INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SANTA CATARINA - CÂMPUS JOINVILLE
DEPARTAMENTO ACADÊMICO DE ENGENHARIA ELÉTRICA
CURSO DE GRADUAÇÂO EM 2024

VALDIR FERREIRA FILHO

ESTUDO E DESENVOLVIMENTO DE REDES NEURAIS ARTIFICIAIS VOLTADAS AO AUXÍLIO DE DIAGNÓSTICO DA DOENÇA LINFOMA DE HODGKIN

JOINVILLE, 2024.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SANTA CATARINA - CÂMPUS JOINVILLE
DEPARTAMENTO ACADÊMICO DE ENGENHARIA ELÉTRICA
CURSO DE GRADUAÇÂO EM 2024

VALDIR FERREIRA FILHO

ESTUDO E DESENVOLVIMENTO DE REDES NEURAIS ARTIFICIAIS VOLTADAS AO AUXÍLIO DE DIAGNÓSTICO DA DOENÇA LINFOMA DE HODGKIN

Trabalho de Conclusão de Curso submetido ao Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina como parte dos requisitos para obtenção do título de Engenheiro em Engenharia Elétrica.

Orientador:
Prof. Rodrigo Coral, Dr. Engenharia

Coorientadora:
Daiani Cristina Savi, Dra. em Microbiologia

JOINVILLE, 2024.

Ferreira Filho, Valdir.

Estudo e desenvolvimento de redes neurais artificiais voltadas ao auxílio de diagnóstico da doença linfoma de Hodgkin/ Valdir Ferreira Filho. – Joinville, SC, 2024.
48 p.

Trabalho de Conclusão de Curso (Graduação) - Instituto Federal de Educação Ciência e Tecnologia de Santa Catarina, Curso de Engenharia Elétrica, Joinville, 2024.
Orientador: Rodrigo Coral
Coorientadora: Daiani Cristina Savi

Redes Neurais Artificiais (RNA). Linfoma de Hodgkin (LH). Mutações Genéticas. I. Instituto Federal de Educação Ciência e Tecnologia de Santa Catarina. II. Título.

Valdir Ferreira Filho

Estudo e desenvolvimento de redes neurais artificiais voltadas ao auxílio de diagnóstico da doença Linfoma de Hodgkin

Este trabalho foi julgado adequado para obtenção do título de Engenheiro Eletricista, pelo Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina, e aprovado na sua forma final pela comissão avaliadora abaixo indicada.

Joinville, 13 de dezembro de 2023.

Prof. Rodrigo Coral, Dr. Eng.
Orientador
Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina

Prof. Daiani Cristina Savi, Dra. em Microbiologia
Coorientadora
Católica de Santa Catarina

Prof. Michael Klug, Dr. Eng.
Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina

Prof. Stefano Romeu Zeplin, Me. Eng.
Instituto Federal de Educação, Ciência e Tecnologia de Santa Catarina

RESUMO

Este estudo aborda a elaboração teórica e prática de redes neurais artificiais (*RNAs*), uma subcategoria da inteligência artificial, focadas no auxílio do diagnóstico do Linfoma de Hodgkin (LH). O projeto engloba desde o marco teórico, detalhando as fundamentações conceituais e técnicas das RNAs, até a criação de modelos específicos. Neste contexto, são selecionadas arquiteturas e algoritmos adequados, empregando métodos de aprendizado e treinamento eficazes, com o objetivo de decifrar padrões complexos em mutações genéticas ligadas ao LH. A análise se baseia em dados genômicos rigorosamente selecionados do banco de dados NCBI, integrando essas informações com dados clínicos para estimar o desenvolvimento da doença. Esta pesquisa documenta todas as fases do projeto, desde a coleta e preparação dos dados até a validação final dos modelos de RNA. Destaca-se ainda, como a intersecção entre a genética, as RNAs e a medicina podem resultar em uma ferramenta diagnóstica promissora. Ao discutir os resultados e suas implicações, o trabalho visa contribuir significativamente para a evolução dos tratamentos personalizados e para a melhoria da precisão diagnóstica, não apenas para o Linfoma de Hodgkin, mas também para outras patologias de natureza genética.

Palavras-chave: Redes Neurais Artificiais (RNA). Linfoma de Hodgkin (LH). Mutações Genéticas.

ABSTRACT

This study addresses the development and practical of artificial neural networks (*ANNs*), a subcategory of artificial intelligence, focused on aiding the diagnosis of Hodgkin's Lymphoma (HL). The project encompasses everything from the theoretical framework, detailing the conceptual and technical foundations of ANNs, to the creation of specific models. In this context, suitable architectures and algorithms are selected, employing effective learning and training methods, with the aim of deciphering complex patterns in genetic mutations linked to HL. The analysis is based on meticulously selected genomic data from the NCBI database, integrating this information with clinical data to predict the development of the disease. This research documents all phases of the project, from data collection and preparation to the final validation of the ANN models. It also highlights how the intersection between genetics, ANNs, and medicine can result in a promising diagnostic tool. By discussing the results and their implications, the article aims to significantly contribute to the evolution of personalized treatments and to improve diagnostic accuracy, not only for Hodgkin's Lymphoma but also for other pathologies of a genetic nature.

Keywords: Artificial Neural Networks (ANNs). Hodgkin's Lymphoma (HL). Genetic Mutations.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RNA	Rede Neural Artificial
DNA	Ácido Desoxirribonucleicos
HRS	Células Reed-Sternberg
IA	Inteligência Artificial
LH	Linfoma de Hodgkin
LP	Linfócitos Predominantes
NCBI	National Center for Biotechnology Information (Centro Nacional de Informações sobre Biotecnologia
RBMFC	Revista Brasileira de Medicina da Família e Comunidade
RNA	Ácido Ribonucleico

SUMÁRIO

1 INTRODUÇÃO

Ao abordar o desenvolvimento de uma ferramenta promissora para apoiar um campo crítico da medicina contemporânea, torna-se essencial detalhar minuciosamente todo o processo e fundamentar-se em um marco teórico robusto para sua efetivação. Este estudo atual tem como propósito contextualizar os conceitos-chave e os principais aspectos, delinear os objetivos visados e, posteriormente, implementar sua aplicação prática.

1.1 Contextualização

Nos últimos anos, assistimos a uma verdadeira transformação no panorama tecnológico, particularmente notável no domínio da inteligência artificial (IA) e das redes neurais artificiais (*RNAs*). Estas inovações têm revolucionado uma ampla gama de setores, incluindo áreas cruciais como saúde e segurança pública, demonstrando a versatilidade e o potencial impactante dessas tecnologias (TEIXEIRA, 2019). As *RNAs*, em especial, têm emergido como ferramentas poderosas, abrindo novos caminhos para soluções inovadoras e eficientes em desafios complexos desses setores (GONÇALVES, 1994).

Um dos maiores obstáculos atualmente na saúde é o tratamento de doenças hematológicas, que incluem distúrbios como hemofilia, linfomas, anemia e policitemia vera, afetando o sangue e órgãos como o baço e gânglios linfáticos (BENEDEK *et al.*, 2016). Essas patologias podem ser benignas ou malignas e têm um impacto substancial na qualidade de vida. Dentro desse espectro, o Linfoma de Hodgkin (LH), um câncer linfático, se destaca pela necessidade de diagnóstico precoce, tratamento e seu impacto nas vida das pessoas, possuindo um registro de centenas de mortes no nordeste brasileiro na última década (BOTENTUIT *et al.*, 2023).

As *RNAs*, uma subcategoria da IA, têm se mostrado promissoras na análise de dados genéticos complexos para o diagnóstico de patologias como o LH. Estudos publicados na Nature evidenciam que algoritmos de IA são capazes de classificar lesões cutâneas com precisão igual ou superior à de profissionais médicos (ESTEVA *et al.*, 2017). A Revista Brasileira de Medicina da Família e Comunidade (RBMFC) ressalta também a importância de ferramentas eficazes como as RNAs para o diagnóstico precoce de doenças genéticas, destacando a necessidade de pesquisa e tratamento contínuos na área (SANTOS *et al.*, 2020).

1.2 Objetivo

No limiar de uma nova era médica, a presente pesquisa se dedica à intersecção entre tecnologia avançada e cuidado à saúde, procurando endereçar um dos

desafios mais persistentes no campo da oncologia: o diagnóstico do Linfoma de Hodgkin. Com a tecnologia da informação elevando-se como um pilar fundamental na luta contra doenças complexas, almeja-se construir uma ponte entre os dados genômicos disponíveis e as necessidades clínicas por meio de modelos de *RNAs* capazes de auxiliar na identificação de mutações genéticas com certa precisão e agilidade.

1.2.1 Objetivo geral

Este estudo visa o desenvolvimento de uma estrutura teórica e prática para a criação de redes neurais artificiais com a finalidade de auxiliar no diagnóstico da doença hematológica Linfoma de Hodgkin. Essa estrutura aborda desde as arquiteturas e algoritmos de *RNAs*, métodos de aprendizado e treinamento, até a integração de dados genéticos humanos. Ao identificar padrões complexos nas mutações genéticas associadas ao LH, a pesquisa busca criar uma ferramenta auxiliar diagnóstica, que poderá servir de base para futuros trabalhos relacionados e, consequentemente, possibilitar tratamentos mais eficazes e personalizados para doenças genéticas.

1.2.2 Objetivos específicos

Um dos cernes da pesquisa vigente é inicialmente o marco teórico e o aprofundamento nos conhecimentos específicos necessários para o desenvolvimento da ferramenta mencionada. A compreensão das *RNAs* e seus cálculos, vem a ser fundamental para o papel do especialista de selecionar com precisão os melhores métodos de criação dos modelos. Junto a isso, além do conhecimento de utilização dos bancos de dados internacionais, vem o estudo sobre a genética associada a patologia selecionada, verificando a fundo os fatores e mutações que impactam a sua evolução.

Outra base deste projeto visa a coleta de sequências genéticas do banco de dados NCBI, acompanhada de uma análise rigorosa para garantir que os dados sejam preparados adequadamente para o processamento. A etapa de alinhamento das sequências genéticas vem a ser um procedimento técnico essencial, visando a uniformidade e a confiabilidade dos dados que alimentarão as redes desenvolvidas.

A construção de modelos de *RNA* envolve a assimilação de conceitos técnicos e a seleção de ferramentas de programação adequadas. O estudo buscará entender as variáveis que influenciam os métodos de treinamento e como esses parâmetros podem ser otimizados para melhorar o desempenho dos algoritmos. Esta etapa é sobre aplicar conhecimento especializado para aprimorar as ferramentas de diagnóstico, não sobre reinventar conceitos, mas sim adaptá-los de maneira eficaz ao contexto da doença em estudo.

Por fim, o projeto procura identificar os desafios práticos na geração de *RNAs* quando utilizadas sequências genéticas como dados de entrada. A intenção é

relatar as dificuldades encontradas de maneira clara, propondo soluções factíveis e fornecendo recomendações que possam ser úteis para pesquisadores que continuarão a trabalhar nesta área. Este objetivo tem a intenção de contribuir de forma concreta para a melhoria contínua da aplicação de *RNAs* em contextos similares.

1.3 Estrutura Documental

Este trabalho foi concebido e estruturado adotando uma abordagem baseada em artigos científicos para compor os capítulos principais da monografia. O capitulo segundo é constituído pelo artigo que se dedica ao marco teórico, fornecendo uma base sólida e abrangente sobre as *RNAs* e sua aplicação no diagnóstico do Linfoma de Hodgkin. O capitulo subsequente, por sua vez, é formado pelo artigo que detalha o desenvolvimento prático do estudo, apresentando metodologias, experimentações e análises realizadas.

A escolha de formato para a escrita foi motivada pela intenção de facilitar a publicação dos artigos em revistas científicas. Acredita-se que este formato não só proporciona uma estrutura clara e bem definida para a monografia, mas também valoriza o conteúdo acadêmico ao prepará-lo para uma eventual publicação. Além disso, permite uma divulgação mais ampla dos resultados e das conclusões alcançadas, contribuindo para o avanço da pesquisa na área das *RNAs* aplicada à saúde.

2 MARCO TEÓRICO PARA O DESENVOLVIMENTO DE REDES NEURAIS ARTIFICIAIS VOLTADAS AO AUXÍLIO DE DIAGNÓSTICO DA DOENÇA LINFOMA DE HODGKIN

Este estudo foca em uma estrutura teórica para o desenvolvimento de redes neurais artificiais (*RNAs*) com o objetivo de auxiliar no diagnóstico da doença hematológica Linfoma de Hodgkin (LH). Será abordado inicialmente as arquiteturas e algoritmos de redes neurais, bem como métodos de aprendizado e treinamento. A investigação esclarece como essas técnicas podem ser aplicadas para identificar padrões complexos em meio às mutações genéticas no DNA humano, demonstrando ferramentas e conceitos para se alcançar o objetivo. Paralelamente, o artigo explora a doença Linfoma de Hodgkin, examinando os genes envolvidos e coletando informações fornecidas por meio do banco de dados NCBI. A integração para predizer o possível desenvolvimento de LH com as *RNAs* têm o potencial de criar uma ferramenta robusta e eficiente. Ao estabelecer essa base teórica multidisciplinar, o artigo lança as fundações para futuros desenvolvimentos de sistemas que buscam melhorar o diagnóstico de Linfoma de Hodgkin e de outras doenças predominantemente genéticas, o que pode, por consequência, alcançar tratamentos mais eficazes e personalizados.

2.1 Introdução ao Marco Teórico

Ao longo das últimas décadas, avanços tecnológicos têm redefinido dramaticamente nossas formas de viver e trabalhar. José Ernesto, com um mestrado em administração de empresas pela Universidade de São Paulo, explora o vasto impacto dessas mudanças, enfatizando setores como saúde e segurança pública (GONÇALVES, 1994). Paralelo a isso, a inteligência artificial (IA), especialmente no que se refere a redes neurais artificiais (*RNAs*), emerge como um dos domínios mais inovadores dentro do campo tecnológico (TEIXEIRA, 2019).

Apesar do grande desenvolvimento da tecnologia dos últimos anos, existem alguns pontos críticos principalmente para a área da saúde, em que se possui uma enorme dificuldade, uma delas é na prevenção, tratamento e descoberta de fatores genéticos relacionados a doenças hematológicas, que segundo estudos atinge centenas de pessoas apenas no norte do Brasil (FREITAS *et al.*, 2021). As doenças hematológicas são aquelas que afetam diretamente o sangue e os órgãos que estão ligados à sua produção, como a medula óssea. Essas doenças podem ser benignas ou malignas, e incluem condições como anemias, trombocitopenia, leucemias, linfomas, mieloma múltiplo, entre outras. Determinadas doenças trazem um impacto elevado na qualidade de vida dos pacientes, com possibilidades de causar à morte se não forem tratadas adequadamente.

Incluído na gama de doenças hematológicas, o Linfoma de Hodgkin é uma forma de câncer que se origina no sistema linfático, uma parte crucial do sistema

imunológico. Caracterizado pela presença na maioria dos casos 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. Entretanto, a doença ainda representa um desafio significativo em termos de mortalidade. Uma pesquisa recente revelou que, entre os anos de 2011 e 2020, um total de 1.141 brasileiros da região nordeste faleceram devido ao Linfoma de Hodgkin, destacando a necessidade contínua de pesquisas e intervenções eficazes para combater esta patologia. (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 (como a citada Linfoma de Hodking) na atenção primária da saúde, permitindo assim, um diagnóstico precoce. Além disso, é ressaltado a necessidade e utilização 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).

Dentro dos aspectos apontados, é notável a capacidade da inovação tecnológica relacionada a IA e as *RNAs* para com o desenvolvimento de ferramentas que possam a vir auxiliar no diagnóstico de doenças hematológicas. Um exemplo disso, artigos 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). Essa precisão é alcançada através do treinamento de redes neurais convolucionais em grandes conjuntos de imagens, dados laboratoriais e genéticos, permitindo que os algoritmos aprendam a reconhecer padrões e características específicas associadas a diferentes tipos de doenças (NETO; SILVA; NOGAROLI, 2020; MARTíNEZ, 2021).

Nessas condições, o projeto atual propõe uma fundamentação teórica sobre os métodos de aplicações para o desenvolvimento de *RNAs* relacionadas ao auxílio na identificação de padrões genéticos da doença Linfoma de Hodgkin. Outra questão levantada, é o apontamento de estudos para validação biológica do processo, além de explicações sobre a busca de informações em bancos de dados genéticos do NCBI. O marco teórico deseja alcançar o objetivo de facilitar trabalhos futuros relacionados ao assunto em questão e demais temas da área, se tornando um forte auxílio para produção de ferramentas inovadores e com enorme potencial. Como exemplo, o estudo do departamento de medicina do Japão, juntamente com outras instituições espalhadas no mundo, apontaram 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).

2.2 Fundamentação Teórica

A essência de uma pesquisa inicial relacionado a produção de *RNAs*, com o intuito de auxiliar no diagnóstico médico de doenças como o Linfoma Hodgkin, está conectado diretamente com a complexidade e possibilidades do tema. O universo que cerca as *RNAs* podem ser amplas e necessitam de uma contextualização para compreender como de fato é o seu funcionamento e sua base matemática. Somente com este entendimento, é possível direcioná-las para a aplicação visando solucionar problemas.

Além disso, uma boa preparação dos dados de entrada é fundamental para garantir a eficácia e validação biológica da rede desenvolvida, evitando possíveis complicações em seu treinamento. Juntamente a isso, é apresentada a genética voltada ao LH, determinando os genes relevantes e informações essenciais para sua compreensão. O mecanismo de estudo deste projeto vem para ajudar na base teórica sobre os diversos temas comentados acima, oferecendo um caminho estruturado para futuras explorações e descobertas na área.

2.2.1 Uma análise das 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 redes neurais artificiais, elas se diferem significativamente daqueles que utilizam processamento convencional, especialmente no que tange à modelagem dos fenômenos. 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).

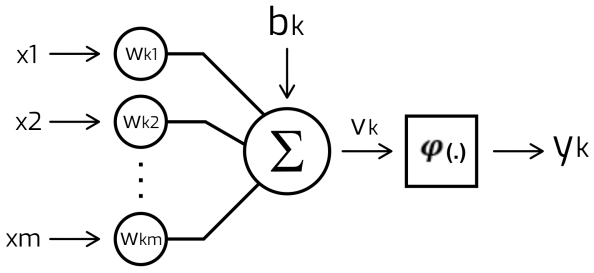
Dentre os benefícios observados das *RNAs*, se encontra a habilidade de mapear não-linearidades, realizar mapeamentos de entrada/saída, adaptar-se a pequenas modificações nas condições iniciais, e generalizar, fornecendo saídas adequadas para entradas não presentes durante o treinamento. Essas características 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 (CORAL *et al.*, 2015).

2.2.1.1 Redes Feedforward

As redes neurais feedforward, também conhecidas como redes neurais multicamadas (MLP, do inglês Multi-Layer Perceptron), são uma classe de redes neurais artificiais 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 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) obtém-se a função de ativação ($\phi(\cdot)$) 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 de múltipla 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.

2.2.1.2 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 se baseiam no supervisionado e não supervisionado, possuindo diferenças para as 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).

2.2.1.3 Função 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 será apontando as principais funções de ativação utilizadas em *RNAs* e comentado brevemente a sua utilização (JAGTAP *et al.*, 2019).

- Sigmóide (Função Logística)

A função logística $\sigma(x)$ 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.

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

- 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.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2}$$

2.2.1.4 Métodos de Treinamento

O treinamento de *RNAs* é uma tarefa complexa que envolve não apenas a otimização de uma série de parâmetros, mas também a seleção cuidadosa do algoritmo de aprendizado adequado ao problema em questão. Enquanto os métodos de aprendizado se referem à teoria subjacente que orienta as técnicas, os métodos de treinamento se concentram na aplicação prática de algoritmos para ajustar os pesos da rede e otimizar as suas funções. Será abordado alguns dos métodos mais comuns para o treinamento de redes neurais, demonstrando seus cálculos matemáticos para facilitar uma maior compreensão e auxiliar na 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, 2017). Trabalhos 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 $\mathcal{J}(\omega)$, onde a diferença entre os valores reais y e as previsões $\hat{y}^{(i)}$ é avaliada sobre todos os m exemplos do conjunto de treinamento.

$$\mathcal{J}(\omega) = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2 \tag{3}$$

Já o cálculo do gradiente 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:

$$\nabla \mathcal{J}(\omega) = \left[\frac{\partial \mathcal{J}}{\partial \omega_1}, \frac{\partial \mathcal{J}}{\partial \omega_2}, \ldots, \frac{\partial \mathcal{J}}{\partial \omega_n} \right] \tag{4}$$

Para finalizar, é realizada a atualização dos pesos na direção oposta ao gradiente, gerando assim a fórmula simplificada:

$$\omega_{\text{novo}} = \omega_{\text{antigo}} - \alpha \nabla \mathcal{J}(\omega) \tag{5}$$

Sendo:

ω = valor inicial dos pesos

α = taxa de aprendizado

- Levenberg-Marquardt (LM)

O método de Levenberg-Marquardt 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 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 agressiva e se assemelha mais ao método de Newton. Se uma atualização aumenta o erro, λ é aumentado, tornando a atualização mais próxima com o gradiente descendente.

$$\omega_{\text{novo}} = \omega_{\text{antigo}} - (J^T J + \lambda I)^{-1} J^T e \tag{6}$$

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

2.2.1.5 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 treinamento 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.

2.2.2 Ferramenta de Programação

Um dos passos importantes para o desenvolvimento de uma rede neural artificial, 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*, o MATLAB se destaca como uma alternativa viável. Este software é uma escolha atraente 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).

2.2.3 Genética da Doença Linfoma de Hodgkin

O estudo sobre a genética do Linfoma de Hodgkin abre novas perspectivas para entender os mecanismos moleculares que contribuem para a diversidade clínica e as respostas ao tratamento em pacientes. Esse segmento se dedica a explorar os elementos genéticos cruciais tanto para a origem quanto para o diagnóstico da doença LH.

2.2.3.1 Origem Celular

As células tumorais Hodgkin e Reed-Sternberg (HRS), que são características marcantes na doença de Linfoma de Hodgkin, têm sua origem nas células B, que são um tipo de célula do sistema imunológico responsável pela produção de anticorpos. No entanto, é intrigante que, apesar de sua origem, as células HRS frequentemente não retêm muitas das características genéticas e funcionais típicas das células B. Este fenômeno, onde as células HRS perdem na maioria o programa de expressão gênica típico das células B, foi evidenciado em diversos estudos (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.

2.2.3.2 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 Linfoma de Hodgkin Clássico (cHL). 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).

2.2.3.3 Tabela dos Genes

O artigo "Molecular biology of Hodgkin lymphoma"de Marc A. Weniger apresenta uma tabela (tabela 1) que lista várias alterações genéticas em células HRS e LP (Linfócitos Predominantes, mutação de menor frequência nas células B) associadas ao Linfoma de Hodgkin. 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 (WEINIGER; KÜPPERS, 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.

Tabela 1 – 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
	JAK2ª	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-L2ª	Evasão imunológica	Ganhos/amplificações	30
	XPO1	RNA nuclear e exportação de proteínas	SNVs (codon 571), gains	20
	ARID1A	Remodelação da cromatina	SNVs, indels	25
	JMJD2Cª	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

Adaptado de Weniger e Kuppers (2021).

2.2.4 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 ao surgimento de uma nova necessidade - a necessidade de locais adequados para armazenar, avaliar e compartilhar essas informações. Surgem, então, os bancos de dados genéticos especializados.

2.2.4.1 Centro Nacional de Informações sobre Biotecnologia (NCBI)

Estabelecido há mais de três décadas, o Centro Nacional de Informações sobre Biotecnologia dos Estados Unidos, conhecido como NCBI (COORDINATORS, 2015), desempenha um papel vital no mundo dos bancos de dados genéticos. Sua missão sempre foi reunir, categorizar e tornar acessíveis informações genéticas e biomoleculares para especialistas e pesquisadores de todos os cantos do mundo.

O NCBI abriga uma série de bancos de dados relacionados, e cada um tem seu próprio papel a desempenhar, como o GenBank, RefSeq e dbSNP, que são alguns dos bancos de dados chave que a plataforma mantém. Sendo um vasto depósito de sequências de nucleotídeos e proteínas, o GenBank armazena informações genéticas de muitos tipos de organismos. Esse banco de dados é atualizado todos os dias e os pesquisadores podem adicionar suas próprias sequências, garantindo que o conteúdo genético sempre seja recente.

Além dos bancos de dados genéticos, o NCBI fornece outros recursos valiosos, como a PubMed, uma base de dados bibliográfica que cobre literatura nas áreas biomédica e de ciências da vida, e o OMIM (Online Mendelian Inheritance in Man), um catálogo detalhado de genes humanos e fenótipos associados a doenças genéticas.

2.2.5 Gerando 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 utilizados 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).

2.2.5.1 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 *RNA* em tarefas de classificação binária 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* são evidenciadas em diversos estudos e aplicações, incluindo a identificação e categorização de pacotes de rede em contextos de segurança cibernética (ABDULLAH; AL-ASHOOR, 2020).

No âmbito deste trabalho, a classificação binária será utilizada para determinar se a entrada de dados haverá ou não a doença Linfoma de Hodgkin, levando a saída ser determinada de maneira supervisionada, em que é apontado o resultado desejado. Com este método, asseguramos uma saída na forma de percentual entre dois valores fundamentais, proporcionando uma visão clara da probabilidade associada à presença do câncer em foco na entrada analisada. Além disso, já é possível apontar com base nas funções de ativação demonstradas, que a Sigmóide (Função Logística) é a melhor para projeto, visto que sua faixa de atuação acontece justamente entre os valores de 0 e 1.

2.2.5.2 Seleção de Dados

O processo de seleção de dados para o treinamento da rede neural artificial se inicia na observação dos genes presentes na tabela 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 é muito relevante visto que abre a oportunidade de trabalhar a sequência genética dos mesmos para com os dados de entrada da rede desenvolvida.

No banco de dados NCBI, é perceptível a sua capacidade de encontrar informações genéticas e não é diferente com os genes apontados na tabela 1. 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 Linfoma de Hodgkin. As alterações ou mutações sofridas dentro da sequência são de extrema importância para determinar se haverá possibilidade da pessoa obter ou não a enfermidade, e esse vem a ser o grande

propósito da *RNA* desenvolvida.

2.2.5.3 Sequência Genética

Uma sequência genética é uma série ordenada de nucleotídeos, que são os blocos de construção fundamentais do DNA e RNA (ácido ribonucleico), e é codificada por quatro bases nitrogenadas: **adenina (A)**, **citosina (C)**, **guanina (G)** e **timina (T)** no DNA, ou **uracila (U)** no RNA. 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 como 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 (indels) e podem ocorrer em qualquer gene (ZIEGENHAIN; SANDBERG, 2021).

O processo de treinamento das redes neurais artificiais exigem que seus dados de entrada possuam uma padronização de tamanho, haja visto os métodos de cálculos utilizados, que precisam de matrizes com dimensões idênticas (MOORE *et al.*, 2023). Para resolver o problema comentado sobre as diferenciações dos genes, o processo de alinhamento é 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).

2.2.5.4 Treinamento

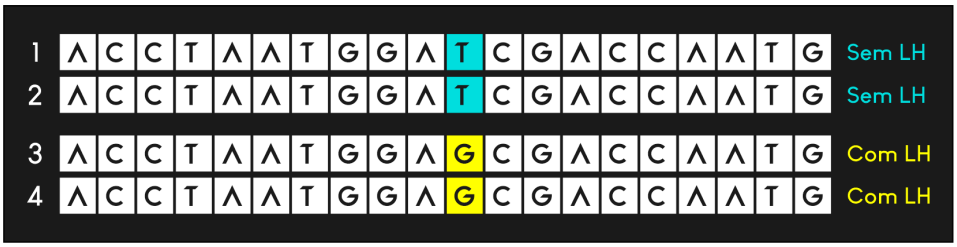
Com os dados de entrada selecionados e alinhados, o próximo passo é a programação e aplicação dos conceitos matemáticos das redes neurais artificiais. O MATLAB proporciona uma plataforma com linguagem própria de programação, com a ferramenta NN Toolbox para realização do treinamento e definição de parâmetros. 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.

A rede neural treinada é aprimorada para discernir padrões distintivos nas sequências genéticas associadas à presença ou ausência do Linfoma de Hodgkin. Essa análise comparativa é crucial, pois permite identificar regiões genômicas que diferem consistentemente entre indivíduos saudáveis e aqueles afetados pela doença. Tal identificação tem potencial para localizar mutações específicas responsáveis pelo LH, fornecendo uma base probabilística para a predição da doença em novas sequências genéticas. A Figura 2 ilustra de maneira simplificada este processo de detecção. Nela, são apresentadas quatro sequências de DNA, sendo duas de portadores do LH e duas de indivíduos sem a doença. As bases nitrogenadas que variam entre os dois grupos são realçadas, simbolizando os pontos críticos de divergência que podem ser chave na determinação genética do LH.

Figura 2 – Exemplo de comparação das sequências genéticas para identificação de mutações do Linfoma de Hodgkin.



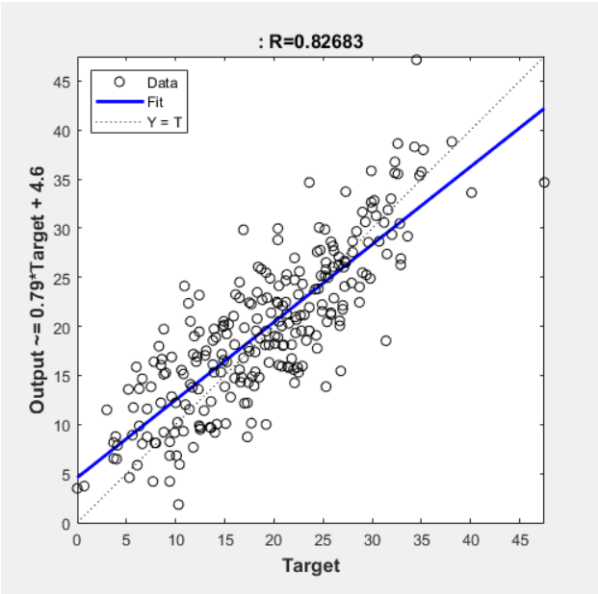
Fonte: Autor.

2.2.5.5 Métricas de Avaliação

Ao se deparar com a *RNA* finalizada, é necessário conseguir compreender a sua confiabilidade em relação ao treinamento, e para isso foi definido duas principais métricas de avaliação: análise do gráfico de regressão dentro do MATLAB e a validação dos dados de teste nunca vistos antes pela rede. No primeiro processo, envolve a comparação dos valores previstos pela rede com os valores reais dos dados de teste ou validação. A métrica resultante dessa comparação fornece uma medida quantitativa da precisão da rede neural em suas previsões. A figura 3 demonstra o gráfico no MATLAB com o eixo Y (output) possuindo os valores previstos, enquanto o eixo X (target) trás os

valores reais. A proximidade dos pontos de dados a esta linha ideal reflete a precisão da *RNA*, quanto mais próximos os pontos estiverem da linha, mais precisa é a previsão da rede e mais próximo de 1 é o valor de R (MATHWORKS, 2024).

Figura 3 – Exemplo do gráfico de regressão no MATLAB.



Fonte: Mathworks (2024).

Na fase de validação, a rede neural será submetida a um conjunto de dados de teste previamente separados e não expostos durante o treinamento, cujos resultados finais são conhecidos antecipadamente. Essa etapa é crucial para avaliar a capacidade de generalização da rede: após o treinamento, esses dados são introduzidos como um teste para determinar a eficácia da rede em generalizar o aprendizado para novos exemplos. A precisão da rede é medida pela sua habilidade em prever corretamente a presença ou ausência do LH. Quanto mais próximas as previsões da rede estiverem dos resultados reais, mais eficiente e confiável será considerada, indicando uma generalização bem-sucedida (GOODFELLOW; BENGIO; COURVILLE, 2017).

2.2.6 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.

2.2.6.1 Quantidade de dados

Ao levar em consideração a utilização de sequenciamento genético como base para entrada de dados no treinamento, notamos uma informação extremamente extensa, visto que normalmente o tamanho desses genes possuem milhares de bases

nitrogenadas. Para um melhor desempenho do treinamento e estabelecer um equilíbrio dos dados, encontrar uma quantidade significativa de casos que podem ser utilizados, vem a ser o ideal para o desenvolvimento. No entanto, nem sempre é uma tarefa fácil, por mais que os bancos de dados disponibilizados pelo NCBI sejam vastos e com muitos casos, a filtragem por informações específicas podem trazer uma dificuldade nesse sentido.

2.2.6.2 Tempo de treinamento

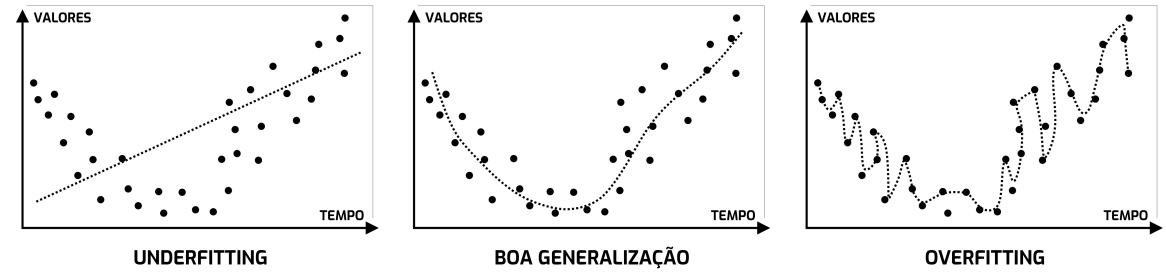
É notável que as equações demonstradas para o desenvolvimento de uma rede neural artificial pode ser muito complexo, ainda mais quando se trata de dados tão extensos e com tanta informação. Por conta disso, dependendo da configuração e do tamanho do comitê de redes desejado, o processo de treinamento pode exigir uma performance razoável de poder de processamento ou então um tempo longo de trabalho. Isso vem a ser uma complicação, visto que o projeto necessita normalmente de diversos treinamentos diferentes e variáveis, tentando assim encontrar o melhor caminho para uma rede performar de maneira desejável.

2.2.6.3 Overfitting e Underfitting

O overfitting 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 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 (RUSSELL; 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.



Adaptado da ABRACD (BRANCO, 2020).

2.3 Considerações Finais

No contexto da elaboração teórica para o desenvolvimento de uma rede neural artificial destinada a identificar padrões genéticos que possam auxiliar no diagnóstico do Linfoma de Hodgkin, este estudo delineou meticulosamente os pilares essenciais para viabilizar a implementação prática deste empreendimento científico. A jornada exploratória iniciou desde uma imersão na compreensão e detalhamento matemático das *RNAs*, passando por uma análise da ferramenta disponível para sua operacionalização. Além disso, foi realizada uma pesquisa focada na genética associada à referida patologia, na seleção e tratamento de dados e, crucialmente, na estruturação e refinamento dos dados de entrada.

Alguns aspectos potencialmente relevantes, tais como o mapeamento dos dados, a exploração de outras plataformas de programação e a consideração de métodos de treinamento alternativos, não foram profundamente explorados neste projeto. No entanto, a estrutura fornecida aqui estabelece um marco teórico robusto e uma fundação sólida para futuras investigações e implementações práticas. A aspiração deste trabalho é, portanto, catalisar a engenharia da *RNA* proposta, incentivando a busca por um banco de dados que seja não apenas adequado, mas também otimizado, para assim conduzir treinamentos com percepções significativas para o diagnóstico e compreensão do Linfoma de Hodgkin. A antecipação é que as sementes plantadas por esta pesquisa teórica possam germinar em futuros projetos, florescendo em inovações práticas e contribuições valiosas para a ciência e a medicina.

3 DESENVOLVIMENTO DE REDES NEURAIS ARTIFICIAIS VOLTADAS AO AUXÍLIO DE DIAGNÓSTICO DA DOENÇA LINFOMA DE HODGKIN

Este estudo descreve o desenvolvimento prático de redes neurais artificiais (*RNAs*) para o diagnóstico auxiliar da doença Linfoma de Hodgkin (LH). Detalha-se a construção de modelos de *RNAs*, incluindo a seleção de arquiteturas e algoritmos específicos, além dos métodos de aprendizado e treinamento utilizados. A aplicação dessas técnicas visa a identificação de padrões em mutações genéticas associadas ao LH, com base em dados genômicos obtidos do banco de dados NCBI. O trabalho apresenta o processo de integração das *RNAs* com dados clínicos para tentar prever o desenvolvimento do LH, destacando o potencial de criar uma ferramenta diagnóstica avançada e eficiente. O artigo documenta as etapas práticas do projeto, desde a coleta de dados até a validação do modelo, e discute as implicações dos resultados para o diagnóstico do LH, com o intuito de contribuir para o aprimoramento de tratamentos personalizados e aumentar a precisão diagnóstica para esta e outras doenças genéticas.

3.1 Introdução do desenvolvimento e aplicação

A medicina, em sua essência, é uma ciência que busca incessantemente compreender e tratar as complexidades do corpo humano. Ao longo dos séculos, a medicina tem se beneficiado de avanços em diversas áreas do conhecimento, desde a biologia até a física. No entanto, nas últimas décadas, a tecnologia e, mais especificamente, o crescimento da inteligência artificial (IA), tem se mostrado aliadas poderosas na busca por diagnósticos mais precisos e tratamentos mais eficazes (GONÇALVES, 1994)(MIRANDA et al., 2022).

Um dos desafios na área da saúde atualmente é o tratamento de enfermidades hematológicas. Elas incluem distúrbios como hemofilia, linfomas, anemia e policitemia vera, influenciando o sangue e os órgãos responsáveis pela sua produção, como o baço e os gânglios linfáticos. Essas patologias variam em gravidade, podendo ser não malignas ou cancerígenas, e têm um impacto profundo na qualidade de vida dos indivíduos, chegando a ser letais caso não recebam a intervenção médica adequada (BENEDEK *et al.*, 2016). O Linfoma de Hodgkin (LH), um tipo de câncer do sistema linfático marcado pelas células de Reed-Sternberg, atinge indivíduos de todas as idades e representa um desafio em mortalidade, evidenciado pela morte de centenas de pessoas no nordeste do Brasil entre 2011 e 2020, um acontecimento que realça a urgência em pesquisas e tratamentos na área (BOTENTUIT *et al.*, 2023).

Nesse contexto, as redes neurais artificiais (*RNAs*) surgem como uma solução promissora. Essa metodologia, que é uma subcategoria da IA, têm a capacidade de analisar grandes volumes de dados genéticos em um curto período de tempo, ofere-

cendo indicações valiosas que podem ser cruciais para o diagnóstico do LH. Estudos recentes, têm mostrado que, quando treinadas adequadamente, as *RNAs* podem identificar mutações genéticas específicas com uma precisão satisfatória (LEOSHCHENKO *et al.*, 2019).

O presente artigo é uma continuação do trabalho do próprio autor, visando ampliar e aplicar o marco teórico que foi desenvolvido em seu estudo anterior, focando na implementação prática de uma *RNA* específica para o diagnóstico do LH. Através da coleta e análise de dados genéticos reais, este trabalho visa não apenas demonstrar a eficácia e viabilidade da *RNA* na área médica, mas também discutir os desafios e limitações de sua aplicação.

3.2 Metodologia

No âmbito do desenvolvimento de uma *RNA* capaz de exercer um papel de auxílio no diagnóstico da doença hematológica do LH, o embasamento do marco teórico do estudo citado é de extrema relevância. Com estudo o anterior, vem a ser possível obter os referenciais necessários do processo prático para se alcançar os objetivos definidos. O presente trabalho irá delinear brevemente os principais conceitos e ferramentas essenciais para a sua evolução, reforçando a sua relevância. Além disso, será apresentado um detalhamento para a criação da *RNA*, indicando quais seus principais pontos de construção e como afetam o seu funcionamento.

3.2.1 Conceitos

Para uma melhor compreensão da operação da ferramenta que auxiliará no diagnóstico da patologia, é interessante contextualizar primeiro os principais temas relacionados a ela, antes de explorar mais profundamente seu progresso.

3.2.1.1 Redes Neurais Artificiais

As *RNAs* são sistemas de IA que simulam a arquitetura e os mecanismos do cérebro humano. Consistem em camadas de neurônios artificiais interligados que interpretam e aprendem informações. O processo de treino ajusta os pesos sinápticos dos neurônios para aprimorar a exatidão das predições e decisões da rede. Esse refinamento demanda o conhecimento de especialistas para selecionar e aplicar conjuntos de dados de treinamento apropriados. Distintamente dos métodos tradicionais que se apoiam em modelos matemáticos estabelecidos, as *RNAs* aprendem diretamente de dados reais, o que lhes confere vantagem em enfrentar questões complexas em que formulações matemáticas podem ser insuficientes. As *RNAs* destacam-se pela habilidade em captar relações não-lineares, ajustarem-se diante de variações de condi-

ções e generalizarem de maneira eficaz a partir das informações aprendidas durante o treinamento (HAYKIN, 1999).

Nesse universo, as redes feedforward ocupam um lugar de destaque. Essas são redes que processam informações de uma maneira unidirecional, sem qualquer forma de loop ou retorno. Além disso, são formadas por neurônios dispostos em camadas, com camadas ocultas entre as camadas de entrada e saída. Os neurônios destas camadas ocultas são cruciais, pois ajudam a detectar e definir características e funções que identificam o conjunto de treinamento. Redes feedforward são amplamente usadas em diversas tarefas de aprendizado de máquina, desde a classificação até o reconhecimento de padrões (HAYKIN, 1999).

O treinamento de *RNA* é uma operação complexa que necessita da otimização de vários parâmetros. Para tornar esse processo mais eficiente e eficaz, foram desenvolvidos algoritmos de treinamento, sendo um deles o método Levenberg-Marquardt (LM). Essa é uma técnica avançada de treinamento de *RNAs*, destacando-se pela eficácia em problemas não-lineares. Ele combina as vantagens do método de Newton com a robustez do gradiente descendente, proporcionando uma convergência mais rápida. Durante o treinamento, o LM ajusta dinamicamente o parâmetro de amortecimento, alternando entre as abordagens do método de Newton e do gradiente descendente. Essa alternância otimiza a atualização dos pesos da rede, tornando o processo de aprendizagem mais eficiente (YU; WILAMOWSKI, 2011; HAGAN; MENHAJ, 1994).

3.2.1.2 Linfoma de Hodgkin

O linfoma de Hodgkin é um tipo de câncer hematológico que se origina nos linfócitos, células especializadas que são um componente vital do sistema imunológico, encarregado de defender o corpo contra infecções e outras ameaças. Essa doença distingue-se de outros linfomas pela presença única de células tumorais denominadas células Hodgkin e Reed-Sternberg (HRS). Essas células, quando observadas sob um microscópio, são geralmente maiores do que as células normais e possuem uma aparência distintiva, frequentemente com dois núcleos. O linfoma de Hodgkin pode afetar pessoas de qualquer idade, mas é mais comum em adultos jovens e em pessoas acima dos 55 anos (CELLINI *et al.*, 2023).

O estudo da genética associada ao Linfoma de Hodgkin visa compreender os mecanismos moleculares que influenciam a diversidade clínica e as respostas terapêuticas nos pacientes. As células HRS, ícones desta doença, originam-se das células B, responsáveis pela produção de anticorpos (SCHWERING *et al.*, 2003). Outro ponto a se ressaltar sobre a patologia, é o Vírus Epstein-Barr (EBV), ao qual também desempenha um papel fundamental na genética do linfoma, principalmente na infecção

de células HRS. As proteínas expressas pelo EBV em casos positivos para o vírus contribuem diretamente para a patogênese do linfoma (PORTIS *et al.*, 2003).

Estudos recentes, como o de Weniger e Kuppers (2021), identificaram várias alterações genéticas associadas ao Linfoma de Hodgkin, destacando que não existe uma única mutação genética responsável pela doença e mostrando a presença dos Linfócitos Predominantes. A identificação dessas mutações e o apontamento dos genes responsáveis, é fundamental para avanços do estudo, visando a sua utilização como base para coleta de dados no treinamento da *RNA*.

3.2.1.3 Banco de Dados - NCBI

O National Center for Biotechnology Information (NCBI) é uma divisão da National Library of Medicine (NLM), que por sua vez faz parte do National Institutes of Health (NIH) dos Estados Unidos. Fundado em 1988, o NCBI foi estabelecido como uma resposta nacional à necessidade de organizar e disseminar informações biomédicas e biotecnológicas para pesquisadores, profissionais de saúde e o público em geral. Desde a sua criação, o NCBI tem desempenhado um papel fundamental no desenvolvimento de novas tecnologias de informação que facilitam a compreensão da genética e da biologia molecular. Uma de suas contribuições mais notáveis foi a criação do GenBank, o banco de dados público de sequências de DNA, que serve como uma biblioteca de sequências genéticas para pesquisadores de todo o mundo.

Atualmente, o NCBI continua a ser uma instituição vital para a comunidade científica global. Ele oferece uma variedade de bancos de dados e ferramentas que são essenciais para a pesquisa biomédica, incluindo o PubMed, uma base de dados de referências e resumos de artigos de pesquisa em ciências da vida e tópicos biomédicos. Além disso, o NCBI mantém o BLAST, uma ferramenta de alinhamento de sequências que permite aos pesquisadores comparar sequências de proteínas e nucleotídeos com as sequências no GenBank. Através desses recursos e muitos outros, o NCBI facilita o acesso a dados biomédicos, ajudando a impulsionar descobertas científicas e a avançar na medicina personalizada, contribuindo assim para a saúde e o bem-estar em todo o mundo (RICHA *et al.*, 2018).

3.2.2 Coleta, preparação e treinamento dos dados

O processo de desenvolvimento da *RNA* neste trabalho, se inicia na coleta de informações genéticas, logo após parte para a transformação dos genes coletados em dados de entrada viáveis para a *RNA*, iniciando assim a programação e treinamento, em que se aplicam os métodos de cálculos e fazem a obtenção dos resultados.

3.2.2.1 Extraindo Dados Genéticos

A coleta e preparação dos dados são etapas fundamentais em qualquer projeto de pesquisa, especialmente quando se trata de genética e biologia molecular. O objetivo principal desta fase é garantir que os dados coletados sejam precisos, relevantes e prontos para o processo subsequente.

Inicialmente, foi realizada uma análise aprofundada dos genes responsáveis por mutações no Linfoma de Hodgkin com base no artigo "Molecular biology of Hodgkin lymphoma"de Weniger e Kuppers (2021). Esse artigo é uma referência valiosa, visto que fornece informações detalhadas sobre os genes associados ao LH e suas respectivas mutações.

Após identificar os genes de interesse, a próxima etapa foi realizar pesquisas individuais para cada gene no NCBI, utilizando o banco de dados de nucleotídeos. Esse banco de dados em específico é uma coleção abrangente de sequências de DNA e RNA de várias fontes e espécies, fornecendo um rico banco de informações genéticas (SELAMA *et al.*, 2013). A estratégia de pesquisa adotada foi combinar o termo "doença"com o "gene"e o "ser vivo"de interesse. Por exemplo, para o gene BCL6, a pesquisa foi formulada como "Lymphoma BCL6 homosapiens".

Ao realizar a busca, dentro de cada caso, é possível verificar diversas informações associadas à pessoa responsável que sofreu com as mutações genéticas. Cada registro no banco de dados oferece uma riqueza de informações, incluindo a definição do gene, os autores que o publicaram, a tradução da proteína associada e, mais crucialmente, a sequência de DNA do gene em questão.

A sequência de DNA é, em essência, o código genético que determina as características de um organismo. É composta por quatro nucleotídeos diferentes: **adenina (A)**, **citosina (C)**, **guanina (G)** e **timina (T)**. A ordem específica desses nucleotídeos na sequência de DNA é o que codifica informações genéticas e determina a função e expressão de um gene (ALBERTS *et al.*, 2002). A análise dessas sequências é fundamental para entender as mutações e suas implicações na doença.

A quantidade de dados encontrados no NCBI varia consideravelmente dependendo do gene em questão. Além da disponibilidade de informações no banco de dados, outro aspecto notável é a variação no tamanho dos dados da sequência de DNA em um mesmo gene. Esta diferença de tamanho não é arbitrária, mas sim uma consequência de vários fatores, incluindo a complexidade do gene, eventos de inserções conhecidos como polimorfismos (ZIEGENHAIN; SANDBERG, 2021).

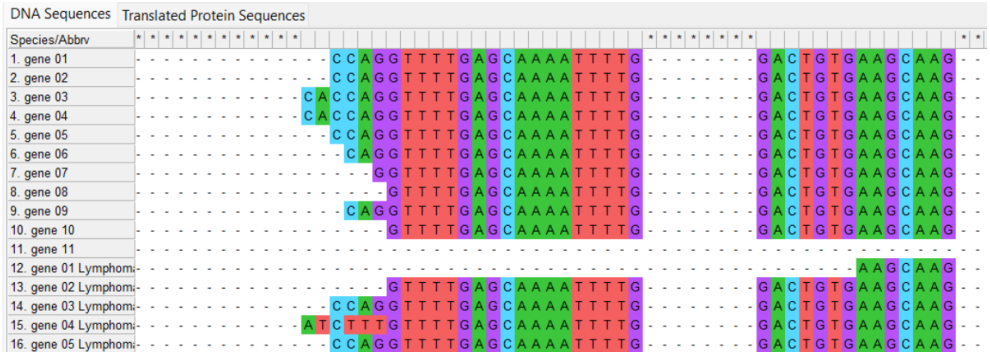
Essa variação no tamanho dos dados apresenta um desafio significativo quando se trata de análise envolvendo *RNAs*. As redes operam com dados de entrada em formato de matrizes e exigem que esses dados tenham dimensões consistentes para realizar cálculos matemáticos eficientes. A inconsistência no tamanho dos dados

genéticos pode, portanto, impedir o processamento eficaz e a análise subsequente.

3.2.2.2 Alinhamento Genético

Para superar o obstáculo de variação da extensão do gene, recorreu-se ao software MEGA (Molecular Ecolutionary Genetics Analysis), uma ferramenta amplamente reconhecida na bioinformática para alinhamento de sequências (KUMAR; TAMURA; NEI, 2004). O alinhamento é um processo pelo qual sequências de DNA são organizadas de maneira que nucleotídeos de diferentes sequências, mas de funções similares, se alinhem corretamente (PUGACHEVA; KOROTKOV; KOROTKOV, 2016). Durante esse processo, lacunas são introduzidas nas sequências para garantir o alinhamento adequado, e estas são representadas pelo símbolo **-**. As lacunas indicam a ausência de nucleotídeos naquela posição específica do gene e são essenciais para manter a extensão entre os genes idênticas sem danificar geneticamente o dado. Abaixo, na figura 4, é possível visualizar o alinhamento dos nucleotídeos com o preenchimento das lacunas no software MEGA versão 11 e após serem exportados para o MATLAB na figura 5.

Figura 5 – Alinhamento das sequências de DNA de um gene no MEGA.



Autor.

Figura 6 – Alinhamento das sequências de DNA de um gene no MATLAB.

Autor.

No entanto, um desafio adicional surge durante o alinhamento. Muitas sequências obtidas do NCBI podem não alinhar corretamente devido à presença de genes misturados ou informações genéticas irrelevantes. Esse acontecimento pode resultar na exclusão de muitos dados genéticos, reduzindo ainda mais o já limitado conjunto de informações por gene. Essa filtragem, embora necessária para garantir a precisão da análise, limita a quantidade de informação disponível para treinamento da *RNA*.

Após o meticuloso processo de alinhamento, os dados estão prontos para serem introduzidos para o treinamento. A quantidade de dados disponíveis vem a ser relativamente pequena e a dimensão de cada entrada vasta, muitas vezes trabalhando na casa dos milhares de nucleotídeos. Esta riqueza de informação, oferece uma dificuldade para o processo de treinamento, visto que essa discrepância pode levar a resultados que não tem um poder de generalização adequado, por conta disso, o foco na definição dos parâmetros e a etapa de teste para as *RNAs* será indispensável.

3.2.3 Programação

O treinamento de uma *RNA* é uma etapa crucial para garantir sua eficácia na identificação de padrões e na realização de predições. No contexto deste estudo, a sequência genética, composta pelos nucleotídeos **A**, **C**, **G** e **T**, foi enriquecida com a adição do caractere **-**, representando lacunas no alinhamento, como mencionado anteriormente. Dentre as definições de parâmetros para o funcionamento da rede, a função de ativação mais pertinente a ser aplicada é a Sigmóide, visando uma saída binária, ao imaginar a possibilidade de se obter ou não o LH, o que levaria a não necessidade de utilizar valores negativos trazidos pela função Tangente Hiperbólica.

Com as afirmações anteriores, notamos que a realização dos cálculos nas *RNAs* operam com dados numéricos, portanto é imperativo transformar as sequências genéticas em um formato que a rede possa entender e processar. A aplicação de mapeamento dos dados entra em cena. No contexto genético, o mapeamento envolve atribuir valores numéricos específicos a cada nucleotídeo e ao caractere de lacuna. Esta transformação é realizada por meio de código no MATLAB (plataforma versátil de computação numérica e simbólica, possuindo ferramentas especializadas para treinamento de *RNAs* (MATHWORKS, 2023)), que mapeia cada caractere da sequência genética para um valor numérico entre 0 e 1 (função logística), o resultado é possível observar na figura 7.

Figura 7 – Mapeamento das sequências de DNA de um gene no MATLAB.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.2500	0	0	0	0	0	0	0	0.2500	0.2500	0.2500	0.2500	0.2500
2	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
3	0.2500	0	0	0	0	0	0	0	0.2500	0.2500	0.2500	0.2500	0.2500
4	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
5	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
6	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
7	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
8	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
9	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
10	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
11	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
12	1	0	0	0	0	0	0	0	1	1	1	1	1
13	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
14	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
15	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
16	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
17	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
18	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
19	0.7500	0	0	0	0	0	0	0	0.7500	0.7500	0.7500	0.7500	0.7500
20	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000
21	0.5000	0	0	0	0	0	0	0	0.5000	0.5000	0.5000	0.5000	0.5000

Autor.

É crucial enfatizar a importância do mapeamento de todos os conjuntos de dados - seja de entrada, teste ou validação. A consistência deste processo garante que a RNA possa interpretar e processar os dados de maneira eficaz, levando a resultados mais precisos. Após esse processo e com os dados devidamente mapeados, inicia-se a fase de treinamento da RNA.

3.2.3.1 Treinamento

O processo de treinamento de uma *RNA*, especialmente quando se lida com um volume de entrada tão extenso como sequências genéticas, é uma tarefa computacionalmente intensiva. Cada rede, ao ser treinada, requer uma série de iterações e ajustes para alcançar a melhor generalização possível dos dados fornecidos. Isso significa que, para cada rede individual, pode-se esperar um tempo considerável de treinamento até que ela esteja pronta. O objetivo final é que a rede possa identificar e generalizar padrões nos dados de treinamento, permitindo que faça predições em dados não vistos anteriormente.

A definição dos parâmetros que compõem a arquitetura para o desenvolvimento da *RNA* ressalta a importância do especialista que está comandando o treinamento. Esses parâmetros são compostos por diversos influenciadores, sendo alguns deles a quantidade de épocas do sistema (vezes em que ocorrem interações com os cálculos da rede), quantidade máxima de convergências, objetivo a ser alcançado pelo gradiente (um valor próximo de zero), o método do treinamento e outros fatores.

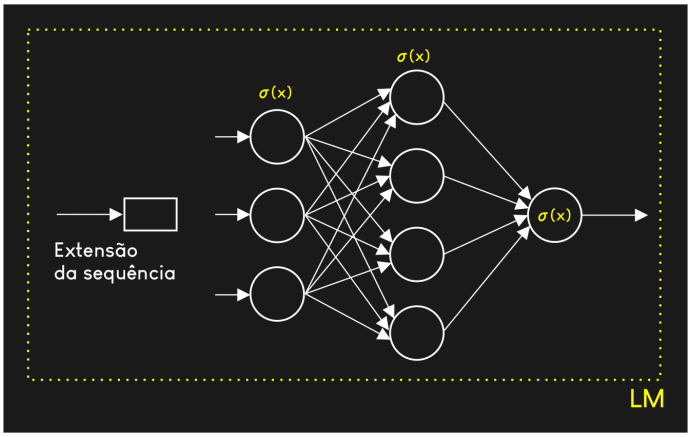
Mais um aspecto importante definido no código de treinamento é a escolha manual dos dados de validação, um processo valioso para evitar possíveis overffiting e underfitting na *RNA* (GAVRILOV *et al.*, 2018). A validação ocorre durante os cálculos, e serve como um termômetro para avaliar a generalização do modelo, isto é, sua capacidade de performar bem em dados não vistos anteriormente. Como a rede em

questão possui uma quantidade limitada de dados, deixar essa seleção de dados de validação de maneira aleatória tende a ser problemático, pois haveria alta probabilidade de pegar informações de um mesmo conjunto. Ao pensar nisso, é selecionado de forma aleatória uma sequência de cada caso (com e sem LH), e definidos como os dados de validação do treinamento.

Um dos conceitos utilizados no processo é o comitê de redes, sendo uma estratégia adotada para mitigar algumas das incertezas e variações inerentes ao treinamento das *RNAs*. Ao treinar várias redes e reunir seus resultados, é possível obter uma visão mais holística e equilibrada do problema em questão (PENZ, 2011). Para consolidar os resultados de todas as redes treinadas, aplica-se uma média aritmética simples. Essa abordagem tem a vantagem de suavizar os resultados, reduzindo o impacto de redes individuais que, por razões como inicialização aleatória de pesos, podem ter tido uma performance atípica ou subótima. Ao fazer isso, o comitê serve como um balizador, garantindo que as predições finais sejam mais estáveis e menos suscetíveis a distorções ou decorrentes do processo de treinamento.

A configuração escolhida para os modelos de *RNAs* é exemplificada na figura 8 e tem sua arquitetura em uma topologia feedforward com duas camadas ocultas: a primeira contendo 3 neurônios e a segunda 4 neurônios, todas utilizando a função de ativação logsig (sigmoid). Esta configuração, juntamente com a técnica de treinamento Levenberg-Marquardt e a execução de 50 épocas, equilibra a necessidade de capturar padrões genéticos complexos sem causar overfitting. As *RNAs* são treinadas em um comitê de 10 redes, processando dados genéticos dos genes BCL6, JMJD2C e TNFAIP3, com tamanho de sequências (entradas) de 8187, 5413 e 16446 bases nitrogenadas, respectivamente. Única variação dos parâmetros ocorreu em relação ao gene TNFAIP3, que por conta da sua extensa sequencia e limitações de processamento, necessitou uma redução de neurônio para 2 na primeira camada oculta.

Figura 8 – Arquitetura das RNAs treinadas.



Autor.

Após o treinamento ser concluído, a etapa decisiva é a aplicação dos dados de teste. Para isso, é requerido introduzir o conjunto de dados separados para teste dentro da *RNA* treinada. Ao alimentar a rede com essas informações, ela produz uma saída que varia no mesmo alcance do valor de entrada, entre 0 e 1. Esta saída representa a probabilidade estimada da sequência genética possuir mutações associadas à patologia Linfoma de Hodgkin. Como exemplo, uma saída próxima de 1 indica uma alta probabilidade do dado genético em questão estar associado à doença, enquanto uma saída próxima de 0 sugere o contrário. Esta informação é o objetivo a ser alcançado pelo desenvolvimento do trabalho, visto que poderá permitir aos pesquisadores e profissionais de saúde utilizá-la de forma auxiliar perante o risco do indivíduo desenvolver LH.

3.3 Resultados e Discussões

Ao longo da análise das sequências genéticas com a preparação para os dados de entrada, foi observado que muitos dos genes apontados no estudo de Weiniger e Kuppers (2021), não possuem uma quantidade significativa de casos para serem implementados na rede. Dentre os genes que se destacaram, foram separados os dados de pessoas sem ligação ao LH e outros dados envolvendo a patologia (sem LH/com LH). Lembrando que essa quantidade final e a eliminação dos demais genes, ocorreu justamente por conta do processo de alinhamento, em que foram observados muitos dados inconsistentes, e por conta de uma análise biológica, foi necessário removê-los.

Cada gene foi segmentado, reservando uma parcela de dados para os testes, que consistem em dados inéditos para a rede, utilizados para avaliar a eficácia da generalização. A seguir, apresentamos os resultados dos três genes principais, que dispunham de um volume suficiente de informações para o treinamento. As tabelas 2, 3 e 4 mostram a quantidade total de dados coletados e alinhados do NCBI, a porção utilizada no treinamento e a fração destinada aos testes.

Nota-se que, nos casos em que restou apenas um dado para teste, procedeu-se a um segundo treinamento, alternando os dados de treino para deixar uma sequência de DNA diferente para o teste, o que reforça assim a eficiente generalização da *RNA*. Para facilitar a visualização, as sequências genéticas foram numeradas de 1 até o total de sequências disponíveis, e ao lado, adicionou-se a sigla RT (rede treinada) seguida do número identificador da rede. Ademais, os valores numéricos de saída da rede de 0 a 1 foram passados para porcentagem, visando também um melhor entendimento.

Tabela 2 – Resultado gene TNFAIP3

Rede Treinada - Gene TNFAIP3	
Status da Doença - Total de dados/Dados treinados	
Sem Linfoma de Hodgkin - 9/8	
Sequência 09 - RT1	45.4%
Sequência 02 - RT2	46.2%
Com Linfoma de Hodgkin - 11/8	
Sequência 09 - RT1	82.3%
Sequência 10 - RT1	82.1%
Sequência 11 - RT1	60.4%
Sequência 09 - RT2	67.9%
Sequência 10 - RT2	67.1%
Sequência 11 - RT2	66.7%

Autor.

Tabela 3 – Resultado gene JMJD2C

Rede Treinada - Gene JMJD2C	
Status da Doença - Total de dados/Dados treinados	
Sem Linfoma de Hodgkin - 10/8	
Sequência 09 - RT1	44.5%
Sequência 10 - RT1	43.1%
Com Linfoma de Hodgkin - 10/8	
Sequência 09 - RT1	55.1%
Sequência 10 - RT1	56.4%

Autor.

Tabela 4 – Resultado gene BCL6

Rede Treinada - Gene BCL6	
Status da Doença - Total de dados/Dados treinados	
Sem Linfoma de Hodgkin - 11/8	
Sequência 09 - RT1	3.5%
Sequência 10 - RT1	3.5%
Sequência 11 - RT1	3.7%
Sequência 09 - RT2	24.4%
Sequência 10 - RT2	24.2%
Sequência 11 - RT2	25.2%
Com Linfoma de Hodgkin - 9/8	
Sequência 06 - RT1	63.6%
Sequência 02 - RT2	88.7%

Autor.

3.3.1 Análise dos Resultados

O primeiro passo para se obter uma análise boa em relação aos resultados, é válido ressaltar o significado das porcentagens presentes. Quando o resultado da rede aponta um valor abaixo de 50%, indica que a sequência analisada está mais propensa a não conter a mutação genética que leva ao LH, por outro lado, acima dessa porcentagem temos um caso com maior probabilidade de obter a mutação da patologia. Observando assim as tabelas, verifica-se que apesar da pouca quantidade de dados

coletados ao longo do processo de alinhamento, os resultados das sequências nunca vistas pela rede generalizaram bem, sempre apontando corretamente entre a sequência ter ou não a mutação do LH.

O gene BCL6 foi o que obteve a melhor resposta, possuindo valores mais próximos da extremidade e garantindo uma maior certeza sobre o seu resultado. Enquanto o JMJD2C e o TNFAIP3 por mais que obtiveram sucesso no apontamento da maior probabilidade entre os dois conjuntos, suas saídas ficaram em torno de 40% a 60%, o que representa uma maior incerteza perante a mutação do gene. Sendo assim, ao observar as *RNAs* treinadas no processo deste estudo, o gene BCL6, dentro das condições de quantidade de informação e disponibilidade de processamento, poderia ter um maior potencial de ser utilizado como auxílio ao diagnóstico da patologia LH.

3.3.1.1 Obstáculos

Ao longo deste estudo, torna-se claro que uma das maiores dificuldades enfrentadas diz respeito à quantidade limitada de dados disponíveis para cada gene. Por mais vasto que seja o banco internacional NCBI, a filtragem pelos genes específicos e o alinhamento genético fez com que ocorre-se uma redução significativa das informações para treinamento. Na pesquisa atual, focada em sequências genéticas de DNA, que geralmente apresentam um alto volume de entradas, é crucial maximizar a coleta de informações para alcançar a resposta mais precisa possível da *RNA*. Esta necessidade poderia ser mais facilmente atendida em futuras investigações, mediante colaborações com universidades e centros de pesquisa especializados na coleta de dados. Tais parcerias permitiriam a obtenção de um conjunto mais amplo de sequências, através de um estudo mais aprofundado sobre o alinhamento e a exploração de outros bancos de dados genéticos.

Outro aspecto que também impactou no desenvolvimento da rede, foi o seu tempo de treinamento para cada comitê de rede gerada. Apesar de usar máquinas com um nível de processamento relativamente bom, o tempo médio para a geração de um conjunto de *RNA* foi de 8 a 12 horas de processamento. Levando em conta que foram necessárias dezenas de tentativas para ajuste dos parâmetros, o tempo total de treinamento foi algo que dificultou maiores investigações e testes. Além disso, o gene TNFAIP3 por ser composto do maior volume de entradas por sequência de DNA, teve que ser treinado com parâmetros inferiores aos demais genes, visto que a memória computacional necessária ultrapassa a disponível.

Ao relacionar um estudo futuro, em que a quantidade de informações de que sequências será notavelmente superior, um poder computacional também deverá evoluir paralelamente. Uma opção viável seria o alocamento de serviços de computação em nuvem, em que é contratado processadores com base na sua necessidade, gerando

assim um maior aproveitamento com o ganho de tempo para o treinamento das *RNAs*. Assim, com um potencial computacional maior, será possível trabalhar com parâmetros internos de treinamento mais robustos, podendo alcançar assim resultados ainda mais precisos e confiáveis.

3.4 Considerações Finais

Este estudo representa um marco importante no campo da bioinformática e da medicina personalizada, demonstrando o potencial das *RNAs* no diagnóstico auxiliar do Linfoma de Hodgkin. Através da integração de dados genômicos e clínicos, o projeto destacou a capacidade das *RNAs* em identificar padrões complexos em mutações genéticas, potencializando a precisão diagnóstica, apesar da quantidade de informações limitadas disponíveis.

A seleção cuidadosa de arquiteturas e algoritmos de *RNA*, aliada aos métodos de aprendizado e treinamento, resultou em modelos capazes de interpretar dados genéticos com certa eficiência. Isso não apenas oferece uma nova perspectiva para o diagnóstico do LH, mas também abre caminho para aplicações semelhantes em outras doenças genéticas, reforçando a importância da inteligência artificial na medicina moderna.

Os resultados obtidos, especialmente com o gene BCL6, apontam para uma possível viabilidade prática das *RNAs* como ferramentas diagnósticas. Entretanto, os desafios enfrentados, como a limitação de dados e a exigência de processamento computacional intenso, evidenciam a necessidade de avanços contínuos tanto na coleta de dados quanto na infraestrutura tecnológica. Por conta disso, futuras pesquisas devem focar no aumento da base de dados genéticos e na melhoria dos algoritmos de *RNA*, visando aprimorar ainda mais a precisão e a confiabilidade dos diagnósticos. A colaboração entre geneticistas, médicos e engenheiros de dados é essencial para vencer os desafios atuais e explorar completamente as capacidades das *RNAs* na área médica.

O progresso alcançado no uso de *RNAs* para auxiliar o diagnóstico o Linfoma de Hodgkin não só representa um avanço significativo no campo da medicina personalizada, mas também serve como uma prova do impacto transformador que a inteligência artificlial e suas subcategorias podem ter na melhoria da saúde e do bem-estar humano.

4 CONCLUSÃO

Este trabalho demonstrou a eficácia e o potencial das *RNAs* na identificação de padrões genéticos associados ao Linfoma de Hodgkin, marcando um avanço significativo na aplicação da bioinformática e da inteligência artificial na medicina personalizada. A pesquisa aprofundada sobre a genética da patologia, combinada com uma análise rigorosa das arquiteturas, algoritmos e métodos de treinamento das *RNAs*, resultou em modelos capazes de discernir complexas mutações genéticas, contribuindo para um diagnóstico mais preciso da doença.

Apesar dos desafios enfrentados, como a limitação na quantidade de dados e as demandas de processamento computacional, os resultados obtidos, particularmente com o gene BCL6, indicam a viabilidade das *RNAs* como ferramentas auxiliares no diagnóstico do LH. Esta investigação não apenas fornece novas descobertas para o diagnóstico dessa patologia específica, mas também abre caminho para futuras aplicações em outras doenças genéticas, enfatizando o papel crucial da inteligência artificial na evolução da medicina.

Contudo, para maximizar o potencial das *RNAs*, é imprescindível um avanço contínuo na coleta e no processamento de dados genéticos, além da melhoria dos algoritmos de *RNA*. A colaboração interdisciplinar entre profissionais da área é fundamental para superar os obstáculos atuais e explorar integralmente as capacidades das *RNAs* no campo médico. O progresso obtido neste estudo não apenas realça o impacto transformador da inteligência artificial na saúde, mas também estabelece um modelo para futuras pesquisas e desenvolvimentos na área da medicina personalizada.

REFERÊNCIAS

ABDULLAH, S. A.; AL-ASHOOR, A. 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. 25

AHMAD, S.; GROMIHA, M. M. Netasa: neural network based prediction of solvent accessibility. *Bioinformatics*, v. 18, n. 6, p. 819–824, 2002. 21

ALBERTS, B. *et al. Molecular Biology of The Cell*. 4th. ed. [S.l.]: Garland Science, 2002. 35

BENEDEK, I. *et al.* Acute coronary syndromes in patients with hematological disorders. *Journal of Cardiovascular Emergencies*, v. 2, n. 4, p. 159–168, 2016. 11, 31

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 2nd. ed. Indian Branch: Pearson Education, 2006. 17

BOTENTUIT, R. C. R. *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. 11, 15, 31

BRANCO, H. *Overfitting e underfitting em Machine Learning*. 2020. ABRACD (Associação Brasileira de Ciência de Dados). Acessado em: ago 2023. Disponível em: https://abracd.org/overfitting-e-underfitting-em-machine-learning. 30

CARON, M. *et al. Deep Clustering for Unsupervised Learning of Visual Features*. 2018. ArXiv preprint arXiv:1807.05520. 18

CARRERAS, J.; HAMOUDI, R. Artificial neural network analysis of gene expression data predicted non-hodgkin lymphoma subtypes with high accuracy. *Mach. Learn. Knowl. Extr.*, v. 3, p. 720–739, 2021. 15

CELLINI, A. *et al.* Tackling the dysregulated immune-checkpoints in classical hodgkin lymphoma: bidirectional regulations between the microenvironment and hodgkin/reed-sternberg cells. *Frontiers in Oncology*, v. 13, p. 1–13, 2023. 33

COORDINATORS, N. R. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 2015. 24

CORAL, R. *Método Para Estimar A Capacidade De Refrigeração De Compressores Herméticos Integrável À Linha De Produção*. Tese (Relatório de pós-doutorado) — Universidade Federal de Santa Catarina, Florianópolis, 2014. 17

CORAL, R. *et al.* Development of a committee of artificial neural networks for the performance testing of compressors for thermal machines in very reduced times. *Metrology and Measurement Systems*, v. 22, n. 1, p. 79–88, 2015. 17

ESTEVA, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, v. 542, n. 7639, p. 115–126, 2017. 11, 15

FREITAS, L. F. D. *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. 14

GAVRILOV, A. D. *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. 29, 38

GONÇALVES, J. E. L. Os impactos das novas tecnologias nas empresas prestadoras de serviços. *Revista de Administração de Empresas*, v. 34, n. 1, p. 1–19, 1994. 11, 14, 31

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2017. 16, 19, 28

HAGAN, M. T.; MENHAJ, M. B. Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, v. 5, n. 6, 1994. 20, 33

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2nd. ed. Indian Branch: Pearson Education, 1999. 17, 18, 21, 33

HERTEL, C. 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. *Natura*, v. 21, p. 4908–4920, 2002. 22

JAGTAP, A. D. *et al.* Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *J. Comput. Phys*, 2019. 18

KOVÁCS, Z. L. *Redes Neurais Artificiais: Fundamentos e Aplicações*. 4th. ed. São Paulo: Livraria da Física, 2006. 16

KUMAR, S.; TAMURA, K.; NEI, M. Mega3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, v. 5, n. 2, p. 150–163, 2004. 36

LEE, J. *et al. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent*. Vancouver: [s.n.], 2019. ArXiv preprint arXiv:1902.06720. 19

LEOSHCHENKO, S. D. *et al.* Modification and parallelization of genetic algorithm for synthesis of artificial neural networks. *Radio Electronics, Computer Science, Control*, v. 4, n. 7, p. 15, 2019. 32

MARTíNEZ, R. H. *Modelo piloto de aprendizaje automático para la predicción de mortalidad por Leucemia Mieloide Aguda*. 87 p. Dissertação (Dissertação de Mestrado) — Universidad Internacional de La Rioja (UNIR), Bucaramanga, 2021. 15

MATHWORKS. *Deep Learning Solutions*. 2023. Disponivel em: www.mathworks.com/solutions/deep-learning. Acessado em: set 2023. 37

MATHWORKS. *Body Fat Estimation*. 2024. Disponivel em: www.mathworks.com/help/deeplearning/ug/body-fat-estimation.html. Acessado em: out 2023. 28

MAZUMDER, D. H.; VEILUMUTHU, R. An enhanced feature selection filter for classification of microarray cancer data. *ETRI Journal*, 2018. 25

MOORE, N. S. *et al.* Graph neural networks and applied linear algebra. *ArXiv*, v. 1, 2023. 26

NETO, M. K.; SILVA, R. da G.; NOGAROLI, R. Inteligência artificial e big data no diagnóstico e tratamento da covid-19 na américa latina: Novos desafios À proteção de dados pessoais. *Revista Brasileira De Direitos Fundamentais & Justiça*, v. 14, n. 1, p. 149–178, 2020. 15

PENZ, C. A. *Procedimentos Para Prover Confiabilidade Ao Uso De Inteligência Artificial Em Ensaios De Desempenho De Compressores Herméticos De Refrigeração*. 180 p. Tese (Tese de doutorado) — Universidade Federal de Santa Catarina, Florianópolis, 2011. 21, 39

PORTIS, T. *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. 22, 34

PUGACHEVA, V.; KOROTKOV, A.; KOROTKOV, E. Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming. *Stat. Appl. Genet. Mol. Biol.*, v. 15, n. 5, p. 381–400, 2016. 26, 36

RICHA, A. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*, v. 46, 2018. 34

RUSSELL, S.; NORVIG, P. *Inteligência Artificial*. 5th. ed. Rio de Janeiro: Elsevier Editora, 2004. 29

SANGER, F.; NICKLEN, S.; COULSON, A. R. Dna sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. Usa*, v. 74, n. 12, p. 5463–5467, 1977. 24

SANTOS, C. S. D. *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. 11, 15

SCHWERING, I. *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. 22, 33

SELAMA, O. *et al.* The world bacterial biogeography and biodiversity through databases: A case study of ncbi nucleotide database and gbif database. *Hindawi Publishing Corporation*, p. 11, 2013. 35

TCHÓRZEWSKI, J.; WIELGO, A. 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. 22

TEIXEIRA, J. de F. *O que é inteligência artificial*. 3rd. ed. São Paulo: E-galáxia, 2019. 11, 14

TIWARI, N. *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. 21

WEINIGER, M. A.; KÜPPERS, R. Molecular biology of hodgkin lymphoma. *Leukemia*, v. 35, p. 968–981, 2021. 23

YU, H.; WILAMOWSKI, B. M. *The Industrial Electronics Handbook: Intelligent Systems. Levenberg–Marquardt Training*. 1st. ed. [S.l.]: CRC Press, 2011. 20, 33

ZENKE, F.; GANGULI, S. Superspike: Supervised learning in multilayer spiking neural networks. *Neural Computation*, Massachusetts Institute of Technology, v. 30, p. 1514–1541, 2018. 18

ZIEGENHAIN, C.; SANDBERG, R. Bamboozle removes genetic variation from human sequence data for open data sharing. *Nature*, p. 1–10, 2021. 26, 35