RTC REVISTA TÉCNICO-CIENTÍFICA DO IFSC
PUBLICAÇÃO DO IFSC
INSTITUTO FEDERAL Santa Catarina

Marco teórico para o desenvolvimento de redes neurais artificiais voltadas ao auxílio de diagnóstico da doença Linfoma de Hodgkin

Theoretical Framework for the Development of Artificial Neural Networks Aimed at Supporting the Diagnosis of Hodgkin's Lymphoma

Valdir Ferreira Filho | https://orcid.org/0009-0002-1710-516X
Rodrigo Coral | https://orcid.org/0000-0001-7788-4842
Daiani Cristina Savi | https://orcid.org/0000-0003-4794-2823

RESUMO

Este estudo apresenta uma fundamentação teórica para o desenvolvimento de redes neurais artificiais (RNAs) aplicadas ao auxílio diagnóstico do Linfoma de Hodgkin (LH), uma doença hematológica de origem genética. Inicialmente, são abordadas as principais arquiteturas e algoritmos de redes neurais, bem como os métodos de aprendizado e treinamento empregados na identificação de padrões complexos. O trabalho discute como essas técnicas podem ser utilizadas para reconhecer mutações no DNA humano, integrando conceitos de inteligência artificial e genética. Paralelamente, examina-se o LH, com ênfase nos genes envolvidos e nas informações obtidas a partir do banco de dados NCBI. A integração desses dados genéticos com RNAs demonstra potencial para o desenvolvimento de uma ferramenta computacional robusta e eficiente. Ao propor essa base teórica multidisciplinar, o estudo estabelece os alicerces para pesquisas futuras voltadas ao aprimoramento do diagnóstico do LH e de outras doenças de base genética, contribuindo para tratamentos mais precisos e personalizados.
Palavras-chave: Redes neurais artificiais (RNAs); Linfoma de Hodgkin (LH); genética.

ABSTRACT

This study presents a theoretical framework for the development of artificial neural networks (ANNs) aimed at supporting the diagnosis of Hodgkin's Lymphoma (HL), a hematological disease with a genetic basis. It initially discusses the main architectures and algorithms of neural networks, as well as learning and training methods applied to the identification of complex patterns. The study explores how these techniques can be used to detect mutations in human DNA, integrating concepts from artificial intelligence and genetics. In parallel, HL is examined, with emphasis on the genes involved and on the information obtained from the NCBI genetic database. The integration of this genetic data with ANNs demonstrates strong potential for developing a robust and efficient computational tool. By proposing this multidisciplinary theoretical foundation, the article establishes the groundwork for future research focused on improving the diagnosis of HL and other genetically based diseases, thereby contributing to more accurate and personalized treatments.
Keywords: Artificial neural networks (ANNs); Hodgkin's Lymphoma (HL); genetics.

Recebido em: 11/11/2023. Aprovado em: 31/10/2025.
Avaliado pelo sistema duplo-anônimo. Publicado conforme as normas da ABNT.
DOI: https://doi.org/10.35700/2316-8382.2025.v15.3657

1 INTRODUÇÃO

Nas últimas décadas, a inteligência artificial (IA) tem se consolidado como uma das tecnologias mais transformadoras no campo da saúde, especialmente pela capacidade de analisar grandes volumes de dados clínicos, genéticos e de imagem com alta precisão (Topol, 2019). Entre suas vertentes mais promissoras estão as redes neurais artificiais (RNAs), modelos computacionais inspirados no funcionamento do cérebro humano, amplamente aplicados em tarefas de diagnóstico, prognóstico e descoberta de padrões biológicos complexos (Malaver; Boos; Azevedo, 2019).

A convergência entre IA e biomedicina tem impulsionado avanços significativos no diagnóstico precoce de doenças oncológicas e hematológicas, nas quais o reconhecimento automatizado de padrões morfológicos e histológicos pode auxiliar a tomada de decisão clínica e contribuir para o desenvolvimento de terapias mais precisas (Bejnordi *et al.*, 2017; Esteva *et al.*, 2017). No entanto, ainda persistem desafios relevantes na área da saúde, especialmente na prevenção, tratamento e identificação de fatores genéticos associados às essas patologias, que afetam milhares de pessoas em todo o país. Um estudo recente demonstrou que infecções virais e alterações genéticas desempenham papel determinante no surgimento de diversas doenças desse grupo, impactando diretamente a morbimortalidade da população (Freitas *et al.*, 2021).

Incluído na gama de doenças hematológicas, o Linfoma de Hodgkin (LH) é uma forma de câncer que se origina no sistema linfático, uma parte crucial do sistema imunológico. Caracterizado pela presença de células anormais conhecidas como células de Reed-Sternberg, este tipo de linfoma pode afetar pessoas de todas as idades. O diagnóstico e tratamento precoces são fundamentais para melhorar as taxas de sobrevivência (Weniger; Küppers, 2021). Entretanto, a doença ainda representa um desafio significativo em termos de mortalidade. Entre 2011 e 2020, 1.141 brasileiros da região Nordeste faleceram em decorrência do LH, evidenciando a necessidade contínua de pesquisas e intervenções mais eficazes. (Botentuit *et al.*, 2023).

Um estudo divulgado pela Revista Brasileira de Medicina da Família e Comunidade (RBMFC) traz a importância da descoberta de doenças genéticas na atenção primária da saúde, permitindo assim, um diagnóstico precoce. Além disso, é ressaltado a necessidade de ferramentas que auxiliam nesse processo para reconhecimento de fatores importantes das doenças citadas, desde aquelas comumente encontradas como também as mais raras (Santos *et al.*, 2020).

A aplicabilidade das RNAs têm se destacado no desenvolvimento de ferramentas capazes de auxiliar no diagnóstico de doenças hematológicas. Um exemplo disso, estudos publicados na Nature, mostraram que algoritmos de IA são capazes de identificar e classificar lesões cutâneas com precisão comparável ou até mesmo superior à de médicos de diversas áreas (Esteva *et al*, 2017). De forma semelhante, o estudo do departamento

de medicina do Japão, juntamente com outras instituições espalhadas no mundo, apontou a eficácia das RNAs para análise da expressão genética do Linfoma Não Hodgkin, em que busca entender qual o estado dos genes em determinadas condições (Carreras; Hamoudi, 2021).

Diante desse panorama, o presente trabalho propõe a construção de um marco teórico voltado ao desenvolvimento de RNAs para auxiliar o diagnóstico do LH. O estudo consolida as bases conceituais e metodológicas necessárias à futura implementação de sistemas inteligentes aplicados à detecção de padrões genéticos da doença. São abordados os princípios estruturais e matemáticos das RNAs, os métodos de aprendizado e treinamento. Na seção seguinte, discute-se a genética do LH e o uso de bancos de dados, como o NCBI, para coleta e padronização das informações. Na última seção, apresentam-se os principais desafios e limitações no preparo dos dados e no treinamento das redes, delineando diretrizes teóricas que poderão orientar pesquisas futuras e o desenvolvimento de ferramentas diagnósticas automatizadas.

2 FUNDAMENTAÇÃO TEÓRICA

A presente fundamentação teórica busca integrar dois eixos centrais do estudo: o campo computacional, representado pelas RNAs, e o campo biomédico, voltado à genética do LH. Essa integração permite compreender como técnicas de aprendizado de máquina podem contribuir para a análise de padrões genéticos associados à doença, estabelecendo as bases conceituais que sustentam o desenvolvimento deste marco teórico.

2.1 Uma análise de redes neurais artificiais na perspectiva deste marco teórico

As RNAs são um tipo de modelo de IA inspirado no funcionamento do cérebro humano, que obteve seu início se relacionando com a natureza booleana e no estabelecimento de uma estrutura matemática para um neurônio (Kovács, 2006). Sendo composta por camadas de neurônios artificiais interconectados, as RNAs processam informações e aprendem a partir de dados. Cada neurônio é responsável por trabalhar um conjunto de informações de entrada, aplicando funções matemáticas a esses dados para gerar saídas correspondentes. (Goodfellow; Bengio; Courville, 2017).

Esse tipo de modelo é capaz de aprender a partir de exemplos, ajustando os pesos das conexões entre neurônios para melhorar a precisão das respostas na saída da rede. Isso é feito via um processo chamado de treinamento, que pode envolver ou não, a apresentação de dados de entrada e as saídas desejadas. Vale ressaltar que para todo o processo, existe por trás um especialista, cujo objetivo é garantir o bom funcionamento

da sua RNA criada, através do fornecimento de dados coerentes com o que se realmente se necessita, para não ocorrerem distúrbios no seu desenvolvimento.

Quando se fala de projetos de inovações tecnológicas que empregam as RNAs, elas se diferem significativamente daqueles que utilizam processamento convencional, especialmente no que tange à modelagem dos fenômenos em estudo. Enquanto o processamento convencional se baseia em modelos matemáticos explícitos dos fenômenos físicos, as RNAs, por outro lado, utilizam dados diretamente do mundo real, estabelecendo um modelo implícito do caso em análise e oferecendo uma solução viável para problemas complexos, em que a modelagem matemática se mostra impraticável (Haykin, 1999).

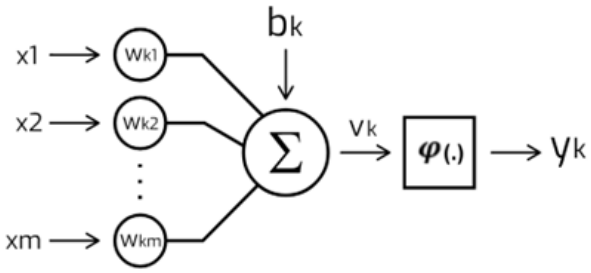
Dentre os benefícios observados das RNAs, especialmente as do tipo *feedforward*, se encontra a habilidade de mapear não-linearidades, realizar mapeamentos de entrada/saída, adaptar-se a pequenas modificações nas condições iniciais, e generalizar, fornecendo saídas adequadas para entradas não presentes durante o treinamento. Essas características fazem a técnica de modelagem eficaz, sendo capaz de descrever modelos complexos com precisão e adequação, conforme evidenciado por diversos estudos e aplicações práticas na literatura.

2.1.1 Redes *Feedforward*

As redes neurais *feedforward*, também conhecidas como redes neurais multicamadas (MLP), são uma classe de RNAs que processam informações em uma única direção, sem formação de ciclos. Essas redes têm sido amplamente utilizadas para resolver uma variedade de problemas de aprendizado de máquina, como classificação, regressão, e reconhecimento de padrões, por conta disso, será o foco no desenvolvimento do estudo (Bishop, 2006).

A composição de uma rede *feedforward* é dada por neurônios organizados em camadas. Cada neurônio (Figura 1) recebe sinais de entrada (xm) de neurônios da camada anterior, onde ocorre a otimização dos pesos sinápticos (wkm), realizando um somatório dessas informações aplicando a elas o bias (bk). Após esse processo, através do campo local induzido do neurônio (vk) chegasse a função de ativação $(\varphi(.))$, responsável por gerar o sinal de saída (yk) (Haykin, 1999).

Figura 1 - Modelo matemático de um neurônio



Fonte: Adaptado de Coral (2014)

Dentre as opções de estruturas disponíveis para a realização das redes neurais *feedforward*, a múltiplas camadas permite uma capacidade ainda maior de resolução de questões com padrões inseparáveis linearmente. Essa topologia é caracterizada por possuir uma ou mais camadas ocultas de neurônios, situadas entre as camadas de entrada e saída, desempenhando um papel crucial na extração de estatísticas de ordem elevada.

Esses neurônios ocultos, que possuem a mesma topologia dos neurônios de saída, atuam como detectores de características, enfatizando funções que definem o conjunto de treinamento à medida que o processo de aprendizagem progride. Enquanto os neurônios da camada de entrada servem primariamente como elementos de sinapses, os neurônios nas camadas ocultas processam e armazenam informações, desempenhando um papel vital na operação de redes com múltiplas camadas.

2.1.2 Método de Aprendizado

A existência dos métodos de aprendizado para a realização do treinamento das RNAs, vem a ser um dos pilares para o desenvolvimento dessa ferramenta. Os principais métodos de utilização se baseiam no supervisionado e não supervisionado, possuindo diferenças significantes para aplicação a quais são submetidas.

Quando o método de aprendizado é um modelo treinado em conjunto de dados rotulados, em que cada exemplo de entrada é emparelhado com uma saída desejada, se nomeia supervisionado (Haykin, 1999). É uma abordagem predominante para treinar redes neurais, especialmente em tarefas que exigem alta precisão. O objetivo é minimizar a diferença entre as saídas apresentadas pelo modelo e as saídas reais, ajustando os pesos da rede durante o treinamento. A utilização eficaz dessas redes em tarefas complexas tem sido uma área de pesquisa ativa, e métodos como o *SuperSpike* têm sido propostos para abordar os desafios associados ao treinamento de tais redes (Zenke & Ganguli, 2018).

Diferente do aprendizado anterior, o aprendizado não supervisionado trabalha com dados não rotulados, visando descobrir padrões intrínsecos nos dados, como agrupamentos ou relações de similaridade. Um dos métodos mais comuns de aprendizado

não supervisionado é o *clustering* (agrupamento), que tem sido adaptado para treinamento *end-to-end* em grandes conjuntos de dados visuais (Caron *et al.*, 2018).

2.1.2 Classificação Binária

A classificação binária em RNAs é uma técnica computacional que visa categorizar dados de entrada em uma de duas categorias possíveis, frequentemente denotadas como 0 e 1. A eficácia da classificação RNA em tarefas binárias é notavelmente evidenciada pela sua capacidade de aprender e generalizar padrões complexos a partir dos dados de treinamento, permitindo que ela faça previsões precisas sobre dados não vistos. A aplicação prática e a relevância da classificação binária em RNAs é observada em diversos projetos, como na identificação e categorização de pacotes de rede em contextos de segurança cibernética (Abdullah; Al-Ashoor, 2020).

2.1.3 Funções de Ativação

Sendo um dos componentes mais críticos no desenvolvimento de RNAs, as funções de ativação ajudam a rede a aprender a partir dos dados de entrada e a fazer aproximações complexas. Elas introduzem não-linearidades no sistema, permitindo que a rede aprenda a partir de erros e generalize para novos dados. Abaixo serão apontadas as principais funções de ativação utilizadas em RNAs e comentado brevemente a sua utilização (Jagtap; Karniadakis, 2019).

- Sigmóide (Função Logística)

A função logística $(\sigma(x))$, caracterizada na Equação 1, é uma das funções de ativação mais antigas, mapeando qualquer valor de entrada para um número entre 0 e 1, tornando-a útil para modelos que estimam probabilidades. No entanto, ela sofre do problema do desaparecimento do gradiente, tornando-a menos popular para redes profundas.

Equação 1

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

- Tanh (Tangente Hiperbólica)

Semelhante a sigmóide, a tangente hiperbólica $(tanh(x))$ possui o diferencial de trabalhar com um intervalo entre -1 e 1, utilizada frequentemente em camadas ocultas de redes neurais multicamadas (Equação 2).

Equação 2

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

2.1.4 Métodos de Treinamento

Quando se fala de treinamento de RNAs define-se como uma tarefa complexa que envolve a otimização de uma série de parâmetros, pensando nisso, diversos algoritmos de treinamento foram desenvolvidos para melhorar a eficiência e a eficácia desse processo. Será abordado alguns dos métodos mais comuns para o treinamento de redes neurais, demonstrando seus cálculos matemáticos para facilitar uma maior compreensão e ajudar em uma futura análise de escolha.

- Gradiente Descendente

O gradiente descendente é um algoritmo de otimização que busca encontrar o mínimo local de uma função de custo, atualizando iterativamente os parâmetros do modelo na direção que minimiza o erro, permitindo que o modelo aprenda a partir dos dados (Goodfellow; Bengio; Courville, 2016). Estudos recentes mostram que, em redes neurais de grande porte, o gradiente descendente pode ser representado como um modelo linear, simplificando assim a dinâmica de aprendizado (Lee *et al.*, 2019). Utilizamos a função de custo de erro quadrático médio (MSE), denotada por $J(\omega)$, onde a diferença entre os valores reais γ^{i} e as previsões Y^{i} é avaliada sobre todos os m exemplos do conjunto de treinamento, como observado na Equação 3.

Equação 3

$$J(\omega) = \frac{1}{m}\sum_{i=1}^{m} (\gamma^{i} - Y^{i})^{2}$$

Já o cálculo do gradiente (Equação 4) utiliza-se a função de custo em relação aos pesos, esse gradiente é um vetor que contém todas as derivadas parciais em relação a cada peso, assim como pode se observar na equação abaixo:

Equação 4

$$\nabla J(\omega) = \left[\frac{\partial J}{\partial \omega 1}, \frac{\partial J}{\partial \omega 2}, \cdots, \frac{\partial J}{\partial \omega n}\right]$$

Para finalizar, a Equação 5 realiza a atualização dos pesos na direção oposta ao gradiente, gerando assim a fórmula simplificada:

Equação 5

$$\omega_{novo} = \omega_{antigo} - \alpha \nabla J(\omega)$$

Sendo:

ω = valor inicial dos pesos

α = taxa de aprendizado

- Levenberg-Marquardt (LM)

O método de Levenberg-Marquardt (LM) é uma técnica de otimização avançada usada para treinar RNAs, sendo especialmente útil para problemas não-lineares e é conhecido por sua eficiência e precisão. O LM é uma combinação dos métodos de Newton e gradiente descendente, buscando aproveitar o melhor de ambos: a rapidez de convergência do método de Newton e a robustez do gradiente descendente (Yu; Wilamowski, 2011; Hagan; Menhaj, 1994).

O cálculo por trás do método visa uma nova equação (6) em termos de atualização dos pesos, em que o parâmetro λ é ajustado dinamicamente durante o treinamento. Se uma atualização reduz o erro, λ é diminuído, tornando a atualização mais parecida com o método de Newton. Se uma atualização aumenta o erro, λ é aumentado, tornando a atualização mais próxima com o gradiente descendente.

Equação 6

$$\omega_{novo} = \omega_{antigo} - (J^T J + \lambda I)^{-1} J^T e$$

Sendo:

J = matriz jacobiana das derivadas parciais da função de erro em relação aos pesos.

e = vetor de erros entre as saídas previstas e reais

λ = parâmetro de amortecimento

I = matriz identidade

2.1.5 Comitê de Redes

A aplicação de RNAs na resolução de problemas complexos exige uma abordagem meticulosa, evitando soluções subótimas e garantindo a minimização efetiva do erro de aprendizagem (Penz, 2011; Haykin, 1999). Desafios como a incerteza de alcançar um mínimo global para a função de erro e a variabilidade inerente ao processo de

aprendizagem demandam atenção especial. Uma estratégia para atenuar a aleatoriedade no treinamento envolve a utilização de um comitê de RNAs, combinando respostas de diversas redes para produzir uma resposta única e mais robusta (Ahmad; Gromiha, 2002; Haykin, 1999).

O método da média simples, que calcula a média aritmética das respostas de várias redes treinadas sob condições semelhantes, é comumente empregado para combinar saídas de RNA. Penz (2011) destaca a aplicação de comitês de redes em diversas áreas e sublinha que, ao exceder trinta RNAs em um comitê, os benefícios na redução de erro começam a ser marginalmente comparados ao custo adicional de treinamento.

2.2 Genética da doença Linfoma de Hodgkin

O LH, conforme descrito na introdução, é uma neoplasia do sistema linfático caracterizada pela presença das células de Reed-Sternberg (HRS), consideradas o principal marcador histológico da doença. Essas células derivam de linfócitos B que sofreram profundas alterações genéticas e funcionais, levando-as a perder características típicas de diferenciação normal e a adquirir um comportamento tumoral (Weniger; Küppers, 2021; Schwering *et al.*, 2003).

A compreensão desses mecanismos celulares e moleculares é essencial para elucidar as causas e os fatores de progressão do LH. A seguir, são abordados os principais aspectos genéticos envolvidos em sua origem, com destaque para as células Hodgkin e HRS e as vias de sinalização associadas à patogênese da doença.

2.2.1 Origem Celular

As células HRS, já mencionadas como o principal marcador histológico do LH, apresentam uma origem derivada das células B, mas com alterações genéticas tão profundas que as afastam de suas funções imunológicas originais. Estudos demonstram que essas células perdem grande parte do programa de expressão gênica típico das células B e passam a ativar genes incomuns a esse tipo celular, adquirindo propriedades que favorecem a proliferação e a evasão imunológica (Schwering *et al.*, 2003).

Essa alteração no perfil de expressão gênica sugere uma reprogramação celular significativa durante a transformação maligna, em que as células HRS não apenas desativam genes associados à função normal das células B, mas também ativam genes que normalmente não são expressos nesse tipo celular. O desvio no perfil de expressão gênica pode ser um mecanismo pelo qual as células HRS adquirem características que favorecem o crescimento e a sobrevivência tumoral, como a capacidade de evadir o sistema imunológico e promover um ambiente inflamatório favorável ao tumor (Hertel *et al.*, 2002).

A compreensão detalhada dessas mudanças na expressão gênica e suas implicações funcionais é vital para desenvolver estratégias mais eficazes e direcionadas para o tratamento e diagnóstico da doença de Hodgkin.

2.2.2 Papel do Vírus Epstein-Barr (EBV)

O Vírus Epstein-Barr (EBV) é um herpesvírus humano que infecta principalmente as células B e é conhecido por ser um fator etiológico em várias doenças, incluindo alguns tipos de câncer como o próprio LH. O EBV desempenha um papel significativo na genética da doença, especialmente na infecção de células HRS, em que proteínas de membrana latente (LMP1 e LMP2a) são expressas em casos EBV+ e têm funções que mimetizam sinais celulares normais, contribuindo para a patogênese do linfoma (Portis *et al.*, 2003).

2.2.3 Tabela dos Genes

O artigo "Molecular biology of Hodgkin lymphoma" de Marc A. Weniger apresenta uma tabela (Figura 2) que lista várias alterações genéticas em células HRS a LP (Linfócitos Predominantes) associadas ao LH. A tabela categoriza os genes com base em suas vias de sinalização ou funções principais, o tipo de alteração genética e a frequência aproximada dessas alterações em casos da patologia (Weniger; Kuppers, 2021).

A tabela sugere que não há uma única alteração genética que define todos os casos de LH. Em vez disso, várias vias de sinalização são afetadas por múltiplas alterações genéticas. Isso destaca a complexidade da genética da doença e sugere que a disfunção de várias vias de sinalização, em vez de genes individuais, é crucial para a patogênese da mesma. O apontamento dos genes que sofrem com essas mutações é de extrema importância para o desenvolvimento e preparação de dados relacionados à RNA, no entanto, vale lembrar que o estudo não mostra todos os pontos de mutações possíveis relacionados à doença, e sim, aqueles que foi conseguido identificar.

Figura 2 - Lesões genéticas em células HSR e LP

	Gene	Caminho ou principal função	Tipo de alteração genética	Frequência aproximada (%)
Células HRS	NFKBIA	NF-κB	SNVs, indels	10–20
	NFKBIE	NF-κB	SNVs, indels	10
	TNFAIP3	NF-κB	SNVs, indels	40
	REL	NF-κB	Ganhos/amplificações	50
	MAP3K14	NF-κB	Ganhos/amplificações	25
	BCL3	NF-κB	Ganhos/amplificações	20
	JAK2[a]	JAK/STAT	Ganhos/amplificações	30
	SOCS1	JAK/STAT	SNVs, indels	40
	STAT6	JAK/STAT	SNVs, ganhos	30
	PTPN1	JAK/STAT	SNVs, indels	20
	CSF2RB	JAK/STAT	SNVs	20
	ITPKB	JAK/STAT	SNVs	15
	GNA13	JAK/STAT	SNVs	20
	B2M	Evasão imunológica	SNVs, indels	30
	MHC2TA	Evasão imunológica	Translocações SNVs	15
	PD-L1, PD-L2[a]	Evasão imunológica	Ganhos/amplificações	30
	XPO1	RNA nuclear e exportação de proteínas	SNVs (codon 571), gains	20
	ARID1A	Remodelação da cromatina	SNVs, indels	25
	JMJD2C[a]	Regulador epigenético	Ganhos/amplificações	30
Células LP	BCL6	Fator de transcrição	Translocações	35
	SOCS1	JAK/STAT	SNVs, indels	40
	SGK1	-	SNVs	50
	JUNB	Fator de transcrição	SNVs	50
	DUSP2	-	SNVs	50
	REL	NF-κB	Ganhos	40

Fonte: Adaptado de Weniger e Kuppers (2021)

3 METODOLOGIA

Parte da essência deste marco teórico está na construção de uma base metodológica que sustente a aplicação de RNAs em contextos biomédicos complexos, como o diagnóstico do LH. A amplitude e a profundidade conceitual que envolvem as RNAs exigem uma abordagem estruturada, capaz de contextualizar seu funcionamento e fundamentos matemáticos para que possam ser direcionadas à solução de problemas reais. Esta seção apresenta o percurso adotado para consolidar o referencial metodológico do estudo, contemplando a revisão da literatura, a escolha das ferramentas computacionais, a definição das fontes genéticas utilizadas como base teórica e as considerações éticas e de privacidade que orientam o uso responsável de dados biomédicos.

3.1 Estratégia de revisão e seleção da literatura

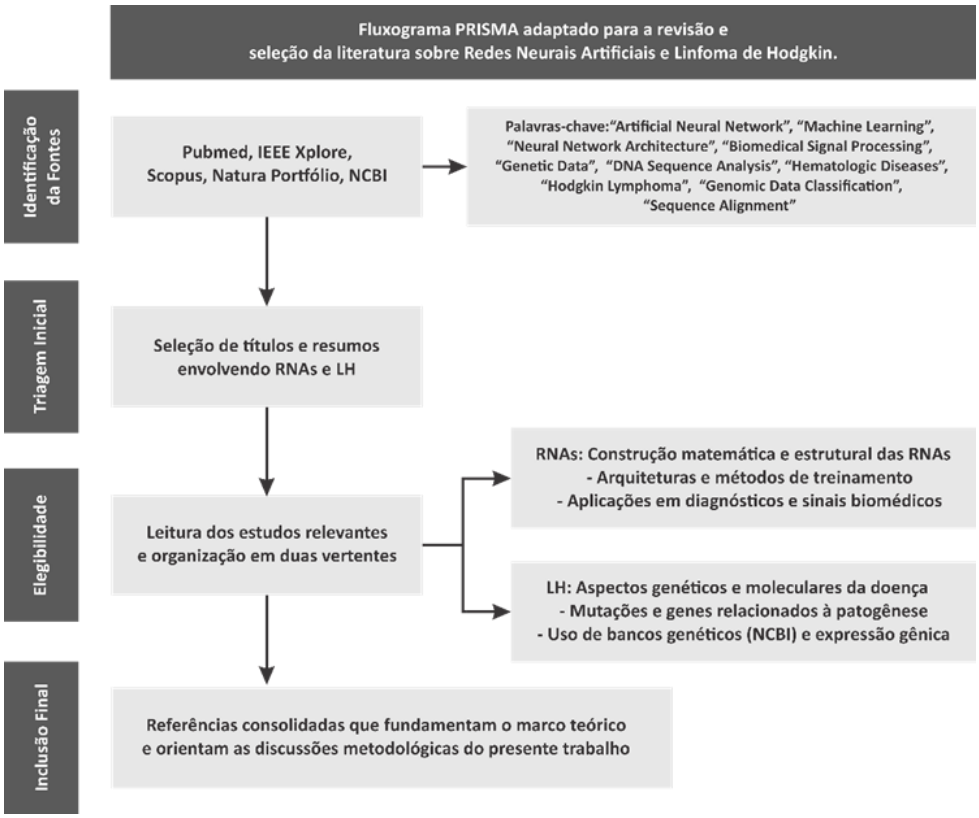
A construção deste marco teórico baseou-se em uma revisão sistematizada da literatura, desenvolvida com o objetivo de reunir estudos que abordassem o uso de RNAs

aplicadas à área biomédica, bem como pesquisas relacionadas à base genética e aos aspectos clínicos e moleculares do LH.

As buscas foram realizadas nas bases PubMed, IEEE Xplore, Scopus, Nature Portfolio e NCBI, selecionadas por sua relevância e complementaridade. Enquanto o IEEE Xplore concentra publicações voltadas à inteligência artificial, aprendizado de máquina e processamento de sinais, as bases PubMed, Scopus e Nature reúnem estudos biomédicos e genéticos, e o NCBI fornece acesso direto a dados e bancos de sequências genéticas, fundamentais para a análise molecular do LH.

O processo de revisão e seleção da literatura está representado na Figura 3, que apresenta o fluxograma PRISMA adaptado para este estudo. O modelo foi utilizado de forma conceitual, de modo a ilustrar o percurso metodológico adotado desde a identificação das fontes até a consolidação das referências que fundamentam o marco teórico e orientam as discussões metodológicas deste trabalho.

Figura 3 – Fluxograma PRISMA para revisão e seleção da literatura



Fonte: Autor.

3.2 Ferramenta de programação

Um dos passos importantes para o desenvolvimento de uma RNA, tende a ser a definição da linguagem e plataforma que será utilizada para a sua produção. A programação de uma RNA é um processo que envolve várias etapas, desde a escolha da

arquitetura da rede até o treinamento e a validação do modelo. O objetivo é criar um modelo que possa aprender a partir de dados, fazer previsões ou tomar decisões sem ser explicitamente programado. Além disso, é crucial entender o problema que está sendo resolvido, isso ajudará a determinar o tipo de rede neural mais adequada para realizar a tarefa.

Dentre as diversas opções de ferramentas e linguagens de programação disponíveis atualmente para trabalhar com RNAs, uma das predominantes e amplamente utilizadas é o MATLAB. Este programa se tornou uma das melhores escolhas para pesquisadores, engenheiros e desenvolvedores, tanto em ambientes acadêmicos quanto industriais. O que o torna atraente são seus ecossistemas ricos e bem desenvolvidos, que oferecem uma variedade de bibliotecas, frameworks e toolboxes especificamente projetados para facilitar a pesquisa e produção em aprendizado de máquina (Tiwari, *et al.*, 2022).

O MATLAB oferece suporte a uma variedade de ferramentas e bibliotecas especializadas, permitindo aos pesquisadores e profissionais explorar e desenvolver soluções de ponta. Entre as ferramentas disponíveis, destacam-se a Deep Learning Toolbox e a Neural Network Toolbox (NNTool). A combinação de ferramentas proporciona um ambiente poderoso e flexível, facilitando o desenvolvimento, treinamento e implantação de modelos de redes neurais. Com isso, os usuários são capazes de abordar uma ampla variedade de problemas e aplicações em diversos campos, desde reconhecimento de padrões e visão computacional até processamento de linguagem natural e modelagem de robôs (Tchórzewski; Wielgo, 2021).

3.3 Banco de dados genéticos

O universo genético tem experimentado um crescimento monumental desde que as técnicas de sequenciamento de DNA foram introduzidas nos anos 70 (Sanger; Nicklen; Coulson, 1977). Essa expansão, alimentada tanto pela evolução das tecnologias de sequenciamento como pela ascensão das práticas de biologia molecular e genômica, deu lugar a uma enxurrada de dados genéticos. Naturalmente, isso levou à emergência de uma nova necessidade - a necessidade de locais adequados para armazenar, avaliar e compartilhar essas informações. Surgem, então, os bancos de dados genéticos especializados.

3.3.1 Centro Nacional para Informação sobre Biotecnologia (NCBI)

Estabelecido há mais de três décadas, o Centro Nacional de Informações de Biotecnologia dos Estados Unidos, NCBI (NCBI, 2016), desempenha um papel essencial na

estrutura metodológica deste estudo por reunir, categorizar e disponibilizar informações genéticas e biomoleculares de forma aberta e continuamente atualizada.

O NCBI hospeda uma série de bancos de dados inter-relacionados, entre eles o GenBank, o RefSeq e o dbSNP, todos fundamentais para a análise de sequências genéticas e identificação de variações associadas a doenças. O GenBank, por exemplo, é atualizado diariamente e contém milhões de registros de sequências de DNA e proteínas provenientes de diversos organismos, o que o torna uma fonte primária de dados de treinamento para modelos de redes neurais aplicados à genômica. Já o RefSeq oferece conjuntos de referência revisados por especialistas, o que é crucial para reduzir ruídos e inconsistências nas análises de aprendizado supervisionado. Além dos bancos de dados de sequências, o NCBI fornece ferramentas integradas, como o BLAST (Basic Local Alignment Search Tool), amplamente utilizado para alinhamento de sequências e detecção de similaridades genéticas.

Diversos estudos recentes demonstram a relevância do NCBI como fonte de dados genéticos para análises baseadas em aprendizado de máquina. Por exemplo, Carreras e Hamoudi (2021) utilizaram dados do GenBank e do RefSeq para treinar redes neurais capazes de distinguir subtipos de linfomas não Hodgkin. Esse aspecto reforça a adequação da metodologia proposta neste estudo, que se apoia no NCBI como principal fonte de dados genéticos padronizados e validados, fundamentais para o desenvolvimento e teste da RNA voltada à detecção de padrões relacionados ao LH.

3.4 Considerações éticas e de privacidade

O uso de dados genéticos em pesquisas biomédicas e em diagnósticos mediados por IA envolve questões éticas que exigem atenção quanto à privacidade, anonimização e consentimento. Mesmo quando se utiliza bancos públicos e abertos, como o NCBI, é fundamental reconhecer que as sequências genéticas são portadoras de informações potencialmente identificáveis, o que demanda a adoção de boas práticas de governança de dados. Segundo Wang *et al.* (2017), a integração entre genótipos e fenótipos em múltiplas bases pode expor participantes a riscos de reidentificação, evidenciando a necessidade de políticas rigorosas de confidencialidade e anonimização no compartilhamento de dados genômicos.

Além da privacidade individual, as informações genéticas possuem caráter familiar e populacional, o que amplia os riscos e responsabilidades éticas associados ao seu uso. Como observam Takashima *et al.* (2018), a genética extrapola o indivíduo, atingindo parentes e grupos étnicos, e por isso deve ser tratada de forma a garantir o respeito à autonomia, à justiça e à equidade no compartilhamento de dados. Essas preocupações têm sido refletidas em diretrizes internacionais como as da Global Alliance for Genomics

and Health (GA4GH, 2016), que reforçam a importância do uso responsável e transparente de dados genômicos, bem como da rastreabilidade e do controle sobre reuso dos dados.

No contexto deste marco teórico, considera-se que pesquisas futuras baseadas neste referencial deverão empregar exclusivamente dados públicos, desidentificados e provenientes de repositórios abertos, sem envolvimento direto de seres humanos ou coleta de amostras originais. O uso desse tipo de dado encontra respaldo nas diretrizes éticas internacionais para pesquisa em genômica, como as orientações da Organização Mundial da Saúde (WHO, 2021), que recomendam a minimização dos riscos de identificação e a adoção de medidas que assegurem a confidencialidade e a integridade das informações biológicas. Caso etapas posteriores de validação prática venham a envolver dados provenientes de pacientes reais, será indispensável o cumprimento dos protocolos de ética em pesquisa envolvendo seres humanos, conforme as diretrizes do Conselho Nacional de Saúde (Resolução CNS nº 466/2012) e as normas internacionais de consentimento informado.

4 RESULTADOS E DISCUSSÕES

A etapa de resultados e discussões constitui o elo entre a fundamentação teórica e a aplicação prática dos conceitos apresentados. Nela, busca-se demonstrar como os conhecimentos sobre RNAs e genética do LH convergem para a construção de um modelo computacional capaz de auxiliar na predição da doença.

Esta seção descreve o processo de geração e preparação dos dados de entrada, a metodologia empregada na classificação binária, as estratégias de seleção e alinhamento das sequências genéticas, bem como os desafios encontrados durante o treinamento e a validação das redes. Ao longo da discussão, são enfatizados os aspectos técnicos e biológicos que sustentam a viabilidade do modelo proposto, apontando limitações práticas e caminhos para futuras otimizações.

4.1 Dados de entrada

A seleção meticulosa de dados de entrada é vital na modelagem eficiente das RNAs, especialmente quando se trata de tarefas complexas como a classificação de condições genéticas ou médicas. A preparação adequada e a escolha dos dados que serão alimentados na RNA não apenas influenciam diretamente a precisão e a eficácia do modelo, mas também otimizam o uso de recursos computacionais durante o treinamento e a validação do modelo. Este processo, que envolve a seleção de genes ou outras variáveis relevantes, é crucial para garantir que o modelo possa generalizar bem a partir dos dados de treinamento e realizar previsões precisas em dados novos e não vistos (Mazumder; Veilumuthu, 2018).

4.1.1 Seleção de dados

O processo de seleção de dados para o treinamento da RNA se inicia na observação dos genes presentes na tabela do estudo de Weniger, M. A. (2021), em que aponta os principais genes que de fato ocorrem as mutações genéticas que levam o indivíduo a ter a doença LH. Este estudo apresenta relevância significativa, pois estabelece uma base teórica que possibilita o uso de sequências genéticas como dados de entrada em RNAs, favorecendo a criação de modelos capazes de reconhecer padrões associados ao LH.

No banco de dados NCBI, vem a ser perceptível a sua capacidade de encontrar informações genéticas e não é diferente com os genes apontados na figura 3. No interior da plataforma é possível fazer a filtragem correta e coletar de forma individual, ou em escala, informações do gene desejado. Dentre as opções de dados oferecidos sobre cada caso encontrado no NCBI, a sequência genética é o que se encaixa nos critérios de relevância no quesito da doença LH. As alterações ou mutações sofridas dentro da sequência são de extrema importância para determinar se a pessoa terá ou não a enfermidade, e esse vem a ser o grande propósito do estudo em si.

4.1.2 Sequência genética

Uma sequência genética é uma série ordenada de nucleotídeos, que são os blocos de construção fundamentais do DNA e do ácido ribonucleico (***RNA***), e é codificada por quatro bases nitrogenadas: adenina (A), citosina (C), guanina (G) e timina (T) no DNA, ou uracila (U) no ***RNA*** (Griffiths *et al.*, 2022). Essas sequências são cruciais para a vida, pois carregam as instruções genéticas usadas no crescimento, desenvolvimento, funcionamento e reprodução de todos os organismos vivos. Cada gene, uma unidade de hereditariedade, é uma sequência única de nucleotídeos que codifica para moléculas específicas e funções em um organismo, desempenhando um papel vital na expressão de traços físicos, suscetibilidade a doenças e até mesmo comportamentos.

Quando se fala em utilizar a sequência genética para dado de entrada em uma RNA, deverá ser levado em consideração alguns pontos relevantes e dificuldades que precisam ser solucionadas. Num primeiro momento, por mais que seja trabalhado com um mesmo gene, o seu tamanho da sequência genética diverge entre os casos apresentados, pois cada indivíduo possui peculiaridades. Essas variações são devido a eventos genéticos como inserções, deleções e duplicações, chamadas de polimorfismos de inserção/deleção e podem ocorrer em qualquer gene (Ziegenhain; Sandberg, 2021).

O processo de treinamento das RNAs exige que seus dados de entrada possuam uma padronização de tamanho, haja visto os métodos de cálculos utilizados, que precisam de matrizes com dimensões idênticas. Para resolver o problema comentado sobre as

diferenciações dos genes, o processo de alinhamento vem a ser o caminho mais indicado para a produção desses dados de entrada. O trabalho de alinhamento de genes é uma técnica fundamental em bioinformática que busca estabelecer uma correspondência ótima entre duas ou mais sequências de DNA, ***RNA*** ou proteínas. Este processo é crucial para identificar regiões de similaridade entre sequências genéticas, que podem indicar relações funcionais, estruturais ou evolutivas entre os genes, ou proteínas (Pugacheva, Korotkov; Korotkov, 2016).

Entre os principais algoritmos de alinhamento, destacam-se o Needleman–Wunsch e o Smith–Waterman, considerados referências na área por permitirem o alinhamento global e local, respectivamente. O método de Needleman–Wunsch é amplamente utilizado quando se deseja comparar duas sequências completas, garantindo a correspondência ideal ao longo de toda a extensão. Já o algoritmo de Smith–Waterman é preferido em análises de alinhamento local, pois identifica regiões de maior similaridade entre porções específicas das sequências (Needleman; Wunsch, 1970; Smith; Waterman, 1981).

Na prática, ferramentas computacionais robustas implementam essas abordagens de forma otimizada. O BLAST, desenvolvido pelo NCBI, é amplamente utilizado para comparações rápidas e busca de similaridade entre sequências genéticas, enquanto programas como o Clustal Omega e o MEGA (Molecular Evolutionary Genetics Analysis) são indicados para alinhamentos múltiplos e análises filogenéticas mais detalhadas (Sievers *et al.*, 2011; Tamura *et al.*, 2021). O MEGA, em particular, oferece uma interface acessível para realizar alinhamentos, gerar matrizes padronizadas e explorar relações evolutivas entre genes, sendo uma ferramenta de grande relevância para o preparo e a validação de dados genômicos em estudos como o presente.

4.1.3 Classificação binária para diagnóstico

No âmbito deste estudo, a classificação binária foi adotada como abordagem principal para o diagnóstico assistido do LH. O objetivo consiste em treinar a RNA para distinguir entre duas condições fundamentais: presença (1) ou ausência (0) da doença, a partir das sequências genéticas previamente selecionadas e processadas. Cada conjunto de dados de entrada, correspondente a uma sequência genética, é acompanhado por um rótulo supervisionado que indica sua classe: 1 para casos associados ao LH e 0 para casos saudáveis. Essa rotulação é essencial para que a rede aprenda a associar padrões específicos de mutações genéticas às respectivas categorias durante o processo de treinamento.

Para esse tipo de classificação, a função de ativação Sigmóide mostra-se a mais adequada, pois converte a soma ponderada das entradas da rede em um valor limitado entre 0 e 1, sendo o intervalo ideal para expressar resultados em termos de probabilidade. Além disso, esta função apresenta uma derivada contínua e diferenciável, o que contribui

para a estabilidade e eficiência do processo de retropropagação durante o treinamento (Haykin, 1999).

4.2 Treinamento e resultados esperados

Com os dados de entrada selecionados e alinhados, o próximo passo é a programação e aplicação dos conceitos matemáticos das RNAs. O MATLAB proporciona uma plataforma com linguagem própria de programação, com a ferramenta para realização do treinamento e definição de parâmetros. Além disso, as funções de ativação e métodos de treinamento, como gradiente descendente e LM, já se encontram disponíveis nesta plataforma, facilitando possíveis utilizações de códigos complexos para desenvolver esses cálculos.

Neste momento, o papel do especialista é fundamental, visto ser ele que definirá quais entradas serão utilizadas, quais serão as saídas desejadas, dados de validação, dados de teste e parâmetros da RNA (quantidade de épocas, objetivo, mínimo gradiente, etc.). Todos esses componentes são chaves para uma rede bem desenvolvida e que esteja encontrando de fato um padrão dentro dos dados fornecidos, e não apenas entregando redes que aparentam bom funcionamento, mas não apresentam uma generalização satisfatória com dados externos.

Espera-se que, ao final do processo, a RNA seja capaz de mapear com precisão as relações entre as variações genéticas e a probabilidade de manifestação do LH. A saída, expressa em valores contínuos entre 0 e 1, deverá refletir a probabilidade estimada de presença da doença, em que valores próximos de 1 indicam maior propensão ao desenvolvimento do LH e valores próximos de 0 sugerem ausência da condição.

Durante o processo de treinamento, projeta-se uma convergência estável do erro ao longo das épocas, resultando em uma correlação consistente entre as saídas previstas e os valores esperados. Essa estabilidade é indicativa de uma aprendizagem eficiente e de uma rede adequadamente ajustada aos padrões genéticos relevantes. Além disso, espera-se que a RNA apresente boa capacidade de generalização, demonstrando desempenho satisfatório em dados não utilizados durante o treinamento, o que confirma que o modelo não apenas memorizou exemplos, mas realmente aprendeu as relações subjacentes entre os atributos genéticos e o diagnóstico da doença.

4.3 Possíveis dificuldades

Durante todo o processo de preparação dos dados e treinamento das RNAs, muitos obstáculos podem aparecer ao longo desse caminho, sendo alguns cruciais até mesmo para a viabilidade do projeto. Abaixo será destacado alguns dos problemas que podem ser enfrentados ao longo do desenvolvimento prático do projeto mencionado.

4.3.1 Quantidade de dados

Ao considerar o uso do sequenciamento genético como base de entrada para o treinamento das RNAs, é essencial buscar um bom desempenho e equilíbrio nos dados. Para isso, o ideal é contar com uma quantidade significativa de casos disponíveis para análise. No entanto, nem sempre essa tarefa é simples. Embora os bancos de dados do NCBI sejam vastos e contenham inúmeros registros, a filtragem de informações específicas pode representar um desafio considerável.

4.3.2 Tempo de treinamento

O desenvolvimento de uma RNA envolve cálculos de alta complexidade, especialmente quando se trabalha com grandes volumes de dados genéticos. Em razão disso, o tempo de processamento pode variar conforme a configuração adotada e o tamanho do comitê de redes utilizado. Essa etapa tende a demandar elevado poder computacional ou longos períodos de treinamento, o que representa um desafio significativo para a execução do projeto. Além disso, é comum que sejam necessárias múltiplas execuções experimentais com diferentes combinações de parâmetros, até que se identifique a estrutura de rede mais adequada e com desempenho satisfatório.

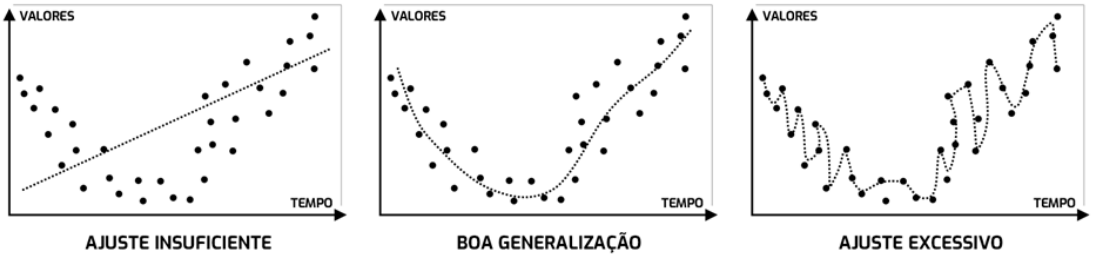
4.3.3 Overfitting e Underfitting

O overfitting (ajuste excessivo) ocorre quando uma rede neural aprende os dados de treinamento tão bem que se torna ineficaz em generalizar para dados não vistos, capturando ruído com os padrões nos dados. Por outro lado, underfitting (ajuste insuficiente) é quando a rede não aprende adequadamente os padrões nos dados de treinamento, resultando em um desempenho pobre tanto nos dados de treinamento quanto nos de teste. Ambos os cenários (Figura 4) são indesejáveis e comprometem a capacidade do modelo de fazer previsões precisas em dados novos e não vistos (Gavrilov *et al.*, 2018).

Esses fenômenos ocorrem principalmente ao ter uma discrepância entre a quantidade de dados em relação ao seu tamanho, ou então, a falta de validação dos dados durante o treinamento. Para mitigar esses problemas, uma das estratégias comumente adotadas em RNAs, é a técnica de validação cruzada, combinada com conjunto de testes (Russel; Norvig, 2004). Esse método divide os dados em subconjuntos, onde alguns são usados para treinamento e outros para validação e teste. Assim, é possível avaliar periodicamente o desempenho do modelo durante o treinamento e também verificar sua capacidade de generalização ao final. Tendo em vista esses desafios, é fundamental

enfatizar a relevância de um cuidadoso pré-processamento dos dados. Esse trabalho meticuloso, conduzido por especialistas, é crucial para garantir a robustez e precisão do modelo, evitando assim resultados indesejados e incoerentes.

Figura 4 - Gráficos de comparação do underfitting e overfitting, com uma rede de boa generalização.



Fonte: Autor

5 CONSIDERAÇÕES FINAIS

O presente estudo consolidou uma base teórica estruturada para o desenvolvimento de uma RNA voltada à identificação de padrões genéticos associados ao LH, delineando o percurso metodológico necessário para sua futura implementação prática. A construção desta estrutura envolveu a integração entre conceitos de IA e fundamentos genéticos, explorando desde os princípios matemáticos das RNAs até os desafios inerentes à preparação e alinhamento de sequências genéticas. Essa abordagem interdisciplinar permitiu estabelecer um modelo conceitual coerente, no qual o potencial das redes neurais se alinha às necessidades emergentes da medicina de precisão.

Embora o trabalho se mantenha em nível teórico, sua relevância reside em antecipar e organizar os elementos críticos para uma aplicação prática bem-sucedida. Aspectos como a seleção criteriosa de genes, a padronização das sequências, a definição dos rótulos supervisionados e o uso de funções de ativação apropriadas foram discutidos de forma integrada, demonstrando que a qualidade das etapas preparatórias é determinante para a performance do modelo final.

O estudo apresenta à formulação de um roteiro técnico e propõe uma visão estruturada de convergência entre biologia molecular e IA, onde o conhecimento computacional é aplicado como instrumento de investigação biomédica. Espera-se que os fundamentos aqui consolidados sirvam de ponto de partida para projetos experimentais futuros, voltados à validação empírica do modelo proposto e à criação de ferramentas preditivas que possam contribuir efetivamente para o prognóstico e compreensão do LH.

REFERÊNCIAS

ABDULLAH, Shubair A.; AL-ASHOOR, Ahmed. An Artificial Deep Neural Network for the Binary Classification of Network Traffic. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 11, n. 1, p. 402-408, 2020. DOI: https://doi.org/10.14569/IJACSA.2020.0110150.

AHMAD, Shandar; GROMIHA, M. Michael. NETASA: neural network based prediction of solvent accessibility. **Bioinformatics**, v. 18, n. 6, p. 819-824, 2002. DOI: https://doi.org/10.1093/bioinformatics/18.6.819.

BEJNORDI, B. E. *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. **JAMA**, v. 318, n. 22, p. 2199–2210, 2017.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. 2. ed. Indian Branch: Pearson Education, 2006.

BOTENTUIT, Raimundo Cezar Ribeiro *et al*. Mortalidade devido ao Linfoma de Hodgkin na região nordeste do Brasil nos anos de 2011-2020. **Research, Society and Development**, v. 12, n. 6, p. 1-9, 2023. DOI: https://doi.org/10.33448/rsd-v12i6.42313.

BRASIL. Conselho Nacional de Saúde. Resolução nº 466, de 12 de dezembro de 2012. Aprova as diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos. **Diário Oficial da União**, Brasília, DF, 13 jun. 2013, Seção 1, p. 59.

CARON, Mathilde *et al*. Deep Clustering for Unsupervised Learning of Visual Features. **ArXiv**, v. 2, 2018. DOI: https://doi.org/10.48550/arXiv.1807.05520

.

CARRERAS, Joaquim; HAMOUDI, Rifat. Artificial neural network analysis of gene expression data predicted non-Hodgkin lymphoma subtypes with high accuracy. **Machine Learning and Knowledge Extraction**, v. 3, p. 720–739, 2021. DOI: https://doi.org/10.3390/make300306.

CORAL, Rodrigo. Método para estimar a capacidade de refrigeração de compressores herméticos integrável à linha de produção. 2014. 238 p. Relatório (Pós-doutorado) – **Programa de Pós-Graduação em Engenharia Mecânica**, Universidade Federal de Santa Catarina, Florianópolis.

FREITAS, Lilian Ferrari de *et al*. Epidemiological and liver biomarkers profile of Epstein-Barr virus infection and its coinfection with cytomegalovirus in patients with hematological diseases. **Biomolecules**, v. 11, n. 8, p. 1–8, 2021. DOI: https://doi.org/10.3390/biom11081151.

ESTEVA, Andre *et al*. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115–126, 2017. DOI: 10.1038/nature21056.

GAVRILOV, Andrei Dmitri *et al*. Preventing model overfitting and underfitting in convolutional neural networks. **International Journal of Software Science and Computational Intelligence**, v. 10, n. 4, p. 19–28, 2018. DOI: https://doi.org/10.4018/IJSSCI.2018100102.

GLOBAL ALLIANCE FOR GENOMICS AND HEALTH (GA4GH). **Framework for Responsible Sharing of Genomic and Health-Related Data**. Toronto: GA4GH, 2016.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge, MA: MIT Press, 2017.

GRIFFITHS, Anthony J. F. *et al*. **Introdução à genética**. 11. ed. Rio de Janeiro: Guanabara Koogan, 2022.

HAGAN, Martin T.; MENHAJ, Mohammad B. Training feedforward networks with the Marquardt algorithm. **IEEE Transactions on Neural Networks**, v. 5, n. 6, 1994. DOI: https://doi.org/10.1109/72.329697.

HAYKIN, Simon. Neural **Networks**: A Comprehensive Foundation. 2. ed. Indian Branch: Pearson Education, 1999.

HERTEL, Cristina B. *et al*. Loss of B cell identity correlates with loss of B cell-specific transcription factors in Hodgkin/Reed-Sternberg cells of classical Hodgkin lymphomas. **Nature**, v. 21, p. 4908–4920, 2002. DOI: https://doi.org/10.1038/sj.onc.1205629.

JAGTAP, Ameya D. *et al*. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. **Journal of Computational Physics**, 2019. DOI: https://doi.org/10.1016/j.jcp.2019.109136.

KOVÁCS, Zsolt L. **Redes neurais artificiais**: fundamentos e aplicações. 4. ed. São Paulo: Livraria da Física, 2006.

LEE, Jaehoon *et al*. Wide neural networks of any depth evolve as linear models under gradient descent. **ArXiv**, Vancouver, v. 4, 2019. DOI: https://doi.org/10.48550/arXiv.1902.06720.

MAZUMDER, Dilwar Hussain; VEILUMUTHU, Ramachandran. An enhanced feature selection filter for classification of microarray cancer data. **ETRI Journal**, 2018. DOI: https://doi.org/10.4218/etrij.2018-0522.

NCBI RESOURCE COORDINATORS. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, 2015. DOI: https://doi.org/10.1093/nar/gkv1290.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology,** v. 48, n. 3, p. 443–453, 1970.

PENZ, Cesar Alberto. **Procedimentos para prover confiabilidade ao uso de inteligência artificial em ensaios de desempenho de compressores herméticos de refrigeração**. 2011. 180 p. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia Mecânica, Universidade Federal de Santa Catarina, Florianópolis, 2011.

PORTIS, Toni *et al*. Epstein-Barr Virus (EBV) LMP2A induces alterations in gene transcription similar to those observed in Reed-Sternberg cells of Hodgkin lymphoma. **Blood**, v. 102, n. 12, 2003. DOI: https://doi.org/10.1182/blood-2003-04-1018.

PUGACHEVA, Valentina; KOROTKOV, Alexander; KOROTKOV, Eugene. Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming. **Statistical Applications in Genetics and Molecular Biology**, v. 15, n. 5, p. 381–400, 2016. DOI: https://doi.org/10.1515/sagmb-2015-0079.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial**. 5. ed. Rio de Janeiro: Elsevier, 2004.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–5467, 1977. DOI: https://doi.org/10.1073/pnas.74.12.5463.

SANTOS, Cleyton Soares dos *et al*. Identificação de doenças genéticas na Atenção Primária à Saúde: experiência de um município de porte médio no Brasil. **Revista Brasileira de Medicina de Família e Comunidade**, v. 15, n. 42, p. 2347, 2020. DOI: https://doi.org/10.5712/rbmfc15(42)2347.

SCHWERING, Ines *et al*. Loss of the B-lineage–specific gene expression program in Hodgkin and Reed-Sternberg cells of Hodgkin lymphoma. **Blood**, v. 101, n. 4, 2003. DOI: https://doi.org/10.1182/blood-2002-03-0839.

SIEVERS, F. *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Molecular Systems Biology**, v. 7, p. 539, 2011.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **Journal of Molecular Biology**, v. 147, n. 1, p. 195–197, 1981.

TAKASHIMA, Kyoko *et al*. Ethical concerns on sharing genomic data including patients' family members. **BMC Medical Ethics**, v. 19, n. 61, p. 1–9, 2018. DOI: https://doi.org/10.1186/s12910-018-0310-5.

TCHÓRZEWSKI, Jerzy; WIELGO, Arkadiusz. Neural model of human gait and its implementation in MATLAB and Simulink Environment using Deep Learning Toolbox. **Studia Informatica**, v. 1–2, n. 25, 2021. DOI: https://doi.org/10.34739/si.2021.25.03.

TIWARI, Nitin *et al*. Mechanical characterization of industrial waste materials as mineral fillers in asphalt mixes: integrated experimental and machine learning analysis. **Sustainability**, v. 14, p. 5946, 2022. DOI: https://doi.org/10.3390/su14105946.

TOPOL, E. J. High-performance medicine: the convergence of human and artificial intelligence. **Nature Medicine**, v. 25, p. 44–56, 2019.

WANG, Shuang *et al*. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. **Annals of the New York Academy of Sciences**, v. 1387, n. 1, p. 73–83, 2017. DOI: 10.1111/nyas.13259.

WENIGER, M. A.; KÜPPERS, R. Molecular biology of Hodgkin lymphoma. **Leukemia**, v. 35, p. 968–981, 2021. DOI: 10.1038/s41375-021-01204-6.

WORLD HEALTH ORGANIZATION (WHO). **Ethics and governance of artificial intelligence for health**. Geneva: WHO, 2021.

YU, Hao; WILAMOWSKI, Bogdan M. **The Industrial Electronics Handbook**: Intelligent Systems. Levenberg–Marquardt Training. 1. ed. CRC Press, 2011.

ZENKE, Friedemann; GANGULI, Surya. SuperSpike: supervised learning in multilayer spiking neural networks. **Neural Computation, Massachusetts Institute of Technology**, v. 30, p. 1514–1541, 2018. DOI: 10.1162/neco_a_01086.

ZIEGENHAIN, Christoph; SANDBERG, Rickard. BAMboozle removes genetic variation from human sequence data for open data sharing. **Nature Communications**, p. 1–10, 2021. DOI: https://doi.org/10.1038/s41467-021-26152-8.