Marlon de Souza

Aitheron: Redes Neurais para classificação de mutações genéticas nos genes BRCA1/BRCA2

Joinville
2025

Marlon de Souza

Aitheron: Redes Neurais para classificação de mutações genéticas nos genes BRCA1/BRCA2

Trabalho apresentado como requisito parcial para obtenção do grau de Bacharel em Engenharia de Software.

Catolica de Santa Catarina
Curso de Engenharia de Software

Joinville
2025

Resumo

O trabalho propõe um pipeline para classificar a patogenicidade de variantes nos genes BRCA1/BRCA2 a partir de um FASTA do "paciente". As variantes são detectadas por alinhamento à referência, anotadas via Ensembl VEP e transformadas em features binárias/numéricas (termos de consequência e atributos posicionais) (RICHARDS et al., 2015) (Ensembl, 2025). O classificador é um MLP multitarefa (uma cabeça por gene) com classificação ordinal (CORAL), que modela as quatro classes ordenadas (Benigno, Possivelmente Benigno, VUS, Patogênico) por meio de logits cumulativos, respeitando a estrutura ordinal do problema. O sistema é calibrado para alto recall na classe patogênica, priorizando a redução de falsos negativos clinicamente críticos (CAO; MIRJALILI; RASCHKA, 2019). A saída por variante inclui a classe final, o risco de ser patogênica P(y=3) e uma medida de confiança baseada em 1−entropia das probabilidades de classe, facilitando triagem e tomada de decisão. A aplicação disponibiliza visualização das anotações e pode integrar contexto estrutural (AlphaFold) para inspeção de posições mutadas (ALPHAFOLD, 2020). Para treinamento e validação utilizamos dados públicos (NCBI/Ensembl/UniProt), e o FASTA do "paciente" é sintético, construído a partir da sequência de referência com mutações in silico que emulam casos reais e permitem avaliar o pipeline ponta a ponta sem expor dados clínicos.

Palavras-chave: recall; variantes; Sequenciamento Genético; MLP; *BRCA1/BRCA2*.

Abstract

The work proposes a pipeline to classify the pathogenicity of variants in the BRCA1/BRCA2 genes from a "patient" FASTA. Variants are detected by alignment to the reference, annotated via Ensembl VEP, and transformed into binary/numeric features (consequence terms and positional attributes) (RICHARDS et al., 2015) (Ensembl, 2025). The classifier is a multitask MLP (one head per gene) with ordinal classification (CORAL), which models the four ordered classes (Benign, Likely Benign, VUS, Pathogenic) through cumulative logits, respecting the problem's ordinal structure. The system is calibrated for high recall in the pathogenic class, prioritizing the reduction of clinically critical false negatives (CAO; MIRJALILI; RASCHKA, 2019). The per-variant output includes the final class, the risk of being pathogenic P(y=3), and a confidence measure based on 1−entropy of the class probabilities, facilitating triage and decision-making. The application provides visualization of annotations and can integrate structural context (AlphaFold) to inspect mutated positions (ALPHAFOLD, 2020). For training and validation we used public data (NCBI/Ensembl/UniProt), and the "patient" FASTA is synthetic, built from the reference sequence with in silico mutations that emulate real cases and allow end-to-end evaluation of the pipeline without exposing clinical data.

Keywords: recall; variants; Genetic Sequencing; MLP; *BRCA1/BRCA2*.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Classificação de patogenicidade de variantes em BRCA1/BRCA2 com abordagem ordinal

O câncer de mama permanece um desafio de alta incidência e complexidade biológica. No contexto hereditário, alterações em BRCA1 e BRCA2 estão associadas a risco aumentado e motivam a classificação de patogenicidade de variantes nesses genes, com impacto direto em aconselhamento genético e tomada de decisão clínica (ROY; CHUN; POWELL, 2012; RICHARDS et al., 2015).

Este trabalho propõe um pipeline de IA para classificar a patogenicidade de variantes em BRCA1 e BRCA2 a partir de um sequenciamento genético (fasta) do paciente. As variantes são detectadas por alinhamento à referência, anotadas via Ensembl VEP e transformadas em features binárias e numéricas. O classificador é um MLP multitarefa (uma cabeça por gene) com classificação ordinal (CORAL), que modela quatro classes ordenadas por meio de logits cumulativos, respeitando a estrutura ordinal do problema. O sistema é calibrado para alto recall na classe patogênica, priorizando a redução de falsos negativos clinicamente críticos (CAO; MIRJALILI; RASCHKA, 2019).

A acurácia estrutural oferecida pela AlphaFold ampliou o acesso a modelos tridimensionais, incluindo BRCA1 e BRCA2, permitindo contextualizar variantes em regiões e domínios da proteína. Neste trabalho, essa visualização é utilizada como apoio à interpretação, enquanto a predição de patogenicidade permanece a tarefa principal (JUMPER et al., 2021).

Para treinamento e validação, utilizamos dados públicos; para testes ponta a ponta, o FASTA do "paciente" é sintético, derivado da sequência de referência com mutações in silico que emulam casos reais, permitindo verificar o pipeline sem expor dados clínicos (Ensembl, 2025).

2 Descrição do Projeto

Este capítulo apresenta a concepção geral do projeto, detalhando o tema de pesquisa, os objetivos que norteiam o estudo e o problema resolvido. São discutidos os fundamentos teóricos multidisciplinares que sustentam a proposta — da genética à inteligência artificial — e delineada a arquitetura técnica que viabiliza a aplicação prática do modelo desenvolvido.

2.1 Tema do Projeto

O projeto situa-se na intersecção entre oncogenética e aprendizado de máquina, com foco na classificação de variantes genéticas associadas ao câncer de mama. O escopo concentra-se nos genes BRCA1 e BRCA2 e na necessidade de priorizar, para fins de triagem, variantes com maior probabilidade de impacto clínico. Diante desse cenário, a questão que orienta o trabalho é: como estruturar um fluxo computacional reprodutível que, a partir de dados públicos e anotações padronizadas, estime de forma ordenada a patogenicidade de variantes em BRCA1/BRCA2, preservando alta sensibilidade para a classe patogênica.

O objetivo geral é demonstrar, de ponta a ponta, um processo de classificação que traduza evidências moleculares em probabilidades por classe, produzindo saídas úteis à priorização. Desdobram-se daí objetivos específicos: normalizar e enriquecer dados públicos com anotações de efeito de variante; projetar um classificador com tratamento explícito da ordem entre rótulos; calibrar limiares para maximizar o recall em patogênicas; e apresentar resultados de forma interpretável, incluindo rótulo final, probabilidade da classe de maior gravidade (patogênica) e uma medida objetiva de confiança. A justificativa decorre do impacto prático da triagem de variantes em aconselhamento genético e manejo clínico, sobretudo diante do volume de achados de significado incerto e do custo de falsos negativos em genes de reparo do DNA (RICHARDS et al., 2015; ROY; CHUN; POWELL, 2012).

Do ponto de vista teórico, o trabalho ancora-se em recomendações clínicas para interpretação de variantes, em ferramentas consolidadas de anotação genômica e em abordagens de modelagem que respeitam a natureza ordinal do desfecho. Metodologicamente, utiliza-se um conjunto de dados públicos para treinamento e validação, além de um FASTA sintético — derivado da sequência de referência e modificado in silico — para exercitar o fluxo completo sem recorrer a dados clínicos reais. O pipeline inicia no alinhamento do FASTA, passa pela detecção de diferenças e pela anotação padronizada, deriva atributos tabulares binários e numéricos e, por fim, treina um modelo multitarefa por gene com regressão ordinal, avaliado por validação cruzada estratificada. As métricas são reportadas com

ênfase em sensibilidade para patogênicas, preservando transparência e reprodutibilidade dos resultados (AHMAD et al., 2023).

2.2 Objetivos do Projeto

O objetivo central é desenvolver e validar um pipeline de IA para classificar a patogenicidade de variantes nos genes *BRCA1/BRCA2* em quatro classes ordenadas (Benigno, Possivelmente Benigno, VUS, Patogênico), utilizando classificação ordinal (CORAL) e priorizando alto recall para a classe patogênica, em consonância com diretrizes de interpretação clínica de variantes (RICHARDS et al., 2015; CAO; MIRJALILI; RASCHKA, 2019).

1. Construção de um pipeline end-to-end a partir de um arquivo FASTA do paciente: alinhamento à referência, detecção de variantes, anotação via Ensembl VEP e engenharia de features binárias/numéricas (Ensembl, 2025).

2. Projeção de um MLP multitarefa (uma cabeça por gene) com CORAL para estimar probabilidades cumulativas e derivar $P(y = 0..3)$, respeitando a ordem natural das classes (CAO; MIRJALILI; RASCHKA, 2019).

3. Definir estratégia de calibração e limiares para maximizar o recall da classe patogênica, reportando também métricas como precisão e F1 por gene.

4. Entregar saídas interpretáveis por variante: classe final, risco patogênico $P(y = 3)$ e medida de confiança baseada em $1-$entropia das probabilidades.

5. Disponibilizar um frontend leve (Streamlit) para upload do FASTA, visualização das anotações e resultados do modelo, com integração de contexto estrutural (AlphaFold) para inspeção de posições mutadas (ALPHAFOLD, 2020).

6. Utilizar dados públicos para treinamento e validação; empregar, nos testes ponta a ponta, um FASTA sintético derivado da referência com mutações in silico que emulam casos reais.

2.3 Problemas a Resolver

O problema a ser enfrentado é a necessidade de classificar, com precisão e confiabilidade, a patogenicidade de variantes genéticas nos genes BRCA1 e BRCA2, de modo a priorizar casos clinicamente relevantes e reduzir a incerteza associada a VUS. Em vez de ampliar o escopo para diagnóstico por imagem ou recomendações terapêuticas, concentramos o esforço na tarefa de transformar dados genômicos e anotações padronizadas em probabilidades de classe úteis à triagem. Essa formulação alinha o projeto às demandas reais de

aconselhamento genético e manejo clínico, onde o custo de falsos negativos é elevado e a interpretabilidade do resultado (classe final, probabilidade de patogênica e uma medida objetiva de confiança) é fundamental para a decisão (RICHARDS et al., 2015).

Para atacar esse problema, estruturamos um fluxo reprodutível que parte de um FASTA do "paciente" (sintético para testes ponta a ponta), realiza alinhamento e detecção de variantes, enriquece com anotações via VEP e deriva atributos binários/numéricos que preservam informação de posição e consequência. Sobre esse conjunto, treinamos um classificador multitarefa por gene e adotamos a regressão ordinal (CORAL) como estratégia técnica para modelar a relação entre rótulos. O sistema reporta, por variante, a classe prevista, a probabilidade de ser patogênica e a confiança associada.

2.4 Estudos científicos

Os genes BRCA1 e BRCA2 são conhecidos como genes supressores tumorais, responsáveis por proteger o organismo contra o desenvolvimento de tumores(ROY; CHUN; POWELL, 2012). Essas funções são exercidas principalmente por meio da produção de proteínas especializadas no reparo do DNA, garantindo a estabilidade do material genético. Ambas as proteínas desempenham papéis essenciais em diferentes etapas da resposta ao dano do DNA (DDR – DNA Damage Response) .

A proteína BRCA1, por exemplo, é considerada pleiotrópica, ou seja, possui múltiplas funções dentro do organismo, atuando principalmente nos pontos de verificação do ciclo celular e no reparo do DNA (DENG, 2006). Por outro lado, a proteína BRCA2 atua como mediadora central da recombinação homóloga, mecanismo responsável pela troca dos nucleotídeos entre as moléculas de DNA para corrigir danos genéticos.

O gene BRCA1 é composto por 24 éxons, regiões codificantes que fornecem as instruções necessárias para a produção dessa proteína. Caso ocorra uma mutação em algum desses éxons, a função da proteína pode ser comprometida, elevando significativamente o risco tumoral. A proteína resultante do gene BRCA1 é nuclear, atuando predominantemente dentro do núcleo celular, e é composta por 1.863 aminoácidos, que representam as unidades estruturais fundamentais das proteínas.

2.4.1 Mutações dos genes BCRA1 e BCRA2

Embora a maioria dos casos de câncer de mama ocorra de maneira esporádica ao longo da vida, estima-se que cerca de 5 a 7% dos casos estejam relacionados à Síndrome de Câncer de Mama e Ovário Hereditários (HBOC – Hereditary Breast and Ovarian Cancer), caracterizada por mutações nos genes BRCA1 e BRCA2. Essas alterações genéticas aumentam consideravelmente o risco de desenvolvimento de câncer de mama e ovário ao longo da vida.

Segundo artigo de Roy R., Chun J. e Powell S.N., publicado na PubMed Central, indivíduos que apresentam mutações no gene BRCA1 têm de 70% a 80% de probabilidade de desenvolver câncer de mama durante sua vida. Já indivíduos com mutações no gene BRCA2 possuem um risco estimado entre 50% e 60% (ROY; CHUN; POWELL, 2012).

2.5 Genoma germinativo

Para o baseline deste estudo empregamos a sequência L78833.1, descrita por (SMITH et al., 1996) como "uma sequência de DNA de 117143 pb que abrange o gene BRCA1, obtida por sequenciamento aleatório de quatro cosmídeos provenientes de uma biblioteca específica para o cromossomo 17 humano" — portanto proveniente de DNA germinativo e representativa do alelo selvagem do gene BRCA1.

Segundo o manual "DDBJ/ENA/GenBank Feature Table Definition, v 11.3" do consórcio INSDC, qualquer mutação ou polimorfismo relevante deve ser anotado com as chaves-de-feature variation ou misc difference (International Nucleotide Sequence Database Collaboration, 2024). A completa ausência dessas chaves no registro L78833.1 confirma que a entrada corresponde à sequência de referência ("wild-type") e não a um alelo patogênico.

O presente genoma possui 117.143-bp, ou seja, pares de bases, que são combinações de duas bases nitrogenadas que se ligam por meio de ligações de hidrogênio formando a dupla hélice do DNA. Cada hélice do DNA é composta por um 'esqueleto repetitivo' de desoxirribose (açúcar) ligada a grupos fosfato. A cada molécula de açúcar associa-se uma das quatro bases nitrogenadas — adenina (A), citosina (C), guanina (G) ou timina (T).

Isso é crucial pois a interpretação de variantes depende da comparação base a base com a sequência de referência; mesmo uma única troca de nucleotídeo em *BRCA1* pode gerar assinaturas mutacionais características e aumentar o risco oncológico, como demonstrado por análises de genomas completos em células BRCA1-deficientes (ZÁMBORSZKY et al., 2021).

Ou seja, pode-se pensar nos códons como "sílabas" que o ribossomo lê. O códon `AUG` diz ao ribossomo: comece a proteína e coloque o aminoácido metionina. Se apenas a primeira "letra" for trocada e o códon virar `CUG`, a instrução muda para coloque leucina. Ou seja, uma única troca de base já altera o primeiro tijolo da proteína, o que pode modificar todo o seu formato e, por consequência, a sua função.

2.6 Base de Dados e Confiabilidade

2.6.1 NCBI — ClinVar

Utilizamos exclusivamente o arquivo público *variant_summary.txt.gz* do ClinVar para BRCA1/BRCA2, de onde extraímos rótulos clínicos, histórico/estado de revisão, tipo/ori-

gem de variante e coordenadas em GRCh38, conforme o seu fluxo de download e filtragem. O ClinVar é o repositório de referência internacional para interpretações clínicas de variantes, com submissões de laboratórios/painéis de especialistas, curadoria contínua e histórico de reclassificações documentado pelo NCBI e atualizado regularmente no Nucleic Acids Research (LANDRUM et al., 2016).

2.6.2 Ensembl - VEP

A anotação molecular das variantes é feita via VEP (endpoint /vep/human/id/rsid), de onde derivamos consequence terms, impactos, coordenadas em cDNA/proteína, códons, aminoácidos e seleção de transcrito; complementamos com Lookup (/lookup/id) para obter ENSP/ENST e Sequence (/sequence/id e /sequence/region) para recuperar FASTA de CDS/proteína e, quando necessário, sequência por região em GRCh38 (Ensembl, 2025).

2.6.3 UniProt — Identidades canônicas e curadoria Swiss-Prot

Empregamos a API de ID mapping para resolver ENSP/ENST em acessões UniProtKB e, em seguida, recuperar descrição funcional e features/domínios anotados usados como contexto interpretativo. O UniProtKB/Swiss-Prot é curado por especialistas, com controle de qualidade e histórico de versões, e é reconhecido como o recurso líder e de alta qualidade para informação funcional de proteínas (The UniProt Consortium, 2025).

2.6.4 AlphaFold - Estrutura da Proteína

Para visualização e inspeção contextual das posições mutadas, integramos os modelos 3D previstos para BRCA1/BRCA2 a partir do AlphaFold DB (VARADI et al., 2022).

2.7 Aspectos Éticos e Regulatórios

A base de dados utilizada neste estudo é integralmente pública e previamente anonimizada, circunstância que, conforme o art. 12 da Lei Geral de Proteção de Dados Pessoais (LGPD), a descaracteriza como "dado pessoal identificável" (LGPD, 2018). Nesses termos, todas as etapas analíticas desenvolvidas neste projeto operam exclusivamente sobre informações desidentificadas, afastando riscos à privacidade dos sujeitos de origem. Em consonância, a Resolução nº 510/2016 do Conselho Nacional de Saúde (CNS) dispensa o consentimento individual para pesquisas que empregam dados públicos sem possibilidade de reidentificação (RESOLUçãO. . . , 2016).

3 Especificação Técnica

Este capítulo descreve, de maneira objetiva, os pilares técnicos que sustentam o Aitheron. Primeiro, apresenta-se a stack tecnológica escolhida — das linguagens de programação aos principais frameworks de ciência de dados e engenharia de software.

3.1 Stack Tecnológica

- **Linguagens de Programação**:
  - Python: será a linguagem utilizada para Machine Learning e manipulação de dados;
- **Frameworks e Bibliotecas**:
  - Pytorch: criação e treinamento de Redes Neurais Artificiais;
  - Pandas e NumPy: manipulação e análise de dados em escala;
  - Matplotlib: visualização de resultados (como curvas ROC e matrizes de confusão);
  - FastAPI: criação das rotas backend (Python);
  - Streamlit (frontend): criação da interface para visualização do resultado.

4 Modelo de Inteligência Artificial

4.1 Features e pré-processamento

O conjunto de atributos combina variáveis numéricas e indicadores binários derivados da anotação molecular. As colunas numéricas incluem coordenadas e medidas posicionais (`Start`, `End`, `len_ref`, `len_alt`, `length_change`), um ranque de impacto (`ImpactRank`) e mapeamentos de cDNA/proteína (`CDNAStart`, `CDNAEnd`, `Protein_Start`, `Protein_End`). As variáveis binárias abrangem todas as flags com prefixo `Is*` (por exemplo, `IsMissenseVariant`, `IsFrameshiftVariant`, `IsSynonymousVariant`, além de `IsCoding`). O pré-processamento é ajustado exclusivamente nos dados de treino para evitar vazamento: valores numéricos ausentes são imputados com constante, padronizados por *z-score*, e as binárias são consumidas como 0/1 tal como geradas; categorias adicionais via *one-hot* permanecem opcionais e, nesta versão, não foram necessárias. Esse desenho privilegia interpretabilidade (flags explícitas por efeito/consequência) e estabilidade numérica (escala uniforme para o MLP).

4.2 Arquitetura e racional de modelagem

O classificador é implementado como um MLP (Multilayer Perceptron) com duas camadas ocultas ($[256, 128]$) e *dropout* 0,2, compartilhado entre tarefas, seguido de duas cabeças independentes — uma para BRCA1 e outra para BRCA2. Cada cabeça produz $(K-1)$ logits cumulativos para K=4 classes (Benigno, Possivelmente Benigno, VUS, Patogênico), conforme a formulação ordinal do CORAL. A escolha por MLP deve-se à robustez em dados tabulares mistos (numéricos + binários), simplicidade de calibração e baixo custo computacional, preservando flexibilidade para incorporar novas *features*. A abordagem ordinal é usada como *meio* para respeitar a ordem natural entre classes e induzir fronteiras consistentes entre limiares, reduzindo inconsistências locais em faixas de decisão adjacentes.

4.3 Função de perda, desbalanceamento e otimização

O aprendizado ordinal emprega `BCEWithLogitsLoss` por logit cumulativo, com *pos_weight* específico de cada cabeça, calculado a partir da distribuição das metas cumulativas por gene. Esse reponderamento mitiga desbalanceamentos entre os limiares (e, indiretamente, entre classes). A otimização utiliza AdamW ($lr = 10^{-3}$, $weight\_decay = 10^{-5}$), com *batch size* 256 e critério de *early stopping* por AUROC médio de validação. O backbone com-

partilhado aprende uma representação comum, enquanto as cabeças se especializam nas distribuições de BRCA1 e BRCA2, reduzindo variância e evitando dois modelos totalmente separados.

4.4 Validação, escolha do vencedor e *thresholding*

A avaliação segue StratifiedKFold (10 *folds*) estratificado por (gene, rótulo), prevenindo que a distribuição de classes por gene se deteriore em alguma partição. Em validação, convertem-se as cumulativas em $P(y{=}0..3)$ por cabeça e extrai-se $P(y{=}3)$ (risco patogênico) para relatórios auxiliares binários (patogênico vs. não). O *fold* vencedor é escolhido pelo maior AUROC médio entre as duas cabeças. A calibração operacional dos limiares é diferenciada: para BRCA1, busca-se *recall* alvo elevado (p.ex., 0,99) selecionando, entre os limiares que atingem esse patamar, aquele que maximiza a precisão; para BRCA2, o limiar é escolhido maximizando *F1*. Os limiares do *fold* vencedor são persistidos para uso no treinamento final.

4.5 Treinamento final e artefatos

Após a seleção do *fold* vencedor, os pesos são utilizados para inicializar o treinamento em todos os dados rotulados, com número de épocas derivado da mediana de épocas "vencedoras" entre os *folds* (mais uma época adicional). Ao término, são salvos: (i) os pesos do modelo final, (ii) o pré-processador ajustado no conjunto completo e (iii) as métricas consolidadas, incluindo limiares por cabeça e matrizes de confusão finais.

4.6 Inferência e saídas do sistema

Em produção, o pré-processador persistido transforma o *DataFrame* de entrada (já anotado) sem *refit*. O MLP gera logits cumulativos por cabeça, convertidos em probabilidades por classe $[P_0, P_1, P_2, P_3]$. A classe final é o $\arg\max$ dessas probabilidades; o risco patogênico é $P(y{=}3)$, comparado ao limiar calibrado do gene; e a confiança é computada como $1-$ entropia normalizada das probabilidades (com *clipping* numérico). Esse desenho entrega, por variante, três elementos práticos para triagem: rótulo clínico previsto, probabilidade de patogênica e medida objetiva de confiança — alinhando o modelo ao objetivo operacional de priorização sensível a falsos negativos.

5 Considerações de Segurança do Software

Esta secção sintetiza os requisitos e controles de segurança que devem ser adotados em todas as fases do ciclo de vida do Aitheron, com ênfase na proteção de dados sensíveis de saúde, na resiliência operacional dos serviços de IA e na conformidade com normas nacionais e internacionais.

5.1 Objetivos de Segurança

- **Integridade**: garantir que informações e modelos de IA não sejam alterados de maneira indevida.
- **Disponibilidade**: assegurar funcionamento contínuo do sistema com $\geq$99,5 % de *uptime* (RNF03).
- **Privacidade**: cumprir a LGPD e boas-práticas globais (HIPAA, GDPR) para dados de saúde.
- **Explicabilidade e Confiabilidade**: prover artefatos que sustentem a interpretação clínica dos resultados de IA (RNF06).

5.2 Governança e Conformidade

Políticas de Segurança: adotar um Sistema de Gestão de Segurança da Informação baseado na ISO/IEC 27001.

Classificação de Dados: mapear dados em três níveis (*restrito*, *interno*, *público*) e aplicar controles compatíveis.

Regulatórios: LGPD (Brasil), HIPAA (EUA) e RDC 657/2022 (Anvisa) para software como dispositivo médico.

6 Requisitos do Projeto

Nesta seção, são descritos os requisitos necessários para a construção do ambiente de desenvolvimento e execução do projeto, bem como os requisitos funcionais e não funcionais que orientam o comportamento e a qualidade do sistema.

6.1 Requisitos Funcionais (RF)

RF01 – Interface de Inserção e Consulta O sistema deve fornecer uma interface (frontend) que permita aos usuários fornecer o sequenciamento génetico (FASTA) do paciente.

RF02 – Implementação de Modelos de IA O sistema deve implementar algoritmos de Inteligência Artificial baseados em redes neurais para classificar variantes genéticas nos genes BRCA1/2 a partir de um sequenciamento genético.

RF03 – Métricas de Desempenho O sistema deve disponibilizar métricas de avaliação dos modelos de IA, como Acurácia, Precisão, Recall e AUC-ROC, permitindo a validação e o acompanhamento da evolução do treinamento.

RF04 – Relatórios de Diagnóstico e Tratamento O sistema deve gerar relatórios que incluam diagnósticos indicativos (por exemplo, tumor benigno ou patogênico), confiança, probabilidade de patogenicidade e demais.

6.2 Requisitos Não Funcionais (RNF)

RNF01 – Desempenho O sistema deve responder às solicitações de consulta de dados clínicos e resultados de inferência em até 25 segundos em condições normais de uso.

RNF02 – Modularidade O sistema deve ser projetado de forma modular, permitindo a substituição ou atualização de componentes sem comprometer o restante da aplicação.

RNF03 – Disponibilidade O sistema deve estar disponível pelo menos 99,5% do tempo, garantindo aos profissionais de saúde acesso ininterrupto aos dados e resultados em regime de produção.

RNF04 – Registro de Logs Todas as operações críticas devem ser registradas em logs para fins de auditoria e rastreamento de atividades.

RNF05 – Portabilidade de Implantação O sistema deve poder ser implantado em diferentes infraestruturas (on-premises ou em serviços de nuvem), utilizando contêineres ou outras tecnologias que facilitem o deploy.

RNF06 – Explicabilidade do Modelo de IA Sempre que possível, o sistema deve fornecer, junto ao resultado da inferência, indicadores que auxiliem na interpretação das decisões tomadas pelo modelo de IA.

7 Plano de Testes

7.1 Plano de Testes — escopo e organização

Objetivo e escopo

Validar, de forma reprodutível, o desempenho do classificador de patogenicidade em BRCA1/BRCA2 e a utilidade clínica das saídas (rótulo, P(y=3), confiança), combinando avaliação quantitativa em validação cruzada com validação biomédica qualitativa especializada.

Fontes de dados para teste

i) Conjunto rotulado público filtrado por gene e assembly. ii) Arquivo FASTA sintético para teste ponta a ponta do fluxo FASTA→anotação→predição, sem uso de dados clínicos reais.

Protocolo de validação

i) Validação cruzada estratificada por (gene, rótulo), com 10 folds, sem vazamento entre treino e validação. ii) Seleção do fold vencedor pelo AUROC médio entre cabeças. iii) Reajuste final em todos os dados com inicialização pelo fold vencedor e preservação de limiares calibrados por cabeça.

Métricas e justificativas

Serão reportadas, por cabeça, as métricas: AUROC e AUPRC (discriminação), recall e precision (trade-off sensibilidade versus falso positivo), F1 (equilíbrio), balanced accuracy (robustez ao desbalanceamento), MCC (correlação preditiva robusta) e accuracy. A matriz de confusão binária (patogênico vs. não) será apresentada para leitura direta de erros tipo I/II. Os limiares operacionais seguem política distinta por gene: BRCA1 prioriza sensibilidade alvo elevada; BRCA2 maximiza F1.

Resultados a incorporar

Inserir as métricas finais fornecidas por cabeça e as respectivas matrizes de confusão:

- BRCA1: métricas agregadas e matriz de confusão
- BRCA2: métricas agregadas e matriz de confusão

As tabelas serão incluídas no formato:

- Tabela 1: métricas por cabeça (BRCA1 e BRCA2)
- Tabela 2: matrizes de confusão por cabeça

Validação biomédica especializada

Revisão por especialista externa: Mayara Thais Moreira (Mestrado em Biotecnologia, Universidade de Caxias do Sul, 2023). Escopo da validação: coerência biológica das anotações, plausibilidade dos rótulos previstos, leitura da priorização de risco e consistência da visualização estrutural (AlphaFold) das posições mutadas. Resultado esperado: parecer textual atestando a correção biomédica do pipeline e da apresentação das evidências.

Reprodutibilidade e auditoria

Versionamento dos artefatos do modelo, registro dos limiares por cabeça, armazenamento das métricas por fold e relatório final em JSON. Seeds controladas, salvamento do pré-processador ajustado e scripts de execução para replicação do experimento.

Critérios de aceitação

O modelo é considerado apto para triagem se: i) atingir AUROC e AUPRC superiores a patamares definidos para cada cabeça; ii) cumprir o alvo de sensibilidade em BRCA1 com precisão mínima associada; iii) manter MCC elevado e matrizes de confusão com falso negativo residual; iv) receber parecer favorável da validação biomédica.

Riscos e mitigação

Possível variação entre folds será tratada por relatório com média e desvio padrão. Eventuais assimetrias de classe serão acompanhadas por balanced accuracy e MCC. Casos fronteira serão analisados qualitativamente com visualização estrutural e inspeção das features relevantes.

7.2 Resultados e validação

7.2.1 Desempenho do Modelo

Os resultados a seguir foram obtidos no protocolo descrito na seção de plano de testes, com seleção do fold vencedor por AUROC médio entre cabeças e calibração de limiares distinta por gene (BRCA1 voltado a alta sensibilidade; BRCA2 voltado a F1). As métricas

são reportadas no cenário binário patogênico vs. não, a partir de P(y=3) por cabeça do modelo ordinal.

Tabela 1 – Métricas por cabeça (patogênico vs. não)

Cabeça	AUROC	AUPRC	Recall	Precision	F1	Bal.Acc	MCC	Acc	Thr
BRCA1	0.9942	0.9971	0.9898	0.9187	0.9529	0.9088	0.8555	0.9352	0.07836
BRCA2	0.9958	0.9975	0.9946	0.9793	0.9869	0.9785	0.9631	0.9830	0.15000

Leitura rápida: AUROC e AUPRC próximos de 1 indicam que o modelo separa bem casos patogênicos dos não patogênicos; o recall elevado confirma a prioridade operacional em reduzir falsos negativos clínicos; a manutenção de precisão, F1 e MCC altos sugere equilíbrio entre sensibilidade e especificidade e correlação global consistente, inclusive sob desbalanceamento (refletido pela balanced accuracy).

Tabela 2 – Matriz de confusão — BRCA1

	Prev. Não patogênico	Prev. Patogênico
Verdadeiro Não patogênico	TN = 909	FP = 189
Verdadeiro Patogênico	FN = 22	TP = 2136

Tabela 3 – Matriz de confusão — BRCA2

	Prev. Não patogênico	Prev. Patogênico
Verdadeiro Não patogênico	TN = 1382	FP = 54
Verdadeiro Patogênico	FN = 14	TP = 2555

Nas matrizes, observa-se baixa quantidade de falsos negativos (22 em BRCA1; 14 em BRCA2) em relação ao total de verdadeiros patogênicos, compatível com os limiares e com o objetivo de priorização segura. O número de falsos positivos permanece controlado, preservando precisão e reduzindo sobrecarga de revisão.

7.2.2 Validação biomédica especializada

A validação qualitativa foi conduzida por Mayara Thais Moreira (Mestrado em Biotecnologia, Universidade de Caxias do Sul, 2023), que revisou periodicamente o fluxo de dados e o painel de resultados. O escopo revisado abrangeu: seleção e filtragem do conjunto rotulado (ClinVar BRCA1/BRCA2), anotação molecular via Ensembl VEP e complementos UniProt, inferência e apresentação das saídas (classe prevista, P(y=3), medida de confiança), além da visualização estrutural baseada em modelos do AlphaFold. Segundo parecer verbal, o pipeline e suas evidências estão coerentes para uso em triagem acadêmica de variantes, considerando as limitações usuais de dados públicos e a ausência

de validação clínica prospectiva. A citação no texto foi autorizada em uma conversa realizada no dia 10/11/2025.

7.2.3 Exemplo de saída do sistema

A Figura 1 ilustra uma execução típica do pipeline sobre um FASTA sintético (derivado da referência com mutações in silico). À esquerda, a posição mutada é realçada no modelo 3D (AlphaFold), com indicação de domínio/feature quando aplicável. Este exemplo é representativo do fluxo de triagem: a decisão final combina a probabilidade de patogênica com a confiança e o contexto estrutural; O painel à direita resume a anotação molecular (VEP/UniProt) e a classificação prevista por cabeça do modelo (rótulo final, P(y=3) e confiança).

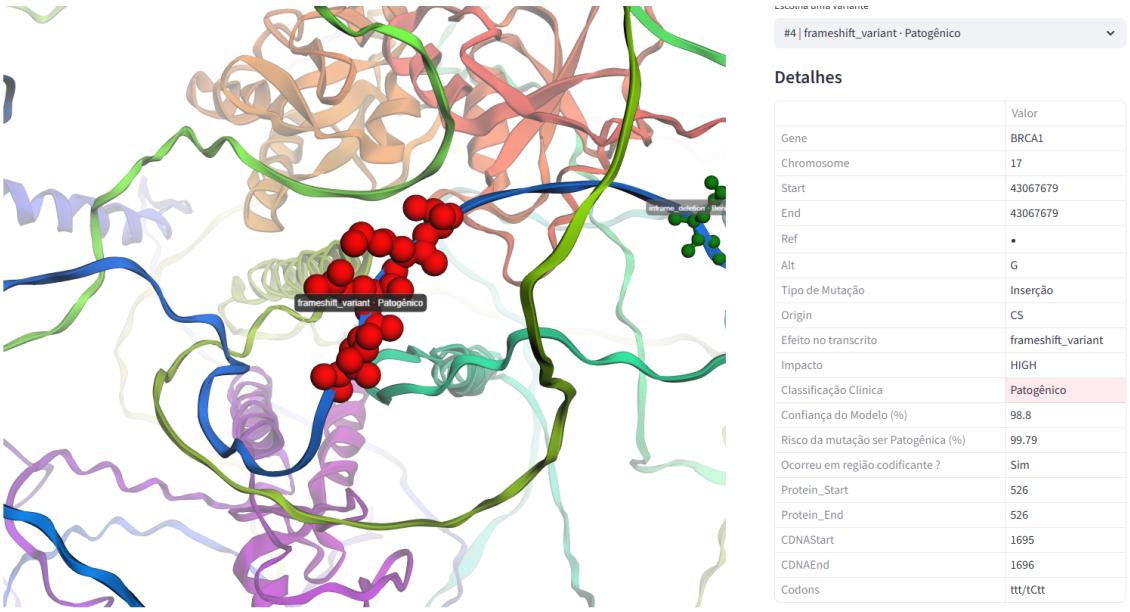

Figura 1 – Exemplo de retorno do sistema

REFERÊNCIAS

AHMAD, R. M. et al. A review of genetic variant databases and machine learning tools for predicting the pathogenicity of breast cancer. *Briefings in Bioinformatics*, v. 25, n. 1, p. bbad479, 12 2023. ISSN 1477-4054. Disponível em: <https://doi.org/10.1093/bib/bbad479>.

ALPHAFOLD. *AlphaFold Protein Structure Database*. 2020. Disponível em: <https://alphafold.ebi.ac.uk/>.

CAO, W.; MIRJALILI, V.; RASCHKA, S. Rank consistent ordinal regression for neural networks. *arXiv preprint arXiv:1901.07884*, 2019. Disponível em: <https://arxiv.org/abs/1901.07884>.

DENG, C.-X. Brca1: Cell cycle checkpoint, genetic instability, dna damage response and cancer evolution. *Nucleic Acids Research*, v. 34, n. 5, p. 1416–1426, 2006. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC1390683/>.

Ensembl. *Ensembl REST API — Variant Effect Predictor (VEP)*. 2025. Documentação oficial da API. Disponível em: <https://rest.ensembl.org>.

International Nucleotide Sequence Database Collaboration. *The DDBJ/ENA/GenBank Feature Table Definition*. 2024. Disponível em: <https://www.insdc.org/submitting-standards/feature-table/>.

JUMPER, J. et al. Highly accurate protein structure prediction with alphafold. *Nature*, v. 596, n. 7873, p. 583–589, 2021. Disponível em: <https://www.nature.com/articles/s41586-021-03819-2>.

LANDRUM, M. J. et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, v. 44, n. D1, p. D862–D868, 2016. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26582918/>.

LGPD. 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.

RESOLUçãO nº 510, de 7 de abril de 2016. 2016. Disponível em: <https://www.gov.br/conselho-nacional-de-saude/pt-br/acesso-a-informacao/atos-normativos/resolucoes/2016/resolucao-no-510.pdf>.

RICHARDS, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, v. 17, n. 5, p. 405–424, 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/25741868/>.

ROY, R.; CHUN, J.; POWELL, S. N. Brca1 and brca2: Different roles in a common pathway of genome protection. *Nature Reviews Cancer*, 2012. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4972490/>.

SMITH, T. M. et al. Complete genomic sequence and analysis of 117 kb of human dna containing the gene brca1. *Genome Research*, v. 6, n. 11, p. 1029–1049, 1996. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/8938427/>.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, v. 53, n. D1, p. D609–D616, 2025. Disponível em: <https://academic.oup.com/nar/article/53/D1/D609/7902999>.

VARADI, M. et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, v. 50, n. D1, p. D439–D444, 2022. Disponível em: <https://academic.oup.com/nar/article/50/D1/D439/6430488>.

ZÁMBORSZKY, J. et al. Brca1 deficiency specific base substitution mutagenesis is dependent on translesion synthesis and regulated by 53bp1. *Nature Communications*, v. 12, p. 7352, 2021. Disponível em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8752635/>.