INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA (ILACVN)

BIOTECNOLOGIA

MÉTODO DE AVALIAÇÃO EXPERIMENTAL *IN SILICO* BASEADO EM ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE PEPTÍDEOS ANTICÂNCER

ISABELLA CAROLINE SACHINI LORENA

Foz do Iguaçu
2023

INSTITUTO LATINO-AMERICANO DE CIÊNCIAS DA VIDA E DA NATUREZA (ILACVN)

BIOTECNOLOGIA

MÉTODO DE AVALIAÇÃO EXPERIMENTAL *IN SILICO* BASEADO EM ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE PEPTÍDEOS ANTICÂNCER

ISABELLA CAROLINE SACHINI LORENA

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Ciências da Vida e da Natureza da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Biotecnologia.

Orientador: Prof. Dr. Willian Zalewski

Foz do Iguaçu
2023

ISABELLA CAROLINE SACHINI LORENA

MÉTODO DE AVALIAÇÃO EXPERIMENTAL *IN SILICO* BASEADO EM ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA A PREDIÇÃO DE PEPTÍDEOS ANTICÂNCER

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Ciências da Vida e da Natureza da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Biotecnologia

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Willian Zalewski
UNILA

_____

Prof. Dr. Joylan Nunes Maciel
UNILA

_____

Prof. Dr. Marcelo Nepomoceno Kapp
UNILA

_____

Prof. Dr. Michel Rodrigo Zambrano Passarini
UNILA

Foz do Iguaçu, 01de Novembro de 2023.

À Elisangela Sachini Lorena

À Jonato Nunes Lorena

AGRADECIMENTOS

Não sou muito boa com as palavras mas...lá vai.. Agradeço, primeiramente, a Deus, pois até aqui ele me sustentou com o seu infinito amor. E me propiciou momentos incríveis que eu jamais poderia imaginar.

Agradeço imensamente ao meu orientador Willian Zalewski por toda paciência; atenção e confiança durante a realização da pesquisa. Por ter me dado a honra de compartilhar muitos momentos de aprendizagem e troca de conhecimentos, desde a iniciação científica até o presente momento. Conhecimentos que me fizeram me desenvolver tanto profissionalmente assim como me fizeram me tornar uma pessoa melhor.

Agradeço ao meu pai Jonato Lorena por toda as vezes que me deu suporte. O qual sempre acreditou no meu potencial e sempre me ajudou a prosseguir. Que desde a infância cuida de mim com todo o carinho do mundo, carinho até nos pequenos detalhes. Agradeço também a minha mãe Elisangela por todos os momentos de suporte, que mesmo em momentos difíceis esteve ao meu lado, me cuidando com todo amor e carinho e que sempre foi guerreira.

Agradeço aos meus avós.

Agradeço também aos meus irmãos Emilly, Luiz Davi, Evillyn, Erick e Eduarda por todos os momentos.

Agradeço, em especial, as minhas sobrinhas Mariane Liz e Heloise Liz que me ensinam todos os dias como é bom viver a vida, além de me ensinarem todos os dias o que é amar.

Agradeço ao meu professor de graduação Raphael Fortes que desde os primeiros anos da graduação e até hoje me motiva a prosseguir nos estudos.

Agradeço aos meus amigos de graduação Angie, Eduardo, Gustavo, Heveli, Grecia, João, Kauanny, Natalia, Luiz, Bernie, Viviana e Yara por todos os momentos de risadas, estudos e loucuras. Agradeço aos meus amigos Camila Santos, Vitor Hugo, Camila Silva, Karla que mesmo depois do ensino médio a amizade continua.

Agradeço a Marília Mendonça, pois as músicas dela foram peças chaves para que eu pudesse escrever esse trabalho.

Ademais, agradeço a todos os colegas dos laboratórios de computação.

“Em seu coração o homem planeja o seu caminho, mas o Senhor determina os seus passos”.
Provérbios 16:9

RESUMO

O Câncer representa uma das principais causas de morte em todo mundo. As formas de tratamento disponíveis causam graves efeitos colaterais e resistência das células cancerígenas. Neste contexto, os Peptídeos Anti Câncer (ACP) constituem uma alternativa promissora de tratamento, uma vez que possuem efeitos como imuno moduladores e indutores da apoptose. No entanto, a identificação e caracterização de um ACP de forma experimental consiste em uma tarefa onerosa em relação ao custo monetário e tempo. Desse modo, estudos *in silico*, em especial, por meio de técnicas de aprendizado de máquina, tem motivado o desenvolvimento de métodos automáticos para auxiliar no processo de identificação. Entretanto, apesar de resultados promissores reportados na literatura, o mecanismo pelo qual os peptídeos apresentam atividade anticancerígena é um tema em aberto que ainda não é adequadamente compreendido. Além desse aspecto, os estudos existentes apresentam métodos diversos em termos das estratégias adotadas, tais como: descritores de características; algoritmos de aprendizado de máquina e método de avaliação. Esse cenário dificulta a análise adequada das abordagens existentes e a comparação com novas propostas. Portanto, neste trabalho, com o intuito de definir um baseline para a literatura, foi proposto um método de avaliação experimental para o qual foram analisadas diferentes estratégias de predição de peptídeos. Para alcançar esse objetivo, foi realizada uma extensa análise por meio da combinação de 18 algoritmos de aprendizado de máquina com 13 descritores de características. Como resultado desse estudo experimental, um total de 2240 modelos foram construídos e avaliados utilizando diferentes métricas. Desse modo, pela análise de resultados estatísticos obtidos, observa-se que é preferível a utilização de descritores que apresentam maior desempenho em termos de acurácia e menor dimensionalidade. Além disso, é possível constatar, que considerando as técnicas avaliadas neste estudo, o desempenho dos modelos construídos possui maior influência em relação ao tipo de descritor de característica utilizado como, por exemplo, os modelos Circular/catboost e CTD/gbc.

Palavras-chave: aprendizado de máquina; *in sílico*; peptídeos anticâncer; predição de sequências; QSAR.

RESUMEN

El cáncer representa una de las principales causas de muerte en todo el mundo. Las formas de tratamiento disponibles causan graves efectos secundarios y resistencia de las células cancerígenas. En este contexto, los Péptidos Anticancerígenos (ACP) representan una prometedora alternativa de tratamiento, ya que poseen efectos como inmunomoduladores e inductores de la apoptosis. Sin embargo, la identificación y caracterización de un ACP de manera experimental resulta en una tarea costosa en términos de tiempo y dinero. Por lo tanto, los estudios *in silico*, en particular a través de técnicas de aprendizaje automático, han impulsado el desarrollo de métodos automáticos para ayudar en el proceso de identificación. No obstante, a pesar de los resultados prometedores reportados en la literatura, el mecanismo a través del cual los péptidos muestran actividad anticancerígena sigue siendo un tema en debate que aún no se comprende adecuadamente. Además, los estudios existentes presentan diversos métodos en términos de las estrategias adoptadas, como descriptores de características, algoritmos de aprendizaje automático y métodos de evaluación. Esta situación dificulta el análisis adecuado de los enfoques existentes y la comparación con nuevas propuestas. Por lo tanto, en este trabajo, con el fin de establecer un punto de referencia en la literatura, se propuso un método de evaluación experimental en el que se analizaron diferentes estrategias de predicción de péptidos. Para lograr este objetivo, se llevó a cabo un análisis exhaustivo combinando 18 algoritmos de aprendizaje automático con 13 descriptores de características. Como resultado de este estudio experimental, se construyeron y evaluaron un total de 2240 modelos utilizando diversas métricas. De esta manera, mediante el análisis de los resultados estadísticos obtenidos, se observa que es preferible utilizar descriptores que exhiban un mayor rendimiento en términos de precisión y menor dimensionalidad. Además, es posible constatar que, al considerar las técnicas evaluadas en este estudio, el rendimiento de los modelos construidos tiene una mayor influencia en relación al tipo de descriptor de características utilizado, como por ejemplo, los modelos Circular/catboost y CTD/gbc.

Palabras clave: aprendizaje automático; *in silico*; péptidos anticancerígenos; predicción de secuencias; QSAR.

ABSTRACT

Cancer represents one of the leading causes of death worldwide. The available treatment options result in severe side effects and cancer cell resistance. In this context, Anti-Cancer Peptides (ACP) offer a promising alternative for treatment, as they possess effects such as immunomodulation and apoptosis induction. However, the identification and experimental characterization of ACPs constitute a costly and time-consuming task. Consequently, *in silico* studies, particularly through machine learning techniques, have spurred the development of automatic methods to aid in the identification process. Nevertheless, despite promising results reported in the literature, the mechanism through which peptides exhibit anticancer activity remains an open and inadequately understood topic. In addition to this aspect, existing studies employ diverse methods in terms of the strategies adopted, including feature descriptors, machine learning algorithms, and evaluation methods. This scenario complicates the proper analysis of existing approaches and comparison with new proposals. Therefore, in this work, with the aim of establishing a baseline in the literature, an experimental evaluation method was proposed to analyze different peptide prediction strategies. To achieve this goal, an extensive analysis was conducted by combining 18 machine learning algorithms with 13 feature descriptors. As a result of this experimental study, a total of 2240 models were constructed and evaluated using various metrics. Thus, through the analysis of obtained statistical results, it is observed that the utilization of descriptors exhibiting higher accuracy performance and loer dimensionality is preferable. Moreover, it is possible to note that, considering the techniques evaluated in this study, the performance of constructed models has a greater influence regarding the type of feature descriptor used, such as the Circular/catboost and CTD/gbc models, for example.

Key words: machine learning; in silico; anticancer peptides; sequence prediction; QSAR.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1D-CNN	1D *Convolutional Neural Networks*
2DCNN	Rede Neural Convolucional Bidimensional
AAC	Composição de aminoácidos
AAindex	Amino Acid Index Database
ACP	Peptídeos Anti-Câncer
ACP-2DCNN	*Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network*
ACP-DA	*Improving the Prediction of Anticancer Peptides Using Data Augmentation*
ACP-DL	*A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation*
ACP-GCN	*The Identification of Anticancer Peptides Based on Graph Convolution Networks*
ACP-MCAM	*Anticancer Peptide Prediction via Multi-Kernel CNN and Attention Model*
ACPNet	*A Deep Learning Network to Identify Anticancer Peptides by Hybrid Sequence Information*
ACPred-BMF	*Bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction*
ACPred-FL	A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides
ADA	*AdaBoost Machine Learning Adaptive Boosting*
AM	Aprendizado de Máquina
AMP	Peptídeos Antimicrobianos
APAAC	Composição de Pseudo-Aminoácidos Anfifílicos
APD	Antimicrobial Peptide Date
Acc	Acurácia
BPF	*Binary profile feature*
Bi-LSTM	*Bi-directional Long ShortTerm Memory*
CAMP	*Collection of Antimicrobial Peptides*
CATB	CatBoost

CD	Conjunto de dados
CL-ACP	*A parallel combination of CNN and LSTM anticancer peptide recognition model*
CNN	*Convolutional Neural Network*
CT	Tríade Conjunta
CTD	Composição/Transição/Distribuição
CTriad	Tríade Conjugada
CV	*K-fold Cross-Validation*
CancerPPD	*A database of anticancer peptides and proteins*
Circular	*Circular Fingerprint*
DADP	*A database of anuran defense peptides*
DDE	Desvio de dipeptídeos da média esperada
DLFF-ACP	*Prediction of ACPs based on deep learning and multi-view features fusion*
DPC	Composição de dipeptídeos
DRAMP	*Data repository of antimicrobial peptides*
DT	*Decision Tree*
DeepACP	*A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm*
Dense Net	*Dense Convolutional Network*
ECFPs	*Extended-connectivity fingerprints*
EMA	Agência Europeia de Medicamentos
ET	*Extra Trees: Extremely Randomized Trees*
F1	Pontuação F1
FDA	*Food and Drug Administration*
FN	Falsos Negativos
FP	Falsos Positivos
GAAC	Composição de aminoácidos agrupados
GBC	*Gradient Boosting Classification*
GCN	*Graph Convolutional Network*
GCNCPR-ACPs	*A Novel Graph Convolution Network Metrod for ACPs prediction*
GPC	*Gaussian Process Classification*
HDNN	*Hybrid Deep Neural Network*

ILACVN	Instituto Latino-Americano de Ciências da Vida e da Natureza
LDA	*Linear Discriminant Analysis*
LGBM	*LightGBM: Light Gradient Boosting Machine*
LR	*Logistic Regression*
ME-ACP	*Multi-view neural networks with ensemble model for identification of anticancer peptides*
MLACP 2.0	*An updated machine learning tool for anticâncer peptide prediction*
MLP	*Multi-Layer Perceptron*
Macckeys	*Macckeys Fingerprint*
Mcc	Correlação de Matthews
Mol2vec	*Mol2VecFingerprint*
NACP	Não Anti-Câncer
NAO	*Noise adding oversampling*
NB	*Naive Bayes*
PAAC	Composição de Pseudo-Aminoácidos
PNL	Processamento de Linguagem Natural
Prec	Precisão
PydPi	*Freely Available Python Package for Chemoinformatics*
QDA	*Quadratic Discriminant Analysis*
QSAR	*Quantitative structure-activity relationship*
Qualc	*Qualitative Properties of amino acids*
Quanc	*Quantitative Properties of amino acids*
RF	*Random Forest*
RNN	*Recurrent Neural Network*
Rec	*Recall*
SSA	*Soft symmetric alignment*
SVM	*Support Vector Machines*
TPC	Composição Tripeptídica
UNILA	Universidade Federal da Integração Latino-Americana
UNiRep	*Unified Representation*
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
XGB	XGBoost: eXtreme Gradient Boosting

Knn	*K-Nearest Neighbors*

SUMÁRIO

1. INTRODUÇÃO

O câncer representa uma das principais causas de morte em todo o mundo. Estimativas indicam que cerca de 19,3 milhões de novos casos de câncer surgem anualmente (18,1 milhões excluindo o câncer de pele não melanoma). Além disso, as diferentes formas de câncer foram responsáveis por aproximadamente 10 milhões de mortes em todo o mundo no ano de 2020. Ademais, segundo os relatórios mais atuais da Organização Mundial de Saúde (2019), os diferentes tipos de câncer são a primeira ou segunda principal causa de morte antes dos 70 anos em cerca de 112 países (SUNG, H. et al., 2021). O câncer engloba uma lista de diferentes neoplasias, as quais possuem características em comum, sendo essas as responsáveis por transformar uma célula considerada normal em uma célula neoplásica. Essas características possibilitam que essas células neoplásicas desafiem as condições e normas do micro ambiente. Por exemplo, essas células podem habilitar a imortalidade replicativa, que facilita o crescimento celular fora dos limites naturais, levando a aquisição de outras características, tais como: a evasão da resposta imune; a ativação de invasão; a metástase; e a indução de angiogênese (indução do desenvolvimento de vasos sanguíneos para nutrição e oxigenação do tumor). Devido a isso, a partir da década de 1940, diversas pesquisas buscam pela “cura do câncer”. Essa busca resultou em formas de tratamento como, por exemplo, a quimioterapia e a radioterapia. Apesar desses tratamentos apresentarem resultados significativos contra os cânceres, existem efeitos colaterais relevantes que devem ser considerados. Um dos principais problemas é que esses tratamentos não são unicamente direcionados à destruição das células anômalas, o que gera a lesão de células saudáveis que estão próximas ao microambiente tumoral. Além desse aspecto, a utilização de drogas quimioterápicas induz a morte das células tumorais por meio da genotoxicidade. Como consequência, essas células podem adquirir propriedades de resistência, que podem manifestar-se de modo intrínseco ou extrínseco perante as diferentes estratégias terapêuticas (FREITAS SAITO et al., 2015).

Dentre as propostas de novos tratamentos, estudos *in vitro* com Peptídeos Anticâncer (ACP) demonstraram a toxicidade seletiva para células cancerígenas; potencialidade de penetrar pela membrana plasmática das células para

posterior lise dessas; efeitos anti-angiogênicos; efeitos imunomoduladores; e indução da apoptose. Além desses aspectos, os ACPs têm se destacado pela alta biocompatibilidade e facilidade de ser sintetizados, o que pode viabilizar o desenvolvimento de uma forma terapêutica direcionada (KURRIKIFF et al., 2019). Apesar desses excelentes atributos, os ACPs apresentam algumas desvantagens. Os métodos de identificação; caracterização e síntese dos ACPs são realizados de forma experimental, ou seja, por meio de análises laboratoriais e com a utilização de métodos, que necessitam de financiamento em equipamentos de última geração, além da utilização de reagentes e mão de obra especializada. Desse modo, atualmente, o processo de identificação de ACPs consiste em uma tarefa onerosa em relação ao custo monetário e de tempo (CHINNADURAI, R. K. et al., 2023).

Nesse contexto, métodos *in silico* têm sido uma alternativa viável para explorar de modo eficiente o grande espaço químico que os peptídeos ocupam e reduzir o número de peptídeos que precisam ser rastreados experimentalmente. Técnicas computacionais baseadas em algoritmos de aprendizado de máquina (MITCHELL, 2017) têm sido exploradas nessa tarefa, em especial, por apresentarem características como: geração automática de modelos sobre grandes conjuntos de dados, generalização e tratamento de dados não lineares. Diversos métodos para a predição de ACPs foram propostos e descritos na literatura utilizando técnicas de aprendizado de máquina (LI, 2017). Esses métodos utilizam diferentes formas de representação e caracterização dos peptídeos, incluindo informações físico-químicas, estruturais, composição e distribuição de aminoácidos (LIANG, 2021). No entanto, apesar de resultados promissores reportados na literatura, o mecanismo pelo qual os peptídeos apresentam atividade anticancerígena é um tema em aberto que ainda não é adequadamente compreendido. Ademais, os estudos existentes apresentam métodos dispersos em termos das estratégias adotadas, tais como: descritores de características; algoritmos de aprendizado de máquina; método de avaliação; e conjuntos de dados. Esse cenário dificulta a análise adequada das abordagens existentes e a comparação com novas propostas.

Desse modo, o objetivo geral deste trabalho consiste em definir um baseline para a literatura e propor um método para avaliar estratégias de aprendizado de máquina aplicadas à identificação de ACPs. Como objetivos específicos deste estudo, é possível elencar:

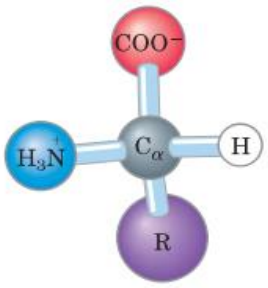
1. Avaliar diferentes descritores de características dos peptídeos que possibilitem oferecer diferentes perspectivas sobre os dados;
2. Analisar distintos algoritmos de aprendizado de máquina, utilizando diferentes métricas de desempenho;
3. Realizar um estudo exploratório dos parâmetros dos algoritmos de aprendizado de máquina, com o intuito de identificar aqueles com melhor desempenho para os modelos.

2. REVISÃO BIBLIOGRÁFICA

2.1. PEPTÍDEOS ANTICÂNCER

Proteínas são macromoléculas biológicas, abundantes na natureza e controlam em grande parte todos os processos que ocorrem dentro de uma célula. Exemplos dessa vasta funcionalidade de proteínas dentro de um organismo podem ser observados em enzimas; hormônios; anticorpos; fibras musculares e entre outros. As proteínas são polímeros biológicos, os quais são formados pela união de aminoácidos. Na natureza existem 20 aminoácidos, que podem ser representados por letras (Aminoácidos: A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y). Dentre esses aminoácidos temos a presença do (carbono ∝), onde os grupamentos amino e carboxila se ligam, além do grupamento R, o qual é diferente para cada um dos aminoácidos (NELSON; COX, 2014). Na Figura 1, pode-se observar a estrutura geral de um aminoácido.

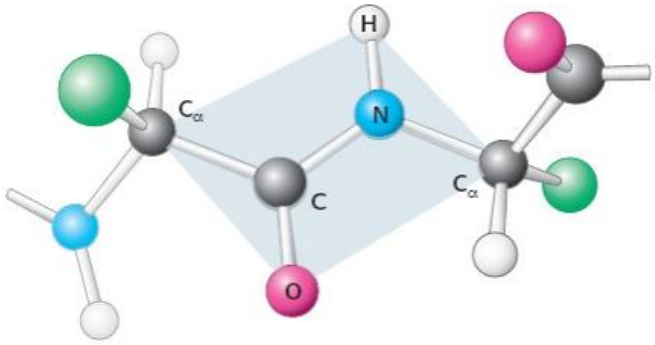
Figura 1 - REPRESENTAÇÃO DA ESTRUTURA GERAL DOS AMINOÁCIDOS



Fonte: NELSON; COX, 2014.

Cada aminoácido presente na proteína é chamado de “resíduo de aminoácido”, uma vez que ao formar a ligação peptídica, ocorre uma reação de desidratação, gerando apenas um “resíduo”. Ao ocorrer a união de dois aminoácidos, temos a formação de um peptídeo. Na Figura 2, podemos visualizar um dipeptídeo, que é composto por dois resíduos de aminoácidos (BERG et al., 2014).

Figura 2 - REPRESENTAÇÃO DE UM DIPEPTÍDEO

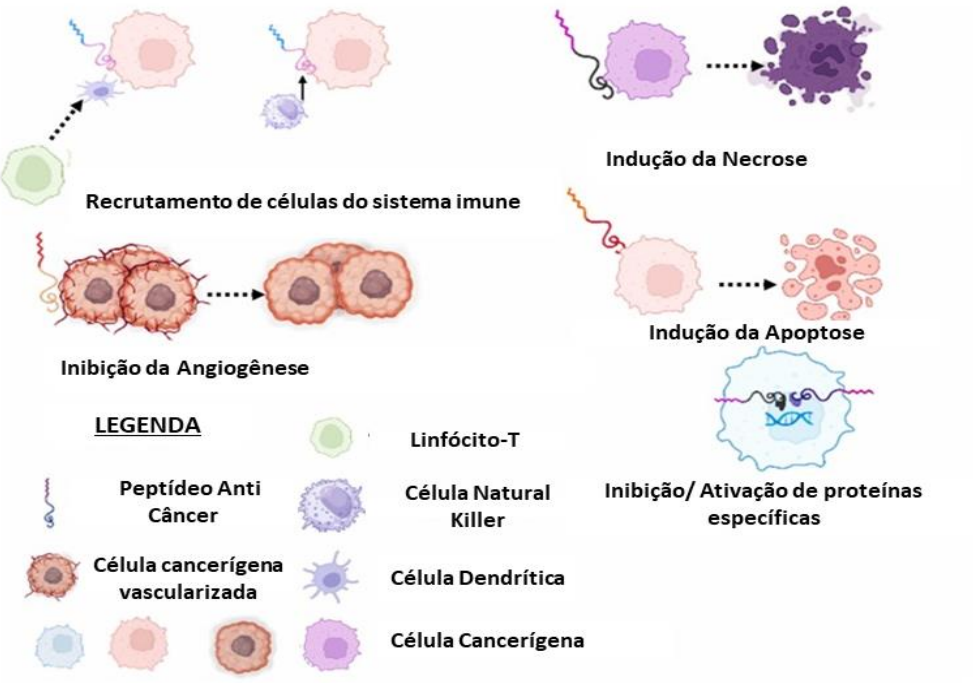


Fonte: BERG et al., 2014

Os peptídeos possuem uma grande aplicabilidade ao considerar o seu potencial terapêutico. Na busca pelo desenvolvimento de novos fármacos, descobriu-se que um conjunto de peptídeos possuem a função de inibir o crescimento microbiano, também chamados de Peptídeos Antimicrobianos (AMP), sendo abundantes da natureza e podendo ser encontrados, como exemplo, em artrópodes e anfíbios. Em especial, uma subclasse dentro desse grupo possui a potencialidade de apresentar atividade anticancerígena. Esse subgrupo é conhecido como Peptídeos Anti Câncer (ACP).

Os ACPs são peptídeos catiônicos que contêm de 5 a 50 resíduos de aminoácidos. Com relação a estrutura de um ACP, temos que esses podem ter uma estrutura linear estendida; uma estrutura α Hélice, mas também uma estrutura folha β. Os ACPs possuem a capacidade de interagir de forma direcionada com a membrana de células cancerígenas, através de uma interação eletrostática e consequente levar à morte celular. Os ACPs também podem induzir a apoptose por meio da lise de membranas mitocondriais. Além disso, podem inibir a via de angiogênese, logo, evitando o desenvolvimento do tumor. Outra característica é a funcionalidade de imunomodulação por meio do recrutamento de células do sistema imunológico. Na Figura 3, pode-se visualizar o modo de ação dos peptídeos Anti Câncer (ACP).

Figura 3 - MODO DE AÇÃO DOS PEPTÍDEOS ANTI CÂNCER



Fonte: MODIFICADO DE CHINNADURAI., et al, 2023.

Devido a possibilidade de aplicação clínica de peptídeos na terapêutica do câncer, o desenvolvimento dessas biomoléculas tem atraído a atenção de pesquisadores e empresas da área. Até o momento deste trabalho, verificou-se que na base de dados do *Drug Bank* [1]existem cerca de 460 compostos relacionados ao câncer, dentre os quais 29 são pertencentes à classe dos peptídeos. Esses 29 compostos estão categorizados em três grupos, os quais são: aprovados, investigados e experimentais. Dentre os aprovados, temos essa aprovação realizada por agências reguladoras como a *Food and Drug Administration*[2] (FDA); e a Agência Europeia de Medicamentos[3] (EMA) (CHINNADURAI et al., 2023). No Quadro 1, é possível observar os 5 medicamentos anti câncer disponíveis atualmente a base de peptídeos.

[1] https://go.drugbank.com/
[2] https://www.fda.gov/
[3] https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/european-medicines-agency-ema_pt

Quadro 1 - MEDICAMENTOS APROVADOS A BASE DE PEPTÍDEOS ANTICÂNCER

Nome genérico do banco de medicamentos	**Indicação**	**Tipo de câncer alvo**	**Status de aprovação**	**Referência**
Tebentafusp	Tratamento do melanoma uveal.	Melanoma uveal	FDA	(HOWLETT et al., 2023)
Buserelina	Usado no tratamento de cânceres responsivos a hormônios.	Câncer de mama	EMA	(BROGDEN et al., 1990)
Plitidepsina	Destinado ao tratamento de diversas formas de câncer.	Mieloma múltiplo	EMA	(ALONSO-ALVAREZ et al., 2017)
Triptorelina	Tratamento paliativo do câncer de próstata avançado	Câncer de próstata	FDA	(FURMAN et al., 2007)
Dactinomicina	A actinomicina é usada para tratar uma ampla variedade de cânceres.	Variedade de câncer	FDA	(KWOK et al., 2017)

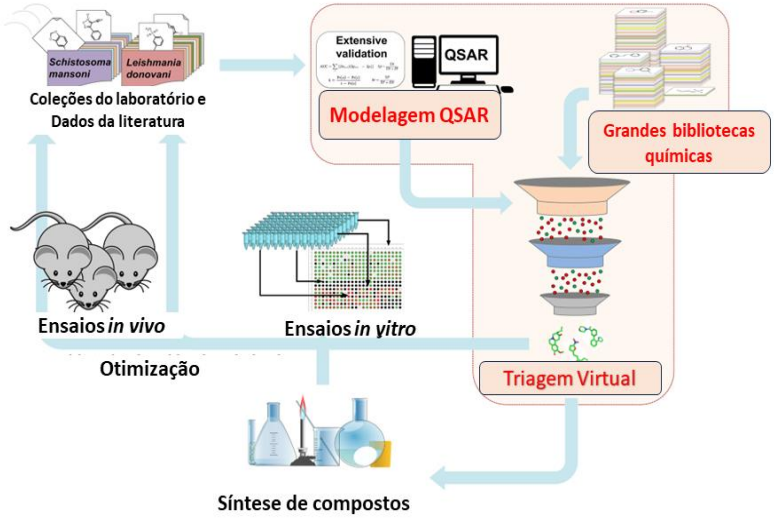
Fonte: MODIFICADO DE CHINNADURAI., et al, 2023.

2.2. QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR)

Os métodos convencionais para a descoberta de um fármaco incluem os processos de identificação e validação; rastreamento e aprimoramento de compostos principais; além do desenvolvimento de ensaios pré-clínicos e clínicos. Estima-se que para a realização desses processos, as indústrias farmacêuticas necessitam, em relação ao tempo, em média 12 anos, e em relação ao valor monetário, cerca de US$1,8 bilhões. Portanto, a descoberta e desenvolvimento de um fármaco é uma tarefa onerosa com relação ao tempo e valor monetário (SHAKER et al., 2021). Essa problemática também está presente no desenvolvimento de fármacos à base de peptídeos anti cancerígenos, pois ainda não se tem o conhecimento de uma característica da estrutura da molécula que indique de modo determinístico a propriedade farmacológica. Nesse sentido, com o intuito de mitigar essas limitações, abordagens *in silico* tem se mostrado como alternativas promissoras.

Uma das abordagens *in silico* propostas, comumente utilizada na literatura, é o *Quantitative structure-activity relationship* (QSAR). Essa estratégia tem como objetivo relacionar a estrutura molecular com a bioatividade da molécula, a nível *in silico*, com base em descritores de características. Basicamente esses descritores possibilitam descrever as propriedades físico-químicas de uma molécula de forma numérica, sob diversas perspectivas. Desse modo, a ideia de estratégias QSAR consiste em utilizar esses descritores de características como parâmetros para a construção de modelos, matemáticos ou computacionais, para inferir a atividade biológica dos compostos analisados. Na Figura 4, podemos visualizar a utilização de estudos QSAR para o desenvolvimento de novos fármacos. (NEVES et al., 2018).

Figura 4 - UTILIZAÇÃO DE FERRAMENTAS QSAR PARA O DESENVOLVIMENTO DE NOVOS FÁRMACOS



Fonte: MODIFICADO DE NEVES., et al (2018).

Dentre os tipos de modelagens existentes, as estratégias computacionais têm atraído a atenção de pesquisadores da área, em especial, pela aplicação de algoritmos de Aprendizado de Máquina (DANISHUDDIN et al., 2016).

Apesar de esforços no desenvolvimento dessas ferramentas, estudos *in silico* para a predição de peptídeos anti cancerígenos são recentes na literatura. A escolha de descritores de QSAR que permitam caracterizar corretamente as sequências peptídicas e a seleção de algoritmos de Aprendizado de Máquina para a predição ainda são um desafio dentro da literatura.

2.3. APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (AM) consiste na criação de hipóteses ou aproximações de funções derivadas a partir de um conjunto de exemplos que permitam caracterizar o domínio ou problema que se deseja tratar. O principal meio para a aquisição de conhecimento, de modo automático, por meio dos algoritmos de AM é baseado na inferência indutiva, na qual novos conhecimentos podem ser derivados a partir de outros previamente conhecidos (MITCHELL, 1997).

No contexto de AM, de acordo com o tipo de tarefa a ser realizada, o aprendizado indutivo pode ser organizado em aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado, os algoritmos de indução são aplicados sobre um conjunto de exemplos do domínio, chamado conjunto de treinamento, no qual o conceito que caracteriza cada exemplo (classe) é conhecido. Caso o domínio desse conceito seja composto por um conjunto de valores nominais, a indução da hipótese consiste em uma tarefa de classificação. Caso o domínio seja um conjunto infinito e ordenado de valores, a tarefa é denominada regressão (REZENDE, 2003).

Na tarefa de classificação, os algoritmos de AM são treinados com um conjunto de dados rotulados, onde cada exemplo é associado a uma classe específica. Isso permite que o algoritmo aprenda a categorizar ou prever as classes de novos dados baseando-se nas informações aprendidas durante o treinamento (REZENDE, 2003). Formalmente, seja E o conjunto de todas os exemplos possíveis de um determinado domínio e seja C = {c1, ..., cw} um conjunto de w classes: para todo Ei ∈ E, Ei pertence a pelo menos uma das classes de c1 até cw, e se Ei pertence a uma classe cj, então Ei não pertence a qualquer outra classe ck, tal que j≠k. Assim, um classificador consiste em uma função f que permite mapear um exemplo Ei ∈ E para uma classe c ∈ C:

$$f : \mathbf{E} \rightarrow \{c1, ..., cw\}$$

2.4. TRABALHOS RELACIONADOS

No contexto deste trabalho, foi realizada uma revisão aprofundada da literatura, considerando os estudos publicados entre os anos de 2020 e 2023, uma vez que esses trabalhos são um aprimoramento de trabalhos de anos anteriores. Desse modo, essa revisão buscou compreender os diferentes métodos aplicados na tarefa de predição *in silico* de peptídeos anticancerígenos. Com base nos estudos avaliados observou-se que os métodos propostos na literatura seguem uma arquitetura comum. Nesses estudos, inicialmente, realiza-se a escolha de um conjunto de dados; posteriormente, é proposta a utilização de descritores de características dos peptídeos, em seguida é proposta a aplicação de algum algoritmo classificador, e por fim é realizada a construção dos modelos e a comparação de desempenho. Especificamente, observou-se que os dois pontos principais de distinção entre os trabalhos avaliados são a definição dos descritores de características e dos algoritmos de indução de modelos. Assim, no contexto dos estudos avaliados nesta revisão da literatura, buscou-se identificar o conjunto de dados utilizado; a dimensionalidade dos conjuntos de treinamento e teste; os principais descritores de características utilizados o algoritmo de aprendizagem utilizado; os métodos de avaliação; e os resultados obtidos (Quadro 2)

O primeiro estudo analisado desta revisão com o título "*Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model",* proposto por (LI QINGWEN et al., 2020), buscou realizar uma experimentação tendo como base o efeito da dimensionalidade na predição de peptídeos, utilizando o MRMD 2.0 para redução da dimensionalidade. No entanto, o estudo demonstrou que a utilização de uma quantidade maior de descritores apresenta melhor desempenho. Embora existam diversos preditores de sequências anticancerígenas utilizando ferramentas de Machine Learning para a predição, observou-se, nessa revisão, que os pesquisadores buscaram a utilização de algoritmos de *Deep Learning* como uma proposta para o aprimoramento da predição. No artigo, "*ACP-GCN: The Identification of Anticancer Peptides Based on Graph Convolution Networks***",** realizado por (RAO et al., 2020), os autores propuseram o preditor "ACP-GCN". Nesse estudo, apenas a técnica *One-hot-encoding* foi utilizada para extração de características para a construção dos modelos foi utilizado o algoritmo *Graph Convolutional Networks* (GCNs), o qual apresentou um resultado superior aos outros classificadores de *Deep Learning* tradicionais. Seguindo a base da experimentação com técnicas de *Deep Learning,* no trabalho "*DeepACP: A*

Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm", proposto por (YU et al. 2020), os autores propuseram a avaliação do desempenho de três arquiteturas, dentre elas a *Convolutional Neural Network* (CNN), *Recurrent Neural Network* (RNN) com *Bidirectional long short-term memory cells* (biLSTMs) e a combinação CNN-RNN, para a construção do preditor "DeepACP" A partir dos resultados obtidos no estudo, verificou-se que a arquitetura de RNN-biLSTMs apresentou desempenho superior. Já no trabalho "*Anticancer peptides prediction with deep representation learning features*" (LV et al. 2021) os autores propuseram o preditor iACP-DRLF, o qual utiliza o algoritmo *Light gradient boosting machine* (LGBM). O diferencial desse estudo está na utilização de dois modelos diferentes de incorporação das sequências, sendo o S*oft symmetric alignment* (SSA) e UniRep. Seguindo a linha de utilização de aprendizagem profunda, no trabalho intitulado "*Development of Predictive Tools for Anti-Cancer Pepetide Candidates using Generative Machine Learning Models*", os autores propuseram (LU; GIBSON, 2021) a construção de um modelo generativo baseado no algoritmo *Long short-term memory networks* (LSTM), com o objetivo de criar um modelo que pudesse aprender o que é ACP e posteriormente gerar sequências candidatas a ser um ACP. Buscando inovar com relação a predição de ACP no trabalho "ACP-DA: *Improving the Prediction of Anticancer Peptides Using Data Augmentation*" (CHEN et al., 2021), os autores propuseram o aumento de quantidade de sequências com a técnica *Noise adding oversampling* (NAO) para a construção da ferramenta preditiva ACP-DA, baseada no algoritmo Multilayer perceptron (MLP), no entanto, os autores reportaram baixo desempenho preditivo. Com a intenção de identificar padrões sem a utilização descritores pré-definidos, o trabalho "CL-ACP: *A parallel combination of CNN and LSTM anticancer peptide recognition model*", proposto por (WANG et al., 2021) apresentaram um método de combinação de "*Long short-term memory networks"* (LSTM) e CNN para aprender diretamente sobre os dados das sequências peptídicas Como resultado, a ferramenta CL-ACP apresentou resultados superiores de predição quando comparado aos algoritmos de *Machine Learning* tradicionais aos algoritmos baseados em redes convolucionais. Tentando aprimorar os estudos a respeito da utilização de algoritmos de *Deep Learning* e a utilização de diferentes descritores de características, no trabalho "*DLFF-ACP: prediction of ACPs based on deep learning and multi-view features fusion",* proposto por (CAO, RUIFEN et al., 2021), buscou-se

construir uma ferramenta de predição chamada DLFF-ACP, na qual foi realizada a extração de características com descritores frequentemente utilizados na literatura e posteriormente os combinaram com uma CNN para aprimorar a extração. No entanto, quando os autores compararam o seu potencial preditor com outras ferramentas, constataram um resultado inferior. Assim como em trabalhos anteriores, o trabalho "*ACPNet: A Deep Learning Network to Identify Anticancer Peptides by Hybrid Sequence Information***",** proposto por (SUN et al. 2022), os autores avaliaram a utilização de diferentes descritores de características organizados em categorias como: recursos selecionados manualmente (exemplo: tamanho da sequência); recursos físicos químicos" (exemplo: escala de pH) e recursos de codificação automática (exemplo: *one-hot-encoding*). Os autores avaliaram também a utilização das arquiteturas *Dense Convolutional Network* (Dense Net) e RNN. Com base nos resultados reportados os autores concluíram que a ACPNet, alcançou desempenho superior, quando comparada com outras ferramentas. No estudo "*Peptide-Based Drug Predictions for Cancer Therapy Using Deep Learning*", proposto por SUN et al. (2022), os autores propuseram a construção de uma ferramenta de predição chamada "Al4ACP", a qual baseou-se na utilização de 6 descritores de de características físico-químicas (método P6). Os autores adotaram uma CNN para a classificação, no entanto, os resultados de desempenho foram reportados como inferiores em comparação a algoritmos de *Machine Learning* tradicionais. Apesar de muitos trabalhos voltarem seus estudos apenas utilizando conjuntos de dados com peptídeos consideradas Anticâncer (ACP) e Não Anti-câncer (NACP), é importante ressaltar que, tratando-se da temática "câncer" temos que esse termo engloba diferentes tipos de neoplasias malignas, as quais podem estar localizados em diferentes partes do organismo como, por exemplo na mama e no sangue. Nesse contexto, no trabalho "*Breast and Lung Anticancer Peptides Classification Using N-Grams and Ensemble Learning Techniques***",** os autores ABBAS et al. (2022) propuseram a avaliação de sequências anticancerígenas com atividade específica, já previamente comprovadas a nível de bancada, para as linhagens celulares de câncer de mama e de pulmão. Nesse estudo, os autores propuseram a utilização de descritores *K-mers*. Os autores avaliaram algoritmos de classificação tradicionais de *Machine Learning*, dentre os quais, o SVM apresentou resultados superiores em termos de desempenho. Em WU et al. (2022), como o trabalho "*Anticancer Peptide Prediction via Multi-Kernel CNN and*

Attention Model", os autores propuseram um modelo de de predição chamado de "ACP-MCAM", o qual utiliza também o algoritmo CNN para aprender automaticamente características a partir das sequências peptídicas. No estudo de ALSANEA et al. (2022), "*To Assist Oncologists: An Efficient Machine Learning-Based Approach for Anti-Cancer Peptides Classification*" os autores propuseram a utilização de diferentes descritores de características. Para a predição de ACPs foi utilizada a técnica de *Ensemble* para combinar os classificadores como *support vector machines*, *random forests* e *naive bayes*. No trabalho "*ME-ACP: Multi-view neural networks with ensemble model for identification of anticancer peptides*", os autores FENG et al. (2022) propuseram a construção do preditor "ME-ACP" que utiliza em sua arquitetura uma combinação dos algoritmos "*Multiple light boosting machine*" (lightGBMs), "*Hybrid Deep Neural Network*" (HDNN) e "*Bi-directional Long ShortTerm Memory*" (Bi-LSTM). Neste estudo, foram extraídas características a nível residual e a nível global das sequências peptídicas. No trabalho "*ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network*", os autores GHULAM et al. (2022) propuseram um novo modelo preditor chamado ACP-2DCNN, o qual possui como classificador uma Rede Neural Convolucional Bidimensional (2DCNN). Além disso, utilizaram a extração de características locais e globais com o método *Dipeptide Deviation from Expected Mean*. Assim como muitos autores da literatura buscaram a criação de um novo modelo preditor, alguns também buscaram o aprimoramento de ferramentas existentes No trabalho "*MLACP 2.0: An updated machine learning tool for anticancer peptide prediction"*, os autores PHAN et al. (2022) propuseram o aperfeiçoamento de um modelo criado anteriormente chamado MLACP 2.0 cujo a melhoria do preditor está na mudança dos extratores de características, no entanto, apesar de apresentar bom desempenho, o modelo proposto possui a limitação de não conseguir prever peptídeos com mais de 50 resíduos de aminoácidos. Em "*ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction*", os autores HAN et al. (2022) propuseram um novo modelo de preditor chamado "ACPred-BMF" No modelo proposto foi utilizada uma bi-LSTM como classificador. Além disso, o estudo propôs um processo de combinação de descritores qualitativos e quantitativos dos peptídeos. E por último nessa revisão, temos o trabalho "*GCNCPR-ACPs: a novel graph convolution network method for ACPs*

prediction”**,** no qual os autores WU et al. (2022) propuseram um modelo nomeado de GCNPR-ACPs, um preditor baseado no algoritmo “*graph convolution network”* (GCN), que explora a utilização de recursos físicos-químicos e métodos de codificação como o*ne-hot-encoding*.

Quadro 2 - DETALHAMENTO DOS TRABALHOS RELACIONADOS

(continua)

Ano de Publicação	Nome do artigo	Conjunto de dados (CD)[1]	Tamanho dos CD de referência/ Treinamento[2]	Principais recursos[3]	Número de recursos	Modelo de aprendizagem[4]	Método de avaliação	Acc[5]
2020	Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model	Dataset de Hajisharifi et al.	138(ACP) e 206(NACP)	Composição de aminoácidos (AAC), tríade conjunta (CT), composição de pseudo-aminoácidos (PAAC), composição de aminoácidos agrupados (GAAC) e C/T/D(Solubilidade, estrutura secundária e hidrofobicidade relativa)	10	SVM	10-fold-cross-validation	92,73%
2020	ACP-GCN: The Identification of Anticancer Peptides Based on Graph Convolution Networks	Dataset de Wei et al.'s	250(ACP) e 250(NACP)	One-hot-enconding	1	GNN	5-fold-cross-validation	77,80%
2020	DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm,	Dataset de ACPred-FL	250(ACP) e 250(NACP)	word2Vec	1	RNN	Treino e teste	82.9%
2021	Anticancer peptides prediction with deep representation learning features	Dataset de AntiCP 2.0	861(ACP) e 861(NACP)	Pretrained SSA e Pretrained UniRep	3	LGBM	5-fold-cross-validation	79,10%
2021	Development of Predictive Tools for Anti-Cancer Peptide Candidates using Generative Machine Learning Models	Dataset de DRAMP;APD;CancePPD e UniProt	667(ACP) e 584(NACP)	Biblioteca PydPi e One-hot-enconding	2050	SVM	Treino e teste	90,04%
2021	ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation	Dataset de ACPred-FL e ACP-DL	376(ACP) e 364(NACP)	BPFs, AAindex e K-mer sparse matrix	9	MLP	5-fold-cross-validation	90%
2021	CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model	Dataset de ACP-DL; Antimicrobial peptide date (APD); collection of anti-microbial peptides(CAMP);database of anuran defense peptides(DADP).	189(ACP) e 350(NACP)	One-hot-enconding	1	CNN+LSTM	5-fold-cross-validation	87,92%

Quadro 2 - DETALHAMENTO DOS TRABALHOS RELACIONADOS

(conclusão)

Ano de Publicação	Nome do artigo	Conjunto de dados (CD)[1]	Tamanho dos CD de referência/ Treinamento[2]	Principais recursos[3]	Número de recursos	Modelo de aprendizagem[4]	Método de avaliação	Acc[5]
2022	MLACP 2.0: An updated machine learning tool for anticancer peptide prediction	-	1084(ACP) e 1084(NACP)	Composição de dipeptídeos(DPC); Desvio de dipeptídeos da média esperada(DDE); Composição de aminoácidos(AAC); Transição e distribuição de composição(CTDC, CTDT e CTDD); one-hot-enconding; word2vec e etc.	17	CNN	10-fold-cross-validation	76,50%
2022	ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction	Dataset de AntiCP 2.0	861(ACP) e 861(NACP)	Binary profile feature (BPF); Quantitative properties of amino acids (Quanc); Qualitative properties of amino acids (Qualc).	15	bi-LSTM	5-fold-cross-validation/ Treino e teste	80.81%
2022	GCNCPR-ACPs: a novel graph convolution network method for ACPs prediction	Dataset de ACPred-FL	250(ACP) e 250(NACP)	One-hot-enconding; node2vec; propriedades físico-química	12	GCN	10-fold-cross-validation/ Treino e teste	84,60%

Fonte: A autora (2023).

Nota:

[1]Conjunto de dados – Banco de dado de onde foram extraídas as sequências de peptídeos.

[2]Tamanho dos CD de referência/ Treinamento – Referente a dimensionalidade do banco de dados utilizados para treinamento dos modelos construídos.

[3]Principais recursos – Referente à descritores de características, sendo esses responsáveis por extrair informações da molécula.

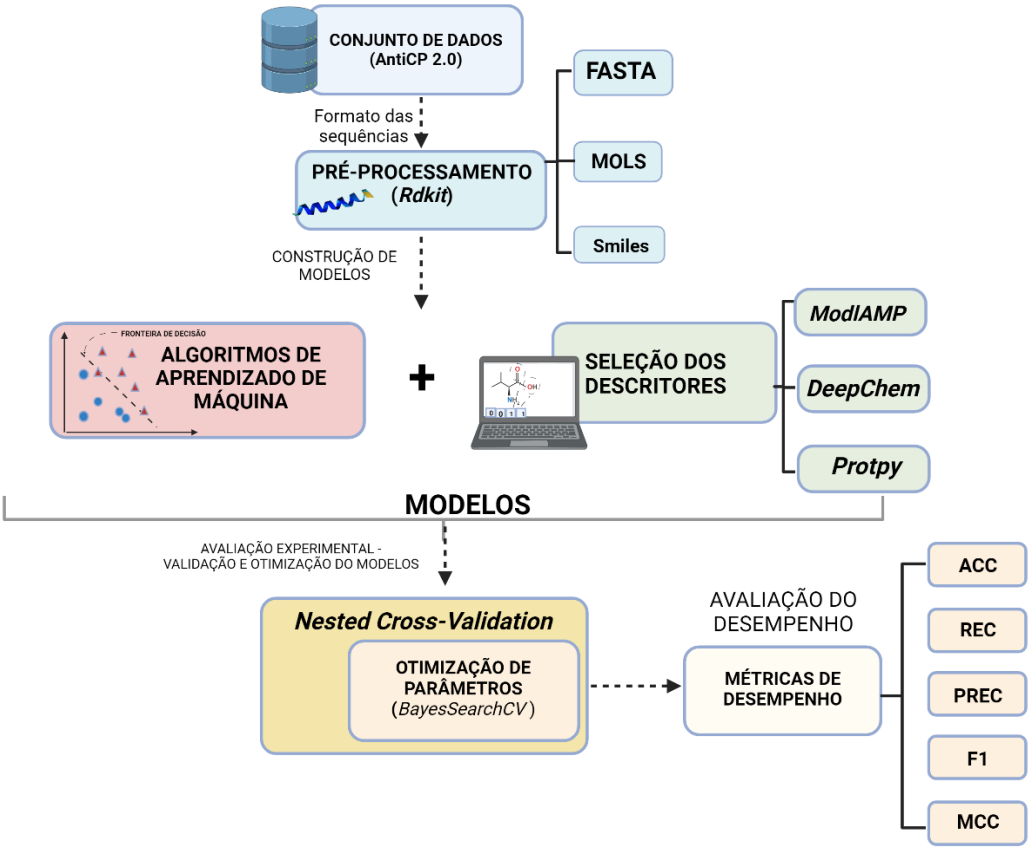
[4]Modelo de aprendizagem – Referente aos algoritmos que tiveram desempenho superior no estudo.

[5]Acc – Referente ao desempenho dos modelos de aprendizagem em termos de acurácia.

3. MATERIAIS E MÉTODO

Nesta seção, será detalhada a arquitetura implementada neste estudo. Essa arquitetura é composta pelas etapas de Conjunto de Dados, Pré-processamento dos Dados, Seleção dos Descritores de Características, Construção de modelos, Avaliação Experimental, Otimização de Parâmetros e a Configuração Experimental e Validação dos Modelos. A Figura 5 contém uma ilustração da arquitetura desse estudo.

Figura 5 – ETAPAS DA ARQUITETURA IMPLEMENTADA



Fonte: A autora, 2023.

3.1. CONJUNTO DE DADOS

Neste estudo, foi utilizado o conjunto de dados proveniente do banco de dados AntiCP 2.0[4]. Os dados disponibilizados pelo AntiCP 2.0 são de acesso

[4] https://webs.iiitd.edu.in/raghava/anticp2/download.php

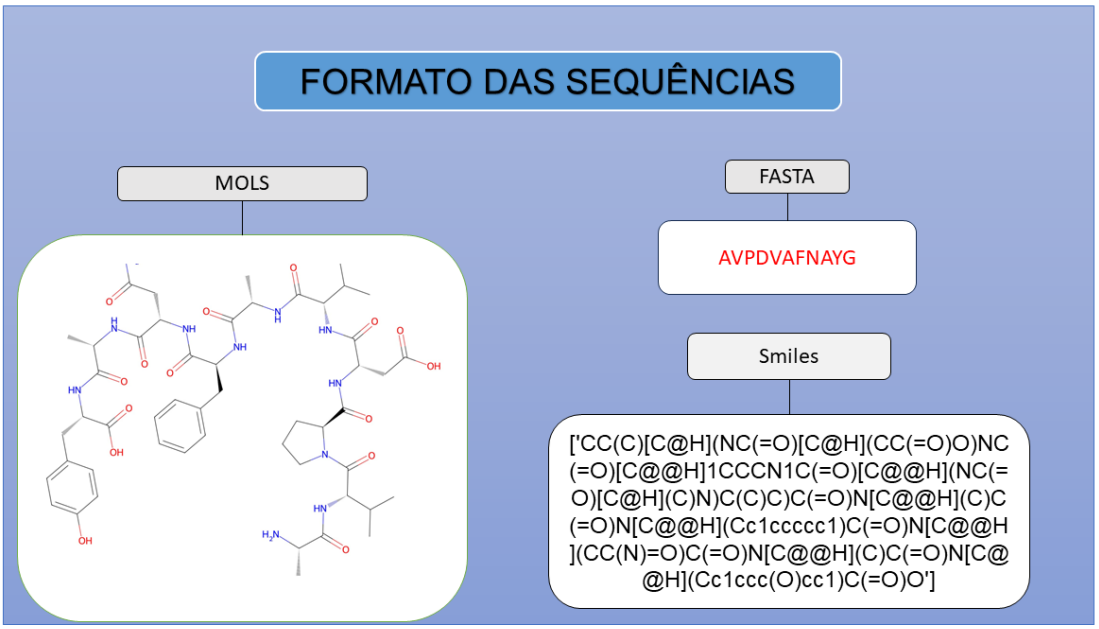
público e foram construídos utilizando sequências peptídicas coletadas de trabalhos da literatura, tais como: ACP-DL, ACPP, ACPred-FL, AntiCP, iACP e CancerPPD. Especificamente, neste estudo, foram selecionados os dados do conjunto denominados *Main Dataset* do AntiCP 2.0, o qual contém 1722 sequências peptídicas de tamanhos variados, entre 2 a 50 resíduos de aminoácidos, representadas em formato FASTA. Desse total, 861 sequências são Peptídeos Anti-Câncer (ACP), que neste estudo são considerados como a classe positiva, e 861 são Peptídeos Antimicrobianos (AMP), que não apresentam atividade Anticâncer e que neste estudo são considerados a classe negativa. É importante ressaltar que os dados disponibilizados pela AntiCP 2.0 foram previamente validados experimentalmente (AGRAWAL et al., 2021).

3.2. PRÉ-PROCESSAMENTO DOS DADOS

Nesta etapa de tratamento dos dados, foram realizadas modificações dos dados coletados da AntiCP 2.0, considerando os tipos de descritores e os tipos de algoritmos de classificação a serem utilizados. Desse modo, com relação aos descritores, foram utilizadas as representações FASTA; Smiles e MOLS. O formato FASTA foi aplicado para os descritores provenientes dos módulos *Modlamp e Protpy*. O formato Smiles e MOLS foram utilizados para descritores provenientes do módulo *DeepChem*. Neste caso, as sequências FASTA foram convertidas para Smiles e MOLS por meio do módulo RDkit[5]. Os descritores de características utilizados por meio desses módulos permitem extrair informações físico-químicas e estruturais das moléculas, organização e distribuição dos aminoácidos. Na Figura 6, estão representados os diferentes formatos utilizados neste trabalho, onde todos os formatos utilizados são uma representação da mesma sequência peptídica, sendo essa sequência pertencente à classe dos peptídeos anti câncer.

[5] https://www.rdkit.org/docs/source/rdkit.Chem.rdmolfiles.html

Figura 6 - FORMATO DAS SEQUÊNCIAS



Fonte: A autora, 2023.

3.3. DESCRITORES DE CARACTERÍSTICAS

Nesta seção, detalhamos sobre os pacotes e descritores de características utilizados neste estudo. Desse modo, ressaltamos que, utilizamos uma diversidade de descritores de características, provenientes de 3 pacotes distintos disponíveis na linguagem *Python* 3. A utilização desses descritores está vinculada a necessidade de representar uma molécula considerando sua forma estrutural assim como suas propriedades químicas e/ou biológicas em uma forma numérica ou simbólica. Os descritores utilizados neste estudo foram selecionados após uma análise dos trabalhos relacionados (Quadro 2).

3.3.1. ModlAMP

O pacote *ModlAMP*[6] oferece diferentes ferramentas para desenvolvimento de pesquisas *in silico* para facilitar a descoberta e o design de novos AMPs sintéticos. A biblioteca disponibiliza funções para calcular uma variedade de diferentes propriedades físico-química das moléculas, além de descritores de

[6] https://modlamp.org/

peptídeos baseados em resíduos de aminoácidos. Especificamente, neste trabalho, utilizamos os descritores provenientes da classe *PeptideDescripto*r, os quais podem ser visualizados no Quadro 4. Neste trabalho foi utilizada uma combinação de todos os descritores.

Quadro 3- DESCRITORES FÍSICO-QUÍMICOS PROVENIENTES DO PACOTE MODLAMP

(continua)

Descritores	**Descrição do descritor**
AASI	Índice de seletividade de aminoácidos
ABHPRK	escala de características físico-químicas internas do modlabs
Argos	escala de aminoácidos de hidrofobicidade de Argos
Bulkiness	escala de volume da cadeia lateral de aminoácidos
Charge_phys	carga de aminoácidos em pH 7,0 - carga de histidina +0,1.
Charge_acid	carga de aminoácidos em pH ácido - carga de histidina +1,0
Cougar	seleção interna do modlabs de descritores globais de peptídeos
Eisenberg	Escala de consenso de Eisenberg
Ez	Energia de inserção da bicamada lipídica
Flexibility	Flexibilidade da cadeia lateral
Grantham	composição da cadeia lateral de aminoácidos, polaridade e volume molecular
Gravy	escala de aminoácidos de hidrofobicidade GRAVY
Hopp-woods	escala de hidrofobicidade de aminoácidos de Hopp-Woods
ISAECI	Área de superfície isotrópica – índice de carga eletrônica
Janin	Hidrofobicidade
Kytedoolittle	Hidrofobicidade
Levitt_alpha	propensão α-helicoidal
MSS	Forma e tamanho topológico da cadeia lateral
MSW	Principais componentes das propriedades dos resíduos estéricos e 3D
PepArc	Características farmacofóricas modlabs hidrofobicidade, polaridade, carga positiva, carga negativa, prolina
Pepcats	Características farmacofóricas binárias
Polarity	Polaridade de aminoácidos
PPCALI	Principais componentes das propriedades selecionadas da cadeia lateral

Quadro 3 - DESCRITORES FÍSICO-QUÍMICOS PROVENIENTES DO PACOTE MODLAMP

(conclusão)

Descritores	Descrição do descritor
Refractivity	Valores relativos de refratividade
T_scale	PCA com base na cadeia lateral de aminoácidos com valores do GRID
TM_tend	Propensão transmembrana
Z3	Escala Z tridimensional original
Z5	Escala Z pentadimensional estendida

Fonte: A autora, 2023.

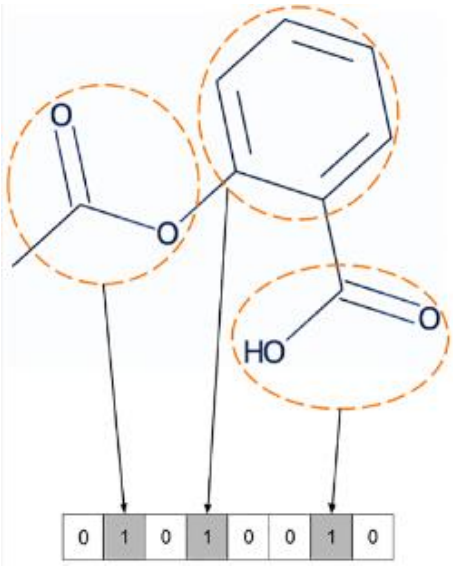
3.3.2. DeepChem

Desenvolvido por pesquisadores de Stanford, o pacote *DeepChem*[7], possui diversas funcionalidades que possibilitam o desenvolvimento de novos medicamentos; assim como análises moleculares de modo geral (ALTAE-TRAN et al., 2017). O pacote disponibiliza diferentes descritores que realizam uma análise mais detalhada da estrutura da molécula. Além disso, possui diferentes descritores de características físico-químicas e descritores que descrevem a molécula 2D (RAMSUNDAR, 2022). Os descritores utilizados neste trabalho provenientes do *DeepChem* são:

I. ***MACCSKeys Fingerprint*:** Esse descritor é baseado na identificação de subestrutura chaves, sendo 166 chaves pré-definidas (CERETO-MASSAGUÉ et al., 2015). Essas chaves indicam a presença de grupos funcionais ou presença de heteroátomos como, por exemplo, elementos da família 7A (RDKit, 2023). Caso exista a presença é identificado como 1 e caso a ausência 0. Na Figura 7 é possível visualizar como o descritor descreve uma molécula.

[7] https://deepchem.readthedocs.io/en/latest/index.htm

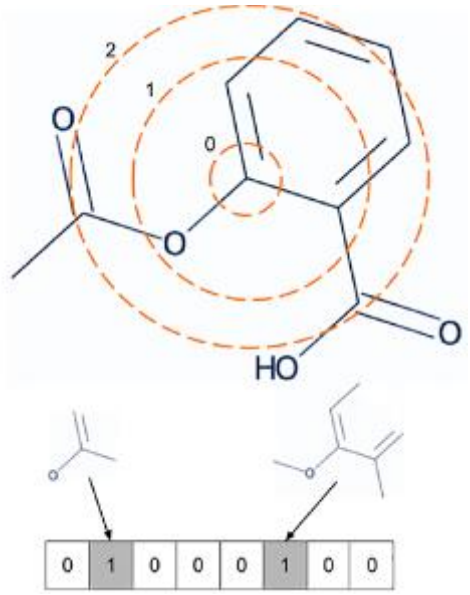
Figura 7 - MACCKEYS FINGERPRINT



Fonte: BAPTISTA et al., 2023.

II. ***Circular Fingerprint*:** Essa estratégia também conhecida por *Extended-connectivity fingerprints* (ECFPs), consiste em uma forma de representação molecular, na qual as moléculas são divididas em fragmentos circulares únicos e separados. A análise ocorre a partir do átomo inicial, e pode ampliar a análise quando em fragmentos de raio superior, desse modo, levando em consideração os átomos vizinhos (BAPTISTA et al., 2022). Nesse caso, cada raio equivale a 1 a distância de ligações dos átomos, ou seja, raio 1 significa que está distante por uma ligação do átomo. Cada fragmento circular gera um valor numérico, como exemplo, representando a presença de ligação dupla (ROGERS; HAHN, 2010). Na Figura 8 é possível visualizar a forma como esse descritor descreve a molécula de forma estrutural.

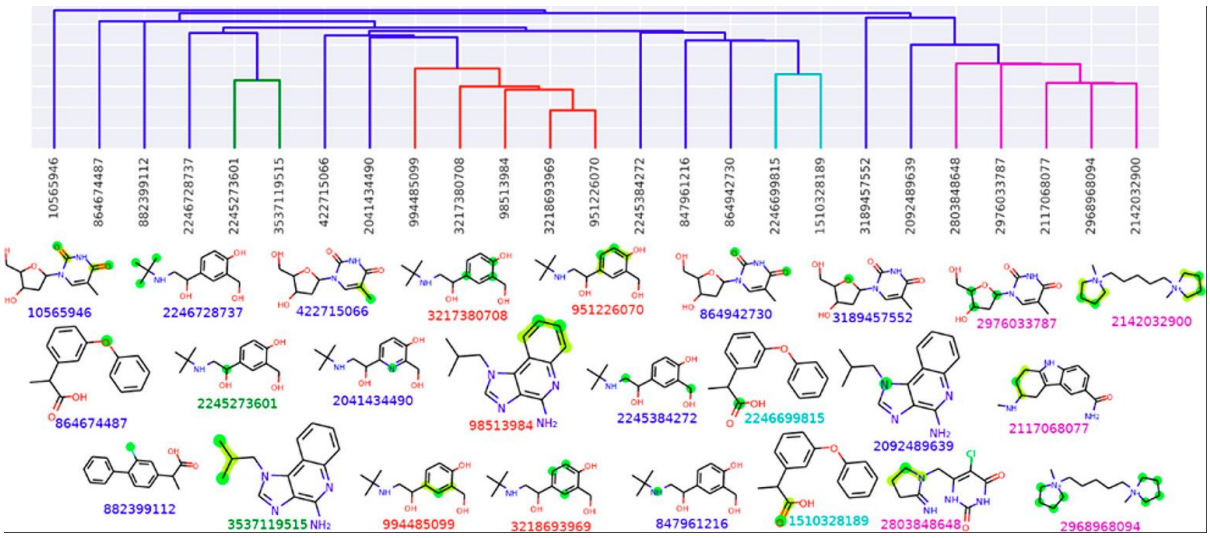
Figura 8 - CIRCULAR FINGERPRINT



Fonte: BAPTISTA et al., 2022.

III. ***Mol2VecFIngerprint*:** Refere-se a uma estratégia baseado em técnicas de processamento de linguagem natural (PNL). Esse método, considera subestruturas compostas pré-definidas derivadas do algoritmo de Morgan como “palavras” e compostos como “sentenças” (JAEGAR et al., 2018). Na Figura 9 é possível observar, por meio de um dendrograma, é possível observar a relação entre os vetores e como ocorre a descrição da molécula considerando suas subestruturas.

Figura 9 - DENDROGRAMA BASEADO NA RELAÇÃO ENTRE OS VETORES NO DESCRITOR MOL2VECFINGERPRINT



Fonte: JAEGAR et al., 2018.

IV. FASTA2SEQ: Este descritor realiza a transformação de uma sequência peptídica no formato FASTA para um vetor numérico. Na Figura 10 é possível observar como ocorre essa transformação.

Figura 10 - DESCRITOR FASTA2SEQ

Fonte: A autora, 2023.

V.SMILES2SEQ: Similar ao FASTA2SEQ, esse realiza a transformação de uma sequência no formato Smiles para um vetor numérico.

3.3.3. Protpy

O pacote *Protpy*[8], disponível para Python 3, possui descritores físico-químicos e estruturais que permitem expressar a composição e organização dos aminoácidos nas sequências peptídicas.

I. Codificação da Composição de Aminoácidos (AAC): Realiza o cálculo da frequência de cada tipo de aminoácido em uma sequência de peptídeo. As frequências para todos os 20 aminoácidos naturais, ou seja, a frequência dos aminoácidos (“ACDEFGHIKLMNPQRSTVWY”) podem ser calculadas como:

$$f(t) = \frac{N(t)}{N}, t \in \{A, C, D, \dots, Y\}$$

Onde $N(t)$, é o número de aminoácidos do tipo t, enquanto N, é o comprimento de uma sequência de peptídeos.

II. Composição Tripeptídica (TPC): A Composição de Tripeptídeos (TPC) apresenta 8000 descritores, definidos como:

$$f(r, s, t) = \frac{N_{rst}}{N - 2}, \qquad r, s, t \in \{A, C, D, \dots Y\}$$

Onde N_{rst} é o número de tripeptídeos representados pelos tipos de aminoácidos r, s e t

III. Composição de Dipeptídeos (DPC): A Composição de Dipeptídeos fornece 400 descritores. Ela é definida como:

$$D(r, s) = \frac{N_{rs}}{N-1,}, r, s \in \{A, C, D, \dots Y\}$$

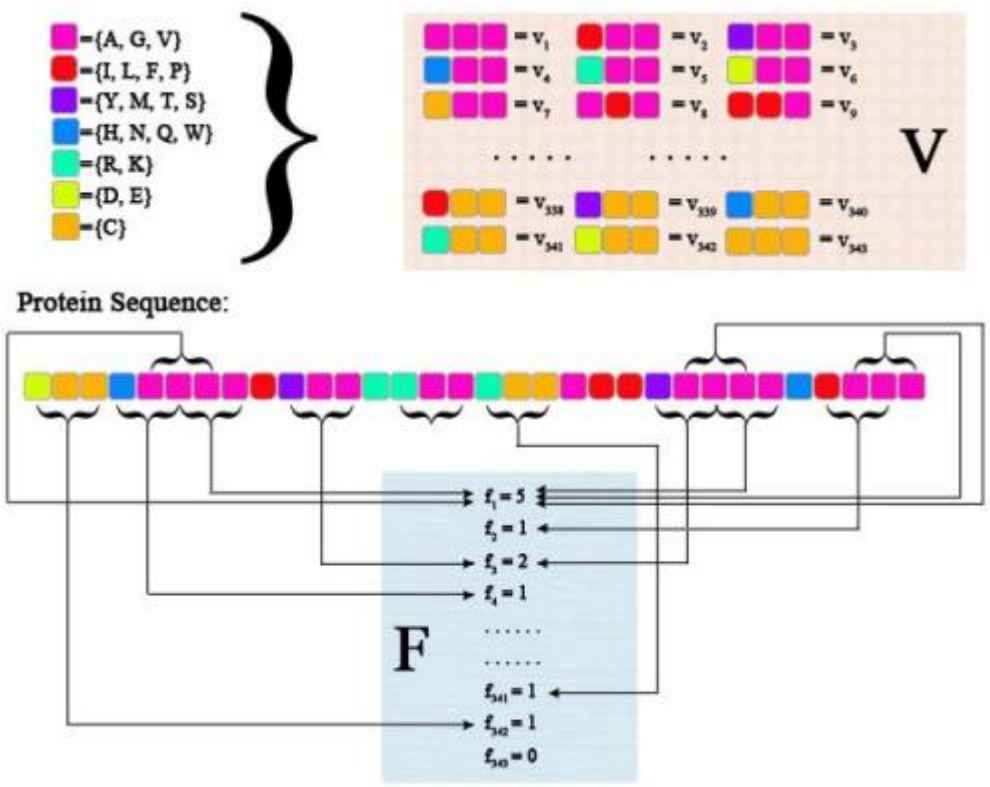
Onde N_{rs} é o número de dipeptídeos representados pelos tipos de aminoácidos r e s.

[8] https://pypi.org/project/protpy/#tests

IV. **Composição/Transição/Distribuição (CTD):** As características CTD representam os padrões de distribuição de aminoácidos com base em alguma propriedade estrutural ou físico-química específica em uma sequência do peptídeo. Foram previamente utilizados 13 tipos de propriedades físico-químicas para calcular essas características. Isso inclui hidrofobicidade, volume de Van der Waals normalizado, polaridade, polarizabilidade, carga, estruturas secundárias e acessibilidade ao solvente. Esses descritores são calculados de acordo com os seguintes procedimentos: (i) A sequência de aminoácidos é transformada em uma sequência de certas propriedades estruturais ou fisico-químicas dos resíduos; (ii) Vinte aminoácidos são divididos em três grupos para cada uma das sete diferentes propriedades físico-químicas com base nos principais agrupamentos dos índices de aminoácidos de Tomii e Kanehisa (TOMII et al., 1996).

V. **Tríade Conjugada (CTriad):** O descritor de Tríade Conjugada (CTriad) considera as propriedades de um aminoácido e seus aminoácidos vizinhos, considerando três aminoácidos contínuos como uma única unidade. Primeiro, a sequência de proteína é representada por um espaço binário *(V, F)*, onde *V* denota o espaço vetorial das características da sequência, e cada característica *Vi* representa um tipo de tríade; *F* é o vetor numérico correspondente a *V*, onde *fi*, o valor da *i*-ésima dimensão de *F*, é o número de vezes que o tipo *Vi* aparece na sequência do peptídeo. Para os aminoácidos que foram catalogados em sete classes, o tamanho de *V* deve ser igual a 7 x 7 x 7=343. De acordo com isso, $i = 1, 2, 3, \ldots, 343$. Um exemplo ilustrado desse esquema de codificação é fornecido na Figura 11.

Figura 11 - ESQUEMA DE CODIFICAÇÃO DO DESCRITOR CTRIAD



Fonte: CHEN et al., 2023.

VI. Composição de Pseudo-Aminoácidos (PAAC): Temos que $H_1^0(i), H_2^0(i), M^0(i)$ para $i = (1, 2, 3, \ldots 20)$ são os valores originais de hidrofobicidade, os valores originais de hidrofilicidade e as massas originais das cadeias laterais dos 20 aminoácidos naturais, respectivamente. Eles são convertidos para as seguintes quantidades por meio de uma conversão padrão:

$$H_1(i) = \frac{H_1^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^0(i)}{\sqrt{\frac{\sum_{i=1}^{20}(H_1^0(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^0(i))^2}{20}}}$$

Onde, $H_2^0(i)$ e $M^0(i)$ são normalizados e da mesma forma, Em seguida, uma função de correlação pode ser definida como:

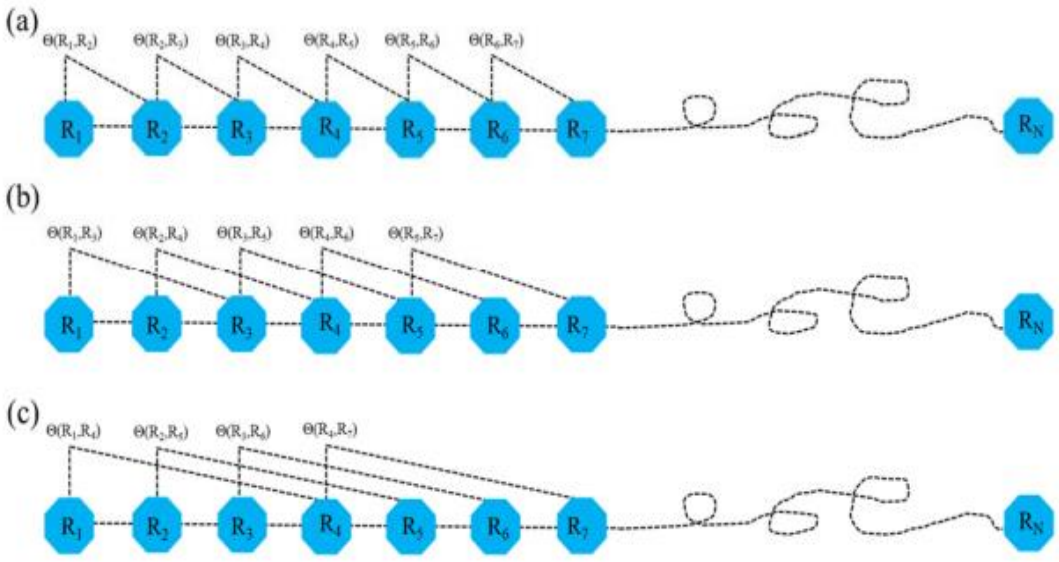
$$(R_i, R_J) = \frac{1}{3}\{[H_1(R_i) - H_1(R_j)]^2 + [H_2(R_i) - H_2(R_j)]^2 + [M(R_i) - M(R_j)]^2\}$$

Esta função de correlação é na verdade um valor médio para as três propriedades dos aminoácidos: valor de hidrofobicidade, valor de hidrofilicidade e massa da cadeia lateral. Para uma propriedade de aminoácido, a correlação pode ser definida como:

$$(R_i, R_J) = [H_1(R_i) - H_1(R_j)]^2$$

Onde, $H(R_i)$ é a propriedade do aminoácido $R_i$após a padronização. Na Figura 12 é representado um esquema de acoplamento de ordem de sequência, onde na primeira classe *a* temos um cálculo de correlação considerando 1° resíduo adjacente. Na classe *b* temos um cálculo de correlação considerando 2° resíduo adjacente. Na classe *c*, temos o cálculo de correlação considerando 3° resíduo adjacente.

Figura 12 - ESQUEMA DA FUNCIONALIDADE DO DESCRITOR PAAC



Fonte: CHEN et al., 2023.

VII. Composição de Pseudo-Aminoácidos Anfifílicos (APAAC): Similar ao PAAC. Porém realiza o agrupamento de Pseudo-Aminoácidos de caráter Anfifílico.

3.4. CONSTRUÇÃO DE MODELOS

Neste trabalho, utilizamos 17 algoritmos tradicionais de aprendizado de máquina, além de 1 algoritmo de aprendizado de máquina profundo para a tarefa de classificação. Os parâmetros de cada algoritmo utilizado, neste estudo, estão especificados no Quadro 5 do Apêndice 1. Os algoritmos utilizados estão especificados no Quadro 4.

Quadro 4 - ALGORITMOS DE APRENDIZADO DE MÁQUINA

(continua)

Sigla	Nome	Link	Referência
ADA	AdaBoost *Machine Learning Adaptive Boosting*	https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html	(FREUND; SCHAPIRE, 1997)
CATB	*CatBoost*	https://catboost.ai/	(JABEUR et al., 2021)
DT	*Decision Tree*	https://scikit-learn.org/stable/modules/tree.html	(BREIMAN, 2001)
ET	*Extra Trees: Extremely Randomized Trees*	https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html	(GEURTS et al., 2006)
GBC	*Gradient Boosting Classification*	https://scikit-learn.org/stable/modules/tree.html	(RASMUSSEN; WILLIAMS, 2006)
GPC	*Gaussian Process Classification*	https://scikit-learn.org/stable/modules/gaussian_process.html#gaussian-process-classification-gpc	(SUBASI et al., 2020)
kNN	*K-Nearest Neighbors*	https://scikit-learn.org/stable/modules/neighbors.html#classification	(COVER; HART, 1967)
LDA	*Linear Discriminant Analysis*	https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers	(HASTIE; TIBSHIRANI; FRIEDMAN, 2001)
LGBM	*LightGBM: Light Gradient Boosting Machine*	https://lightgbm.readthedocs.io/en/stable	(KE et al., 2017)
LR	*Logistic Regression*	https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression	(RUSSEL; NORVIG, 2013)

Quadro 4 - ALGORITMOS DE APRENDIZADO DE MÁQUINA

(conclusão)

Sigla	Nome	Link	Referência
MLP	*Multi-Layer Perceptron*	https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier	(HAYKIN, 2001)
NB	*Naive Bayes*	https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes	(RUSSEL; NORVIG, 2013)
QDA	*Quadratic Discriminant Analysis*	https://scikit-learn.org/stable/modules/lda_qda.html#mathematical-formulation-of-the-lda-and-qda-classifiers	(HASTIE; TIBSHIRANI; FRIEDMAN, 2001)
RF	*Random Forest*	https://scikit-learn.org/stable/modules/ensemble.html#random-forests	(BREIMAN et al., 1984)
Ridge	*Ridge*	https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression-and-classification	(HOERL; KENNARD, 1970)
SVM	*Support Vector Machines*	https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html	(CORTES; VAPNIK, 1995)
XGB	*XGBoost: eXtreme Gradient Boosting*	https://xgboost.readthedocs.io/en/stable	(CHEN; GUESTRIN, 2016)
1D-CNN	*1D Convolutional Neural Networks*	https://www.sktime.net/en/stable/api_reference/auto_generated/sktime.classification.deep_learning.CNNClassifier.html	(Zhao et al., 2017)

Fonte: A autora, 2023.

A arquitetura de 1D-CNN é utilizada principalmente para o processamento de dados unidimensionais, onde essa aplica convoluções em uma única dimensão de dados. De modo geral, uma 1D-CNN realiza um “deslizamento” de um filtro (kernel) unidimensional ao longo de um dado, de modo que seja calculado produtos escalares e seja criado mapas de recursos (FACELI et al., 2021, 2ª edição). Neste estudo foi utilizada a implementação 1D-CNN disponível na biblioteca *sktime*[9].

3.5. AVALIAÇÃO EXPERIMENTAL

[9] https://www.sktime.net/en/stable/api_reference/classification.html

Neste trabalho, foi realizada uma ampla avaliação experimental envolvendo os descritores de características apresentados na seção 3.3 e todos os algoritmos listados na seção 3.4. Para delimitar essa fase, os experimentos foram organizados de acordo com o tipo de informação a ser tratada pelos algoritmos de AM. Considerando este cenário, o algoritmo CNN foi avaliado com os descritores Mol2Vec, FASTA2SEQ e SMILES2SEQ, os quais possibilitam reter informações de dependência entre os atributos. A aplicação do algoritmo CNN juntamente com esses descritores tem como objetivo explorar a existência de relações sequenciais/espaciais entre as variáveis de entrada. Por outro lado, os algoritmos tradicionais de aprendizado de máquina tratam as variáveis de entrada como independentes entre si. Isso significa que, durante o processo de indução do modelo, não há consideração explícita das relações de dependência posicional/espacial entre as diferentes variáveis de entrada. Nesse sentido, todos os algoritmos de aprendizado tradicionais foram avaliados em relação a todos os descritores.

3.5.1. Métricas de desempenho

Para a avaliação dos modelos de classificação desenvolvidos neste estudo, empregamos diversas métricas de desempenho. Dentre elas, acurácia (Acc); precisão (Prec); recall (Rec) (usualmente conhecida por sensibilidade); medida F1 (F1) e Correlação de Matthews (Mcc). Neste trabalho, os modelos foram construídos considerando a classificação de duas classes, ou seja, uma classificação binária. Desse modo, as métricas são calculadas com base nos parâmetros: Verdadeiros Positivos (VP); Verdadeiros Negativos (VN); Falsos Positivos (FP); Falsos Negativos (FN). VP é correspondente ao número de exemplos da classe positiva (anticâncer) que foram classificadas corretamente; VN é correspondente ao número de exemplos da classe negativa (não anticâncer) que foram classificadas corretamente; FP refere-se ao número de exemplos que foram classificados como positivos, mas que são pertencentes à classe negativa; FN refere-se ao número de exemplos que foram classificados como negativos, mas são pertencentes a classes positiva. O total de exemplos é dado pela equação $n = VP + VN + FP + FN$. A descrição de cada métrica utilizada neste estudo pode ser observada abaixo (FACELI et al., 2021, 2ª edição).

- **Acurácia (Acc)**: É obtida somando o número de exemplos corretamente classificadas para todas as classes e, em seguida, dividindo pelo número total de exemplos

$$acc = \frac{(VP + VN)}{n}$$

- **Precisão (Prec)**: É referente a proporção de exemplos classificados corretamente como positivos com relação a todos classificados como positivos. De modo geral, refere-se a capacidade do preditor atribuir corretamente uma classe ao exemplo que realmente é pertencente a essa classe.

$$prec = \frac{VP}{(VP + FP)}$$

- **Recall (Rec):** É uma medida que expressa a proporção de objetos que foram corretamente identificados como pertencentes a uma classe específica em relação à soma de todos os objetos que realmente fazem parte dessa classe. Em termos simples, avalia a capacidade do modelo em capturar todos os objetos que genuinamente pertencem a uma classe.

$$rec = \frac{VP}{(VP + FN)}$$

- **Pontuação F1 (F1)**: A Pontuação F1 é uma métrica que combina precisão e recall por meio de uma média harmônica, destacando a importância de um equilíbrio entre ambas as medidas. Isso a torna particularmente útil quando se busca um compromisso entre identificar corretamente os verdadeiros positivos e evitar falsos positivos.

$$f1 = \frac{(2 \times prec \times rec)}{(prec + rec)}$$

- **Correlação de Matthews (Mcc):** É uma medida útil para avaliar o desempenho de classificadores binários, levando em consideração tanto os verdadeiros positivos quanto os verdadeiros negativos, bem como os erros de classificação

(falsos positivos e falsos negativos). Essa medida varia entre +1 e -1. Um coeficiente de 1 indica uma predição perfeita; 0 não é melhor do que uma predição aleatória; e -1 indica discordância total entre predição e observação. (CHICCO et al., 2020).

$$mcc = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VP + FN)}}$$

3.6. OTIMIZAÇÃO DE PARÂMETROS

Neste estudo, com o propósito de determinar os melhores parâmetros para cada um dos algoritmos de classificação utilizados, foi proposta a aplicação da abordagem de Otimização Bayesiana de Hiperparâmetros (SNOEK et al., 2012). A ideia desse método consiste em identificar o valor máximo de uma função desconhecida, neste contexto, o desempenho de um algoritmo de aprendizado de máquina, através de um número mínimo de iterações. Inicialmente, é formulada uma hipótese inicial sobre a função e, posteriormente, essa hipótese é continuamente atualizada para formar uma distribuição posterior que incorpora as informações obtidas a partir dos dados observados. A cada iteração, a função é otimizada para selecionar o próximo ponto de consulta, visando alcançar a maior melhoria possível em relação à melhor observação até então. Esse processo iterativo é repetido até que um critério de parada seja satisfeito, que pode ser um número pré-determinado de iterações ou quando as melhorias se tornam insignificantes. Neste estudo, implementamos essa abordagem utilizando a função *BayesSearchCV* da biblioteca *scikit-optimize*, configurada com um total de 10 iterações e utilizando a acurácia como métrica de desempenho. Os parâmetros utilizados para cada um dos 18 algoritmos estão descritos na Quadro 6 no Apêndice 1

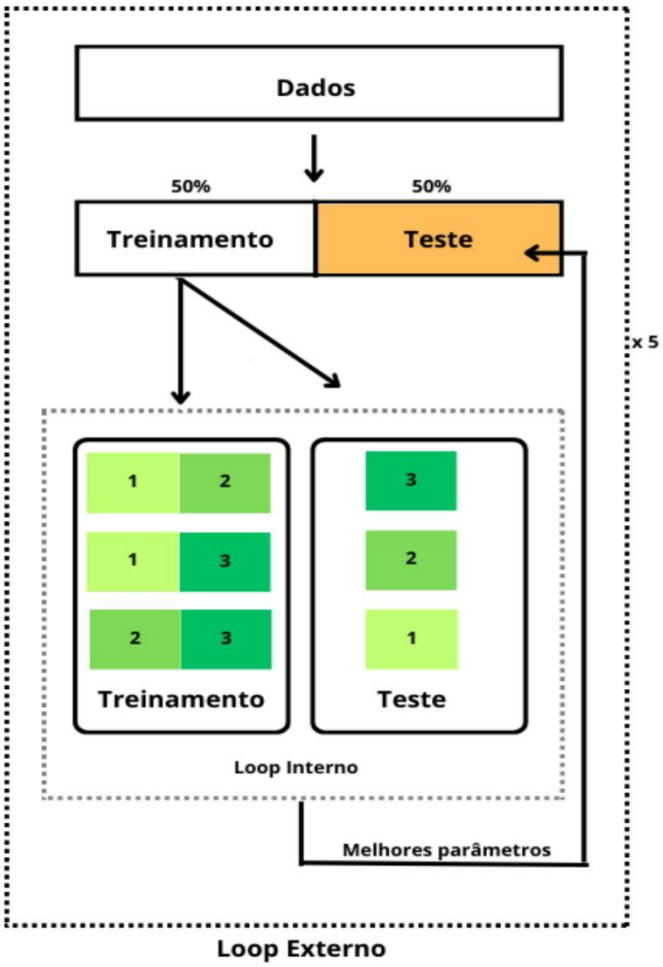
3.7. CONFIGURAÇÃO EXPERIMENTAL E VALIDAÇÃO DOS MODELOS

Neste trabalho, foi utilizada a técnica de *K-fold Cross-Validation* (CV). Essa técnica de análise consiste na divisão em k partições, das quais os dados de k-1 partições são utilizados no treinamento de um preditor, e posteriormente os dados são testados nas partições restantes. Esta operação é repetida k vezes, utilizando em

cada ciclo uma partição diferente, a fim de obter no final uma média de desempenho. No entanto, existe a possibilidade de o modelo utilizar os mesmos dados empregados para o ajuste de parâmetros na avaliação do desempenho. Logo, gerando uma estimativa potencialmente enviesada do desempenho preditivo do modelo (CAWLEY; TALBOT, 2010). Uma alternativa para evitar esse problema é a utilização da técnica *Nested Cross-Validation*. Essa estratégia consiste em aplicar uma dupla validação cruzada. Desse modo, um processo de validação cruzada de *cv_outer* partições (loop externo) é utilizado para dividir os dados em conjuntos de treinamento e teste. Sobre o conjunto de treinamento é aplicado um segundo processo de validação cruzada de *cv_inner* partições (loop interno), com o propósito de aprimorar os parâmetros do modelo. Posteriormente, utilizando os melhores parâmetros estimados, o modelo é construído e avaliado, em termos de desempenho, com base nos dados do conjunto de teste (PARVANDEH et al., 2020). Em ambos os processos foi aplicada a técnica de estratificação, para que cada partição tivesse a mesma distribuição de classes que o conjunto de dados original.

Neste trabalho, os parâmetros utilizados para a estratégia *Nested Cross-Validation* foram *cv_inner*=3, e *cv_outer*=2. Essa estratégia de validação cruzada foi repetida 5 vezes, totalizando 10 estimativas de desempenho para cada avaliação realizada. A configuração experimental utilizada neste trabalho pode ser visualizada por meio de uma representação esquemática apresentada na Figura 13.

Figura 13 - ORGANIZAÇÃO EXPERIMENTAL



Fonte: A autora, 2023

.

Com base na configuração experimental apresentada e considerando as avaliações propostas neste estudo, um total de 224 resultados foram produzidos a partir de 2240 modelos avaliados nos conjuntos de teste. Ainda, considerando o processo de estimativa dos melhores parâmetros, um total de 8960 modelos foram construídos.

Para a execução dos experimentos foi utilizada a plataforma de computação em nuvem Google Cloud com ambiente operacional Linux Debian 5.10.191-1 x86_64 GNU/Linux. A configuração do hardware utilizada incluiu servidores de 8 núcleos Intel(R) Xeon(R) CPU @ 3.10GHz com 32 GB de memória RAM e servidores de 16 núcleos Intel(R) Xeon(R) CPU @ 2.20GHz com 64 GB de memória RAM. Os códigos utilizados neste trabalho, encontram-se disponibilizados no repositório *acp_tcc.*[10]

[10] https://github.com/isabellaloren4/acp_tcc.git

4. RESULTADOS E DISCUSSÃO

Com base na revisão da literatura de trabalhos relacionados, identificou-se a existência de numerosas pesquisas que propuseram variados métodos de predição automática de peptídeos anti câncer. No entanto, pelo melhor de nosso conhecimento, não identificamos nenhum estudo que tenha consolidado uma análise abrangente, englobando uma variedade de descritores e algoritmos de aprendizado de máquina. Nesse contexto, neste trabalho foi proposta uma ampla avaliação experimental para a definição de um *baseline* para a literatura. Assim, para atingir esse objetivo, foram analisadas as combinações de diversos classificadores e descritores por meio de um estudo exploratório a respeito dos parâmetros dos algoritmos de aprendizado de máquina.

A problemática fundamental dos estudos propostos na literatura e também deste trabalho está relacionada a forma como os dados são tratados e combinados com os algoritmos de aprendizado de máquina. Nesse sentido, para que os dados biológicos e químicos abordados neste estudo possam ser analisados por meio de métodos matemáticos e/ou computacionais é necessário que os dados sejam transformados. Essa transformação é realizada pela aplicação de descritores, os quais permitem representar as moléculas em termos de suas características estruturais e/ou físico-químicas para uma forma numérica ou simbólica. No entanto, considerando os estudos de representação molecular existentes, ainda não está claro qual é a melhor forma de representar uma molécula *in silico* para a construção de um modelo de predição. Além desse aspecto, a definição de um método de representação também está relacionada aos algoritmos de indução dos modelos. Isso deve-se ao fato de que, ao criar um modelo de classificação, existem fatores como a organização dos dados e o ajuste de parâmetros, que são essenciais para a criação de um modelo generalista. Ou seja, um modelo que é capaz de aprender os padrões dentro de um conjunto de treinamento e consegue aplicar o reconhecimento desses padrões sobre um conjunto de dados de teste (VAMATHEVAN et al., 2019).

Neste contexto, com o objetivo de construir *baseline* para a literatura, neste trabalho foi realizada a construção de modelos por meio da combinação de descritores e classificadores. Nesse sentido, com relação aos descritores, foram utilizados descritores provenientes do *Protpy* e *ModlAMP* que analisam as

características físico-químicas das moléculas e descritores proveniente do *DeepChem* que analisa a molécula de maneira estrutural. Ademais, com relação aos classificadores, utilizou-se algoritmos de aprendizado de máquina tradicionais que são frequentemente utilizados na literatura e um algoritmo de aprendizado de máquina profundo (cnn).

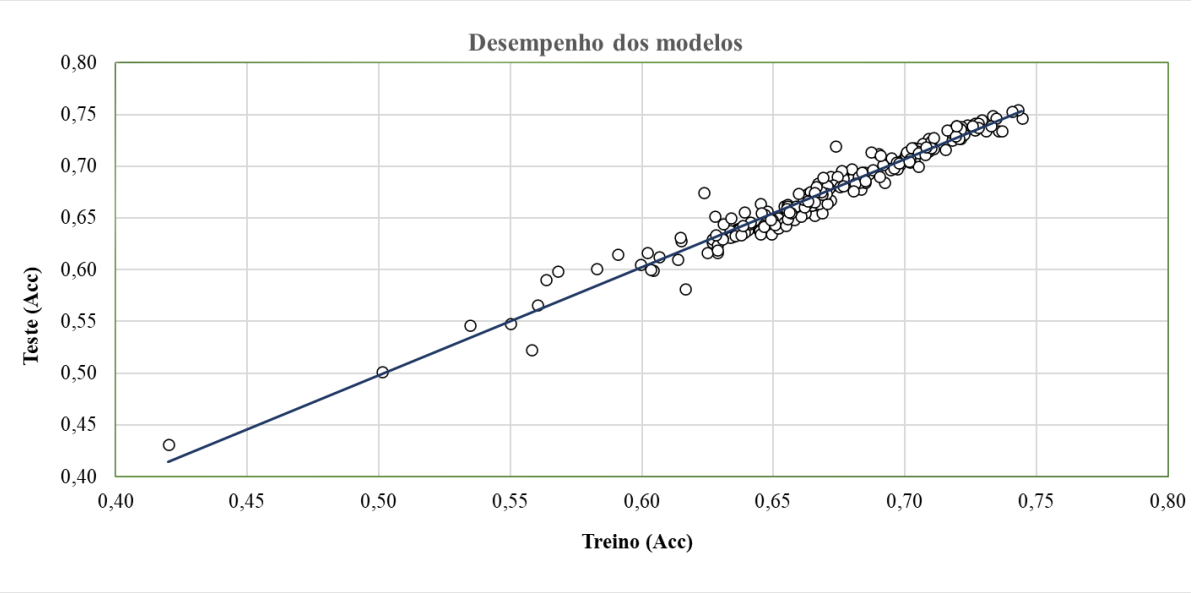
Em relação aos dados analisados neste estudo, foi utilizado o conjunto de dados públicos provenientes do AntiCP 2.0. A escolha desse conjunto de dados foi realizada devido a frequente utilização na literatura, e pelo fato desse banco de dados ser de acesso público. Esse conjunto é composto por 861 sequências de ACP e 861 sequências de AMP, portanto, trata-se de um banco de dados balanceado. Ademais, o banco de dados possui uma quantidade de exemplos significativa, quando comparado com outros conjuntos de dados utilizados em trabalhos anteriores.

4.1. RESULTADOS DA AVALIAÇÃO EXPERIMENTAL

Com relação aos resultados obtidos a partir da avaliação experimental, inicialmente, foi realizada uma análise sobre a capacidade preditiva dos modelos. Na Figura 14 é apresentado um gráfico que expressa a relação de desempenho entre os modelos construídos no conjunto de treinamento – Treino (Acc) – e avaliados no conjunto de teste – Teste (Acc), em termos de acurácia. Por meio da análise desses dados é possível observar que os modelos construídos após a otimização de parâmetros apresentaram resultados aproximados tanto no conjunto de treino quanto no teste. Portanto, constata-se que os modelos alcançaram o objetivo de aprender a reconhecer os padrões presentes nos dados na etapa de treino e posteriormente refletir essa capacidade no conjunto de teste.

Figura 14 - DESEMPENHO DOS MODELOS COM RELAÇÃO AOS VALORES DE ACURÁCIA

OBTIDOS NO CONJUNTO DE TREINAMENTO E NO CONJUNTO DE TESTE

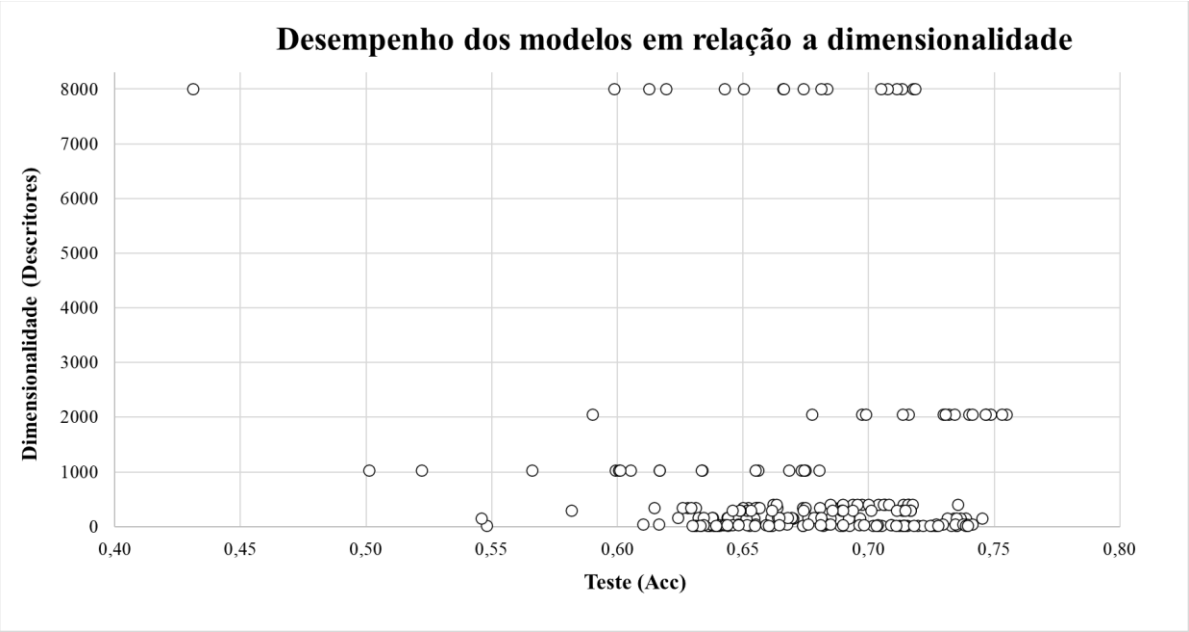


Fonte: A autora, 2023.

Ademais, neste trabalho, os modelos construídos foram analisados em função do desempenho preditivo em diferentes níveis de dimensionalidade dos descritores. Com base nos resultados representados na (Figura 15), observou-se que a dimensionalidade dos descritores não apresentou uma alteração no desempenho dos modelos, em termos de acurácia. Os modelos construídos com a combinação dos descritores de dimensões superiores a 1000, apresentam valores de desempenho na etapa de teste dentro do intervalo de 0,60 a 0,75, os quais são similares aos valores obtidos dos modelos construídos em combinação com descritores de dimensão inferior a 1000. Assim, baseando-se nesta análise, é possível sugerir que, para os algoritmos de classificação com desempenhos similares, é indicada a utilização de descritores com baixa dimensionalidade.

Figura 15 - DESEMPENHO DOS MODELOS – TESTE (ACC) – EM RELAÇÃO A DIMENSIONALIDADE

DOS DESCRITORES



Fonte: A autora, 2023.

Com relação aos resultados obtidos a partir da avaliação experimental, foi realizada uma análise sobre a influência da escolha dos algoritmos na capacidade preditiva. A Tabela 1 é uma sumarização do desempenho dos modelos no conjunto de teste – Teste (Acc), utilizando a média e o desvio padrão da acurácia dos resultados. As colunas representam os valores para os 13 descritores, e as linhas os valores para os 18 algoritmos de aprendizado de máquina. Os valores presentes na Tabela 1 foram analisados de forma a encontrar os melhores algoritmos com relação aos descritores. Desse modo, na Tabela 2, as colunas representam os valores de desempenho utilizando métricas de desempenho como: a acurácia no conjunto de treino – Treino (Acc); acurácia no conjunto de teste – Teste (Acc); recall (Rec); precisão (Prec); pontuação F1 (F1); Correlação de Matthews (Mcc).

Tabela 1 - DESEMPENHO DOS MODELOS NO CONJUNTO DE TESTE

Desempenho dos modelos no conjunto de teste -Teste (acc)-(μ±σ)													
-	Descritores												
Algoritmos	**AAC**	**APAAC**	**Circular**	**CTD**	**Ctriad**	**DPC**	**fasta2seq**	**macckeys**	**modlAMP**	**mol2vec**	**PAAC**	**smiles2seq**	**TPC**
ada	0,68±0	0,66±0,02	0,70±0,01	0,69±0,02	0,63±0,02	0,66±0,02	0,66±0,02	0,63±0,01	0,66±0,02	0,66±0,01	0,67±0,01	0,63±0,01	0,64±0,02
catboost	0,74±0,01	0,72±0,01	0,75±0,01	0,73±0,02	0,69±0,01	0,72±0,02	0,73±0,01	0,68±0,02	0,71±0,02	0,72±0,01	0,74±0,01	0,67±0,01	0,68±0,02
dt	0,65±0,02	0,66±0,03	0,68±0,02	0,65±0,02	0,63±0,01	0,66±0,02	0,67±0,02	0,64±0,01	0,63±0,01	0,65±0,02	0,65±0,02	0,60±0,02	0,62±0,01
et	0,74±0,02	0,73±0,01	0,75±0,02	0,73±0,02	0,68±0,02	0,71±0,02	0,73±0,02	0,68±0,01	0,70±0,03	0,71±0,02	0,74±0,02	0,67±0,01	0,67±0,02
gbc	0,73±0,02	0,71±0,02	0,75±0,02	0,75±0,01	0,69±0,02	0,72±0,01	0,74±0,01	0,67±0,01	0,71±0,02	0,72±0,01	0,72±0,01	0,68±0,01	0,68±0,02
gpc	0,71±0,01	0,69±0,02	0,72±0,01	0,55±0	0,63±0,01	0,72±0,01	0,68±0,02	0,68±0,01	0,65±0,02	0,69±0,01	0,71±0,02	0,50±0	0,61±0,05
knn	0,72±0,02	0,69±0,02	0,71±0,01	0,7±0,01	0,66±0,01	0,71±0,02	0,68±0,02	0,66±0,01	0,68±0,02	0,69±0,02	0,63±0,01	0,66±0,01	0,67±0,06
lda	0,64±0,01	0,63±0,01	0,73±0,01	0,67±0,01	0,66±0,02	0,70±0,02	0,65±0,01	0,64±0,02	0,65±0,02	0,65±0,01	0,72±0,02	0,63±0,01	0,71±0,02
lightgbm	0,73±0,02	0,70±0,02	0,74±0,02	0,74±0,01	0,66±0,01	0,70±0,02	0,74±0,01	0,66±0,01	0,70±0,02	0,71±0,02	0,72±0,02	0,67±0,02	0,65±0,02
lr	0,64±0,01	0,64±0,01	0,73±0,01	0,67±0,01	0,65±0,02	0,69±0,02	0,64±0,01	0,64±0,01	0,64±0,01	0,65±0,02	0,63±0,01	0,61±0,02	0,71±0,02
mlp	0,70±0,02	0,71±0,01	0,73±0,01	0,69±0,02	0,67±0,01	0,70±0,01	0,68±0,02	0,67±0,01	0,66±0,02	0,57±0,08	0,70±0,02	0,62±0,01	0,72±0,02
nb	0,66±0,01	0,55±0,05	0,70±0,02	0,65±0,01	0,63±0,02	0,68±0,01	0,61±0,01	0,62±0,02	0,64±0,02	0,67±0,05	0,64±0,01	0,60±0,01	0,71±0,02
qda	0,67±0,02	0,64±0,01	0,59±0,01	0,68±0,01	0,61±0,02	0,74±0,01	0,65±0,01	0,66±0,01	0,65±0,02	0,69±0,01	0,66±0,01	0,57±0,02	0,43±0,01
rf	0,73±0,02	0,72±0,01	0,75±0,02	0,74±0,01	0,67±0,02	0,71±0,02	0,73±0,02	0,68±0,01	0,70±0,02	0,71±0,02	0,73±0,01	0,67±0,01	0,67±0,03
ridge	0,64±0,01	0,64±0,01	0,73±0,01	0,67±0,01	0,65±0,02	0,71±0,02	0,64±0,02	0,63±0,02	0,64±0,02	0,65±0,02	0,63±0,01	0,60±0,02	0,72±0,02
svm	0,64±0,01	0,64±0,01	0,75±0,02	0,67±0,01	0,65±0,02	0,70±0,01	0,62±0,02	0,66±0,01	0,64±0,02	0,65±0,01	0,63±0,01	0,62±0,02	0,70±0,03
xgboost	0,72±0,01	0,69±0,02	0,74±0,02	0,74±0,01	0,66±0,02	0,69±0,01	0,74±0,01	0,64±0,01	0,69±0,02	0,70±0,01	0,71±0,02	0,65±0,01	0,60±0,01
cnn	-	-	-	-	-	-	0,68±0,02	-	-	0,69±0,02	-	0,52±0,05	-

Fonte: A autora (2023).
LEGENDA: μ - Média; σ – Desvio Padrão.

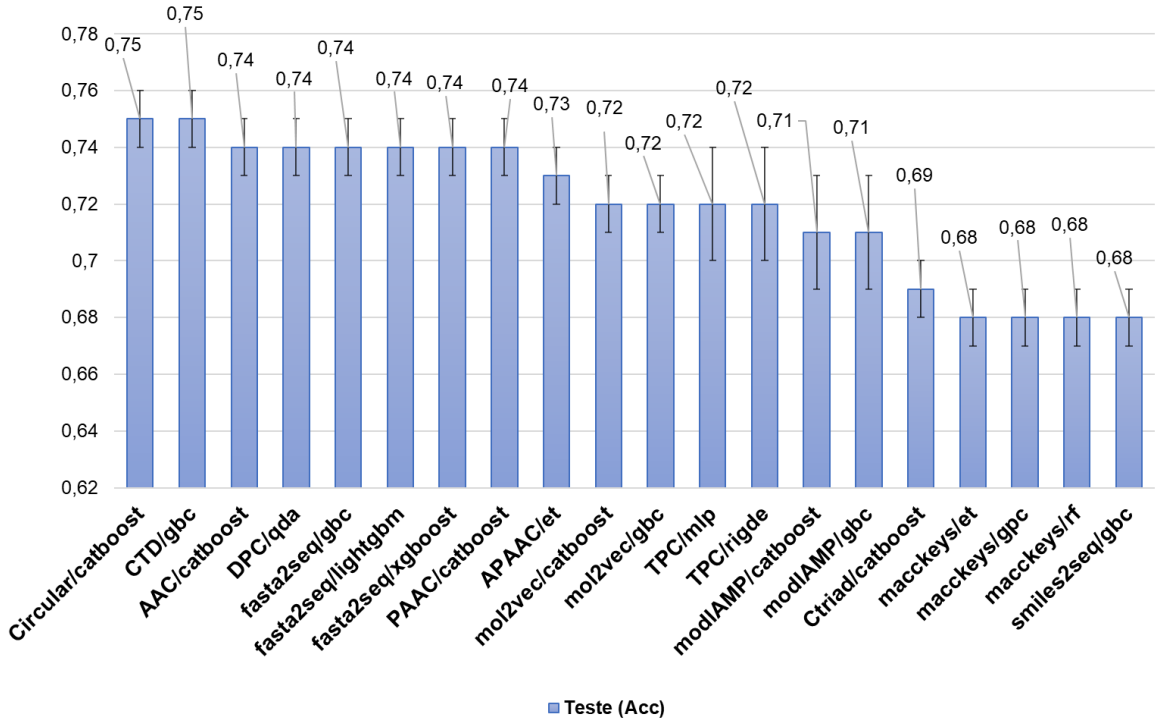
Tabela 2 - MELHORES ALGORITMOS COM RELAÇÃO AOS DESCRITORES

Melhores algoritmos com relação aos descritores (μ±σ)						
Modelos	**Teste (acc)**	**Treino (acc)**	**Rec**	**Prec**	**f1**	**mcc**
AAC/catboost	0,74±0,01	0,73±0,01	0,68±0,02	0,76±0,03	0,72±0,01	0,47±0,03
APAAC/et	0,73±0,01	0,72±0,02	0,66±0,02	0,78±0,03	0,71±0,01	0,47±0,03
Circular/catboost	0,75±0,01	0,69±0,01	0,71±0,03	0,77±0,02	0,74±0,01	0,50±0,02
CTD/gbc	0,75±0,01	0,73±0,02	0,71±0,04	0,76±0,02	0,74±0,02	0,49±0,03
Ctriad/catboost	0,69±0,01	0,68±0,03	0,68±0,02	0,69±0,01	0,69±0,02	0,38±0,02
DPC/qda	0,74±0,01	0,72±0,03	0,67±0,03	0,77±0,02	0,72±0,02	0,48±0,03
fasta2seq/gbc	0,74±0,01	0,73±0,02	0,72±0,03	0,75±0,02	0,73±0,02	0,48±0,03
fasta2seq/lightgbm	0,74±0,01	0,73±0,02	0,73±0,03	0,73±0,01	0,74±0,02	0,48±0,03
fasta2seq/xgboost	0,74±0,01	0,73±0,02	0,79±0,02	0,72±0,01	0,75±0,01	0,48±0,02
macckeys/et	0,68±0,01	0,67±0,02	0,59±0,04	0,72±0,03	0,64±0,02	0,36±0,03
macckeys/gpc	0,68±0,01	0,67±0,02	0,60±0,03	0,71±0,02	0,65±0,02	0,36±0,02
macckeys/rf	0,68±0,01	0,67±0,02	0,59±0,03	0,72±0,03	0,65±0,01	0,37±0,03
modlAMP/catboost	0,71±0,02	0,70±0,02	0,67±0,03	0,73±0,03	0,70±0,02	0,42±0,05
modlAMP/gbc	0,71±0,02	0,70±0,02	0,67±04	0,72±0,02	0,70±0,02	0,42±0,03
mol2vec/catboost	0,72±0,01	0,71±0,01	0,69±0,02	0,73±0,02	0,71±0,01	0,43±0,03
mol2vec/gbc	0,72±0,01	0,71±0,02	0,70±0,03	0,73±0,01	0,71±0,02	0,43±0,03
PAAC/catboost	0,74±0,01	0,73±0,01	0,68±0,02	0,77±0,02	0,72±0,01	0,48±0,03
smiles2seq/gbc	0,68±0,01	0,67±0,02	0,64±0,03	0,70±0,02	0,67±0,02	0,36±0,03
TPC/mlp	0,72±0,02	0,71±0,02	0,70±0,03	0,73±0,03	0,71±0,02	0,44±0,04
TPC/ridge	0,72±0,02	0,71±0,02	0,69±0,03	0,74±0,03	0,71±0,02	0,44±0,04

Fonte: A autora (2023).
LEGENDA: μ - Média; σ – Desvio Padrão.

Na Figura 16 estão representados os resultados de acurácia no conjunto de teste para cada combinação descritor/classificador. Por meio da análise desses resultados é possível observar que 6 modelos construídos com o algoritmo *catboost* estão entre os melhores modelos. Portanto, indicando que esses modelos conseguem obter um desempenho superior de classificação. Além disso, observa-se que o melhor modelo é aquele o *Circular/catboost*, com acurácia média de desempenho no conjunto de teste – Teste (Acc) de 0,75. Ademais, observa-se que os modelos construídos com combinação dos descritores *macckeys* e *smile2seq* apresentaram resultados inferiores quando comparados com os demais modelos.

Figura 16 - MELHORES ALGORITMOS COM RELAÇÃO AOS DESCRITORES



Fonte: A autora, 2023.

Na Tabela 3 estão apresentados os melhores resultados, considerando os valores de acurácia no conjunto de teste. Nas linhas estão representadas a combinação entre descritor/classificador e nas colunas os valores de desempenho por meio das métricas utilizadas

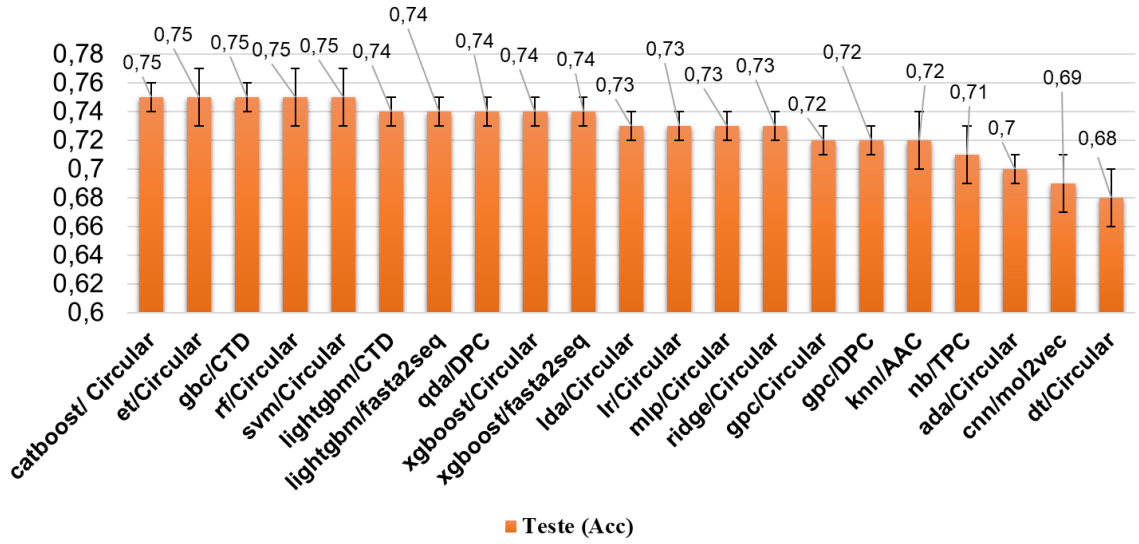
Tabela 3 - MELHORES DESCRITORES COM RELAÇÃO AOS ALGORITMOS

Melhores descritores com relação aos algoritmos (μ±σ)						
Combinação	**Teste (acc)**	**Train (acc)**	**rec**	**prec**	**F1**	**mcc**
ada/Circular	**0,70±0,01**	0,69±0,02	0,67±0,04	0,71±0,02	0,69±0,02	0,40±0,02
catbosst/ Circular	**0,75±0,01**	0,74±0,02	0,71±0,03	0,77±0,02	0,74±0,01	0,50±0,02
dt/Circular	0,68±0,02	0,68±0,03	0,65±0,04	0,69±0,02	0,67±0,03	0,36±0,05
et/Circular	**0,75±0,02**	0,73±0,03	0,69±0,04	0,77±0,02	0,73±0,02	0,50±0,03
gbc/CTD	0,75±0,01	0,73±0,03	0,71±0,04	0,76±0,02	0,74±0,02	0,49±0,03
gpc/Circular	**0,72±0,01**	0,70±0,03	0,74±0,03	0,71±0,01	0,72±0,02	0,43±0,03
gpc/DPC	0,72±0,01	0,70±0,04	0,67±0,02	0,74±0,01	0,70±0,01	0,44±0,02
knn/AAC	0,72±0,02	0,71±0,01	0,68±0,03	0,73±0,02	0,70±0,02	0,43±0,04
lda/Circular	**0,73±0,01**	0,72±0,04	0,73±0,02	0,73±0,02	0,73±0,01	0,46±0,02
lightgbm/CTD	0,74±0,01	0,72±0,03	0,71±0,03	0,75±0,01	0,73±0,02	0,48±0,03
lightgbm/fasta2seq	0,74±0,01	0,73±0,02	0,73±0,03	0,73±0,01	0,74±0,02	0,48±0,03
lr/Circular	**0,73±0,01**	0,72±0,03	0,73±0,02	0,73±0,02	0,73±0,01	0,46±0,03
mlp/Circular	**0,73±0,01**	0,74±0,02	0,73±0,03	0,73±0,02	0,73±0,01	0,47±0,02
nb/TPC	0,71±0,02	0,70±0,02	0,69±0,09	0,72±0,05	0,70±0,02	0,42±0,03
qda/DPC	0,74±0,01	0,72±0,03	0,67±0,03	0,77±0,02	0,72±0,02	0,48±0,03
rf/Circular	**0,75±0,02**	0,73±0,05	0,69±0,04	0,78±0,02	0,73±0,02	0,50±0,03
ridge/Circular	**0,73±0,01**	0,72±0,05	0,73±0,02	0,73±0,02	0,73±0,01	0,46±0,03
svm/Circular	**0,75±0,02**	0,74±0,05	0,73±0,03	0,77±0,03	0,75±0,01	0,51±0,03
xgboost/Circular	**0,74±0,01**	0,73±0,05	0,78±0,03	0,73±0,02	0,75±0,02	0,48±0,04
xgboost/fasta2seq	0,74±0,01	0,73±0,02	0,79±0,02	0,72±0,01	0,75±0,01	0,48±0,02
cnn/mol2vec	0,69±0,02	0,69±0,02	0,66±0,05	0,70±0,02	0,68±0,02	0,37±0,03

Fonte: A autora, 2023.
LEGENDA: μ - Média; σ – Desvio Padrão

Na Figura 17 são representados os resultados da combinações descritor/classificador considerando os melhores descritores com relação aos algoritmos, em termos de acurácia no conjunto de teste.

Figura 17 - MELHORES DESCRITORES COM RELAÇÃO AOS ALGORITMOS



Fonte: A autora, 2023.

Por meio de uma análise detalhada dos valores presentes na Figura 17, pode-se observar que considerando as 5 melhores combinações, temos a presença de apenas um modelo que utiliza o descritor CTD e 4 modelos com o descritor *Circular*. Ademais, quando analisamos a Tabela 3, pode-se observar que os valores de desempenho (em negrito) para 11 combinações que utilizam o descritor *Circular* apresentaram desempenho – Teste (Acc) – com valores superiores a 0,70. Portanto, indicando a possibilidade de desempenho superior de classificação para aqueles modelos que são construídos utilizando o descritor *Circular* e CTD.

4.2. ANÁLISE DE SIGNIFICÂNCIA ESTATÍSTICA

Os resultados dos modelos apresentados na Tabela 3 (melhores descritores) foram submetidos a uma análise comparativa utilizando o teste t pareado corrigido[11] de Nadeau e Bengio (NADEAU; BENGIO, 2003), com nível de significância em 0,05. Esse teste específico foi escolhido devido à natureza não independente das amostras produzidas pela aplicação do processo de validação cruzada. A hipótese nula nessa análise é que os desempenhos, em termos de acurácia, dos modelos são iguais. Como resultado da aplicação desse teste verificou-se que há uma diferença

[11] https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_stats.html#comparing-two-models-frequentist-approach

estatisticamente significativa (t=5,66 e p-valor=0,03) entre os resultados dos modelos Circular/catboost e CTD/gbc Para todas as demais comparações realizadas não foi possível rejeitar a hipótese nula.

Em relação aos resultados dos modelos apresentados na Tabela 3 (melhores descritores), não foi possível rejeitar a hipótese nula para nenhuma das comparações realizadas. Desse modo, pela análise de resultados estatísticos obtidos, é possível indicar que é preferível a utilização de descritores que apresentam maior desempenho em termos de acurácia e menor dimensionalidade. Portanto, é possível constatar, que considerando as técnicas avaliadas neste estudo, o desempenho dos modelos construídos possui maior influência em relação ao tipo de descritor de característica utilizado.

5. CONCLUSÃO E TRABALHOS FUTUROS

Os ACPs constituem uma alternativa promissora para o tratamento de diferentes tipos de câncer, destacando-se pela toxicidade seletiva para as células cancerígenas; potencialidade de penetrar a membrana plasmática das células; efeitos anti-angiogênicos; efeitos imunomoduladores; além de induzir a apoptose. No entanto, identificá-los experimentalmente ainda é um processo oneroso em relação ao custo e ao tempo. Métodos *in silico*, especialmente com o uso de algoritmos de aprendizado de máquina, possibilitam estratégias para explorar e analisar ACPs de modo automático, reduzindo a necessidade de testes experimentais. Contudo, na literatura, ainda não há um consenso sobre a representação ideal de uma molécula e sobre qual combinação de algoritmo e descritor é mais eficaz para a predição de ACPs.

Portanto, neste trabalho, com o intuito de definir um *baseline* para a literatura, foi proposto um método de avaliação experimental para o qual foram analisadas diferentes estratégias de predição de peptídeos. Para alcançar esse objetivo, foi realizada uma ampla avaliação experimental por meio da combinação de 18 algoritmos de aprendizado de máquina com 13 descritores de características. Como resultado desse estudo experimental, um total de 224 resultados foram produzidos, a partir de um total de 2240 modelos avaliados nos conjuntos de teste.

Com base na análise dos resultados, foi possível constatar que as estratégias Circular/catboost e CTD/gbc apresentaram os melhores desempenhos, em termos de acurácia, na predição. Contudo, não foi possível verificar influência estatisticamente significativa entre os algoritmos de aprendizado de máquina e entre a maioria dos descritores, avaliados neste estudo.

Em suma, as principais contribuições deste trabalho são:

- Definição de um baseline para a literatura de predição de ACPs, por meio da análise de 18 algoritmos de aprendizado de máquina;
- Avaliação de 13 descritores, que abrangem características físico-químicas, estruturais, composição e distribuição dos aminoácidos;
- Proposição de método de avaliação experimental para mitigar o viés sobre o conjunto de dados e para a estimativa adequada dos melhores parâmetros dos algoritmos de aprendizado.

Como perspectivas para trabalhos futuros, é possível elencar:

- Estudar a combinação dos diferentes descritores de características apresentados neste trabalho; bem como aplicar técnicas específicas para seleção de atributos relevantes;
- Utilizar estratégias de combinação de diferentes classificadores;
- Construção de modelos utilizando diferentes algoritmos de *Deep Learning* como, por exemplo, *Graph Convolutional Network* (GCN);
- Desenvolver estudos exploratórios utilizando a avaliação de modelos constituídos com a combinação de descritores considerando a molécula em representações bidimensionais e tridimensionais;
- Explorar a predição utilizando banco de dados de sequências de peptídeos anticâncer específicos para algum tipo de câncer como, por exemplo, câncer de mama.

REFERÊNCIAS

ABBAS, A. R.; MAHDI, B. S.; FADHIL, O. Y. Breast and Lung Anticancer Peptides Classification Using N-Grams and Ensemble Learning Techniques. **Big Data and Cognitive Computing**. p. 40, 2022. Disponível em:https://doi.org/10.3390/bdcc6020040

AGRAWAL, Piyush et al. *AntiCP 2.0: an updated model for predicting anticancer peptides*. ***Briefings in Bioinformatics***, v. 22, n. 3, maio de 2021, p. bbaa153. Disponível em: https://doi.org/10.1093/bib/bbaa153.

ALSANEA, Majed et al. To Assist Oncologists: An Efficient Machine Learning-Based Approach for Anti-Cancer Peptides Classification. **Sensors**, v. 22, n. 11, p. 4005, 2022. Disponível em: https://doi.org/10.3390/s22114005.

ALTAE-TRAN, H. et al. Low Data Drug Discovery with One-Shot Learning. **ACS Central Science**, v. 3, n. 4, p. 283-293, 2017. DOI: 10.1021/acscentsci.6b00367. Acesso em: 18 set. 2023

BAPTISTA, D.; CORREIA, J.; PEREIRA, B.; ROCHA, M. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. ***Journal of Integrative Bioinformatics***, v. 19, n. 3, 2022. DOI: https://doi.org/10.1515/jib-2022-0006.

BERG, Jeremy Mark; TYMOCZKO, John L.; STRYER, Lubert; GATTO JR., Gregory J. (Revisão Técnica por Deborah Schechtman). Bioquímica de Stryer. 7. ed. Rio de Janeiro: **Guanabara Koogan**, 2014. ISBN: 978-8-5277-2387-9.

BREIMAN, L. **Florestas Aleatórias**. Machine Learning, v.45, p.5–32, 2001

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R.; STONE, C. **Classification and Regression Trees**. Belmont, CA: Wadsworth, 1984.

BROGDEN, R. N.; BUCKLEY, M. M.; WARD, A. Buserelin. A review of its pharmacodynamic and pharmacokinetic properties, and clinical profile. ***Drugs***, v. 39, n. 3, p. 399–437, 1990. DOI: 10.2165/00003495-199039030-00007.

CAWLEY, Gavin C.; TALBOT, Nicola L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. **Journal of Machine Learning Research**, v. 11, p. 2079–2107, jul. 2010. Disponível em: https://jmlr.org/papers/v11/cawley10a.html

CAO, Ruifen et al. DLFF-ACP: prediction of ACPs based on deep learning and multi-view features fusion. **PeerJ**, vol. 9, e11906, 3 August 2021. Disponível em: https://peerj.com/articles/11906/. DOI: http://dx.doi.org/10.7717/peerj.11906.

CERETO-MASSAGUÉ, A.; OJEDA, M. J.; VALLS, C.; MULERO, M.; GARCIA-VALLVÉ, S.; PUJADAS, G. Molecular fingerprint similarity search in virtual screening.

Methods, v. 71, p. 58-63, 2015. ISSN 1046-2023. Disponível em: https://doi.org/10.1016/j.ymeth.2014.08.005. Acesso em: 20 de set. de 2023

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v.13, n.1, p.21-27, 196

CHEN, Xian-gan; ZHANG, Wen; YANG, Xiaofei; LI, Chenhong; CHEN, Hengling. ACP-DA: Improving the Prediction of Anticancer Peptides Using Data Augmentation. **Frontiers in Genetics**, v. 12 , 2021. ISSN 1664-8021. DOI: 10.3389/fgene.2021.698477. Disponível em: https://www.frontiersin.org/articles/10.3389/fgene.2021.698477

CHINNADURAI, R. K. et al. Current research status of anti-cancer peptides: Mechanism of action, production, and clinical applications. **Biomedicine & Pharmacotherapy**, v. 164, p. 114996, 2023. ISSN 0753-3322. Disponível em: https://doi.org/10.1016/j.biopha.2023.114996.

CHICCO, D.; JURMAN, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. ***BMC Genomics***, v. 21, n. 6, 2020. Disponível em: https://doi.org/10.1186/s12864-019-6413-7.

DANISHUDDIN, Asad U. Khan. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. ***Drug Discovery Today***, vol. 21, no. 8, pp. 1291-1302, 2016. ISSN 1359-6446. Disponível em:https://doi.org/10.1016/j.drudis.2016.06.013

FACELI, Katti et al. ***Inteligência artificial: uma abordagem de aprendizado de máquina*****.** 2ª edição. Rio de Janeiro: LTC, 2021. Acesso em: 14 set. 2023.

FENG, Guanwen et al. ME-ACP: Multi-view neural networks with ensemble model for identification of anticancer peptides. **Computers in Biology and Medicine**, v. 145, p. 105459, 2022. ISSN 0010-4825. Disponível em: https://doi.org/10.1016/j.compbiomed.2022.105459

FREITAS SAITO, Renata . et al. **Fundamentos de Oncologia Molecular.** São Paulo: Editora Atheneu, 2015. ISBN 978-85-388-0684-4.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119-139, 1997. DOI: 10.1006/jcss.1997.1504.

GEURTS, P.; ERNST, D.; WEHENKEL, L. **Extremely randomized trees**. *Machine Learning*, p. 3–42, 2006

GHULAM, Ali et al. ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. **Chemometrics and Intelligent Laboratory Systems**, v. 226, p. 104589, 2022. ISSN 0169-7439. Disponível em: https://doi.org/10.1016/j.chemolab.2022.104589.

HAN, B.; ZHAO, N.; ZENG, C. et al. ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. **Scientific Reports**, v. 12, p. 21915, 2022. Disponível em: https://doi.org/10.1038/s41598-022-24404-1.

HOERL, A.E.; KENNARD, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, v.12, n.1, p.55-67, 1970.

HOWLETT, S., CARTER, T. J., SHAN, H. M., & Nathan, P. D. (2023). Tebentafusp: um tratamento inovador para melanoma uveal metastático**. Therapeutic Advances in Medical Oncology**, 15, 17588359231160140. https://doi.org/10.1177/17588359231160140

JABEUR, S. B. et al. CatBoost model and artificial intelligence techniques for corporate failure prediction. ***Technological Forecasting and Social Change***, v. 166, 2021, p. 120658

JAEGAR, Sabrina; FULLE, Simone; TURK, Samo. Mol2vec: Abordagem de Aprendizado de Máquina Não Supervisionado com Intuição Química. ***Journal of Chemical Information and Modeling***, v. 58, n. 1, p. 27-35, 2018. DOI: 10.1021/acs.jcim.7b00616

KURRIKOFF, K.; APHKHAZAVA, D.; LANGEL, Ü. **The future of peptides in cancer treatment. Current Opinion in Pharmacology**, v. 47, p. 27-32, 2019. ISSN 1471-4892. Disponível em: https://doi.org/10.1016/j.coph.2019.01.008

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. **In Advances in Neural Information Processing Systems**, 2017. Disponível em https://dl.acm.org/doi/10.5555/3294996.3295074.

LI QINGWEN, et al. Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model. **Frontiers in Bioengineering and Biotechnology**, v. 8, 2020. Disponível em: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00892. DOI: 10.3389/fbioe.2020.00892. ISSN: 2296-4185

LIANG, X. et al. (2021). Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. **Briefings in Bioinformatics**, 22(4), bbaa312. Disponível em: https://doi.org/10.1093/bib/bbaa312.

LU, Michael A.; GIBSON, Tina. Development of Predictive Tools for Anti-Cancer Peptide Candidates using Generative Machine Learning Models. **Journal of Young Investigators**, v. 39, n. 5, Maio 2021. Disponível em: https://www.jyi.org/2021-may/2021/5/1/development-of-predictive-tools-for-anti-cancer-peptide-candidates-using-generative-machine-learning-models.

LV, Zhibin; CUI, Feifei; ZOU, Quan; ZHANG, Lichao; XU, Lei. Anticancer peptides prediction with deep representation learning features. **Briefings in Bioinformatics**,

v. 22, n. 5, p. bbab008, 2021. ISSN 1477-4054. Disponível em: https://doi.org/10.1093/bib/bbab008.

MITCHELL, T. M. **Machine Learning**. Boston, USA: McGraw-Hill, 1997.

MÜLLER, A.T. et al. modlAMP: Python for antimicrobial peptides. **Bioinformatics**, v. 33, n. 17, p. 2753-2755, set. 2017. Disponível em : https://doi.org/10.1093/bioinformatics/btx285

NELSON, David L.; COX, Michael M. (Revisão Técnica por Carlos Termignoni et al.). **Princípios de Bioquímica de Lehninger**. 6. ed. Porto Alegre: Artmed, 2014. ISBN 978-85-8271-073-9

NEVES, Bruno J.; BRAGA, Rodolpho C.; MELO-FILHO, Cleber C.; MOREIRA-FILHO, José Teófilo; MURATOV, Eugene N.; ANDRADE, Carolina Horta. QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery**. Frontiers in Pharmacology**, v. 9, 2018. Disponível em: https://www.frontiersin.org/articles/10.3389/fphar.2018.01275.

PARVANDEH, Saeid; YEH, Hung-Wen; PAULUS, Martin P; MCKINNEY, Brett A. Consensus features nested cross-validation. ***Bioinformatics***, v. 36, n. 10, p. 3093-3098, maio 2020. Disponível em: https://doi.org/10.1093/bioinformatics/btaa046.

PHAN, Le Thi et al. MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. **Computational and Structural Biotechnology Journal**, v. 20, p. 4473-4480, 2022. ISSN 2001-0370. Disponível em: https://doi.org/10.1016/j.csbj.2022.07.043.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. CatBoost: unbiased boosting with categorical features. In: **Advances In Neural Information Processing Systems**, 2017. Disponível em: https://dl.acm.org/doi/abs/10.5555/3327757.3327770

RAMSUNDAR, B. deepchem.io. Disponível em: https://github.com/deepchem/deepchem. Acesso em: 18 set. 2023

RAO, Bing; ZHANG, Lichao; ZHANG, Guoying. ACP-GCN: The Identification of Anticancer Peptides Based on Graph Convolution Networks. **IEEE** Access, v. 8, p. 176005-176011, 2020. Disponível em: https://ieeexplore.ieee.org/document/9207973

RDKit. MACCSkeys.py. Copyright (C) 2001-2011 Greg Landrum and Rational Discovery LLC. Disponível em: https://github.com/rdkit/rdkit-orig/blob/master/rdkit/Chem/MACCSkeys.py. Acesso em: 20 de set. de 2023

REZENDE, S. **O. Sistemas Inteligentes: Fundamentos e Aplicações**. Barueri, Brasil: Manole, 2003.

ROGERS, David; HAHN, Mathew. Extended-Connectivity Fingerprints. ***Journal of Chemical Information and Modeling***, v. 50, n. 5, p. 742-754, 2010. DOI: 10.1021/ci100050t.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial: Uma Abordagem Moderna**. 3. ed. Rio de Janeiro: Pearson, 2013.

SUBASI, A. (2020). **Machine learning techniques. In: Practical Machine Learning for Data Analysis Using Python**, p. 91-202. DOI: 10.1016/b978-0-12-821379-7.00003-5.

SUNG, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries**.** CA: **A Cancer Journal for Clinicians**, v. 71, n. 3, p. 209-249, maio 2021.. Disponível em: https://doi.org/10.3322/caac.21660. PMID: 33538338

SUN, M., YANG, S., HU, X., & ZHOU, Y. (2022). ACPNet: A Deep Learning Network to Identify Anticancer Peptides by Hybrid Sequence Information. **Molecules (Basel, Switzerland**), 27(5), 1544. Disponível em: https://doi.org/10.3390/molecules27051544.

SUN, YIH-YUN et al. "Peptide-Based Drug Predictions for Cancer Therapy Using Deep Learning." **Pharmaceuticals**, vol. 15, no. 4, artigo 422, 2022. Disponível em: https://doi.org/10.3390/ph15040422.

SNOEK, Jasper; LAROCHELLE, Hugo; ADAMS, Ryan P. Practical Bayesian Optimization of Machine Learning Algorithms. In: **Advances In Neural Information Processing Systems**, 25., 2012, Lake Tahoe. Proceedings. 2012. p. 2951-2959

SHAKER, Bilal; AHMAD, Sajjad; LEE, Jingyu; JUNG, Chanjin; NA, Dokyun. In silico methods and tools for drug discovery. ***Computers in Biology and Medicine***, v. 137, p. 104851, 2021. ISSN 0010-4825. Disponível em: https://doi.org/10.1016/j.compbiomed.2021.104851.

TOMII, K.; KANEHISA, M. Analysis of amino acid indices and mutation matrices for sequence comparison and protein structure prediction. **Protein Eng**, v. 9, p. 27-36, 1996

VAMATHEVAN, J. et al. **Applications of machine learning in drug discovery and development.** *Nat Rev Drug Discov*, v. 18, p. 463-477, 2019. DOI: https://doi.org/10.1038/s41573-019-0024-5.

WANG, H.; ZHAO, J.; ZHAO, H. et al. CL-ACP: a parallel combination of CNN and LSTM anticancer peptide recognition model. **BMC Bioinformatics**, v. 22, p. 512, 2021. DOI: 10.1186/s12859-021-04433-9

WU, X.; ZENG, W.; LIN, F.; XU, P.; LI, X. Anticancer Peptide Prediction via Multi-Kernel CNN and Attention Model. **Frontiers in Genetics**, v. 13, p. 887894, 2022. Disponível em: https://doi.org/10.3389/fgene.2022.887894.

WU, X.; ZENG, W.; LIN, F. GCNCPR-ACPs: a novel graph convolution network method for ACPs prediction. **BMC Bioinformatics**, v. 23, Supl 4, p. 560, 2022. Disponível em: https://doi.org/10.1186/s12859-022-04771-2.

YU, Lezheng; JING, Runyu; LIU, Fengjuan; LUO, Jiesi; LI, Yizhou. DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm. Molecular Therapy - **Nucleic Acids**, v. 22, p. 862-870, 2020. ISSN 2162-2531. DOI: 10.1016/j.omtn.2020.10.005.

ZHANG, Lu; TAN, Jianjun; HAN, Dan; ZHU, Hao. **From machine learning to deep learning: progress in machine intelligence for rational drug discovery.** *Drug Discovery Today*, v. 22, n. 11, 2017, p. 1680-1685. ISSN 1359-6446. Disponível em: https://doi.org/10.1016/j.drudis.2017.08.010.

ZHAO, et al. Convolutional neural networks for time series classification. Journal of **Systems Engineering and Electronics**, v. 28, n. 1, 2017.

APÊNDICE 1 – PARÂMETROS DOS ALGORITMO

Quadro 5 - PARÂMETROS DOS ALGORITMOS

(continua)

Algoritmos	Parâmetros
ada	n_estimators': Integer(10, 500), 'learning_rate': Real(1e-3, 1, prior='log-uniform'), 'algorithm': Categorical(['SAMME', 'SAMME.R'])
catboost	'learning_rate': Real(1e-3, 1, prior='log-uniform'), 'iterations': Integer(10, 500), 'depth': Integer(3, 10), 'l2_leaf_reg': Real(1, 10, prior='uniform'), 'border_count': Integer(1, 255), 'bagging_temperature': Real(0, 1, prior='uniform'), 'random_strength': Real(1e-9, 10, prior='log-uniform')
cnn	n_epochs': Categorical([100,1000]), 'batch_size': Categorical([32,64,128]), 'kernel_size': Integer(3, 10)
dt	criterion': Categorical(['gini', 'entropy']), 'splitter': Categorical(['best', 'random']), 'max_depth': Integer(3, 30), 'min_samples_split': Integer(2, 10), 'min_samples_leaf': Integer(1, 10), 'max_features': Real(0.1, 1.0, prior='uniform')
et	n_estimators': Integer(10, 500), 'criterion': Categorical(['gini', 'entropy']), 'max_depth': Integer(3, 30), 'min_samples_split': Integer(2, 10), 'min_samples_leaf': Integer(1, 10), 'max_features': Real(0.1, 1.0, prior='uniform'), 'bootstrap': Categorical([True, False]), 'class_weight': Categorical(['balanced', 'balanced_subsample', None])
gbc	n_estimators': Integer(10, 500), 'learning_rate': Real(1e-3, 1, prior='log-uniform'), 'max_depth': Integer(3, 10), 'min_samples_split': Integer(2, 10), 'min_samples_leaf': Integer(1, 10), 'max_features': Real(0.1, 1.0, prior='uniform'), 'subsample': Real(0.1, 1.0, prior='uniform')

Quadro 5 - PARÂMETROS DOS ALGORITMOS

(continuação)

Algoritmos	Parâmetros
gpc	optimizer': Categorical(['fmin_l_bfgs_b', None]), 'n_restarts_optimizer': Integer(0, 10), 'max_iter_predict': Integer(100, 1000)
knn	n_neighbors': Integer(1, 50), 'weights': Categorical(['uniform', 'distance']), 'algorithm': Categorical(['auto', 'ball_tree', 'kd_tree', 'brute']), 'p': Integer(1, 5)
lda	solver': Categorical(['lsqr', 'eigen']), 'shrinkage': Real(0, 1, prior='uniform'), 'tol': Real(1e-6, 1e-4, prior='log-uniform')
lightgbm	learning_rate': Real(1e-3, 1, prior='log-uniform'), 'n_estimators': Integer(10, 500), 'num_leaves': Integer(2, 100), 'max_depth': Integer(3, 10), 'min_child_samples': Integer(1, 50), 'min_child_weight': Real(1e-5, 1e-3, prior='log-uniform'), 'subsample': Real(0.1, 1.0, prior='uniform'), 'colsample_bytree': Real(0.1, 1.0, prior='uniform'), 'reg_alpha': Real(0, 1, prior='uniform'), 'reg_lambda': Real(0, 1, prior='uniform')
lr	C': Real(1e-4, 1e4, prior='log-uniform'), 'fit_intercept': Categorical([True, False]), 'solver': Categorical(['newton-cg', 'liblinear', 'sag', 'saga']
mlp	learning_rate':Categorical(['constant','invscaling','adaptive']), 'learning_rate_init': Real(1e-4, 1e-1, prior='log-uniform'), 'max_iter': Integer(1000,1001)
nb	var_smoothing': Real(1e-10, 1e-1, prior='log-uniform')
qda	var_smoothing': Real(1e-10, 1e-1, prior='log-uniform') 'reg_param': Real(0, 1, prior='uniform'), 'store_covariance': Categorical([True, False]), 'tol': Real(1e-5, 1e-1, prior='log-uniform')

Quadro 5 - PARÂMETROS DOS ALGORITMOS

(conclusão)

Algoritmos	Parâmetros
rf	n_estimators': Integer(10, 500), 'criterion': Categorical(['gini', 'entropy']), 'max_depth': Integer(3, 30), 'min_samples_split': Integer(2, 10), 'min_samples_leaf': Integer(1, 10), 'max_features': Real(0.1, 1.0, prior='uniform'), 'bootstrap': Categorical([True, False]), 'class_weight': Categorical(['balanced', 'balanced_subsample', None
ridge	alpha': Real(1e-4, 1e4, prior='log-uniform'), 'fit_intercept': Categorical([True, False]), 'solver': Categorical(['auto', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga']
svm	C': Real(1e-6, 1e+6, prior='log-uniform'), 'loss': Categorical(['hinge', 'squared_hinge']), 'tol': Real(1e-6, 1e-2, prior='log-uniform')
xgboost	learning_rate': Real(0.01, 0.3, prior='uniform'), 'n_estimators': Integer(50, 500), 'max_depth': Integer(3, 10), 'min_child_weight': Integer(1, 10), 'gamma': Real(0, 1, prior='uniform'), 'subsample': Real(0.5, 1, prior='uniform'), 'colsample_bytree': Real(0.5, 1, prior='uniform'), 'reg_alpha': Real(0, 1, prior='uniform'), 'reg_lambda': Real(1, 3, prior='uniform'), 'scale_pos_weight': Real(1, 5, prior='uniform')

Fonte: A autora (2023).

APÊNDICE 2 – DIMENSÃO DOS DESCRITORES

Tabela 4 - DIMENSÃO DOS DESCRITORES

Descritor	Dimensão
AAC	20
APAAC	22
circular	2048
CTD	147
Ctriad	343
DPC	400
Fasta2seq	50
macckeys	167
modlamp	28
mol2vec	300
PAAC	21
Smiles2seq	1031
TPC	8000

Fonte: A autora, 2023.