Seção 3.3.3.

As demais métricas do estudo, são referentes ao resultado do conjunto de teste.

4.2.1. Modelos individuais

Tabela 8. Desempenho dos modelos individuais com RING.

Câncer	AC^1%	DP	AC^2%	AUC%	TPR%	TNR%
UCEC	75,6	0,13	68,4	64,2	90,9	37,5
STAD	71,0	0,24	71,4	50,0	100,0	0,0
SKCM	85,0	0,12	83,3	50,0	100,0	0,0
COAD	69,3	0,21	50,0	40,0	80,0	0,0

Abreviações: AC% - Acurácia; DP - Desvio padrão; AUC% - Área sob a curva ROC; TPR% - Taxa de Verdadeiros Positivos; TNR%- Taxa de Verdadeiros Negativos.

Referente a acurácia média obtida pela validação cruzada stratificada nos modelos individuais, observa-se que o modelo que obteve maior valor foi o modelo sem RING para

Tabela 9. Desempenho dos modelos individuais sem RING.

Câncer	AC[1]%	DP	AC[2]%	AUC%	TPR%	TNR%
UCEC	67,9	0,04	42,1	44,8	27,0	62,5
STAD	88,0	0,09	71,4	50,0	100,0	0,0
SKCM	70,0	0,24	66,6	40,0	80,0	0,0
COAD	69,0	0,15	50,0	40,0	80,0	0,0

o câncer STAD com 88,0%. Além disso o modelo obteve também uma acurácia 71,4% no conjunto teste reservado. Todavia, devido à limitação e desequilíbrio nos dados foi optado por utilizar a métrica Área sob a Curva ROC (AUC) para verificação dos resultados. Essa escolha se deve à capacidade discriminatória mais robusta dessa métrica em contextos com dados desequilibrados, conforme destacado por [Halimu et al. 2019] em seus estudos.

Com isso, os resultados nas Tabelas 8 e 9 demonstram que os modelos individuais obtiveram uma maior AUC no conjunto de teste quando os dados incluíam informações das RINGs. O modelo com maior AUC foi o modelo para o câncer UCEC com uma taxa de 64,2%. O modelo equivalente sem RING obteve 44,9%.

Vale ressaltar, que uma AUC acima de 50% indica que o modelo possui poder discriminatório entre as classes do rótulo (deletéria e não deletéria) e não faz seleções aleatórias, o que ocorreria se a AUC fosse inferior a 50%. Dessa forma, os modelos sem RING obtiveram uma capacidade discriminatória não muito interessante uma vez que dos quatro modelos, três deles (UCEC, SKCM e COAD) estão abaixo de 50%.

Ao analisar as métricas TPR (Taxa de Verdadeiros Positivos) e TNR (Taxa de Verdadeiros Negativos), que mensuram a capacidade do modelo em classificar corretamente positivos e negativos, destacamos que o modelo para o tecido UCEC com a inclusão de RING atingiu uma sensibilidade elevada, apresentando um TPR de 90,9% nas amostras de teste. Esse resultado mostra uma potencial capacidade de detectar verdadeiros positivos em dados não utilizados durante o treinamento.

Por outro lado, o modelo para o mesmo tipo de câncer sem RING obteve um TPR de 27,3%. Em termos gerais, os modelos que incorporaram RING demonstraram uma capacidade interessante em classificar mutações verdadeiramente deletérias. Destaca-se o desempenho dos modelos STAD e SKCM, alcançando uma taxa de 100% (TPR), identificando todas as mutações que eram verdadeiramente deletérias.

Embora não tenham alcançado os resultados mais expressivos quando comparados aos modelos mencionados na Seção 2 deste estudo, é importante ressaltar que, dadas as limitações dos dados de treinamento, as informações provenientes das RINGs parecem ter exercido um impacto positivo na capacidade preditiva dos modelos individuais. Isso se manifesta na detecção de verdadeiros positivos, destacando a relevância das RINGs como um componente de potencial interessante.

4.2.2. Modelos generalistas

Tabela 10. Desempenho dos modelos generalistas com RING.

Descrição	AC[1]%	DP	AC[2]%	AUC%	TPR%	TNR%
Gen. sem BLCA	83,8	0,04	75,0	75,0	90,9	50,0
Gen. sem BRCA	80,9	0,02	100	100	100	100
Gen. sem CESC	81,4	0,04	92,3	83,3	100	66,6
Gen. sem COAD	82,6	0,02	75,0	67,5	88,8	46,1
Gen. sem ESCA	84,0	0,04	87,5	75,0	100	50,0
Gen. sem GBM	86,9	0,03	90,9	75,0	100	50,0
Gen. sem HNSC	84,0	0,04	77,7	67,8	85,7	67,8
Gen. sem KIRC	86,1	0,03	100	100	100	100
Gen. sem LGG	84,26	0,02	80,0	75,0	100	50
Gen. sem LIHC	83,76	0,05	100	100	100	100
Gen. sem LUAD	86,1	0,02	100	100	100	100
Gen. sem LUSC	87,2	0,04	88,8	83,3	100	66,6
Gen. sem OV	87,2	0,04	75,0	80,0	100	60,0
Gen. sem PAAD	84,7	0,02	83,3	87,5	100	75,0
Gen. sem PRAD	84,1	0,02	75,0	73,3	80,0	66,6
Gen. sem SKCM	84,2	0,06	70,4	70,4	90,9	50
Gen. sem STAD	83,8	0,04	70,9	67,2	80,0	54,5
Gen. sem UCEC	84,9	0,04	74,4	72,8	88,4	57,1

Abreviações: AC% - Acurácia; DP - Desvio padrão; AUC% - Área sob a curva ROC; TPR% - Taxa de Verdadeiros Positivos; TNR%- Taxa de Verdadeiros Negativos.

Nas Tabelas 10 e 11, são apresentados os resultados dos modelos generalistas com e sem RINGs. O modelo generalista mais preciso, avaliado pela acurácia média dos subconjuntos obtidos pela validação cruzada, destacou-se nos modelos testados com os cânceres LUSC e OV com RING, atingindo ambos a marca de 87,2%. A taxa de desvio padrão baixa sugere que as acurácias dos subconjuntos são relativamentes consistentes e não variam muito em relação à média [Fonseca 2023].

Em contraste, os modelos análogos sem dados RINGs obteve uma acurácia média de 81,9% e 82,1% para os modelos testados com os cânceres LUSC e OV. De maneira geral, para os modelos generalistas, a acurácia média dos modelos com RING obtida na

Tabela 11. Desempenho dos modelos generalistas sem RING.

Descrição	AC[1]%	DP	AC[2]%	AUC%	TPR%	TNR%
Gen. sem BLCA	80,1	0,02	50,0	50,0	100	0
Gen. sem BRCA	78,7	0,03	85,7	91,6	83.3	100
Gen. sem CESC	78,4	0,02	84,6	78,3	90,0	66,6
Gen. sem COAD	81,6	0,08	72,5	65,6	85,1	46,1
Gen. sem ESCA	81,3	0,05	75,0	50	100	0
Gen. sem GBM	82,3	0,02	72,7	44,4	88,8	0
Gen. sem HNSC	81,8	0,03	88,8	75,0	100	50
Gen. sem KIRC	86,1	0,03	100	100	100	100
Gen. sem LGG	82,9	0,01	80,0	75,0	100	50,0
Gen. sem LIHC	81,0	0,04	100	100	100	100
Gen. sem LUAD	79,3	0,03	100	100	100	100
Gen. sem LUSC	81,9	0,06	88,8	83,3	100	66,6
Gen. sem OV	82,1	0,02	50,0	60,0	100	20,0
Gen. sem PAAD	80,2	0,03	50,0	62,5	100	25,0
Gen. sem PRAD	80,0	0,04	87,5	83.3	100	66,6
Gen. sem SKCM	79,4	0,02	73,0	53,4	81,8	25,0,
Gen. sem STAD	79,5	0,05	77,4	72,7	90,0	54,5
Gen. sem UCEC	82,1	0,05	60,6	56,8	92,3	21,4

validação cruzada não obteve tanta variação quando comparados com os modelos correspondentes sem RING. A maior diferença ocorreu nos modelos treinados sem o câncer LUAD, no qual a taxa de acurácia média é de 86,1% com RING e 79,3% sem RING, representando uma diferença de 6,8 pontos percentuais. Isso sugere que para o modelo treinado sem o câncer LUAD as informações de RING apresentam uma importância.

Em termos gerais, os modelos generalistas não demonstraram uma sensibilidade superior na detecção de verdadeiros positivos (TPR), quando submetidos aos dados de testes reservados, em comparação com os modelos individualistas. Ao analisar as matrizes de confusão dos modelos que alcançaram as maiores acurácias médias (AC[1]), conforme mostrado nas Figuras 4 e 5, observa-se que os acertos relacionados aos verdadeiros positivos não apresentam variação significativa entre os modelos com RING e sem RING obtendo os mesmo valores de acertos de mutações deletérias.

Até então, é crucial enfatizar que a discussão realizada não sugere uma superio-

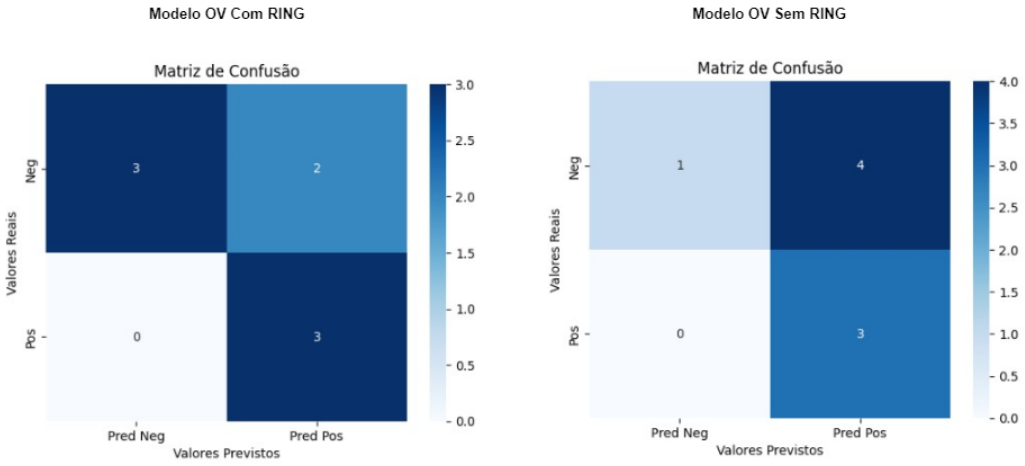

Figura 4. Matriz de confusão modelos genelaristas testados com câncer OV.

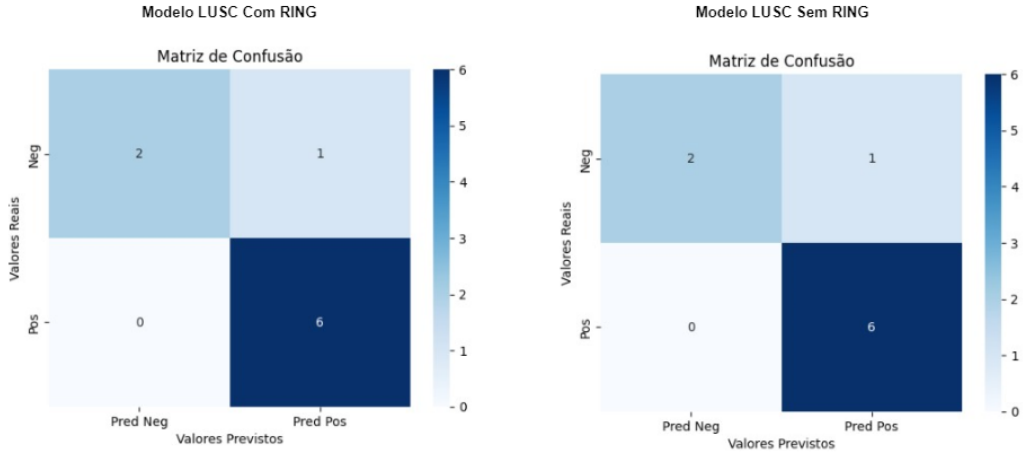

Figura 5. Matriz de confusão modelos genelaristas testados com câncer LUSC.

ridade inerente dos modelos com RING sobre os modelos sem RING. Observamos uma possível sensibilidade nos modelos individuais que incorporam essas informações, especialmente na detecção de verdadeiros positivos. Entretanto, ao considerar uma perspectiva mais abrangente, como exemplificado pelos modelos generalistas, não foi evidenciada uma sensibilidade significativa quando comparados aos modelos análogos com e sem RING.

Nesse contexto, essas informações revelam um impacto na sensibilidade dos modelos individuais com RING, especialmente na detecção de verdadeiros positivos. Dessa maneira, esse aspecto pode ser um fator relevante na identificação de mutações missenses associadas à classificação “deletéria”.

5. Conclusão

Neste estudo, exploramos o potencial das RINGs em modelos de aprendizado de máquina para a predição de mutações missenses associadas ao câncer. Apesar dos desafios decorrentes da disponibilidade limitada de dados sobre RINGs no ClinVar, observamos relações relevantes mesmo com a restrição de dados para treinamento. Embora nosso modelo não tenha atingido os níveis mais elevados de acurácia e precisão quando comparado a estudos correlatos, como o AllelePred [Sobahy et al. 2022], que alcançou uma notável acurácia de 98% nos dados de teste, conseguimos destacar uma percepção mais acentuada nos conjuntos de teste dos modelos individuais com informações de RINGs, evidenciada por taxas como 90,9% no modelo UCEC, 100% no modelo STAD, 100% no modelo SKCM e 80% no modelo COAD. Nos modelos generalistas, as variações foram menos marcantes, exceto no treinamento do modelo sem o câncer LUAD, onde a ausência de informações de RING resultou em uma diferença de 6,8 pontos percentuais na acurácia média de treinamento durante a validação cruzada, em comparação com o modelo generalista análogo com RING.

Modelos com RING

Descrição	AC¹%	DP	AC²%	AUC%	TPR%	TNR%
Gen. sem BLCA	83,8	0,04	75,0	75,0	90,9	50,0
Gen. sem BRCA	80,9	0,02	100	100	100	100
Gen. sem CESC	81,4	0,04	92,3	83,3	100	66,6
Gen. sem COAD	82,6	0,02	75,0	67,5	88,8	46,1
Gen. sem ESCA	84,0	0,04	87,5	75,0	100	50,0
Gen. sem GBM	86,9	0,03	90,9	75,0	100	50,0
Gen. sem HNSC	84,0	0,04	77,7	67,8	85,7	67,8
Gen. sem KIRC	86,1	0,03	100	100	100	100
Gen. sem LGG	84,26	0,02	80,0	75,0	100	50
Gen. sem LIHC	83,76	0,05	100	100	100	100
Gen. sem LUAD	86,1	0,02	100	100	100	100
Gen. sem LUSC	87,2	0,04	88,8	83,3	100	66,6
Gen. sem OV	87,2	0,04	75,0	80,0	100	60,0
Gen. sem PAAD	84,7	0,02	83,3	87,5	100	75,0
Gen. sem PRAD	84,1	0,02	75,0	73,3	80,0	66,6
Gen. sem SKCM	84,2	0,06	70,4	70,4	90,9	50
Gen. sem STAD	83,8	0,04	70,9	67,2	80,0	54,5
Gen. sem UCEC	84,9	0,04	74,4	72,8	88,4	57,1

Modelos sem RING

Descrição	AC¹%	DP	AC²%	AUC%	TPR%	TNR%
Gen. sem BLCA	80,1	0,02	50,0	50,0	100	0
Gen. sem BRCA	78,7	0,03	85,7	91,6	83.3	100
Gen. sem CESC	78,4	0,02	84,6	78,3	90,0	66,6
Gen. sem COAD	81,6	0,08	72,5	65,6	85,1	46,1
Gen. sem ESCA	81,3	0,05	75,0	50	100	0
Gen. sem GBM	82,3	0,02	72,7	44,4	88,8	0
Gen. sem HNSC	81,8	0,03	88,8	75,0	100	50
Gen. sem KIRC	86,1	0,03	100	100	100	100
Gen. sem LGG	82,9	0,01	80,0	75,0	100	50,0
Gen. sem LIHC	81,0	0,04	100	100	100	100
Gen. sem LUAD	79,3	0,03	100	100	100	100
Gen. sem LUSC	81,9	0,06	88,8	83,3	100	66,6
Gen. sem OV	82,1	0,02	50,0	60,0	100	20,0
Gen. sem PAAD	80,2	0,03	50,0	62,5	100	25,0
Gen. sem PRAD	80,0	0,04	87,5	83.3	100	66,6
Gen. sem SKCM	79,4	0,02	73,0	53,4	81,8	25,0,
Gen. sem STAD	79,5	0,05	77,4	72,7	90,0	54,5
Gen. sem UCEC	82,1	0,05	60,6	56,8	92,3	21,4

Câncer	AC¹%	DP	AC²%	AUC%	TPR%	TNR%
UCEC	75,6	0,13	68,4	64,2	90,9	37,5
STAD	71,0	0,24	71,4	50,0	100,0	0,0
SKCM	85,0	0,12	83,3	50,0	100,0	0,0
COAD	69,3	0,21	50,0	40,0	80,0	0,0

Câncer	AC¹%	DP	AC²%	AUC%	TPR%	TNR%
UCEC	67,9	0,04	42,1	44,8	27,0	62,5
STAD	88,0	0,09	71,4	50,0	100,0	0,0
SKCM	70,0	0,24	66,6	40,0	80,0	0,0
COAD	69,0	0,15	50,0	40,0	80,0	0,0

Figura 6. Tabela com todos os modelos reunidos

Essas descobertas oferecem perspectivas promissoras para pesquisas futuras que buscam integrar dados de RINGs, especialmente na construção de modelos mais especializados para a detecção de tipos específicos de câncer na área da saúde, onde a sensibilidade é um indicador crucial, conforme mencionado por [Monaghan et al. 2021]. Os resultados deste experimento indicam uma possível correlação positiva entre a presença

de informações de RINGs e a sensibilidade, especialmente em modelos mais direcionados e individualizados, ressaltando assim o potencial significativo desses dados. Essa observação sugere que a inclusão de informações de RINGs pode ser uma estratégia valiosa para aprimorar a capacidade de detecção em modelos mais focalizados em um contexto clínico específico.

Portanto, ressalta-se a necessidade de ampliar a disponibilidade de dados de RINGs, mapeados e devidamente validadas, para que possa somar em pesquisas futuras. Essa ampliação permitirá que próximas investigações integrem esses dados com outras informações, potencialmente contribuindo para o desenvolvimento de preditores mais robustos e eficazes. Essa abordagem aprimorada pode ser crucial para avançar o diagnóstico precoce e a medicina direcionada, promovendo melhorias significativas na prática clínica.

Referências

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., and Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143:106839.

Dakshit, S., Dakshit, S., Khargonkar, N., and Prabhakaran, B. (2023). Bias analysis in healthcare time series (baht) decision support systems from meta data. *Journal of Healthcare Informatics Research*, pages 1–29.

Fonseca, F. V. d. (2020). Comparação de redes de interação de resíduos (rins) como uma forma de avaliar a variação conformacional de proteínas. Master's thesis, Universidade Federal do Rio Grande do Norte.

Fonseca, W. R. d. (2023). Medida de dispersão.

Gyulkhandanyan, A., Rezaie, A. R., Roumenina, L., Lagarde, N., Fremeaux-Bacchi, V., Miteva, M. A., and Villoutreix, B. O. (2020). Analysis of protein missense alterations by combining sequence-and structure-based methods. *Molecular Genetics & Genomic Medicine*, 8(4):e1166.

Halimu, C., Kasem, A., and Newaz, S. S. (2019). Empirical comparison of area under roc curve (auc) and mathew correlation coefficient (mcc) for evaluating machine learning algorithms on imbalanced datasets for binary classification. In *Proceedings of the 3rd international conference on machine learning and soft computing*, pages 1–6.

Johnson, J. M. and Khoshgoftaar, T. M. (2021). Encoding techniques for high-cardinality features and ensemble learners. In *2021 IEEE 22nd international conference on information reuse and integration for data science (IRI)*, pages 355–361. IEEE.

Malhotra, S., Alsulami, A. F., Heiyun, Y., Ochoa, B. M., Jubb, H., Forbes, S., and Blundell, T. L. (2019). Understanding the impacts of missense mutations on structures and functions of human cancer-related genes: a preliminary computational analysis of the cosmic cancer gene census. *PLoS One*, 14(7):e0219935.

McDermott, M., Wang, S., Marinsek, N., Ranganath, R., Ghassemi, M., and Foschini, L. (2019). Reproducibility in machine learning for health. *arXiv preprint arXiv:1907.01463*.

Monaghan, T. F., Rahman, S. N., Agudelo, C. W., Wein, A. J., Lazar, J. M., Everaert, K., and Dmochowski, R. R. (2021). Foundational statistical principles in medical research: sensitivity, specificity, positive predictive value, and negative predictive value. *Medicina*, 57(5):503.

Nguyen, T. B., Myung, Y., de Sá, A. G., Pires, D. E., and Ascher, D. B. (2021). mmcsmna: accurately predicting effects of single and multiple mutations on protein–nucleic acid binding affinity. *NAR Genomics and Bioinformatics*, 3(4):lqab109.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., and Tzovara, A. (2021). Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2(10).

Pargent, F., Pfisterer, F., Thomas, J., and Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, 37(5):2671–2692.

Pereira, D. C. B. G. (2020). Classificação de mutações associadas ao câncer, por meio de algoritmos de aprendizagem de máquina. Master's thesis, Universidade Federal do Rio Grande do Norte.

Petrosino, M., Novak, L., Pasquo, A., Chiaraluce, R., Turina, P., Capriotti, E., and Consalvi, V. (2021). Analysis and interpretation of the impact of missense variants in cancer. *International Journal of Molecular Sciences*, 22(11):5416.

Sobahy, T. M., Motwalli, O., and Alazmi, M. (2022). Allelepred: A simple allele frequencies ensemble predictor for different single nucleotide variants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(1):796–801.

Tong, S.-Y., Fan, K., Zhou, Z.-W., Liu, L.-Y., Zhang, S.-Q., Fu, Y., Wang, G.-Z., Zhu, Y., and Yu, Y.-C. (2022). Mvppt: a highly efficient and sensitive pathogenicity prediction tool for missense variants. *Genomics, Proteomics & Bioinformatics*.

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., and Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *Journal of Applied Science and Technology Trends*, 1(2):56–70.

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., and Si, Y. (2018). A data-driven design for fault detection of wind turbines using random forests and xgboost. *Ieee Access*, 6:21020–21031.

Zhang, N., Lu, H., Chen, Y., Zhu, Z., Yang, Q., Wang, S., and Li, M. (2020). Prempri: Predicting the effects of missense mutations on protein–rna interactions. *International journal of molecular sciences*, 21(15):5560.