CURSO DE CIÊNCIA DA COMPUTAÇÃO – TCC (RES_020/2016 – 2024_2)	
( ) PRÉ-PROJETO (X) PROJETO	ANO/SEMESTRE: 2024/2

RECONHECIMENTO DE PADRÕES EM DADOS GENÔMICOS ASSOCIADOS À PREDISPOSIÇÃO AO CÂNCER DE MAMA

Ana Caroline Cipriani dos Santos

Prof. Aurélio Faustino Hoppe - Orientador

1 INTRODUÇÃO

O câncer de mama é considerado um dos principais causadores da mortalidade entre mulheres no contexto global, com mais de 2 milhões de novos diagnósticos registrados a cada ano. Na perspectiva de Arnold *et al.* (2022), estima-se que, até 2040, o número de novos casos anuais poderá ultrapassar 3 milhões, resultando em cerca de 1 milhão de óbitos relacionados à doença. De acordo com Conceição *et al*. (2022), a detecção precoce e a prevenção são fundamentais para reduzir a mortalidade e melhorar a qualidade de vida das pacientes. No entanto, ainda segundo os autores, a complexidade do câncer de mama e a variabilidade individual tornam desafiador o desenvolvimento de estratégias eficazes de prevenção e tratamento.

Segundo pesquisas científicas, como as conduzidas por Quazi (2022) e Chowdhury (2024), diversas doenças podem ser diagnosticadas precocemente, bem como mitigadas com o auxílio do estudo genético, possibilitando o desenvolvimento de estratégias de prevenção mais eficazes. No contexto do câncer de mama, um dos tipos de câncer mais comuns entre as mulheres, a predisposição genética é um fator importante a ser considerado. Mutações em genes específicos, como o *Breast Cancer 1* e o *Breast Cancer 2* (BRCA1/2), estão diretamente associadas a um aumento significativo no risco de desenvolvimento da doença, o que torna essencial a análise genética na formulação de estratégias personalizadas para o diagnóstico e prevenção (Fu *et al*., 2022).

Os genes BRCA1 e BRCA2 desempenham papéis fundamentais na reparação de danos ao *DeoxyriboNucleic Acid* (DNA) e na manutenção da estabilidade genômica, funções críticas para a prevenção da transformação maligna das células (Wajid; Rehman, 2024). Estudos indicam que entre 5% e 10% dos casos de câncer de mama são de natureza hereditária, com mutações nesses genes responsáveis pela maior parte dos casos (Coelho *et al*., 2018). Considerando a elevada taxa de risco associada a essas mutações, é relevante, do ponto de vista da saúde pública, implementar medidas de diagnóstico precoce e estratégias preventivas personalizadas, como o aconselhamento genético e o rastreamento regular, para mulheres com predisposição hereditária (Guindalini *et al*., 2021).

De acordo com Quazi (2022), a detecção do câncer de mama no ambiente clínico faz uso de diversos métodos, com destaque para a mamografia, a ultrassonografia e a Ressonância Magnética (RM). Marcomini (2013) destaca que a mamografia, amplamente utilizada, oferece uma visualização detalhada de lesões e alterações no tecido mamário, sendo o principal exame de rastreamento. A ultrassonografia complementa essa avaliação, proporcionando uma análise mais profunda de lesões suspeitas, auxiliando na distinção entre patologias benignas e malignas. A RM, por sua vez, é essencial na determinação da extensão tumoral (Silva *et al.*, 2023). Além dos métodos convencionais, a análise de imagens médicas tem se beneficiado da utilização de técnicas avançadas de processamento de imagens e de aprendizado de máquina. Essas tecnologias, aplicadas tanto na análise de mamografias quanto de ultrassonografias, permitem a detecção automática de lesões, aumentando a precisão e a rapidez do diagnóstico, contribuindo para uma detecção precoce e mais eficaz do câncer de mama (Quazi, 2022).

O diagnóstico precoce do câncer de mama em mulheres com predisposição hereditária pode ser aprimorado por meio do aconselhamento genético e do rastreamento regular. O aconselhamento genético é um processo sistemático que avalia detalhadamente a história familiar e médica do indivíduo, visando identificar o risco de desenvolvimento de doenças genéticas, como o câncer de mama. Conduzido por profissionais especializados, como médicos, enfermeiros ou conselheiros genéticos, esse processo envolve a análise de antecedentes familiares e a discussão de testes genéticos, como a pesquisa de mutações no gene BRCA1, fornecendo suporte para a formulação de estratégias personalizadas de prevenção (Teofilo *et al.*, 2024).

Além do aconselhamento genético, o rastreamento regular inclui exames médicos periódicos, como mamografias e ultrassonografias, para a detecção de alterações no tecido mamário. Em alguns casos, testes genéticos adicionais para mutações nos genes BRCA1 e BRCA2 também são realizados. O objetivo combinado dessas estratégias é identificar mulheres com predisposição hereditária ao câncer de mama, oferecendo-lhes opções adequadas de prevenção, como a mastectomia profilática, terapia hormonal e tratamentos personalizados (Kaliks *et al.*, 2009). Além disso, essas intervenções são pertinentes para mulheres já diagnosticadas com câncer, auxiliando na escolha de tratamentos eficazes e na prevenção de recidivas.

Sob essa perspectiva, segundo Sikandar *et al.* (2020), a utilização de técnicas de aprendizado de máquina emergiu como uma ferramenta promissora na análise de dados genômicos. Essas técnicas permitem identificar

padrões em grandes volumes de dados, que poderiam passar despercebidos em análises convencionais, oferecendo novas perspectivas para a predição do risco de doenças. Considera-se que a combinação com redes complexas pode aprimorar ainda mais essa análise, auxiliando na compreensão das interações entre genes e variantes genéticas e permitindo identificar padrões e interações que podem ser críticos para o desenvolvimento da doença. Esta abordagem integrada não apenas aprimora a predição de risco e o diagnóstico precoce, mas também fornece uma compreensão mais profunda dos mecanismos moleculares envolvidos na progressão do câncer de mama.

Diante disso, o presente estudo busca responder à seguinte pergunta de pesquisa: como técnicas de aprendizado de máquina e redes complexas podem ser utilizadas para identificar padrões genéticos específicos associados à predisposição ao câncer de mama em portadores de mutações no gene BRCA1?

1.1 OBJETIVOS

O objetivo principal deste trabalho é identificar padrões genéticos específicos associados à predisposição ao câncer de mama em portadores de mutações no gene BRCA1, utilizando técnicas de aprendizado de máquina e redes complexas.

Os objetivos específicos são:

a) identificar variantes genéticas relevantes do gene BRCA1 associadas ao risco de câncer de mama, por meio da análise de bases de dados genômicas públicas;
b) explorar as interações entre diferentes variantes genéticas, utilizando técnicas de redes complexas para determinar como elas contribuem para o aumento do risco de desenvolvimento do câncer de mama;
c) avaliar a eficácia e a precisão do modelo desenvolvido.

2 TRABALHOS CORRELATOS

Nesta seção, são apresentados trabalhos que possuem características semelhantes aos principais objetivos do estudo proposto. A revisão sistemática realizada buscou identificar pesquisas recentes que utilizem técnicas de aprendizado de máquina para a análise de dados genômicos associados a doenças, dando enfoque ao câncer de mama, especialmente em portadores de mutações no gene BRCA1. Na subseção 2.1, é apresentada a revisão sistemática dos trabalhos selecionados, indicando os assuntos abordados, os filtros utilizados para a busca e os critérios de preferência para a escolha dos trabalhos correlatos. Por fim, na subseção 2.2, é feita uma síntese dos trabalhos correlatos selecionados, destacando suas principais contribuições, metodologias aplicadas, resultados obtidos e suas relações com o estudo proposto. Essa síntese permitiu compreender como cada trabalho se alinha aos objetivos do presente estudo e de que forma eles podem contribuir para o desenvolvimento de novas abordagens na identificação de padrões genéticos associados ao câncer de mama utilizando técnicas de aprendizado de máquina e redes complexas.

2.1 REVISÃO SISTEMÁTICA

Para a pesquisa de trabalhos correlatos, foram utilizados os portais de busca Science Direct, Litmaps com integração ao Semantic Scholar, Google Acadêmico e IEEE Xplore. A seleção dos estudos priorizou os trabalhos mais recentes, com um filtro de publicações entre 2020 e 2024. Conforme demonstra o Quadro 1, as *strings* de busca incluíram termos como “câncer de mama”, “aprendizado de máquina”, “variantes genéticas”, “predisposição genética”, “BRCA1”, “predição”, entre outros. A partir disso, os termos foram traduzidos para o inglês, sendo repetidas as buscas com os mesmos conjuntos, conforme pode ser observado no Quadro 2.

Quadro 1 – Resultado de buscas por termos contidos em artigos em língua portuguesa

Termos de busca	Semantic Scholar	Science Direct	Google Acadêmico	IEEE Xplore	Total
“aprendizado de máquina” + “variantes genéticas” + “predição”	120	-	1.100	-	1.220
“aprendizado de máquina” + “predição” + “dados genômicos”	625	-	317	-	942
“aprendizado de máquina” + “modelos preditivos” + “predisposição genética”	36	-	171	-	207
“aprendizado de máquina” + “BRCA1” + “predição”	114	-	27	-	141
“câncer de mama” + “predição” + “aprendizado de máquina	164	-	464	-	628
“dados genômicos” + “doença” + “predição”	87	-	1.520	-	1.607

Fonte: elaborado pela autora.

O Quadro 1 apresenta a quantidade de artigos que atenderam aos termos de busca em português, considerando o critério de seleção por ano de publicação. A busca resultou em aproximadamente 5 mil

documentos. No entanto, ao realizar a leitura dos títulos dos artigos que mais se correlacionavam, identificou-se que muitos enfatizavam a predição de doenças mais específicas ou na predição de sobrevida relacionado ao câncer de mama. Além disso, não foi possível encontrar artigos em português que utilizassem exclusivamente dados genômicos em suas análises. Dessa forma, tornou-se necessária a ampliação de busca para publicações em inglês, conforme apresentado no Quadro 2, de modo a garantir uma análise abrangente dos estudos voltados à utilização de dados genômicos para predição de doenças.

Quadro 2 – Resultado de buscas por termos contidos em artigos em língua inglesa

Termos de busca	Semantic Scholar	Science Direct	IEEE Xplore	Total
“machine learning” + “genetic variants” + “prediction”	5.530	10.666	44	16.240
“machine learning” + “prediction” + “genomic data”	38.100	9.570	175	47.845
“machine learning” + “predictive models” + “genetic predisposition”	49.500	1.062	26	50.588
“machine learning” + “BRCA1” + “prediction”	88.300	396	1	88.697
“breast cancer” + “prediction” + “machine learning”	8.040	7.642	836	16.518
“genomic data” + “disease” + “prediction”	20.000	33.506	112	53.618

Fonte: elaborado pela autora.

Inicialmente, realizou-se uma triagem para reduzir o número de artigos de mais de 270 mil para cerca de 3 mil, utilizando critérios como a relevância das palavras-chave em relação ao foco deste estudo e a quantidade de citações de cada artigo. Após essa etapa, obteve-se um conjunto aproximado de 80 artigos que demonstraram maior pertinência para a análise.

Em seguida, realizou-se a leitura dos títulos e resumos, onde identificou-se que muitos estudos não estavam relacionados ao tema deste trabalho, abrangendo outras áreas, como predição de respostas a tratamentos ou temáticas muito específicas a uma única doença, e adotando abordagens como genética populacional, justificando a exclusão desses estudos. Foram priorizados artigos que abordavam técnicas de aprendizado de máquina para a análise de dados genômicos associados ao câncer de mama e que possuíam explicações detalhadas sobre as técnicas aplicadas. A filtragem também incluiu a busca por artigos que analisassem variantes genéticas específicas associadas à predisposição ao câncer de mama em portadores de mutações no gene BRCA1. Dessa forma, o Quadro 3 apresenta os trabalhos selecionados que seguiram os critérios mencionados. Esses trabalhos foram escolhidos por suas abordagens metodológicas e aplicabilidade ao contexto da análise genômica e predição de doenças, com o uso de técnicas de aprendizado de máquina. Optou-se por artigos que, além de enfatizarem a aplicação de dados genômicos, priorizassem abordagens mais práticas, que envolvessem a utilização de técnicas computacionais.

Quadro 3 – Síntese dos trabalhos correlatos selecionados

Assunto	Filtro	Referência
Predição genômica de doenças complexas com aprendizado de máquina	Machine learning prediction	Enoma *et al*. (2022)
Predição genômica de doenças com aprendizado profundo	Deep learning disease prediction	Peng *et al.* (2024)
	Genomic data disease prediction	Alzoubi, Alzubi e Ramzan (2023)
Revisão de algoritmos de aprendizado profundo para a predição genômica de doenças	Machine learning genomic data	Koumakis (2020)
Predição genômica e prognóstico de doenças autoimunes com aprendizado de máquina	Machine learning prediction disease	Danieli *et al*. (2024)
Predição de variantes patogênicas do gene BRCA1	BRCA1 prediction	Hidayat *et al.* (2023)
Predição do câncer de mama com aprendizado de máquina com dados de imagem	Breast cancer prediction machine learning	Das *et al*. (2024)
Predição genômica do câncer de mama com aprendizado de máquina	Genomic data disease prediction	Sharma *et al*. (2020)
Análise genômica do câncer de mama com aprendizado de máquina	Breast cancer prediction machine learning	Taheri e Habibi (2024)
	Deep learning breast cancer prediction	Mohamed *et al.* (2023)

Fonte: elaborado pela autora.

Conforme apresentado no Quadro 3, os trabalhos correlatos analisados exploram diferentes abordagens voltadas para a predição de doenças a partir de dados genômicos, com destaque para o uso de técnicas de aprendizado de máquina. Os estudos correlatos têm como tema central a predição genômica de doenças, explorada em diferentes contextos por meio de diversas técnicas de aprendizado de máquina, como aprendizado supervisionado e não supervisionado. Enquanto alguns trabalhos focam em variantes genéticas específicas, há

também uma aplicação significativa dessas técnicas na predição do câncer de mama e na avaliação da patogenicidade de variantes do gene BRCA1. A relevância dessas abordagens para o prognóstico e o diagnóstico precoce é enfatizada, especialmente no contexto de doenças com alta variabilidade genética. A diversidade de tipos de dados genômicos utilizados reforça a amplitude das abordagens metodológicas discutidas. Ao integrar esses dados com técnicas de aprendizado de máquina, os estudos avançam no desenvolvimento de soluções mais robustas e precisas para a predição de doenças.

Enoma *et al*. (2022) exploram o uso de aprendizado de máquina em estudos de associação genômica ampla, com ênfase na identificação de polimorfismos de nucleotídeo único e na predição de riscos para doenças complexas. Seguindo uma linha semelhante, Peng *et al*. (2024) apresentam o DeepRisk, uma abordagem de aprendizado profundo que melhora a precisão da avaliação de risco para doenças comuns, integrando dados de associação genômica ampla com informações genotípicas. Alzoubi, Alzubi e Ramzan (2023) também apresentam uma estrutura de aprendizado profundo para prever o risco de doenças complexas, focando em polimorfismos de nucleotídeo único e utilizando a seleção de informações mútuas em uma rede neural artificial. Koumakis (2020) por sua vez, explora a aplicação de aprendizado profundo na genômica, destacando sua precisão, enquanto aborda os desafios da alta dimensionalidade e da validação clínica. Danieli *et al*. (2024) utilizam aprendizado de máquina no diagnóstico de doenças autoimunes, destacando sua eficácia na identificação de biomarcadores e na previsão de complicações.

Hidayat *et al*. (2023) propuseram um método semi-supervisionado baseado em Modelo de Mistura Gaussiana para prever a patogenicidade de variantes não rotuladas do gene BRCA1, utilizando dados rotulados existentes para aprimorar as previsões. Das *et al*. (2024) disponibilizaram um sistema de aprendizado de máquina que utiliza técnicas avançadas de pré-processamento e seleção de características para melhorar a predição de câncer de mama, enquanto Sharma *et al.* (2020) utilizaram algoritmos supervisionados de aprendizado de máquina com base em perfis de expressão gênica para a predição precoce de câncer de mama. Taheri e Habibi (2024) introduzem uma nova metodologia para identificar genes no câncer de mama, focando em mutações de baixa frequência e análise de redes biológicas, oferecendo novas perspectivas para a personalização de tratamentos. Por fim, Mohamed *et al.* (2023) desenvolveram uma arquitetura de rede neural convolucional para a detecção automática de câncer de mama, utilizando dados de expressão gênica RNA-Seq.

Dentre os trabalhos encontrados, destaca-se o estudo de Alzoubi, Alzubi e Ramzan (2023), que propõe um *framework* de aprendizado profundo para a predição de risco de doenças complexas com base em variações genômicas. Outro trabalho relevante é o de Sharma *et al.* (2020), que desenvolveram modelos de aprendizado de máquina para prever o câncer de mama utilizando dados genômicos. Por fim, o estudo de Hidayat *et al.* (2023) destaca-se por focar na predição de câncer de mama considerando a mesma variante genética analisada no estudo proposto, o gene BRCA1.

2.2 SÍNTESE DOS TRABALHOS CORRELATOS

Nesta subseção, são descritos os três trabalhos correlatos que apresentam maior relação com o estudo proposto. Os trabalhos selecionados possuem características semelhantes aos objetivos principais deste estudo, envolvendo o uso de dados genômicos para predição de riscos de doenças complexas e aplicação de técnicas de aprendizado de máquina. O Quadro 4 apresenta o trabalho de Alzoubi, Alzubi e Ramzan (2023), no qual desenvolveram um *framework* de aprendizado profundo para a predição do risco de doenças complexas utilizando dados de variações genômicas.

Quadro 4 – Deep Learning Framework for Complex Disease Risk Prediction Using Genomic Variations

Referência	Alzoubi, Alzubi e Ramzan (2023)
Objetivos	Desenvolver um *framework* de aprendizado profundo para predição de risco de doenças complexas utilizando dados de variações genômicas.
Principais funcionalidades	Utiliza um *Multi Layer Perceptron* (MLP) como modelo de aprendizado profundo, aplicação da *Joint Mutual Information* (JMI) para seleção de *Single Nucleotide Polymorphisms* (SNPs) e avaliação do modelo em sete conjuntos de dados de referência.
Ferramentas de desenvolvimento	*Frameworks* de aprendizado de máquina e técnicas de seleção de características como o JMI, a linguagem de programação não especificada.
Resultados e conclusões	O modelo foi avaliado em conjuntos de dados do *Wellcome Trust Case-Control Consortium* (WTCCC), *UK National Blood Service* (NBS) *Control Group* e 1958 *British Birth Cohort* (58C), demonstrando superioridade em comparação com técnicas tradicionais de aprendizado de máquina. O modelo proposto alcançou um *F1-score* de 0,94 e uma melhoria de até 22% na *Area Under the Curve* (AUC) em comparação com métodos anteriores em relação à predição de risco de várias doenças complexas.

Fonte: elaborado pela autora.

A principal funcionalidade do trabalho de Alzoubi, Alzubi e Ramzan (2023) é o uso de um MLP como modelo de aprendizado profundo, combinado com a técnica JMI para a seleção de SNPs, o que permite identificar as variações genéticas mais relevantes para a predição de risco. O modelo é avaliado em sete conjuntos de dados de referência e se destaca por alcançar um *F1-score* de 0,94 e uma melhoria de até 22% na AUC em comparação

com métodos anteriores. O uso do MLP combinado com a técnica JMI se destaca pois proporciona um ganho significativo em precisão, superando o desempenho de métodos tradicionais. A escolha por uma abordagem de seleção de características foca nas variações genéticas mais relevantes, otimizando a análise dos dados genômicos. Esta pesquisa se alinha diretamente com o contexto do estudo atual ao explorar o uso de técnicas de aprendizado de máquina na análise genômica para a predição de doenças. A aplicação de um modelo de aprendizado profundo para identificar padrões genéticos relevantes ao risco de doenças complexas exemplifica como essas técnicas podem melhorar o diagnóstico precoce e a prevenção.

No Quadro 5, é descrito o trabalho de Sharma *et al.* (2020) que utiliza técnica de aprendizado de máquina para prever o câncer de mama em estágio inicial, utilizando perfis de expressão gênica, que são medições da atividade de genes em células cancerígenas.

Quadro 5 – Machine Learning Approach for Predicting Breast Cancer Using Genomic Data

Referência	Sharma *et al.* (2020)
Objetivos	Desenvolver um sistema para a predição de câncer de mama em estágio inicial utilizando dados genômicos e algoritmos de aprendizado de máquina supervisionados.
Principais funcionalidades	Utiliza perfis de expressão gênica, implementando quatro algoritmos de aprendizado de máquina, para prever o câncer de mama. Analisar um conjunto de dados genômicos com o intuito de identificar genes associados à doença e avaliar o desempenho dos modelos com base em métricas como acurácia, precisão, *recall* e F1 score.
Ferramentas de desenvolvimento	*Python* 3, utilizado para a implementação dos algoritmos de aprendizado de máquina no ambiente *Google Colab*, a ferramenta *Weka*, utilizada inicialmente para implementar o modelo de *Support Vector Machine* (SVM) e dados de *microarray* para os perfis de expressão gênica.
Resultados e conclusões	O *National Center for Biotechnology Information* (NCBI) foi utilizado como fonte de dados para treinar os modelos. Os modelos de Árvore de Decisão e SVM atingiram as maiores acurácias, de 98% e 97%, respectivamente. Os autores concluíram que o uso de dados genômicos melhora a predição precoce de câncer em comparação com dados clínicos tradicionais.

Fonte: elaborado pela autora.

Sharma *et al.* (2020) desenvolveram uma abordagem de aprendizado de máquina para prever o câncer de mama em estágio inicial, utilizando perfis de expressão gênica, que são medições da atividade de genes em células cancerígenas. Para isso, os autores implementaram quatro algoritmos de aprendizado supervisionado: SVM, *Naive Bayes*, Árvore de Decisão e *K-Nearest Neighbors* (KNN). Entre os principais resultados, destacam-se as taxas de acurácia alcançadas pelos modelos, especialmente o de Árvore de Decisão, que atingiu 98%, e o SVM, que alcançou 97%, demonstrando a eficácia da abordagem proposta na predição precoce do câncer de mama. Os autores concluíram que o uso de dados genômicos melhora significativamente a predição, em comparação com dados clínicos tradicionais, permitindo diagnósticos mais precisos e antecipados. Este estudo se relaciona com o trabalho atual ao utilizar técnicas de aprendizado de máquina para identificar padrões em dados genômicos, contribuindo para a predição precoce de doenças. Além disso, a pesquisa exemplifica como a análise de grandes volumes de dados genéticos pode fornecer uma melhor compreensão dos fatores que influenciam a progressão do câncer, alinhando-se à proposta de melhorar a identificação de predisposições genéticas ao câncer de mama.

Por fim, o Quadro 6 apresenta o trabalho de Hidayat *et al.* (2023) que propuseram um método semi-supervisionado para prever se variantes do gene BRCA1 são benignas ou patogênicas.

Quadro 6 – Utilizing Semi-supervised Method in Predicting BRCA1 Pathogenicity Variants

Referência	Hidayat *et al.* (2023)
Objetivos	Desenvolver um método semi-supervisionado para prever se variantes do gene BRCA1 são benignas ou patogênicas, combinando um modelo de mistura Gaussiana com *embeddings* de proteínas, que são representações numéricas das sequências de proteínas geradas por um modelo de linguagem chamado *Evolutionary Scale Modelling* (ESM-2).
Principais funcionalidades	Classificar variantes do BRCA1 em patogênicas ou benignas; aplicação de *embeddings* de proteínas obtidos do ESM-2 para melhorar a previsão de patogenicidade.
Ferramentas de desenvolvimento	*Python* 3.8, bibliotecas de aprendizado de máquina como *pomegranate* e *scikit-learn,* e *Compute Unified Device Architeture* (CUDA) para processamento acelerado durante o treinamento dos modelos.
Resultados e conclusões	O modelo utilizou dados genômicos obtidos a partir do banco de dados denominado *ClinVar,* atingindo uma AUC de 79,27% e uma acurácia de 71,58% em dados rotulados, demonstrando um desempenho moderado na predição de patogenicidade das variantes de BRCA1. Cerca de 94% das variantes não rotuladas foram classificadas com confiança como benignas ou patogênicas.

Fonte: elaborado pela autora.

A abordagem combina um modelo de mistura Gaussiana, que é uma técnica de aprendizado não supervisionado, com *embeddings* de proteínas obtidos através do ESM-2, que gera representações numéricas das sequências de proteínas. Essas representações ajudam o modelo a compreender melhor as características das variantes do gene BRCA1, melhorando a precisão da previsão de sua patogenicidade. O estudo utiliza dados genômicos extraídos do banco de dados *ClinVar*, que contém informações sobre as variantes genéticas. Entre as principais funcionalidades, destaca-se a capacidade do método de utilizar dados rotulados e não rotulados,

permitindo que o modelo aprenda a partir de um pequeno conjunto de exemplos conhecidos e aplique esse conhecimento a um conjunto maior de dados desconhecidos, o que torna um modelo semi-supervisionado. O trabalho de Hidayat *et al.* (2023) se relaciona diretamente com a problemática abordada no estudo atual, contribuindo para a identificação de padrões genéticos associados à predisposição ao câncer de mama em portadores de mutações no gene BRCA1.

3 PROPOSTA DO MODELO

Nesta seção, é especificada a proposta de desenvolvimento de um modelo computacional que visa explorar a questão de pesquisa, que tem como objetivo identificar padrões genéticos específicos associados à predisposição ao câncer de mama em portadores de mutações no gene BRCA1, utilizando técnicas de aprendizado de máquina e redes complexas. Inicialmente, são apresentadas as justificativas para o desenvolvimento deste estudo, com destaque para sua relevância científica. Em seguida, é descrita a metodologia que será adotada.

3.1 JUSTIFICATIVA

No Quadro 7, encontram-se as características dos trabalhos correlatos selecionados para este estudo, onde os correlatos se encontram dispostos por colunas e as características por linhas. Essa disposição favorece uma comparação objetiva entre os estudos, permitindo identificar as contribuições de cada um e entender como serão solucionados os problemas propostos por este trabalho.

Quadro 7 – Comparativo dos trabalhos correlatos

Trabalhos Correlatos / Características	Alzoubi, Alzubi e Ramzan (2023)	Sharma *et al.* (2020)	Hidayat *et al.* (2023)
Objetivo	Predição de risco de doenças complexas	Predição de câncer de mama em estágio inicial	Predição de patogenicidade de variantes do gene BRCA1
Base de dados	WTCCC	NCBI	*ClinVar*
Método	Aprendizado profundo com MLP	Aprendizado de máquina supervisionado	Aprendizado semi-supervisionado com mistura *Gaussiana* e *embeddings* de proteínas
Métricas de avaliação	*F1-score* e AUC	Acurácia	AUC e acurácia
Limitações	Modelo de MLP pouco especificado	Dados de expressão gênica de alta dimensão pode causar *overfitting*	Dados rotulados limitados podem prejudicar precisão e generalização do modelo

Fonte: elaborado pela autora.

Com base no Quadro 7, pode-se observar que os trabalhos correlatos compartilham algumas semelhanças, apesar de cada um adotar abordagens distintas. O objetivo central de todos os estudos é a predição de doenças por meio de técnicas variadas de aprendizado de máquina. Utilizando bases de dados diversificadas e métodos específicos, cada autor alcançou resultados relevantes que contribuem tanto para o avanço científico na genética quanto para o contexto da ciência da computação, ao aplicar metodologias pertinentes para a análise genômica.

Alzoubi, Alzubi e Ramzan (2023) focaram na predição de doenças complexas, utilizando uma abordagem baseada em aprendizado profundo, especificamente com o modelo MLP, combinado com a técnica JMI para a seleção de SNPs. Essa combinação permitiu identificar variações genéticas que impactam diretamente no risco de diversas doenças. Os autores alcançaram um bom desempenho, medido pelo *F1-score* e pela AUC, o que demonstra a eficiência do modelo em identificar padrões genéticos relevantes.

Por outro lado, Sharma *et al.* (2020) direcionaram seus esforços para a predição de câncer de mama, utilizando algoritmos supervisionados como SVM, *Naive Bayes*, KNN e Árvores de Decisão. Com uma abordagem mais voltada para a prática clínica, o trabalho conseguiu alcançar altas taxas de acurácia, o que reforça a aplicabilidade dos modelos para diagnósticos iniciais. No entanto, apesar do sucesso na predição, o estudo apresenta como limitação a ausência de uma validação clínica completa, o que impede a implementação imediata dos resultados em ambientes médicos.

Hidayat *et al.* (2023), por sua vez, realizaram uma abordagem semi-supervisionada focada na classificação de variantes do gene BRCA1, especificamente na distinção entre variantes patogênicas e benignas. O uso da mistura *Gaussiana* e de *embeddings* de proteínas permitiu uma análise mais detalhada das sequências genéticas. A base de dados utilizada, *ClinVar*, é especializada em variações genéticas e proporciona uma fonte rica para a

predição de doenças hereditárias, como o câncer de mama. O modelo desenvolvido apresentou bons resultados em termos de AUC e acurácia, mas os autores destacam a necessidade de melhorar a precisão, especialmente para variantes incomuns.

Ao comparar as abordagens, percebe-se que Alzoubi, Alzubi e Ramzan (2023) adotaram o aprendizado profundo com um enfoque mais abrangente em doenças complexas, enquanto Sharma *et al.* (2020) e Hidayat *et al.* (2023) direcionaram seus estudos para condições mais específicas, como o câncer de mama. Embora todos tenham utilizado dados genômicos, Hidayat *et al.* (2023) concentraram-se nas variantes do gene BRCA1, ao passo que Sharma *et al.* (2020) e Alzoubi, Alzubi e Ramzan (2023) exploraram bases de dados mais amplas.

De modo geral, os trabalhos correlatos, apesar das diferentes abordagens e técnicas utilizadas, contribuem significativamente para o avanço do campo da análise genômica. As técnicas aplicadas indicam a diversidade metodológica e a amplitude de possibilidades dentro do contexto da predição de doenças genéticas, especialmente no câncer de mama e outras doenças. Essas contribuições fornecem um referencial importante para a continuidade dos estudos na área, servindo como base para a proposta do presente trabalho.

As contribuições deste estudo podem ser categorizadas como teóricas e práticas. Do ponto de vista teórico, o estudo pode avançar o conhecimento sobre a interação entre variantes genéticas relacionadas ao câncer de mama. Além disso, a combinação de aprendizado de máquina e redes complexas pode fornecer novas perspectivas metodológicas para a predição de risco genético, contribuindo para o desenvolvimento de novas abordagens em análises genômicas. Sob a perspectiva prática, a implementação de um modelo preditivo pode ter influência positiva no aconselhamento genético e na tomada de decisões preventivas. Embora o estudo não inclua validação clínica, ele pode ser uma base para futuras implementações em ambientes clínicos.

3.2 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

a) levantamento bibliográfico: estudar conteúdos relacionados à predição de doenças genéticas, em especial o câncer de mama associado ao gene BRCA1, além de técnicas de aprendizado de máquina, redes complexas e os trabalhos correlatos;
b) coleta de dados: os dados genômicos utilizados serão extraídos de bases públicas, focando em variantes do gene BRCA1. A coleta incluirá tanto variantes patogênicas quanto benignas, garantindo uma base diversificada para o treinamento e teste do modelo;
c) pré-processamento: os dados serão tratados e limpos, incluindo a remoção de inconsistências e a normalização das variantes genéticas;
d) definição dos algoritmos de aprendizado de máquina: será feita uma pesquisa e seleção dos algoritmos de aprendizado de máquina mais adequados para a análise dos dados genômicos;
e) implementação dos algoritmos de aprendizado de máquina: após a definição, os algoritmos selecionados serão implementados para detectar padrões genéticos que indicam predisposição ao câncer de mama. O desenvolvimento será feito utilizando Python;
f) análise com redes complexas: a rede de interações entre as variantes genéticas permitirá identificar como diferentes variantes do gene BRCA1 interagem entre si e como essas interações influenciam o risco de desenvolvimento da doença;
g) validação: o modelo será avaliado com métricas como acurácia, precisão, *F1-score* e AUC.

4 REVISÃO BIBLIOGRÁFICA

Nesta seção são apresentados os conceitos que fundamentam o estudo proposto. A subseção 4.1 aborda o câncer de mama. A subseção 4.2 detalha o gene BRCA1. Na subseção 4.3 são apresentadas as principais técnicas de aprendizado de máquina e, por fim, a subseção 4.4 discorre sobre redes complexas.

4.1 CÂNCER DE MAMA

De acordo com o Instituto Nacional de Câncer (2022), o câncer de mama é causado por mutações genéticas que levam ao crescimento desordenado de células no tecido mamário. Entre os diversos tipos de câncer, o câncer de mama se destaca como o mais comum entre as mulheres. É uma doença complexa e heterogênea, apresentando variações significativas em termos de comportamento biológico, resposta ao tratamento e prognóstico. Essa heterogeneidade dificulta tanto a detecção precoce quanto o desenvolvimento de terapias eficazes, exigindo abordagens personalizadas para cada paciente (Testa; Castelli; Pelosi, 2020).

O diagnóstico de câncer de mama envolve não apenas desafios físicos, mas também afeta profundamente a saúde emocional das mulheres. Estudos indicam que muitas pacientes enfrentam problemas de autoestima e imagem corporal, sobretudo após intervenções cirúrgicas como a mastectomia. Além disso, a incerteza quanto à eficácia do tratamento e o medo da recorrência são fatores que contribuem para o estresse psicológico, exigindo suporte psicológico contínuo (Huang *et al.*, 2010).

Além dos desafios relacionados ao diagnóstico e ao tratamento, o câncer de mama continua sendo uma área em constante evolução no que se refere à medicina personalizada. O progresso na identificação de mutações genéticas representa uma revolução na forma como a doença é compreendida e tratada. Avanços científicos nesse campo reforçam a necessidade de abordagens individualizadas, que levam em consideração tanto os fatores genéticos quanto os comportamentais e ambientais. A capacidade de adaptar o tratamento ao perfil de cada paciente não só eleva as chances de sucesso terapêutico, mas também oferece uma perspectiva na luta contra uma doença que afeta mulheres em todo o mundo.

4.2 GENE BRCA1

O termo genômica refere-se ao estudo do genoma, que é o conjunto completo de DNA de um organismo, incluindo todos os seus genes. Segundo o National Human Genome Research Institute (2019), o genoma contém todas as informações necessárias para o desenvolvimento e funcionamento de um organismo. Dados genômicos são informações derivadas da análise do DNA, incluindo as variações genéticas encontradas em uma população. Engreitz *et al*. (2024) ressaltam que esses dados são essenciais para entender como variações genéticas específicas podem influenciar a predisposição de um indivíduo a diversas doenças.

As variantes genéticas podem ser herdadas e impactar a maneira como o corpo responde a fatores ambientais e biológicos, contribuindo para o desenvolvimento de doenças complexas (Wang; Lou; Wang, 2019). Essas variantes, conhecidas como SNPs ou mutações, podem alterar a função dos genes, influenciando processos celulares essenciais. Sendo assim, a interação entre essas variantes genéticas e o ambiente pode intensificar o risco de doenças, destacando a importância de estudos genômicos para identificar predisposições e possibilitar a criação de estratégias preventivas.

Nesse contexto, os estudos genômicos focam principalmente na detecção de variantes que afetam o funcionamento dos genes (Lindenhofer *et al.,* 2024). As mutações em genes como BRCA1 e BRCA2, por exemplo, são conhecidas por aumentar significativamente o risco de câncer de mama e ovário (Han; Yang, 2023). Da mesma forma, outras variantes podem estar associadas a diversas doenças, incluindo as doenças neurodegenerativas como Alzheimer e Parkinson (Zeng *et al.,* 2022).

O gene BRCA1 atua na prevenção do crescimento descontrolado das células, reduzindo a probabilidade de desenvolvimento de tumores. Ele codifica proteínas que participam do reparo de danos ao DNA, um processo essencial para manter a integridade genética das células. Esse mecanismo faz parte da expressão gênica, que inclui a produção de proteínas a partir das informações contidas no gene (Coelho *et al.*, 2018). Mutações no BRCA1 estão frequentemente associadas ao aumento do risco de desenvolvimento de mama triplo-negativo, um subtipo mais agressivo (Fu *et al.*, 2022).

4.3 APRENDIZADO DE MÁQUINA

O aprendizado de máquina é campo específico da inteligência artificial que se baseia na ideia de que sistemas computacionais podem aprender a partir de dados, identificar padrões e tomar decisões com o mínimo de intervenção humana (Rius, 2023). Em vez de serem programados para realizar uma tarefa específica, os algoritmos de aprendizado de máquina são treinados com um conjunto de dados, de modo que possam generalizar seu aprendizado e fazer previsões ou classificações em novos dados. A essência do aprendizado de máquina está em sua capacidade de aprender e melhorar automaticamente com a experiência, tornando-o especialmente útil em áreas com grandes volumes de dados e complexidade, como a genômica.

Com base no estudo de Rius (2023), existem três principais tipos de aprendizado de máquina: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, o algoritmo é treinado com dados rotulados, onde as respostas corretas são fornecidas, permitindo que o modelo faça previsões precisas em novos dados. O aprendizado não supervisionado, por sua vez, trabalha com dados sem rótulos, buscando identificar padrões ocultos ou agrupamentos naturais dentro dos dados. Já o aprendizado por reforço é baseado na interação com um ambiente, onde o algoritmo aprende a tomar decisões por meio de tentativa e erro, recebendo recompensas ou penalidades até alcançar o comportamento desejado.

Os tipos de aprendizado de máquina contam com diversas técnicas que são utilizadas para a otimização de análise de dados. No aprendizado supervisionado, algoritmos como o KNN, *Naive Bayes*, SVM e Árvores de Decisão são amplamente utilizados para tarefas de classificação. O KNN classifica os dados com base na proximidade entre os dados e os de treinamento, enquanto o *Naive Bayes* aplica o teorema de Bayes, que calcula a probabilidade de uma classe com base em informações prévias, considerando a distribuição de cada variável de forma independente. O SVM, por sua vez, busca encontrar um hiperplano que melhor separa as classes dentro de um espaço de alta dimensionalidade, maximizando a margem entre os pontos de diferentes classes (Roberto, 2007). As Árvores de Decisão particionam os dados em subconjuntos, utilizando regras de decisão extraídas dos atributos dos dados para criar uma estrutura hierárquica para a classificação. No aprendizado não supervisionado, algoritmos de agrupamento, como o *K-means*, são utilizados para agrupar dados não rotulados em clusters, identificando padrões e organizando os dados em grupos com base em suas características mais próximas.

No aprendizado de máquina, também são relevantes as redes neurais artificiais, que se destacam em modelar dados por meio de camadas interconectadas que processam informações de maneira hierárquica. Sob a perspectiva de Rauber (2024), a MLP é um exemplo notável, sendo uma rede neural composta por camadas de entrada, ocultas e de saída, utilizada para tarefas de classificação e regressão. Essa rede aprende padrões complexos ao ajustar os pesos das conexões internas através do algoritmo de *backpropagation*, o que aprimora seu desempenho. O aprendizado profundo, por sua vez, expande essa abordagem ao utilizar múltiplas camadas de processamento para extrair características mais abstratas, sendo eficaz em situações que envolvem o reconhecimento de imagens.

Para garantir a eficácia dos modelos de aprendizado de máquina, as formas de validação são cruciais. Conforme Lucas (2024), a validação cruzada é uma das técnicas mais comuns, na qual o conjunto de dados é dividido em partes, permitindo que o modelo seja treinado em um subconjunto e testado em outro, garantindo que o desempenho do modelo não seja afetado por um conjunto específico de dados. Além disso, a utilização de métricas de avaliação, como a acurácia, *F1-score* e a AUC, permitem a análise da capacidade preditiva dos algoritmos.

Na genômica, o aprendizado de máquina é amplamente utilizado para analisar grandes quantidades de dados biológicos e identificar padrões genéticos associados à predisposição a doenças (Sikandar *et al.*, 2020). Algoritmos supervisionados podem ser usados para classificar variantes genéticas como benignas ou patogênicas, auxiliando no diagnóstico precoce de doenças hereditárias. Além disso, técnicas não supervisionadas podem ser aplicadas para agrupar dados genéticos, permitindo a descoberta de novas relações entre genes e doenças. Com essas abordagens, o aprendizado de máquina pode contribuir notavelmente para a área do estudo genômico.

4.4 REDES COMPLEXAS

Redes complexas constituem estruturas teóricas baseadas na teoria dos grafos que descrevem sistemas compostos por elementos interconectados não linearmente. Cada elemento de uma rede é representado por um nó, e as conexões entre eles são representadas por arestas. Esses modelos são usados para descrever sistemas nos quais as interações entre os elementos não são triviais, como em sistemas biológicos, sociais ou tecnológicos (Newman, 2003). A utilidade das redes complexas está em sua capacidade de capturar a interdependência entre elementos e revelar padrões estruturais que ajudam a entender como um sistema se organiza e evolui. Ao identificar os principais componentes e as interações mais influentes, as redes complexas permitem uma análise mais detalhada e compreensiva de sistemas.

As técnicas de análise de redes complexas são variadas e relevantes para entender a estrutura e o comportamento dos sistemas interconectados. Entre elas, destaca-se a análise de centralidade, que, segundo Rocha (2017), identifica os nós mais influentes com base em métricas como grau, intermediação e proximidade, determinando quais elementos desempenham papéis centrais na rede. Além disso, a detecção de comunidades é utilizada para identificar subgrupos de nós mais densamente conectados entre si, e a modularidade serve como métrica para avaliar a qualidade dessas divisões, sendo útil para encontrar padrões em redes.

A validação de redes complexas assegura que as estruturas e padrões observados sejam representativos dos sistemas reais. Uma das abordagens mais comuns é a robustez estrutural, que mede a resiliência da rede ao remover nós ou arestas e observa como isso afeta a conectividade e a funcionalidade da rede (Barbieri, 2010). Além disso, a análise de modularidade permite verificar a presença de comunidades ou subestruturas coesas, avaliando se a rede pode ser dividida em grupos que interagem mais intensamente entre si do que com o restante da rede.

Na genômica, as redes complexas são utilizadas para representar e entender as interações entre diferentes componentes biológicos, como genes e proteínas (Albert, 2005). Ao conectar esses elementos em uma rede, é possível visualizar como eles interagem e influenciam uns aos outros, o que permite uma melhor compreensão dos processos biológicos subjacentes. Essas redes ajudam a identificar padrões que podem estar associados a condições específicas, facilitando a análise de grandes volumes de dados genômicos de maneira mais estruturada e eficiente.

O impacto das redes complexas na genômica vai além da modelagem de interações biológicas, elas revolucionam a forma como interpretamos dados genômicos em larga escala. Ao integrar informações de diferentes níveis, essas redes fornecem uma visão mais integrada e detalhada, permitindo avanços no diagnóstico de doenças (Milano; Cannataro, 2023). A capacidade de identificar padrões ocultos e de destacar relações antes inexploradas faz das redes complexas uma ferramenta relevante para o progresso na compreensão das doenças genéticas.

REFERÊNCIAS

ALBERT, Réka. Scale-free networks in cell biology. **Journal of Cell Science**, v. 118, n. Pt 21, p. 4947–4957, 1 nov. 2005.

ALZOUBI, Hadeel; ALZUBI, Raid; RAMZAN, Naeem. Deep Learning Framework for Complex Disease Risk Prediction Using Genomic Variations. **Sensors**, v. 23, n. 9, p. 4439, jan. 2023.

ARNOLD, Melina. et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. **The Breast**, v. 66, p. 15–23, 1 dez. 2022.

BARBIERI, André L. **Análise de robustez em redes complexas**. 2010. Dissertação (Mestrado em Física Computacional) – Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2010.

CHOWDHURY, Rakibul. H. Intelligent systems for healthcare diagnostics and treatment. **World Journal of Advanced Research and Reviews**, v. 23, n. 1, p. 007–015, 2024.

COELHO, Aline S. et al. Predisposição hereditária ao câncer de mama e sua relação com os genes BRCA1 e BRCA2: revisão da literatura. **Rev. bras. anal. clin**, v. 50, n. 1, p. 17–21, 2018.

CONCEIÇÃO, Matilde S. et al. Perfil dos casos de câncer de mama entre acometidos no Acre período de 2015 a 2019. **Arquivos de Ciências da Saúde da UNIPAR**, v. 26, n. 3, 27 set. 2022.

DANIELI, Maria. G. et al. Machine learning application in autoimmune diseases: State of art and future prospectives. **Autoimmunity Reviews**, v. 23, n. 2, p. 103496, 1 fev. 2024.

DAS, Akhil. K. et al. Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). **Expert Systems with Applications**, v. 242, p. 122673, 15 maio 2024.

ENGREITZ, Jesse. M. et al. Deciphering the impact of genomic variation on function. **Nature**, v. 633, n. 8028, p. 47–57, set. 2024.

ENOMA, David. O. et al. Machine learning approaches to genome-wide association studies. **Journal of King Saud University - Science**, v. 34, n. 4, p. 101847, 1 jun. 2022.

FU, Xiaoyu. et al. BRCA1 and Breast Cancer: Molecular Mechanisms and Therapeutic Strategies. **Frontiers in Cell and Developmental Biology**, v. 10, p. 813457, 2022.

GUINDALINI, Rodrigo. S. C. et al. Detection of inherited mutations in Brazilian breast cancer patients using multi-gene panel testing. **Journal of Clinical Oncology**, v. 36, n. 15_suppl, p. e13610–e13610, 20 maio 2018.

HAN, Jiangxue.; YANG, Yue. Analysis and comparison of BRCA1/2 gene mutations in 310 cases of ovarian cancer. 25 maio 2023. Disponível em: <https://www.researchsquare.com/article/rs-2972334/v1>. Acesso em: 20 set. 2024

HIDAYAT, Alam A. et al. Utilizing semi-supervised method in predicting BRCA1 pathogenicity variants. In: INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND COMPUTATIONAL INTELLIGENCE, 8., 2023. **Procedia Computer Science**, v. 227, p. 36-45, 2023.

HUANG, Guilin. et al. Analysis of the psychological conditions and related factors of breast cancer patients. **The Chinese-German Journal of Clinical Oncology**, v. 9, n. 1, p. 53–57, 1 jan. 2010.

INSTITUTO NACIONAL DE CÂNCER. **Como surge o câncer**. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/como-surge-o-cancer. Acesso em: 20 set. 2024.

KALIKS, Rafael. A. et al. Princípios de prevenção do câncer para o clínico. **Rev. Soc. Cardiol. Estado de São Paulo**, v. 19, n. 4, p. 535-543, out./dez. 2009.

KOUMAKIS, Lefteris. Deep learning models in genomics; are we there yet? **Computational and Structural Biotechnology Journal**, v. 18, p. 1466–1473, 1 jan. 2020.

LINDENHOFER, Dominik. et al. Functional phenotyping of genomic variants using multiomic scDNA-scRNA-seq. bioRxiv, 1 jun. 2024. Disponível em: <https://www.biorxiv.org/content/10.1101/2024.05.31.596895v1>. Acesso em: 22 set. 2024

LUCAS, Hudson B. **A importância da validação cruzada em machine learning**. Ia com Café, 5 ago. 2024. Disponível em: https://iacomcafe.com.br/importancia-validacao-cruzada-machine-learning/. Acesso em: 4 out. 2024.

MARCOMINI, Karem. D. **Aplicação de modelos de redes neurais artificiais na segmentação e classificação de nódulos em imagens de ultrassonografia de mama**. 2013. Dissertação (Mestrado em Ciências, Programa de Engenharia Elétrica) - Universidade de São Paulo, São Carlos.

MILANO, Marianna; CANNATARO, Mario. Network models in bioinformatics: modeling and analysis for complex diseases. **Briefings in Bioinformatics**, v. 24, n. 2, p. bbad016, 1 mar. 2023.

MOHAMED, Tehnan. I. A. et al. A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data. **Scientific Reports**, v. 13, n. 1, p. 14644, 5 set. 2023.

PENG, J. et al. DeepRisk: A deep learning approach for genome-wide assessment of common disease risk. **Fundamental Research**, v. 4, n. 4, p. 752–760, 1 jul. 2024.

QUAZI, S. Artificial intelligence and machine learning in precision and genomic medicine. **Medical Oncology**, v. 39, n. 8, p. 120, 15 jun. 2022.

RAUBER, Thomas W. **Redes neurais artificiais**. Vitória: Universidade Federal do Espírito Santo, 2014. Disponível em: https://www.researchgate.net/publication/228686464_Redes_neurais_artificiais. Acesso em: 4 out. 2024.

RIUS, Alfonso D. D. M. Foundations of artificial intelligence and machine learning. In: **Artificial Intelligence in Finance**. Cheltenham, 2023. p. 2-18.

ROBERTO Junior, A. **Criação de núcleos específicos para determinados problemas de classificação usando máquinas de suporte vetorial (SVM)**. 2007. Dissertação (Mestrado em Ciências, Área de Concentração: Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007.

ROCHA, Wagner A. A. **Aspectos de redes complexas com aplicações em neurociência**. 2017. Dissertação (Mestrado em Matemática Aplicada) – Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, 2017.

SHARMA, Saurabh. et al. Machine Learning Approach for Predicting Breast Cancer Using Genomic Data. Rochester, NY, 8 abr. 2020. Disponível em: https://papers.ssrn.com/abstract=3571724. Acesso em: 7 set. 2024

SIKANDAR, Misba. et al. Analysis for Disease Gene Association Using Machine Learning. **IEEE Access**, v. 8, p. 160616–160626, 2020.

SILVA, Stella. B. V. et al. O papel da Ressonância Magnética no diagnóstico e rastreio do câncer de mama no Brasil. **Research, Society and Development**, v. 12, n. 6, p. e9512641972–e9512641972, 10 jun. 2023.

TAHERI, Golnaz; HABIBI, Mahnaz. Uncovering driver genes in breast cancer through an innovative machine learning mutational analysis method. **Computers in Biology and Medicine**, v. 171, p. 108234, 1 mar. 2024.

TEOFILO, Rhuan. N. F. et al. A Importância Da Mamografia Como Mecanismo Rastreador Precoce Do Câncer De Mama Em Pacientes Com Histórico Familiar: Uma Revisão. **Brazilian Journal of Implantology and Health Sciences**, v. 6, n. 9, p. 3084–3093, 17 set. 2024.

TESTA, Ugo; CASTELLI, Germana; PELOSI, Elvira. Breast Cancer: A Molecularly Heterogenous Disease Needing Subtype-Specific Treatments. **Medical Sciences**, v. 8, n. 1, p. 18, mar. 2020.

WAJID, Usman; REHMAN, Waheed. Genes: BRCA1 & BRCA2. **Scholastic Medical Science**, v. 2, n. 5, p. 1-2, 2024. Disponível em: https://www.scholasticopenaccess.org/SCMS/SCMS-02-0044.pdf. Acesso em: 2 set. 2024.

WANG, Huishan; LOU, Dan; WANG, Zhibin. Crosstalk of Genetic Variants, Allele-Specific DNA Methylation, and Environmental Factors for Complex Disease Risk. **Frontiers in Genetics**, v. 9, 9 jan. 2019.

ZENG, Qian. et al. Evaluation of common and rare variants of Alzheimer's disease-causal genes in Parkinson's disease. **Parkinsonism & Related Disorders**, v. 97, p. 8–14, 1 abr. 2022.