ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL FERNANDES LEAL

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA DESCOBERTAS DE FARMACOGENÔMICA NO TRATAMENTO DO CÂNCER DE ESÔFAGO

Porto Alegre
2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA DESCOBERTAS DE FARMACOGENÔMICA NO TRATAMENTO DO CÂNCER DE ESÔFAGO

GABRIEL FERNANDES LEAL

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. Dr. Rafael Heitor Bordini

Porto Alegre
2022

Ficha Catalográfica

L435a Leal, Gabriel Fernandes

Aplicação de Aprendizado de Máquina para Descobertas de Farmacogenômica no Tratamento do Câncer de Esôfago / Gabriel Fernandes Leal. – 2022.

60f.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientador: Prof. Dr. Rafael Heitor Bordini.

1. Aprendizado de Máquina. 2. Farmacogenômica. 3. Oncologia. 4. Câncer de Esôfago. I. Bordini, Rafael Heitor. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Loiva Duarte Novak CRB-10/2079

GABRIEL FERNANDES LEAL

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA DESCOBERTAS DE FARMACOGENÔMICA NO TRATAMENTO DO CÂNCER DE ESÔFAGO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 29 de Março de 2022.

BANCA EXAMINADORA:

Prof[a]. Dr[a]. Avaliadora Mariana Recamonde Mendoz (INF/UFRGS)

Prof. Dr. Avaliador Duncan Dubugras Alcoba Ruiz (PPGCC/PUCRS)

Prof. Dr. Rafael Heitor Bordini (PPGCC/PUCRS - Orientador)

DEDICATÓRIA

Dedico este trabalho à minha família, em especial minha mãe, Elisângela, minha avó, Terezinha e meu avô, Oralino.

AGRADECIMENTOS

Gostaria de iniciar agradecendo ao meu orientador, o professor Rafael Heitor Bordini, que não só me acolheu ao seu grupo de pesquisa como sempre trabalhou muito para o sucesso desse projeto. Sou grato por termos compartilhado esse período juntos, me fazendo crescer e sendo um ótimo exemplo a seguir. Seu incentivo em diversas etapas foi fundamental e agradeço imensamente pelos conhecimentos, pela paciência, por ter aceito esse desafio e por acreditar em mim.

À professora Renata Vieira, que esteve presente e sempre contribuiu muito com sua dedicação, conhecimentos e ideias para evolução do meu trabalho, ajudando a me guiar durante essa trajetória.

Agradeço também aos colegas que conheci através do AIR, obrigado por me receberem e por todo o apoio.

À professora Fernanda Bueno Morrone e todos os colaboradores do LFA, que contribuíram para o desenvolvimento desse trabalho e com quem pude trocar experiências.

Não poderia deixar de agradecer ao professor Osmar Norberto de Souza pela oportunidade e pelos conhecimentos no período em que trabalhamos juntos.

Um agradecimento especial à minha amiga e colega Flavielle Blanco Marques, que me acompanhou durante esse período e me ajudou em diversos momentos. Muito obrigado pelas experiências que compartilhamos.

Agradeço aos membros do PPGCC, colegas, funcionários e professores que contribuíram ao longo da minha formação.

Agradeço à minha mãe, Elisângela, que é meu grande exemplo, sempre apoiou os caminhos que escolhi e me incentivou. Aos meus avós, Terezinha e Oralino que se dedicaram para meu crescimento e são as pessoas que mais acreditam no meu sucesso. Vocês são o principal motivo de todas as minhas conquistas. Por fim, agradeço aos meus familiares e a todos os amigos que fiz ao longo do caminho e que sempre torceram por mim.

APLICAÇÃO DE APRENDIZADO DE MÁQUINA PARA DESCOBERTAS DE FARMACOGENÔMICA NO TRATAMENTO DO CÂNCER DE ESÔFAGO

RESUMO

A farmacogenômica é a área que estuda como as variações genômicas podem influenciar na resposta aos medicamentos. Através dela é possível explorar e definir os medicamentos mais indicados para diferentes pessoas e seus perfis genéticos, a fim de tornar os tratamentos mais personalizados. Estudos recentes mapeiam a resposta de linhagens celulares relacionadas ao câncer para uma ampla coleção de fármacos utilizados em tratamentos, aplicando técnicas de aprendizado de máquina para tarefas de predição. O objetivo dessa dissertação é desenvolver modelos de redes neurais profundas buscando predizer a resposta de diferentes perfis para 174 fármacos de tratamento do câncer de esôfago. Foram construídos modelos de aprendizagem profunda que, integrando dados do perfil de expressão, mutações e dados clínicos, estimam a resposta de diferentes compostos, com base nos valores de IC_{50}. Foram aplicadas estruturas de autocodificadores para extração de representação dos dados de treinamento, aliado a uma rede neural profunda. O modelo inicial obteve resultados positivos em comparação a trabalhos anteriores e, a partir destes, foram exploradas formas de aprimorar a predição da rede neural. Foi introduzida uma nova arquitetura com a integração dos dados clínicos devido a importância dos fatores de risco relacionados aos casos de câncer de esôfago. Além disso, outra motivação para explorar esses dados é que ainda são mais comuns de serem obtidos na prática clínica. Os modelos apresentaram resultados de 0,74 e 0,72 respectivamente, considerando a métrica de avaliação de erro médio quadrático. Apesar dos resultados positivos, foram identificadas limitações da implementação, especialmente sobre os dados clínicos em relação a sua quantidade e qualidade da informação. Os resultados experimentais mostram que o tema de pesquisa é promissor e podem levar a inovações capazes de melhora na qualidade de vida dos pacientes.

Palavras-Chave: Aprendizado de Máquina, Farmacogenômica, Oncologia, Câncer de Esôfago.

APPLYING MACHINE LEARNING TO THE PHARMACOGENOMICS OF ESOPHAGEAL CANCER

ABSTRACT

Pharmacogenomics is the area that studies how genomic variations influence drug response. Through it studies, it is possible to explore and define the most suitable drugs for different patients and their genetic profiles, in order to make treatments more personalized. Recent studies map the response of cancer-related cell lines to a wide collection of drugs used in treatments, applying machine learning techniques for prediction tasks. Our goal is to develop deep neural network models seeking to predict the response of different profiles to 174 drugs used for the treatment of esophageal cancer. Deep learning models were built to estimate the response of different compounds, based on its IC_{50} values, by integrating expression, mutation and clinical data. Autoencoders were developed to extract the representation of the training data, combined with a deep neural network. The initial model obtained positive results compared to previous work and based on these we explored new approaches to improve the neural network. We introduced a new architecture with the integration of clinical data due to the importance of risk factors related to esophageal cancer cases. Furthermore, another motivation to explore these data is that they are still more common to be obtained in clinical practice. The models presented results of 0.74 and 0.72 respectively, considering the mean squared error evaluation metric. Despite the positive results, implementation limitations were identified, especially regarding clinical data in terms of quantity and quality of information. The experimental results show that the research topic is promising and can lead to innovations capable of improving the quality of life of patients.

Keywords: Machine Learning, Pharmacogenomics, Oncology, Esophageal Cancer.

LISTA DE FIGURAS

2.1 Representação de um modelo matemático simples de um neurônio artificial. Nele os valores de entrada são computados em função da matriz de pesos. Esse resultado é utilizado para atualizar o valor obtido pela função de ativação, que define se haverá a propagação ou não da saída. Modelo adaptado de Norvig e Russell (2014). 27

2.2 Representação gráfica da estrutura de um autocodificador simples, com uma camada de entrada, uma camada de processamento e uma camada de gargalo. 28

5.1 Arquitetura dos modelos de redes de predição desenvolvidos 39

5.2 Distribuição dos valores em escala logarítmica de IC_{50} obtidos pelo GDSC antes da imputação, em vermelho, e após o processo de imputação, em azul. A imputação dos dados gerou uma distribuição próxima a original, apenas com aumento no volume dos valores medianos. 40

5.3 Representação de dispersão da predições utilizando o conjunto de teste com normalização min-max. Observa-se uma concentração dos valores seguindo a distribuição de densidade de IC_{50} e ao próximo da linha central, reforçando o baixo erro apresentado pela rede. 42

5.4 Médias dos resultados da predição do Modelo 1, representado pelas linhas azuis, comparadas à média do conjunto de dados de teste, as estrelas em vermelho. Através do gráfico é possível ter uma noção da distância do resultado para o esperado para cada composto. 43

5.5 Gráfico de comparação da predição para o subconjunto de 25% dos fármacos com menores valores de IC_{50}. Para esse grupo de interesse as predições geraram um erro menor se comparado ao grupo completo de fármacos, uma vez que os maiores erros se concentravam nos fármacos com maiores IC_{50}. 44

5.6 Representação de dispersão das predições utilizando o conjunto de teste com normalização min-max do Modelo 2, para as diferentes avaliações. É possível analisar pelos valores de IC_{50} que o Modelo 2 obteve uma qualidade de predição similar, porém com dispersão maior que o Modelo 1 para as avaliações com os dados clínicos ausentes. 47

5.7 Médias dos valores de predição do Modelo 2, utilizando os dados de expressão, mutação e clínicos para treinamento e teste, comparadas à média do conjunto de dados de teste. Analisando os resultados individualmente é possível perceber que o modelo tende a subestimar os valores de IC_{50}. . . 49

5.8 Distribuição dos valores de IC_{50} para as duas versões dos conjuntos de dados e as predições geradas pelos dois modelos. Em amarelo estão os valores para o Modelo 1, utilizando dados de expressão e mutação e em roxo os valores preditos pelo Modelo 2, com a inclusão dos dados clínicos. Observa-se que o Modelo 2 foi capaz de predizer valores próximos aos do Modelo 1. 51

LISTA DE TABELAS

5.1 Comparação das métricas de avaliação para os diferentes modelos e experimentos. O Modelo 1, que utiliza dados de expressão e mutação, é o que apresenta menor MSE para as predições. O Modelo 2 é referente a arquitetura proposta para integração de dados clínicos, processados pelo *encoder* C. A avaliação 2 do Modelo 2 utiliza a mesma arquitetura e reporta os dados de erro registrados nos experimentos que atribuem os dados clínicos para as linhagens celulares do conjunto de teste, com uma pequena melhora nesse valor. Já a Avaliação 3 do Modelo 2 atribui os dados clínicos durante o treinamento e teste. 50

5.2 Comparação dos conjuntos de dados utilizados para cada rede. 50

LISTA DE SIGLAS

AM – Aprendizado de Máquina

CCLE – Cancer Cell Line Encyclopedia

DFF – Deep Feedforward

DMOG – Dimethyloxallyl Glycine

EC_{50} – Half maximal effective concentration

FPKM – Fragments Per Kilobase of transcript per Million

GDC – Genomic Data Commons Data Portal

GDSC – Genomics of Drug Sensitivity in Cancer Project

HSL – Hospital São Lucas

IA – Inteligência Artificial

IC_{50} – Half maximal inhibitory concentration

INCA – Instituto Nacional do Câncer

LFA – Laboratório de Farmacologia Aplicada

MAF – Mutation Annotation Format

MAE – Mean Absolute Error

NIH – National Institutes of Health

MLP – Multilayer Perceptron

MSE – Mean Squared Error

PAPROP – Programa de Apoio a Projetos de Pesquisa

RELU – Rectified Linear Unit

RNA – Rede Neural Artificial

SNP – Single Nucleotide Polymorphisms

TCGA – The Cancer Genome Atlas Program

SUMÁRIO

1 INTRODUÇÃO 15
1.1 MOTIVAÇÃO 17
1.2 OBJETIVO GERAL 18
1.3 OBJETIVOS ESPECÍFICOS 18
1.4 ORGANIZAÇÃO 19

2 FUNDAMENTAÇÃO TEÓRICA 20
2.1 CÂNCER DE ESÔFAGO 20
2.2 FARMACOGENÔMICA 21
2.2.1 IMPLEMENTAÇÃO DA FARMACOGENÔMICA 22
2.2.2 BANCOS DE DADOS DE FARMACOGENÔMICA 23
2.3 APRENDIZADO DE MÁQUINA 24
2.3.1 *DEEP LEARNING* 26
2.3.2 REDES NEURAIS ARTIFICIAIS 26
2.3.3 *AUTOENCODERS* 28

3 TRABALHOS RELACIONADOS 30

4 METODOLOGIA 32
4.1 CONJUNTO DE DADOS 32
4.1.1 DADOS DE EXPRESSÃO E MUTAÇÃO 32
4.1.2 DADOS CLÍNICOS 33
4.2 CONSTRUÇÃO DOS MODELOS 35
4.2.1 MODELO INICIAL 35
4.2.2 AUTOENCODERS 35
4.2.3 REDE DE PREDIÇÃO 36
4.2.4 INTEGRAÇÃO DOS DADOS CLÍNICOS 36
4.2.5 MÉTRICAS DE AVALIAÇÃO 37

5 RESULTADOS 38
5.1 MODELO INICIAL 38
5.1.1 PRÉ-PROCESSAMENTO DOS DADOS 38
5.1.2 TREINAMENTO DO PRIMEIRO MODELO 40

5.1.3 ANÁLISE QUALITATIVA 41
5.1.4 PREDIÇÃO PARA FÁRMACOS COM BAIXO IC_{50} 43
5.1.5 EXPLORAÇÃO DE RESULTADOS DOS FÁRMACOS ESTUDADOS 44
5.2 IMPLEMENTAÇÃO DE MODELO UTILIZANDO DADOS CLÍNICOS 45
5.3 RESULTADO DO TREINAMENTO 46
5.4 COMPARAÇÃO DOS MODELOS 48

6 CONCLUSÕES E TRABALHOS FUTUROS 52
6.1 LIMITAÇÕES 53
6.2 PERSPECTIVAS 54

1. INTRODUÇÃO

Estudos para desenvolvimento de fármacos e tratamentos são fundamentais no combate e prevenção de diversas doenças. Entretanto, esses não se limitam apenas a descoberta e proposição de novos fármacos mas também na investigação dos efeitos sobre os pacientes. Cada vez mais a área da saúde se preocupa em desenvolver tratamentos especializados para os diferentes perfis de pacientes buscando não só obter sucesso destes como prezar pelo bem estar e a redução de efeitos adversos (Janiaud et al., 2019). Um dos fatores que proporciona essa mudança de perspectiva é a maior compreensão da área de genômica e das ferramentas criadas a partir da bioinformática.

Com a evolução de tecnologias de sequenciamento e diminuição de seu custo, em conjunto com o crescimento da computação e capacidade de processamento, tornou-se possível a coleta de grandes quantidades de dados biológicos (Subramanian et al., 2020). Consequentemente, há uma maior facilidade no desenvolvimento de estudos de genômica, proteômica, transcriptômica e a investigação de seus mecanismos. Isso abre caminho para a busca de novos fármacos e terapias, além de influenciar nas tecnologias utilizadas por outras áreas, além da saúde e da biomedicina.

A bioinformática é definida pelo uso de conhecimentos da ciência da computação e das ciências biológicas, com a utilização de softwares e bancos de dados para armazenamento, análise e simulações aplicadas aos dados obtidos por sequenciamento de genes, dados clínicos ou sequências de proteínas (Pevsner, 2015). Essa área pode ser vista sob duas perspectivas: a primeira que possui ênfase nos dados de sequências de nucleotídeos e aminoácidos; e a segunda, a bioinformática estrutural, que aborda os problemas biológicos de um ponto de vista tridimensional das proteínas.

Dada a natureza desses estudos, são geradas quantidades significativas de dados, em muitos casos, com análises bastante complexas. Isso tende a dificultar a exploração desses dados de forma exclusivamente manual (Nagaraj et al., 2018), o que destaca a importância da criação e aplicação de abordagens da área da computação para a extração de novas informações. Por este motivo, são criadas ferramentas e bases de dados com a intenção de dar suporte aos estudos de relacionados à genômica e outras áreas.

Devido a este cenário, é possível observar a aproximação da área de computação com a biologia e a saúde, empregando cada vez mais alguns métodos computacionais nos estudos de bioinformática. Atualmente, são encontrados trabalhos como os que buscam aplicar técnicas de Aprendizado de Máquina (AM) para tarefas de predição, e que se demonstram promissores na busca por melhores tratamentos e no desenvolvimento de novos medicamentos (Mostavi et al., 2020).

O desenvolvimento dessas técnicas e do aprendizado profundo leva a possibilidade de aplicações sobre diversos problemas biológicos, estudo de mecanismos celulares

e predição de resultados clínicos a partir de um grande conjunto de dados (Libbrecht e Noble, 2015). Ainda assim, a aplicação dessas técnicas possui desafios quando se trata de dados biológicos. Diferentemente de outras áreas, que adotaram rapidamente os modelos de AM, a interpretabilidade de seus resultados precisa estar associada ao seu significado biológico, uma análise por vezes bastante complexa. Além disso, quando se trata de dados genômicos, enfrenta-se também um problema de alta dimensionalidade, uma vez que a natureza desses dados apresenta a característica de baixo número de amostras e alto número de *features* (Hambali et al., 2020).

Exemplos de aplicações são os estudos que mapeiam a resposta de linhagens celulares relacionadas a uma determinada doença para uma ampla coleção de fármacos utilizados em tratamentos. Trabalhos deste tipo fazem parte da área da farmacogenômica, que estuda como as variações genômicas podem influenciar no efeito de medicamentos (Guo et al., 2019). Através de estudos de farmacogenômica é possível explorar e definir o tipo mais adequado de terapia para diferentes pessoas com base em seus perfis genéticos. Para isso, iniciativas recentes trabalham com foco em desenvolver ferramentas e repositórios focados em informações sobre a resposta de fármacos. A partir disso, são desenvolvidas diretrizes e projetos que buscam aproximar esses estudos da prática clínica e explorar novas abordagens para a descoberta de informações sobre a interação entre genes e fármacos e os efeitos nos pacientes.

Ainda que esse tópico não seja novo, nos últimos anos os trabalhos dessa área ganharam maior notoriedade, devido a maior disponibilidade de dados. Além disso, instituições públicas e privadas começam a explorar mais a farmacogenômica, uma vez que avanços desses estudos podem não só trazer benefícios aos pacientes como reduzir o custo de tratamentos a longo prazo.

Uma das áreas da saúde que mais pode se beneficiar desse tipo de estudo é a oncologia. Ainda que os tratamentos contra o câncer tenham avançado muito nos últimos anos, esta doença segue sendo bastante importante a nível mundial (WHO, 2021b). Com o estudo da doença a partir de dados genéticos, é possível investigar melhor suas características e alvos presentes em tumores visando aumentar a eficiência de tratamentos. Aliado a isso, o acesso a dados de resposta de fármacos para linhagens celulares relacionadas ao câncer contribui para estudos de predição. Embora o conhecimento acerca do câncer tenha crescido, ainda são encontrados obstáculos para os tratamentos, especialmente em relação aos seus efeitos adversos. Isso se dá também pelo fator de que a doença possui várias especificidades, uma vez que pode atingir diferentes órgãos (Canzoneri et al., 2019).

Um dos tipos mais graves é o câncer de esôfago, que apresenta uma baixa taxa de sobrevida. Isso se dá devido especialmente a dificuldade no diagnóstico da doença, que apresenta sintomas mais claros apenas em estágios mais avançados, diminuindo o tempo para tratamento (Short et al., 2017). O câncer de esôfago foi selecionado como

foco deste trabalho devido a importância e implicações como um problema de saúde, em especial para a região do Rio Grande do Sul (Fagundes et al., 2006).

Ainda que recente, a farmacogenômica se mostra promissora na busca por melhores tratamentos, apontando uma área promissora para contribuições e inovação. Portanto, através da perspectiva da computação são apresentadas aplicações de técnicas da área de aprendizado de máquina, em destaque as redes neurais profundas, para solução de um problema biológico, de forma a contribuir com estudos na área da farmacogenômica.

Neste trabalho foram explorados modelos de arquiteturas de rede neurais descritas em trabalhos anteriores e foi utilizado desse conhecimento para construir modelos considerando o cenário do câncer de esôfago. O modelo inicial obteve resultados positivos em comparação aos trabalhos anteriores, apresentando uma redução no erro das predições. A partir destes experimentos, foram exploradas formas de aprimorar a rede neural e para isso foi introduzida uma nova arquitetura com a integração de dados clínicos devido a importância dos fatores de risco relacionados aos casos de câncer de esôfago. Durante a implementação foram identificadas limitações, especialmente sobre os dados clínicos em relação a sua quantidade e qualidade da informação. Os resultados experimentais mostram que o tema de pesquisa é promissor e podem levar a inovações capazes de melhora na qualidade de vida dos pacientes.

1.1 Motivação

Seguindo os objetivos da área de farmacogenômica, além do desenvolvimento de novos medicamentos, tenta-se cada vez mais adequar os tratamentos ao perfil dos pacientes. Essa abordagem tem a intenção de trazer benefícios tanto nas decisões clínicas quanto para a saúde, e é conhecida como medicina de precisão personalizada (Puche et al., 2020).

Assim, busca-se utilizar de abordagens computacionais para mapear a grande quantidade de dados que descrevem os perfis genéticos de acordo com as informações de respostas de fármacos disponíveis. Para isso, o projeto aplicou essa estratégia sobre dados relacionados ao câncer de esôfago, devido a sua importância para a região do Rio Grande do Sul. Além disso, há a perspectiva de utilização de dados de pacientes locais, através do Hospital São Lucas da PUCRS (HSL). A análise desses dados e modelos pode melhorar a compreensão da influência de mutações e perfis de expressão na eficácia de fármacos, e assim beneficiar a assistência aos pacientes além de diminuir custos.

Os métodos que utilizam modelos de predição sobre dados biológicos apresentam cada vez mais resultados promissores, o que indica potenciais aplicações em estudos que buscam tratamentos para diferentes doenças. Além disso, trabalhos como este po-

dem impactar não só a qualidade de vida dos pacientes de câncer de esôfago, como também ter um significativo impacto econômico para instituições de saúde. Assim, o trabalho busca contribuir para a área de farmacogenômica, ainda em crescimento, utilizando o potencial das técnicas de aprendizado de máquina e podendo auxiliar no estudo da doença na população local.

1.2 Objetivo Geral

O objetivo geral deste trabalho foi desenvolver modelos de redes neurais profundas sobre dados locais e de repositórios públicos, buscando predizer a resposta de pacientes para diferentes fármacos de tratamento do câncer. Utilizando o câncer de esôfago como foco foi construído um modelo de aprendizagem profunda que, integrando dados do perfil de expressão do indivíduo, informações de mutações e dados clínicos possa predizer a resposta de diferentes tratamentos. O trabalho buscou avaliar os resultados obtidos em relação ao modelo de aprendizagem profunda e ao seu significado biológico, comparando com informações encontradas na literatura sobre os genes e fármacos explorados.

1.3 Objetivos Específicos

Buscando atingir o objetivo geral, inicialmente foram definidos alguns objetivos específicos para execução do trabalho, listados a seguir:

1. Estudar as técnicas e abordagens de IA aplicadas à área de farmacogenômica.
2. Coletar e analisar dados de câncer de esôfago e de resposta a fármacos nas bases de dados biológicos.
3. Implementar modelos de redes neurais capazes de predizer a resposta de fármacos relacionados ao câncer de esôfago.
4. Definir formas de integração de informações clínicas dos pacientes.
5. Comparar os resultados obtidos pelos modelos de predição e realizar possíveis correções.
6. Avaliar os resultados em relação as informações sobre os tratamentos e genes importantes para o estudo do câncer de esôfago.

1.4 Organização

Esse trabalho está dividido em 6 capítulos, organizados como segue:

- O primeiro capítulo apresentou os tópicos que o estudo aborda e as motivações e objetivos do trabalho.
- No Capítulo 2 é descrita a fundamentação teórica, onde são apresentados os principais conceitos relacionados ao tema da pesquisa. São aprofundados os tópicos sobre o câncer de esôfago e sua importância, a área da farmacogenômica e os principais conceitos do aprendizado de máquina e suas aplicações, como as redes neurais profundas.
- A seguir, no Capítulo 3, são explorados os trabalhos relacionados e suas abordagens.
- O Capítulo 4 descreve a metodologia adotada para o estudo e desenvolvimento dos modelos de redes neurais.
- No Capítulo 5 são apresentados os resultados da implementação dos modelos, definindo suas arquiteturas, experimentos realizados e resultados das métricas de avaliação, utilizadas para comparação dos modelos.
- Já no Capítulo 6, são feitas as considerações finais, avaliando as contribuições do trabalho, limitações e as perspectivas de continuidade.
- Por fim, são apresentadas as referências utilizadas durante o desenvolvimento da dissertação.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Câncer de Esôfago

O câncer é uma doença cuja principal característica é o crescimento desordenado de células, podendo levar a formação de tumores e a invasão de tecidos e órgãos (NCI, 2021). O processo de multiplicação celular é necessário para a renovação das células, porém, quando há uma interferência nesse mecanismo, pode haver um crescimento anormal nesse número.

O que causa essa desregulação está associado a fatores genéticos em associação à exposição de agentes do ambiente. Entre esses agentes cancerígenos estão os físicos, químicos ou biológicos, como radiação ionizante, componentes do tabaco e vírus ou bactérias, respectivamente (WHO, 2021a). Essas variações genéticas também podem ocorrer devido a erros no processo de divisão celular ou serem herdadas.

Segundo a Organização Mundial da Saúde (OMS), a doença ainda representa um grande problema de saúde estando em segundo lugar em relação ao número de mortes globalmente (WHO, 2021b). Além disso, há perspectivas de crescimento nos números de novos casos e mortes pelo câncer nos próximos anos, devido a dificuldade no diagnóstico e tratamento de determinados tipos da doença (WHO, 2020).

O câncer de esôfago possui importância mundial, especialmente por apresentar dificuldade no diagnóstico e alta taxa de mortalidade (Arnal et al., 2015). Isso torna ele o sexto maior entre as causas de morte por câncer e o oitavo mais comum no mundo. Estudos reportam que devido a agressividade desse tipo de câncer, a taxa de sobrevida em 5 anos é de aproximadamente 10% (Huang e Yu, 2018). Além disso o diagnóstico tardio se dá pela ausência de sintomas, que tendem a aparecer apenas nos estágios mais avançados da doença.

Os subtipos de câncer de esôfago mais comuns são o carcinoma de células escamosas e o adenocarcinoma, que juntos representam mais de 95% dos casos (Short et al., 2017). O primeiro, mais comum entre os casos de câncer de esôfago, se origina a partir das células epiteliais presentes na região do terço superior e médio do esôfago e está mais associado a ingestão de álcool e bebidas quentes. Já o adenocarcinoma, apesar de representar menor número de casos, vem se tornando mais comum (Fagundes et al., 2016). Os adenocarcinomas se desenvolvem a partir de células secretoras de muco presentes no terço inferior do esôfago e superior do estômago. Esse fato está associado principalmente a uma condição denominada esôfago de Barret (Jonge et al., 2014), que leva a alterações do tecido do esôfago devido a exposição repetida aos ácidos presentes no estômago, causada pelo refluxo.

Entre os principais fatores de risco estão a idade acima de 50 anos, o consumo de álcool, cigarro e a ingestão de bebidas muito quentes, como o chá e o mate, além do excesso de gordura corporal. Segundo o Instituto Nacional do Câncer (INCA), na população brasileira o câncer de esôfago é o sexto mais comum entre homens, com uma incidência duas vezes maior do que em mulheres. Seus principais sintomas, vistos em casos mais avançados, são a disfagia, dificuldade ou dor ao engolir alimentos e bebidas, náuseas e dor no tórax (INCA, 2021).

O estado do Rio Grande do Sul (RS) apresenta significativos índices de mortalidade desse tipo de câncer, especialmente em comparação com o restante do país. Enquanto a taxa de mortalidade no RS é de 8.61 (para cada 100 mil habitantes), a taxa nacional é de apenas 3.66 (Kuiava et al., 2018). Alguns estudos investigam a relação disso com o hábito de consumo de chimarrão, bebida ingerida em altas temperaturas (De Barros et al., 2000; Lubin et al., 2014; Kuiava et al., 2018). Devido a essas características locais da população do Rio Grande do Sul, apresentam-se elevados índices de ocorrência no estado, tornando este um importante foco para o desenvolvimento de estudo (Fagundes et al., 2016).

2.2 Farmacogenômica

A bioinformática é uma área de natureza interdisciplinar, que utiliza principalmente conhecimentos da ciência da computação e das ciências biológicas (Xiong, 2006). Através dela são estudados os mecanismos relacionados ao DNA, RNA, proteínas e como estes influenciam outras subáreas, como a biomedicina, a farmacologia e a imunologia. Esses estudos fazem partes das áreas como a genômica, proteômica, transcriptômica, metabolômica, entre outras. Essas são áreas que utilizam de amostras biológicas para a extração de dados sobre a expressão de genes, transcritos de RNA e quantificação de proteínas. Para isso são aplicadas tecnologias de sequenciamento de nova geração, gerando dados que posteriormente são usados pelos estudos de bioinformática (Conesa e Beck, 2019).

Já a farmacogenômica é considerada a área de estudo voltada para compreender o papel do genoma de indivíduos na resposta a medicamentos (Wake et al., 2019). Através dos estudos de farmacogenômica, são investigados como o perfil genético de uma pessoa, ou grupo de pessoas, responde diante de um tratamento e como é possível definir os melhores medicamentos de acordo com essa informação. Para isso, é correlacionada a expressão genética e as mutações presentes em um indivíduo com fatores de absorção, metabolismo e os efeitos sobre os alvos biológicos dos medicamentos (Relling e Evans, 2015). Esse é um trabalho importante pois em casos de câncer, variações somáti-

cas podem fazer diferença na resposta de tratamento, enquanto em doenças infecciosas determinadas mutações podem interferir na sensibilidade de um patógeno aos fármacos.

2.2.1 Implementação da farmacogenômica

Com o avanço das tecnologias de sequenciamento e a disponibilidade de grandes quantidades de dados nos últimos anos, os estudos de farmacogenômica se tornaram mais populares e, cada vez mais, tenta-se alinhar os seus resultados e descobertas com a prática clínica. Em diversos países já são implantados projetos com o objetivo mapear o genoma e integrá-los a dados clínicos da população (Saunders et al., 2019). A principal motivação é de que este mapeamento gere um acesso maior deste tipo de dados e, consequentemente, haja um avanço nos estudos de terapias para torná-las melhor direcionadas. Esse e outros estudos enfatizam a importância de áreas como a farmacogenômica, uma vez que se beneficia e vai de acordo com o direcionamento que a saúde está tomando, de forma a tornar tratamentos cada vez mais personalizados (Nicholson et al., 2020).

Ainda assim, mesmo com a expansão da área, sua implementação na rotina ainda possui barreiras. Estudos que exploram a aplicação da farmacogenômica afirmam que existem fatores que impedem o aproveitamento dos métodos desta, especialmente devido a limitações tecnológicas e financeiras de determinados países (Klein et al., 2017). Além disso é apontada necessidade de uma infraestrutura de tecnologia da informação segura e capaz de integrar dados que apoie as decisões clínicas dos profissionais de saúde.

A resposta ao tratamento de uma determinada doença pode variar de acordo com diversos fatores, como idade, gênero ou hábitos. O papel da farmacogenômica é investigar essa resposta de acordo com o fator genético para garantir a melhor eficiência possível ou a redução nos efeitos adversos. Isso é importante pois variações genéticas de um indivíduo podem resultar na mudança de estrutura de uma proteína envolvida em um processo metabólico e, consequentemente afetar a qualidade da ligação do composto com a proteína alvo (Goldstein et al., 2003). Essas mudanças estão relacionadas aos conceitos de farmacocinética, estudo dos processos de absorção e metabolismo do fármaco, e farmacodinâmica, estudo da ação do fármaco em relação a efeito e ligação com o alvo.

Um dos exemplos mais conhecidos que enfatiza a importância de estudos desse tipo é o da P450 2D6, enzima codificada pelo gene CYP2D6 e que possui mais de 100 alelos conhecidos (Gopisankar, 2017). Esse gene está envolvido direta ou indiretamente no metabolismo de pelo menos 25% dos fármacos aprovados e utilizados no mercado (Eichelbaum et al., 2006). O problema disso, é que pessoas com diferentes variações desse gene apresentam diferentes taxas de metabolismo, aumentando as chances de apresentar efeitos adversos caso esse fator não seja considerado.

De modo geral, o objetivo dos estudos de farmacogenômica é identificar um gene ou conjunto de genes que possam ser importantes para o funcionamento de um processo metabólico relacionado aos fármacos. Esse tipo de abordagem tem sido aplicado para diversas doenças, mas ganhou mais destaque com os trabalhos voltados à área da oncologia (Vicente et al., 2016).

Para analisar os resultados de um determinado tratamento, os estudos podem levar em consideração características de evolução clínica, volume de expressão de genes relacionados ou, do ponto de vista farmacológico, pelo cálculo de índices IC_{50} e EC_{50} (W Caldwell et al., 2012). O IC_{50} representa a medida de metade da concentração inibitória máxima, em outras palavras, o nível necessário de uma determinada substância para inibir uma função biológica em 50% de sua atividade. Comparativamente, o EC_{50} representa a potência de uma substância considerando a quantidade necessária para atingir metade do efeito biológico máximo após determinado tempo de exposição. O cálculo desses índices são comumente usados na área de farmacologia para indicar a potência de um tratamento e é possível encontrar em repositórios para farmacogenômica dados de resposta de fármacos baseados nestes.

2.2.2 Bancos de dados de farmacogenômica

Os estudos de bioinformática geram grandes quantidades de dados, o que torna complexo a análise manual desse tipo de informação (Nagaraj et al., 2018). Desta maneira, surgem bases de dados que disponibilizam informações de sequências de genomas, proteomas, linhagens celulares, mutações e dados clínicos. A seguir são descritas algumas das principais bases de dados que disponibilizam informações relevantes para o uso em estudos de farmacogenômica.

O *National Institutes of Health* (NIH) é responsável pela fundação do portal *The Cancer Genome Atlas Program* (TCGA), que oferece diversas ferramentas e dados para estudos de bioinformática (Weinstein et al., 2013). O TCGA não é focado exclusivamente na farmacogenômica mas disponibiliza dados de genômica, transcriptômica e proteômica relacionados a mais de trinta tipos de câncer. Através dele é possível obter coleções de perfis genéticos, dados de evolução clínica, além de arquivos que descrevem informações sobre mutações, o *Mutation Annotation Format* (MAF).

A partir do TCGA foi originado o *Genomic Data Commons Data Portal* (GDC), que sumariza informações sobre casos, genes e mutações do câncer (Grossman et al., 2016). Nele é possível explorar e analisar cerca de 84 mil casos, descrições de 23 mil genes e informações de 3,4 milhões de mutações. Essas informações são vinculadas a partir de outros projetos e/ou repositórios que as depositam e disponibilizam para uso público de

outros estudos. Além disso, o GDC oferece ferramentas para a análise dos casos clínicos e comparação do resultado de variações em determinados genes.

O *Genomics of Drug Sensitivity in Cancer Project* (GDSC) é um importante repositório para o estudo de tratamentos do câncer, originado pelo *Cancer Genome Project do Wellcome Sanger Institute* (Yang et al., 2012). Seu objetivo é identificar biomarcadores relacionados ao câncer para categorizar pacientes de acordo com sua probabilidade de resposta a terapias. Para isso, o GDSC disponibiliza informações de linhagens celulares focando na identificação de características genéticas que podem auxiliar na predição de sensibilidade aos fármacos. Atualmente, o GDSC possui dados de mais de 500 compostos, sua relação com diferentes perfis genéticos e a curva de resposta de doses baseados na informação de IC_{50}.

O projeto Cancer Cell Line Encyclopedia (CCLE) (CCLE e GDSC, 2015) é um portal vinculado ao GDC e GDSC que apresenta detalhes de linhagens celulares utilizadas em estudos genéticos do câncer, incluindo aqueles sobre resposta de fármacos. O CCLE disponibiliza acesso público a dados genômicos, visualização e análise de mais de 1.100 linhas de células cancerígenas.

Com o crescimento dos conjuntos de dados de expressão de genes e um maior número de estudos e projetos focados na resposta de fármacos, se tornou possível explorar mais estudos dentro da área de farmacogenômica. Utilizando os dados de expressão e variantes, disponíveis em repositórios como o TCGA e o GDC, em conjunto com os mapeamentos de sensibilidade de fármacos do GDSC e CCLE, é possível aplicar as técnicas de redes neurais profundas para o trabalho de predição para a eficiência dos tratamentos. Além disso, a utilização de outras informações clínicas de evolução e dados locais podem contribuir para esse resultado e um melhor entendimento da relação entre os perfis de pacientes e a resposta a fármacos.

2.3 Aprendizado de Máquina

De acordo com Norvig e Russell (2014), a inteligência artificial (IA) pode ser entendida como uma representação do processo de pensamento ou raciocínio. Seu objetivo é desenvolver técnicas que permitam simular ou expandir a inteligência humana, visando a resolução de problemas. Para atingir esse objetivo e desenvolver novas técnicas, a IA se utiliza de modelos e teorias matemáticas pré-estabelecidas para a interpretação de dados.

Conforme o avanço de tecnologias de processamento e armazenamento de dados e consequente aumento na complexidade de problemas, se faz necessário o desenvolvimento de técnicas capazes de solucionar problemas de forma autônoma, assim originando subáreas como o aprendizado de máquina. Este pode ser descrito como uma

subárea da IA, que busca aprimorar a resolução de tarefas através da experiência e adaptação, segundo descrito por Mitchell (1997).

As técnicas de AM aplicam modelos matemáticos e estatísticos sobre os dados sem exigir instruções explícitas, e sim buscando aprender a representação de sua distribuição (Carvalho et al., 2011). O desenvolvimento dessas técnicas busca ser preciso porém mantendo a capacidade de generalização, influenciada pela estrutura e natureza dos dados (Tan et al., 2016). No AM são utilizados conjuntos de dados que permitam detectar padrões e associações ou tentar compreender o caminho entre os dados de entrada e os dados de saída. Essas duas abordagens são conhecidas como aprendizagens descritiva ou preditiva, respectivamente, e neste trabalho pretende-se encontrar uma solução para um problema biológico focando nos modelos preditivos, através de redes neurais artificiais.

Essas subdivisões também podem ser descritas em termos de aprendizado supervisionado, não-supervisionado e por reforço. No aprendizado supervisionado, ou preditivo, é feita uma busca por uma função, a partir de amostras de treinamento, capaz de designar um rótulo ou classe para novos exemplos (Alpaydin, 2020). Nesse tipo de aprendizado estão inclusas as tarefas de classificação e regressão. Nelas tenta-se mapear um conjunto de entrada x e uma saída y, tarefa que pode ser descrita como uma função $y = g(x|\theta)$ onde $g(.)$ representa o modelo e θ representa seus parâmetros. No caso da classificação, dada a entrada x o algoritmo tem a tarefa de determinar a qual classe k ela pertence, enquanto para a regressão, a partir da entrada x, o modelo tenta predizer por uma função um valor contínuo de saída y.

Já no aprendizado não-supervisionado, o objetivo é descrever um conjunto de dados de entrada, já que suas classes são desconhecidas. Uma vez que não se tem rótulos pré-definidos tenta-se inferir uma estrutura ou similaridade entre esses dados e uma das abordagens mais conhecidas para essa tarefa é o agrupamento (Norvig e Russell, 2014). Nas tarefas de agrupamento, é utilizado um grande conjunto de dados não rotulados que são divididos em grupos a partir da análise de sua similaridade. De forma geral, os dados de um mesmo grupo são mais semelhantes em comparação aos de outros grupos.

No aprendizado por reforço os modelos aprendem através de um sistema de tentativa e erros, com *feedbacks* que levam a adaptação de sua estratégia. Esse sistema baseado em atribuir recompensas e punições busca maximizar a recompensa dentro de um período de tempo (Norvig e Russell, 2014).

Assim, essas abordagens podem ser utilizadas para a solução de problemas de diversas áreas de conhecimento. Isso pode ampliar as descobertas sobre dados biológicos, sendo uma possível solução para o problema da dimensão e volume dos dados envolvidos em estudos de bioinformática, genômica e farmacologia (Pevsner, 2015). Neste trabalho, focaremos nas abordagens de aprendizado supervisionado. A seguir são revisados conceitos relacionados aos algoritmos adotados na metodologia do trabalho.

2.3.1 *Deep Learning*

As abordagens estabelecidas no aprendizado de máquina não são capazes de processar o conjunto de dados na sua forma original. Isso leva a necessidade de um especialista de domínio que terá como tarefa definir as características que serão extraídas para representação do atributo alvo (LeCun et al., 2015).

Com a evolução das tecnologias e o aumento do poder de armazenamento e processamento computacionais, surgem as redes neurais de aprendizado profundo (*deep learning*). Trata-se de redes que podem ser implementadas com uma topologia que pode atingir um número elevado de camadas intermediárias, permitindo o ajuste dos pesos para um maior poder de aprendizado. Isso permite que os modelos não só aprendam uma função que mapeia uma entrada para uma saída mas também possam encontrar uma representação com múltiplos níveis de abstração (LeCun et al., 2015).

As arquitetura de redes de aprendizagem profunda mais comuns são as com propagação para frente, do inglês *Deep Feedforward* (DFF), que também podem ser conhecidas como *multilayer perceptrons* (MLPs). Segundo Goodfellow et al. (2016) são redes neurais artificiais implementadas com pelo menos duas camadas ocultas em que nenhum neurônio retorna sua saída para ele mesmo. Essa pode apresentar muitas aplicações comerciais, como o reconhecimento de objetos em imagens.

Além delas, há os modelos recorrentes, onde são processados em sequência os elementos de entrada e são armazenadas as informações de cada um para ser usado no processamento do elemento seguinte. Isso significa, que essas redes processam não só a informação de entrada mas também são influenciadas pela iteração anterior, funcionando de forma similar a uma memória. Suas principais aplicações são para as tarefas que exigem reconhecer padrões em sequências de dados, como em textos ou sequências de genes (Goodfellow et al., 2016).

2.3.2 Redes Neurais Artificiais

As redes neurais artificiais (RNAs) são implementações inspiradas pelo processamento de informações do cérebro humano e com o objetivo de simular seu aprendizado. Como descrito por Lorena et al. (2000), trata-se de sistemas distribuídos que aplicam uma função matemático, baseado no funcionamento do sistema nervoso central.

As RNAs são compostas por unidades de processamento conectadas que aplicam funções de ativação sobre as entradas. Essas unidades são chamadas de neurônios artificiais e suas conexões, conhecidas como sinapses. Os neurônios podem ser distribuídos

entre camadas e atualmente as redes neurais podem chegar a um grande número camadas permitindo o processamento de funções mais complexas (Goodfellow et al., 2016).

Em 1943, McCulloch e Pitts (1943) introduziram o primeiro modelo matemático de um neurônio artificial. A Figura 2.1 ilustra essa representação matemática de um neurônio, que utiliza um peso que é multiplicado para cada vetor da entrada. Sobre essas entradas é aplicada uma função de ativação, cujo resultado é propagado para as ligações das camadas seguinte, até a camada de saída. Um neurônio artificial pode ser descrito como estrutura composta de sinais de entrada, cujos dados podem ser obtidos do ambiente ou da ativação de outros neurônios; uma matriz de pesos, que descrevem as forças das conexões; nível de ativação, determinado pelos pesos dos sinais de entrada; e função de limiar, usada para o cálculo do estado final, a saída do neurônio (Luger, 2013).

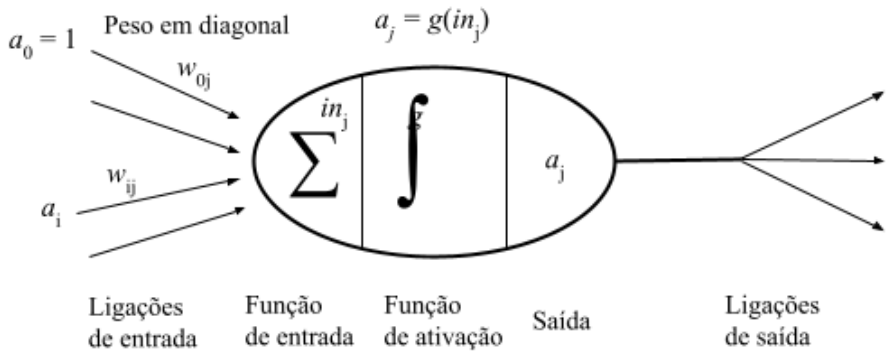


Figura 2.1: Representação de um modelo matemático simples de um neurônio artificial. Nele os valores de entrada são computados em função da matriz de pesos. Esse resultado é utilizado para atualizar o valor obtido pela função de ativação, que define se haverá a propagação ou não da saída. Modelo adaptado de Norvig e Russell (2014).

O aumento no número de camadas e as funções de ativação permitem a computação de funções matemáticas não-lineares, ampliando a possibilidade de criar modelos mais complexos. Quando as RNAs foram propostas, inicialmente eram utilizados modelos simples, com apenas uma camada, também chamado de *perceptron*. Com o aumento da capacidade computacional foi possível aprimorar essas redes e consequentemente o número de possíveis aplicações (Lorena et al., 2000).

As MLPs, adaptações dos *perceptrons*, possuem conexões que podem ser realizadas tanto com propagação para frente (*feedforward*) quanto por *backpropagation*. As redes *feedforward* utilizam conexões em uma única direção, de forma que um determinado nó recebe como entrada a saída do nó anterior, e propaga sua saída para a camada seguinte. Elas são implementadas com uma camada de entrada, camadas ocultas e uma camada de saída, onde nenhum neurônio retorna sua saída para ele mesmo (Goodfellow et al., 2016).

Atualmente diferentes áreas de conhecimento, incluindo a farmacogenômica, buscam explorar essas aplicações, devido especialmente ao seu potencial na descoberta de informações, análise de padrões e das tarefas de predição (Rafique et al., 2021). Neste trabalho, foi investigado como os estudos da importante área da farmacogenômica podem se beneficiar das implementações de redes neurais e aprendizado profundo.

2.3.3 *Autoencoders*

Os *autoencoders* são modelos de aprendizagem profunda não supervisionados formados por um par de codificador e decodificador, que tem como tarefa reduzir a dimensão dos dados de entrada. Esse processo é realizado de forma que a rede recebe um dado complexo e é forçada a comprimir essa informação, na tentativa de fazer sua reconstrução a partir dos atributos de maior importância do conjunto de entrada (Goodfellow et al., 2016). A Figura 2.2 mostra um exemplo de estrutura básica de um *autoencoder*.

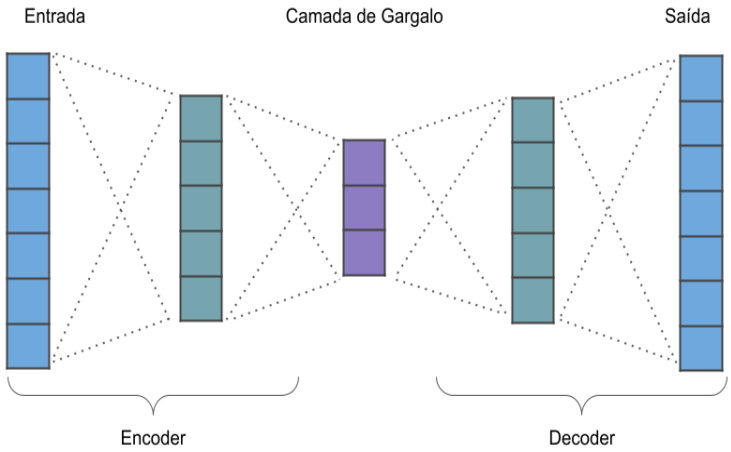


Figura 2.2: Representação gráfica de um autocodificador simples, com uma camada de entrada, uma camada de processamento e uma camada de gargalo. O objetivo é utilizar a subunidade de codificação para extrair uma representação de menor dimensionalidade de um conjunto de dados. A saída é feita para uma camada intermediária (de gargalo) e que envia a informação adiante para a reconstituição nas camadas de decodificação.

As camadas iniciais do modelo consistem em conjuntos de neurônios que utilizam funções de ativação não lineares para a compressão dos dados de entrada. Enquanto isso, as camadas responsáveis pela reconstrução da informação formam o decodificador. Essas estão ligadas por uma interface denominada camada de gargalo, uma camada oculta que restringe a dimensionalidade, chegando uma representação reduzida da informação e que pode ser reconstruída pelas camadas seguintes. Se por um lado o processo de codificação

e decodificação pode gerar um dado menos fiel em comparação ao original, por outro pode remover atributos que não são essenciais para a rede. A perda dessa informação, a diferença entre a reconstrução de saída e o dado de entrada, pode ser medida através da função de *loss*.

Os tipos de autocodificadores mais comuns são chamados de incompletos, formados pela função codificadora $h = f(x)$ e pelo decodificador que tenta produzir uma reconstrução $r = g(h)$. Além deles há variações de implementação como os autoencoders esparsos, de redução de ruídos e *Variational Autoencoders*.

O objetivo desse modelo é reduzir o volume de informações a serem consideradas e elencar os atributos relevantes tanto do ponto de vista dos dados para o aprendizado quanto de fatores biológicos (Rafique et al., 2021). A utilização desse modelo se destaca nos trabalhos de farmacogenômica em especial para lidar com os dados de expressão e de mutações, uma vez que apresentam a informação referente a milhares de genes, aumentando significativamente o volume de atributos para aprendizado.

Além disso, a vantagem de utilizar os autocodificadores é a sua aprendizagem ser não supervisionada, necessitando apenas de um conjunto de dados de treinamento como entrada.

3. TRABALHOS RELACIONADOS

A fim de compreender melhor os trabalhos da área de interesse, foi realizada uma busca por trabalhos relacionados e quais abordagens são aplicadas. Entre eles, destaca-se o trabalho de Chiu et al. (2019) que relaciona o aprendizado de máquina, o tratamento do câncer e a área de farmacogenômica.

Na literatura são encontrados trabalhos recentes que exploram o uso de técnicas de aprendizado de máquina, em especial o aprendizado profundo, como uma nova forma de abordar a área da farmacogenômica. No trabalho de Chiu et al. (2020) é abordado o uso de *deep learning* na classificação de cânceres e seus subtipos, reforçando o potencial deste tipo de aplicações. São descritos os trabalhos desenvolvidos até o momento e a contribuição de dados de linhagens celulares e fármacos para o futuro dos trabalhos de oncologia, tornando cada vez mais próxima a aplicação medicina de precisão.

O DeepDR é um modelo de aprendizagem profunda para predição de resposta de fármacos baseado em informações de mutações e expressão de células cancerígenas ou tumores (Chiu et al., 2019). O trabalho utiliza bases públicas para a coleta de dados de 33 diferentes tipos de câncer e mutações relacionadas a estes. O modelo gerado consiste em três redes neurais profundas para a execução das tarefas de codificação das informações de mutações, de codificação dos dados de perfis de expressão e por fim, predição com base nos dados gerados pelas duas redes anteriores. Assim, o trabalho foi capaz de apontar informações como por exemplo os genes com maior relevância em relação a resistência ou a eficiência para determinados fármacos, assim como características dos grupos de pacientes que apresentam respostas divergentes. O trabalho descreve valores de erro médio quadrático de 1,48 e 1,98 para treino e teste respectivamente.

O modelo DL-ADR de Liang et al. (2016) explora as reações adversas geradas por mutações em dois genes específicos, CYP2D6 e CYP1A2. Para isso, é proposto um modelo de aprendizado profundo que tenta relacionar polimorfismos de nucleotídeo único (do inglês, *single nucleotide polymorphisms* - SNPs) a fatores de risco. O modelo utiliza cadeias de Markov para a classificação das amostras e foi possível detectar 53 possíveis reações adversas, divididas em 14 categorias.

O trabalho desenvolvido por Sakellaropoulos et al. (2019) descreve a criação de um *framework* para a aplicação de algoritmos de *deep learning* para a resposta de terapias contra o câncer. O estudo exemplifica o uso dessa abordagem para o estudo de resposta a fármacos e relata a melhoria encontrada no uso de aprendizagem profunda em comparação com os trabalhos que utilizam outros métodos de AM. Isso vai ao encontro com o relatado em outros trabalhos e mostra o potencial que essa abordagem tem ao explorar relações biológicas e efeitos de fármacos, especialmente para os casso presentes na área de oncologia.

O DeepDSC é uma implementação que utiliza o redes neurais profundas para a predição de IC_{50} baseado em características genéticas de linhagens celulares e informações químicas dos compostos utilizados no tratamento do câncer (Li et al., 2019). Nesse trabalho são aplicados os *autoencoders* para a extração de características genéticas sobre dados do CCLE e do GDSC. Essa informação é combinada com as características dos compostos para a realização da predição da resposta aos fármacos.

Outro exemplo de aplicação de técnicas de aprendizado de máquina pode ser visto no trabalho de Wang et al. (2021), que desenvolve o framework DeepDRK. Este é proposto como um método de predição de resultado do tratamento de pacientes de câncer utilizando métodos de kernel para dados de fármacos e suas combinações. As redes desenvolvidas são treinadas com dados de linhagens celulares e compostos e agrupados através de matrizes de similaridade. O trabalho descreve que a utilização das informações de genômica, de composição química e sobre interações fármaco-receptor são capazes de boas predições além de apontar possíveis reposições de fármacos.

Neste trabalho busca-se aplicar aprendizado de máquina para a área de farmacogenômica, de modo a avançar a capacidade de prever o melhor tratamento para cada paciente através da grande quantidade de casos disponíveis em repositórios públicos. Avaliando os trabalhos relacionados da área, mencionados anteriormente, o trabalho "*Predicting drug response of tumors from integrated genomic profiles by deep neural networks*"foi selecionado como referência para a construção de um modelo preditivo.

O objetivo do trabalho é desenvolver uma rede capaz de estimar a resposta de fármacos, utilizando o câncer de esôfago como foco, devido a sua importância. Foi utilizado o estudo de Chiu et al. (2019) e a arquitetura proposta como ponto de partida, reproduzindo seus experimentos. Foi então construído um modelo de aprendizagem profunda que, dados o perfil de expressão e informações de mutações, busque predizer o valor de IC_{50} de diferentes compostos utilizados nos tratamentos anticâncer. A partir dessa implementação inicial, foram feitas adaptações de acordo com os dados e a adição de novas informações clínicas para explorar as aplicações dos algoritmos de predição no cenário do câncer de esôfago. A análise desses dados e modelos busca melhorar a compreensão da influência de mutações, perfis de expressão e dados clínicos na eficácia de fármacos e, assim, beneficiar a assistência aos pacientes além de poder diminuir custos de tratamentos.

4. METODOLOGIA

Visando desenvolver um modelo capaz de predizer a resposta para fármacos relacionados ao câncer de esôfago, foram estudados trabalhos relacionados, bancos de dados disponíveis e arquiteturas propostas. O modelo de rede descrito por Chiu et al. (2019) foi selecionado e adaptado com foco apenas no câncer de esôfago. A hipótese inicial foi a de que dessa forma seria possível especializar a tarefa de treinamento ao utilizar dados mais semelhantes (em comparação à utilização de informações sobre 33 tipos de câncer) e assim obter uma melhor predição. Além disso, seguindo essa hipótese, poderiam ser integradas mais informações sobre os casos, buscando melhorar esses resultados. Para isso, em uma segunda etapa foram incorporados dados clínicas de pacientes de câncer de esôfago, uma vez que podem apresentar informações sobre fatores de risco que influenciam no caso e no tratamento da doença. Ainda são poucos os trabalhos que adotam essa abordagem, porém além de agregar informações importantes sobre o risco, os dados clínicos são mais acessíveis na rotina clínica em relação a dados de sequenciamento.

Assim, inicialmente foram explorados e coletados dados nas bases do projeto Genomics of Drug Sensitivity in Cancer, Cancer Cell Line Encyclopedia e do The Cancer Genome Atlas Program, relacionados ao câncer de esôfago. Com base nesses dados, foram gerados modelos preditivos utilizando redes neurais profundas.

4.1 Conjunto de Dados

4.1.1 Dados de expressão e mutação

Os dados selecionados nas bases para estudo e treinamento dos algoritmos de AM são conjuntos de dados de expressão e de tumores de linhagens celulares de pacientes de câncer de esôfago. As informações dos casos são disponibilizadas em bases como o GDC, parte do TCGA, através de arquivos no formato XML estruturados para conter as informações clínicas sobre o tratamento, idade no diagnóstico, sobrevida, classificação do tumor, morfologia e informações demográficas como idade, gênero e etnia. Atualmente o GDC possui mais de 1.390 entradas de casos de câncer de esôfago, contendo informações de 49 mil mutações e 20 mil genes relacionados a esse tipo de tumor. Pelo GDC, também podem ser obtidos os dados de caráter molecular dos casos de câncer, disponibilizados através de arquivos de RNA-seq e MAF. O RNA-seq contém os valores correspondentes aos níveis de expressão dos genes, já normalizados no formato FPKM (Fragments Per Kilobase of transcript per Million), enquanto o MAF descreve a presença ou não de mutações nos

genes das amostras de tumores. O GDC contém informações de expressão e mutação para 199 casos de câncer de esôfago.

Os dados utilizados contêm informações sobre os dois principais subtipos de câncer de esôfago, o adenocarcinoma e carcinoma de células escamosas. As diferenças entre esses subtipos podem ser exploradas em futuros experimentos, uma vez que podem influenciar nas características do caso e tipo de tratamento recomendado.

As informações de expressão e mutação foram obtidas para as linhagens celulares de câncer de esôfago através do CCLE. Esses dados foram coletados e relacionados por meio de seu ID com outro dataset, de resposta de fármacos para o tratamento do câncer, disponíveis pelo GDSC. Este que contém o conjunto de valores de resposta de cada linhagem celular para diferentes compostos, no formato CSV. Através dessa base é possível relacionar, além do valor de resposta, a identificação da linhagem celular (a partir de informações do CCLE) e qual o alvo biológico para cada fármaco. Essa curva de resposta é baseada no índice IC_{50} e, a partir da informação da linhagem celular, pode ser relacionada com os dados de expressão e variantes obtidos pelo TCGA.

Foram baixados dados de quantificação de expressão de genes de 67 linhagens celulares referentes ao câncer de esôfago do CCLE, utilizadas nas análises de resposta de fármacos presentes no GDSC. Arquivos MAF, com a informação de mutações nos genes presentes nas linhagens, foram coletados gerando matrizes binárias de mutação, onde os estados são definidos em 1 para mutação e 0 para o tipo selvagem do gene, respectivamente. Aqueles genes sem mutações nas amostras foram eliminados.

Quanto aos dados de resposta de fármacos, foram coletados através do Projeto GDSC as informações de resposta das linhagens analisadas para 174 compostos utilizados no tratamento do câncer. A resposta é medida através do índice IC_{50}, apresentados em μm e representados em escala logarítmica no dataset. Para o pré-processamento, para aqueles dados ausentes foi executado o processo de imputação, calculando os novos valores através de uma média ponderada dos 5 vizinhos mais próximos. Para a tarefa de imputação foi utilizado o KNNImputer, parte da biblioteca Python scikit-learn (Pedregosa et al., 2011). Dentre as 11.658 entradas, 5.298 não possuíam informações, destacando a necessidade de completar esses dados. Com base nessas informações, foi implementada uma ferramenta para predizer os valores de IC_{50} para os diferentes perfis genéticos.

4.1.2 Dados Clínicos

Para além do uso de dados de expressão e mutação, foram integradas as informações clínicas para investigar sua influência na evolução dos pacientes e resposta ao tratamento em comparação com o uso das características genéticas. Uma vez feitos os

experimentos com os dados iniciais, foram feitos experimentos com a inclusão de informações clínicas dos casos como uma possibilidade de aprimoramento da predição.

Inicialmente foi prevista a coleta de dados públicos pelo TCGA e de dados locais através Hospital São Lucas da PUCRS. Foi então selecionado um conjunto de informações clínicas a partir do TCGA sobre 169 casos de câncer de esôfago. Este conjunto de dados é divido nas categorias clínica, de histórico familiar e exposições. São presentes informações do paciente como idade, gênero, idade ao diagnóstico e sobre o estágio da doença, tratamentos e classificação do tumor.

Entre os casos não havia informações sobre o histórico familiar. Já as informações clínicas e de exposição contavam com 151 e 30 atributos, respectivamente. Porém muitos dos atributos não possuíam informações para nenhuma das entradas, reduzindo o volume de informações disponíveis. Entre os 30.589 valores do dataset, 24.702 eram nulos ou não reportados. Assim, após avaliação foram removidas *features* sem nenhuma informação ou preenchidas com valores iguais (como "*Not reported*") em todas as entradas, além de eliminar entradas duplicadas. Em seguida, as variáveis categóricas foram tratadas utilizando as bibliotecas OrdinalEncoder e OneHotEncoder, parte do scikit-learn. Os atributos foram codificados de acordo com seu tipo, nominal ou ordinal, totalizando o número de atributos do dataset em 123.

Isso ocorre por alguns fatores, podendo ser pelo preenchimento não padronizado pelos estudos, mas também pela estrutura apresentar algumas informações que serão mais relevantes para alguns tipos de câncer que para outros. Ainda assim, o ponto positivo é que os principais atributos relacionados aos fatores de risco, como o estágio do tumor, idade e hábitos como fumo, possuem informações para os casos de câncer de esôfago. Por fim, como parte do pré-processamento, os dados categóricos foram codificados para melhor adaptação do treinamento, como no caso de atributos como status de vida do paciente, gênero e classificação do tumor.

Quanto aos dados locais, o trabalho conta com a contribuição do Laboratório de Farmacologia Aplicada (LFA), responsável pela coleta de dados de casos de câncer de esôfago no HSL. Essas atividades são parte de projeto de pesquisa financiado pelo Programa de Apoio a Projetos de Pesquisa (Paprop) e possui aprovação do Comitê de Ética e Pesquisa da PUCRS, sob o número CAAE: 47181421.0.0000.5336.

Os dados presentes nas bases do HSL beneficiarão não somente este estudo como aqueles desenvolvidos pela equipe do LFA. Devido a limitações de tempo de execução do trabalho, a coleta dos dados do HSL ainda se encontra em desenvolvimento, reservando essa etapa prevista inicialmente para trabalhos futuros. Entre as informações sobre cada caso estão: idade, tabagismo, alcoolismo, hábitos de consumo de chimarrão, histologia, estágio da doença, informações de cirurgias realizadas, quimioterapia, radioterapia, pós recidiva, óbito, data de óbito e marcadores para os genes CD39, CD73, P2X7, CD26 e ADA.

4.2 Construção dos modelos

4.2.1 Modelo Inicial

Após coleta e limpeza dos dados, foi desenvolvida uma rede neural seguindo o modelo proposto por Chiu et al. (2019). A RNA foi adaptada para o conjunto de dados de câncer de esôfago e utilizada como ponto de partida para o experimentos e comparação com o trabalho anterior.

A arquitetura da rede construída inicialmente pode ser compreendida em três partes: um *encoder* (M) que processa os dados de mutação; um segundo *encoder* (E) para processar os dados de expressão; e, por fim, uma terceira rede (P) que recebe as informações dos dois *encoders* e utiliza os dados de resposta dos fármacos como alvo e realiza a tarefa de predição. Tanto os *encoders* quanto a rede de predição foram construídos utilizando camadas totalmente conectadas.

As redes neurais foram implementadas utilizando a biblioteca Python Keras 2.4 (Chollet, 2015) e a API TensorFlow 2.5 (Abadi et al., 2015) para a construção e execução do treinamento. A função de ativação utilizada nas camadas de processamento foi a Unidade Linear Retificada (ReLU), e a camada de saída foi definida como linear para os valores de IC_{50} preditos. O método de otimização utilizado foi o Adam com o erro quadrático médio (ou MSE, do inglês Mean Squared Error) sendo usado como a função de *loss* para a atualização dos pesos. A inicialização dos pesos dos *encoders* M, E e da rede de predição P foi definida pela escala de variância uniforme He (He et al., 2015). Como forma a tentar prevenir o *overfitting* das redes, foi utilizado o método de regularização Early Stopping.

4.2.2 Autoencoders

As redes M e E são resultantes da codificação de dois *autoencoders*, utilizados para extrair uma representação dos dados de mutação e expressão, respectivamente. Neste trabalho os *autoencoders* foram utilizados devido ao grande número de atributos que os dataset de mutação e expressão apresentavam.

Assim, as redes foram treinadas utilizando os dados do TCGA por 100 épocas e os parâmetros aprendidos na fase de codificação foram salvos, para uso na execução completa da rede. Os resultados da camada de saída dos codificadores extraídos desse processo são conectados à rede de predição P para obter os valores estimados de IC_{50}.

Os *autoencoders* foram estabelecidos com uma camada de entrada N neurônios (referente ao número de atributos de cada conjunto), 1.024, 256, 64, 256, 1.024 e uma

saída N, que resulta na reconstrução dos dados. Assim, a arquitetura dos codificadores de mutação e expressão são definidas por quatro camadas, referentes ao processo de codificação dos dados. Para M, as camadas contêm 13.864 neurônios de entrada e 1.024, 256 e 64 nas camadas de processamento seguintes. De forma similar, o codificador de expressão E contém quatro camadas com 14.415, 1.024, 256 e 64 neurônios respectivamente.

4.2.3 Rede de predição

Após a etapa de pré-treinamento dos codificadores, a saída de ambas as redes foi ligada a uma terceira rede P do tipo *feedforward*, para a tarefa de predição. A rede completa foi então treinada utilizando as informações dos três conjuntos de dados: de resposta de fármacos, de mutação e de expressão. Para isso foi definida uma divisão de 80, 10 e 10% dos dados para uso como treinamento, validação e teste do modelo, respectivamente.

A rede P possui 5 camadas, sendo a primeira delas a entrada formada pela combinação das saídas dos codificadores e que possui 128 neurônios, três camadas ocultas densamente conectadas, também com 128 neurônios cada, e uma última camada de F neurônios. A camada de saída F representa a quantidade de fármacos para os quais são gerados os valores de IC_{50}, ou seja, 174.

Enquanto os codificadores M e E foram inicializados utilizando os parâmetros obtidos durante o processo de pré-treinamento, a rede P teve seus pesos inicializados aleatoriamente. Da mesma forma, a rede foi configurada para um treinamento de 100 épocas. Devido ao tamanho do conjunto de dados, também foi utilizado o método de Early Stopping, com o hiperparâmetro de tolerância definido em 5.

4.2.4 Integração dos dados clínicos

Após os experimentos iniciais utilizando os dados de expressão e mutação, foi proposta a implementação de mais uma RNA para adicionar as informações clínicas dos pacientes. Essa hipótese será estudada uma vez que, ainda que os dados de genômica sejam de extrema importância e tenham relação direta com a resposta ao tratamento, o mapeamento da expressão dos pacientes ainda não é uma prática amplamente implementada no cotidiano clínico.

A ideia é de que o trabalho possa aproveitar das informações contidas nas bases públicas e os dados clínicos contribuam para a melhor adaptação da rede ao cenário do câncer de esôfago, etapa não explorada nos trabalhos anteriores. Além disso, é possível

explorar a influência de cada informação para a predição e aplicação da rede para casos que não possuem os dados de genômica.

O novo modelo construído utiliza a arquitetura anterior e adiciona um terceiro autoencoder, para o processamento do conjunto de dados clínicos. De forma, similar o autoencoder é treinado utilizando os dados do TCGA e o *encoder* C é extraído com os pesos do treinamento. A saída do codificador é ligada a rede de predição P que passa por uma adaptação, aumentando o número de neurônios de entrada para 160. O *encoder* possui 3 camadas, com 123, 64 e 32 neurônios.

Após sua construção, os modelos foram aplicados para realização do treinamento e predição da resposta aos medicamentos para o conjunto de amostras de teste. Considera-se que um determinado valor de IC_{50} alto indicaria uma possível resposta adversa de um paciente ao composto correspondente.

4.2.5 Métricas de avaliação

Para a avaliação do desempenho dos modelos foram computados os resultados de MSE e o erro médio absoluto (MAE, do inglês *Mean Absolute Error*). As métricas foram utilizados para o treinamento dos autoencoders e das redes de predição para comparação entre modelos e com os trabalhos anteriores.

Através do MSE é calculado o quadrado da média das distâncias observadas entre os valores preditos pela rede e os valores esperados. Assim, busca-se baixos valores de MSE, cuja forma matemática pode ser vista na Equação 4.1. Nela Y_i representa os valores reais e $\hat{Y}_i$ os valores preditos.

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \tag{4.1}$$

Já a outra métrica utilizada, o MAE, apresenta o resultado da média das distâncias entre os valores preditos e os valores reais e é descrito pela Equação 4.2.

$$MAE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} \tag{4.2}$$

5. RESULTADOS

Neste capítulo são relatados os resultados obtidos a partir da implementação de modelos criados para a análise de informações do câncer de esôfago. Neste trabalho foram coletados dados clínicos, de resposta de fármacos, de expressão e mutação de genes de pacientes com câncer de esôfago para o desenvolvimento e treinamento de uma rede neural artificial capaz de predizer valores de IC_{50}.

Como ponto de partida para a construção da RNA foram estudados os trabalhos que aplicam técnicas de aprendizado de máquina para a farmacologia e foi utilizado o modelo proposto por Chiu et al. (2019). No trabalho é investigada a resposta de 256 fármacos para 935 linhagens celulares. O estudo aborda 33 tipos de câncer e define um modelo de rede neural para o processamento dos dados de mutação e expressão destes, de forma a tentar inferir o valor do índice IC_{50} dos diferentes compostos.

Seguindo a abordagem proposta, foi implementada uma rede neural utilizando dois *encoder* para os dados de expressão e mutação e uma rede final que recebe essa informação e prediz os valores de IC_{50} para o conjunto de 174 fármacos. O Modelo 1 da rede foi então treinado e analisado através das métricas de avaliação definidas.

Em uma segunda etapa, após os experimentos iniciais, foi então abordada a integração de dados clínicos, visando trabalhar sobre a hipótese de que estas informações podem contribuir para o aprendizado da rede. Para a construção do Modelo 2, foi desenvolvido um novo *encoder*, que recebe os dados clínicos pré-processados, além de expandir o número de neurônios na camada de entrada da rede de predição P. A representação gráfica completa dos modelos de rede neural está ilustrada pela Figura 5.1 e os detalhes de seus treinamentos são descritos a seguir.

5.1 Modelo inicial

5.1.1 Pré-processamento dos dados

Após a coleta dos dados dos repositórios públicos, foi realizada a etapa de pré-processamento dos conjuntos de dados para o treinamento. Para os arquivos de mutações, foram removidos os atributos correspondentes aos genes que não possuíam mutação para nenhuma das entradas. Assim, o número de genes presentes nas matrizes binárias de mutações do CCLE e do TCGA passou para 13.864. Dentre os arquivos de RNA-seq, de quantificação de expressão, foram eliminados os atributos com todas as entradas vazias ou zeradas, gerando assim conjuntos de dados de expressão de 14.415 genes.

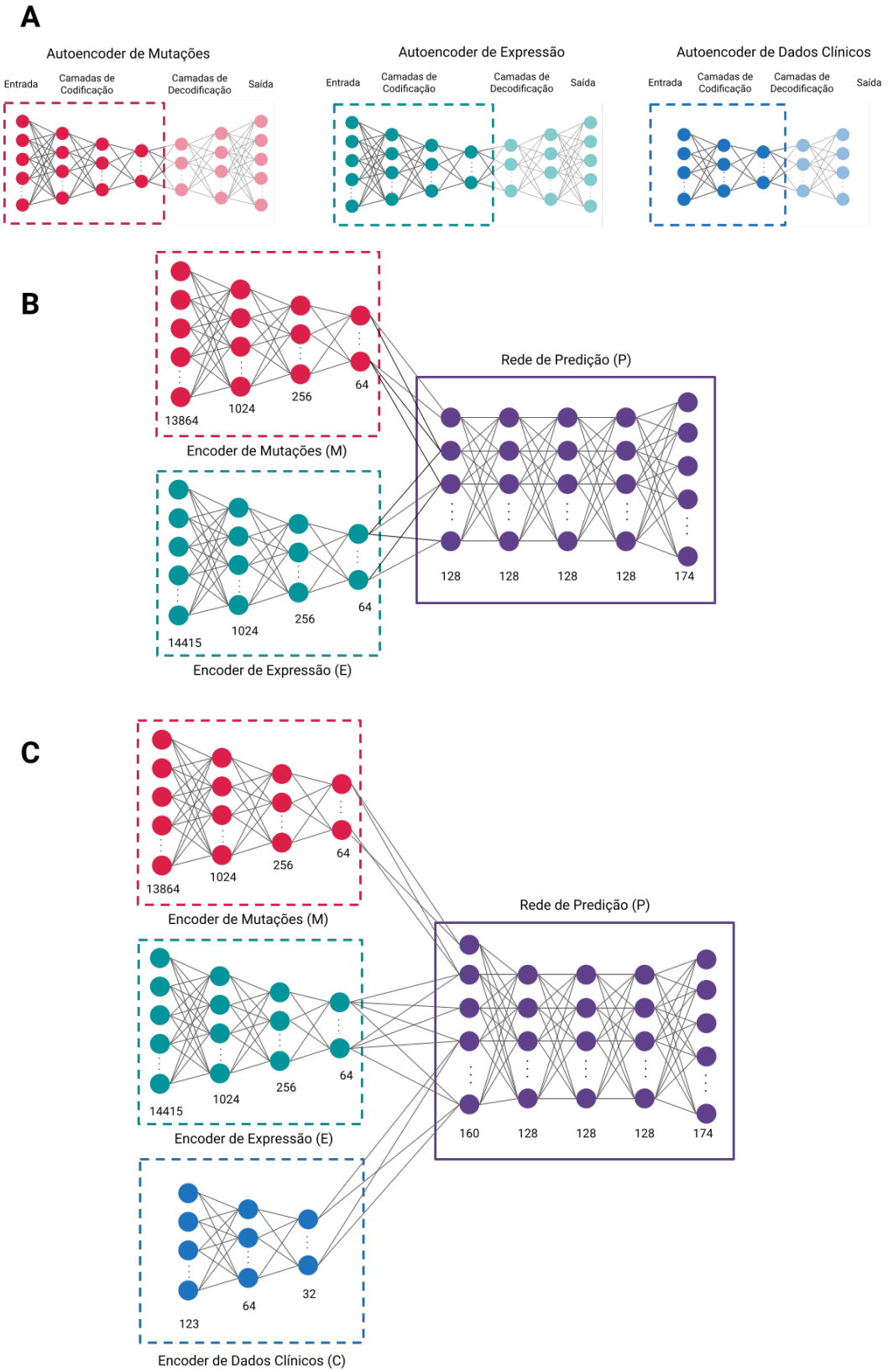


Figura 5.1: Arquitetura dos modelos de redes de predição desenvolvidos. Em (A) estão ilustradas as estruturas dos *autoencoders* M, E e C, cujo as camadas de codificação foram selecionadas após treinamento. Já em (B) é possível ver o modelo inicial desenvolvido para predição a partir do resultado dos codificadores de expressão e mutação para os valores de IC_{50}. Por fim, (C) ilustra o modelo proposto para o processamento dos dados clínicos.

Quanto ao *dataset* de resposta de fármacos, os valores ausentes foram imputados pelo cálculo da média dos 5 vizinhos mais próximos. O conjunto gerado após a imputação dos novos dados apresentou um intervalo de valores de IC_{50} entre -12,54 e 11,90. Para análise do resultado de imputação foi gerado um gráfico de estimativa de densidade comparando a distribuição dos valores observados no conjunto original com os imputados, que pode ser observado na Figura 5.2. É possível visualizar uma estimativa maior para certas regiões dos valores mas, de forma geral, o processo de imputação gerou valores que seguem a distribuição do *dataset* original.

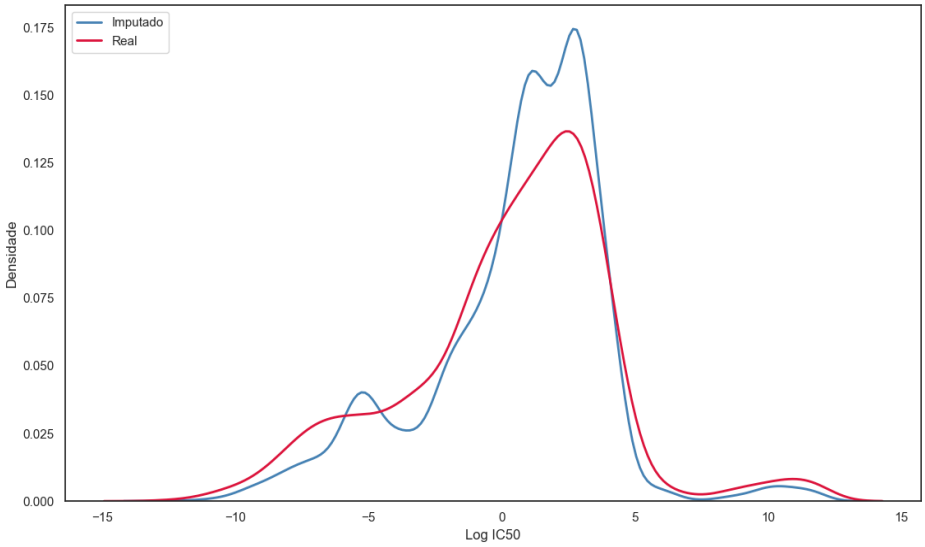


Figura 5.2: Distribuição dos valores em escala logarítmica de IC_{50} obtidos pelo GDSC antes da imputação, em vermelho, e após o processo de imputação, em azul. A imputação dos dados gerou uma distribuição próxima a original, apenas com aumento no volume dos valores medianos.

5.1.2 Treinamento do primeiro modelo

Como parte dos experimentos iniciais, foram treinados os autoencoders de expressão e mutação e a rede completa de predição. O modelo construído para o *autoencoder* de informações de expressão foi treinado utilizando os dados do TCGA. Seu treinamento foi avaliado e obteve um valor de MSE de 0,17. Seguindo os modelos descritos em trabalhos anteriores, o autoencoder foi configurado para realizar 100 épocas de treinamento. Entretanto, devido ao *early stopping* a rede encerrou o treinamento após 78 épocas.

De forma similar, o segundo autocodificador foi treinado com os dados de mutações e foi registrado um resultado de MSE de 0,16. Para essa rede, o treinamento foi executado por apenas 42 épocas antes de atingir a condição do *early stopping*.

Embora essas redes não sejam utilizadas isoladamente, os valores foram considerados para estimar a capacidade dos *encoders* de representação e reconstrução dos dados. Esses modelos foram então salvos, armazenando os pesos computados durante o treinamento para uso posterior.

Nota-se que ambos autocodificadores atingiram valores de erro baixos antes mesmo de executar as 100 épocas de treinamento. Isso pode ser um reflexo das dimensões consideravelmente menores que no trabalho de Chiu et al. (2019). Considerando esse fator, a utilização da técnica de regularização de *early stopping* mostrou-se útil. Assim, foi possível evitar o *overfitting* das redes aos respectivos dados, especialmente no caso do conjunto de mutações que possui um modelo de dados mais simples.

A partir disso, os modelos dos autoencoders foram reutilizados e com a API Functional do Keras foi possível selecionar apenas as camadas de codificação das redes. Isso permite redirecionar o objeto criado com o modelo de camada e seus pesos registrados, concatenando as suas saídas para a entrada da nova rede P, para a tarefa de predição.

Foi realizado então o treinamento da rede completa de predição, utilizando o pré-processamento dos *encoders* para os conjuntos de dados de expressão e mutação das linhagens celulares. Para isso, ao invés de um único atributo, foi definido como o alvo a predição dos valores de IC_{50} para os 174 fármacos.

A performance do treinamento da rede foi registrada, obtendo os valores de MSE de 0,69 e 0,74 com os dados de treino e de teste, respectivamente, enquanto o valor de MAE foi de 0,53 para o treinamento e 0,33 para o teste. Estes testes iniciais são importantes, pois indica não só que a rede foi comparável aos estudos anteriores como há uma melhora nos resultados observados. O trabalho de Chiu et al. (2019), relata valores de MSE de 1,48 e 1,98 para treino e teste respectivamente. Além disso, foram desenvolvidos outros modelos que observaram altos valores de MSE, como os modelos utilizando SVM e regressão linear que obtiveram MSEs de 8,92 e 10,24, respectivamente.

5.1.3 Análise qualitativa

Para visualização dos resultados, foram gerados gráficos para análise da predição. O primeiro método foi representar a distribuição do conjunto total dos valores preditos em relação ao resultado esperado. Isso é demonstrado pela Figura 5.3, que mostra todos os valores preditos de IC_{50} utilizando normalização min-max. O eixo central representa o valor esperado do conjunto de dados de teste. Por meio do gráfico é possível observar que os valores se mantiveram próximos ao eixo central, indicando consistência com os dados reais, sem que estejam completamente ajustados. Além disso, a maioria dos registros está concentrado no centro da distribuição. Porém, além deles destaca-se um grupo de fármacos com valores mais altos, desviando do padrão visto. Esse grupo

apresenta uma dispersão maior, representando um aumento no erro, que pode ser visto ao analisar os resultados individualmente para cada fármaco.

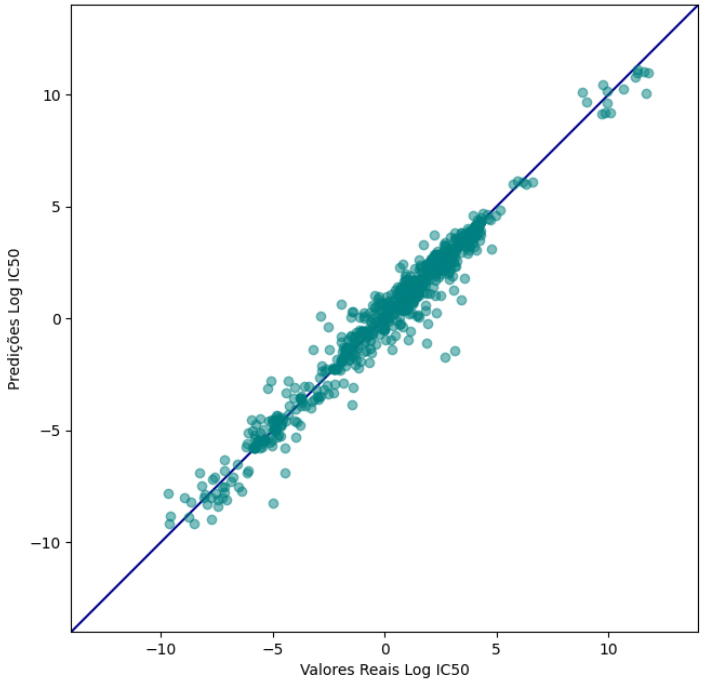


Figura 5.3: Representação de dispersão da predições utilizando o conjunto de teste com normalização min-max. Observa-se uma concentração dos valores seguindo a distribuição de densidade de IC_{50} e ao próximo da linha central, reforçando o baixo erro apresentado pela rede.

Após análise do conjunto completo dos valores de predição, foram analisados os resultados de forma individual. Para isso foi utilizado o gráfico apresentado na Figura 5.4, que serviu para analisar a saída da rede de predição em relação ao valor do conjunto de testes para cada fármaco. Nele as linhas representam as predições, enquanto as estrelas representam o valor real.

Devido a natureza do estudo, o resultado da rede possui um conjunto de saídas para cada entrada, dificultando a visualização individual. Para fins de representação, foi computada a média dos valores de IC_{50} de cada um dos 174 atributos alvos, para o conjunto de teste e de valores preditos respectivamente.

O gráfico representa a comparação da predição dos 174 compostos para o conjunto de entrada do *dataset*. Para cada um destes, percebe-se que o modelo foi capaz de reproduzir o comportamento dos dados de IC_{50}, se aproximando do valor real sem estar sobre-ajustado.

Isso demonstra a importância do fator de tamanho do conjunto de análise da rede, mostrando que a especialização para o aprendizado de um tipo de câncer pode

melhorar a predição de resposta. Esse resultado estar atrelado ao fato de considerarmos conjuntos de dados mais coesos e com valores esperados de IC_{50} com menor variação.

Observa-se que determinados pontos, especialmente os mais baixos, se ajustam melhor ao valor real do conjunto de teste, ou seja para fármacos com IC_{50} menores. Enquanto isso, através do gráfico também é possível notar uma maior distância (ou maior erro) nos casos de fármacos com valores de resposta mais altos, conforme mencionado anteriormente.

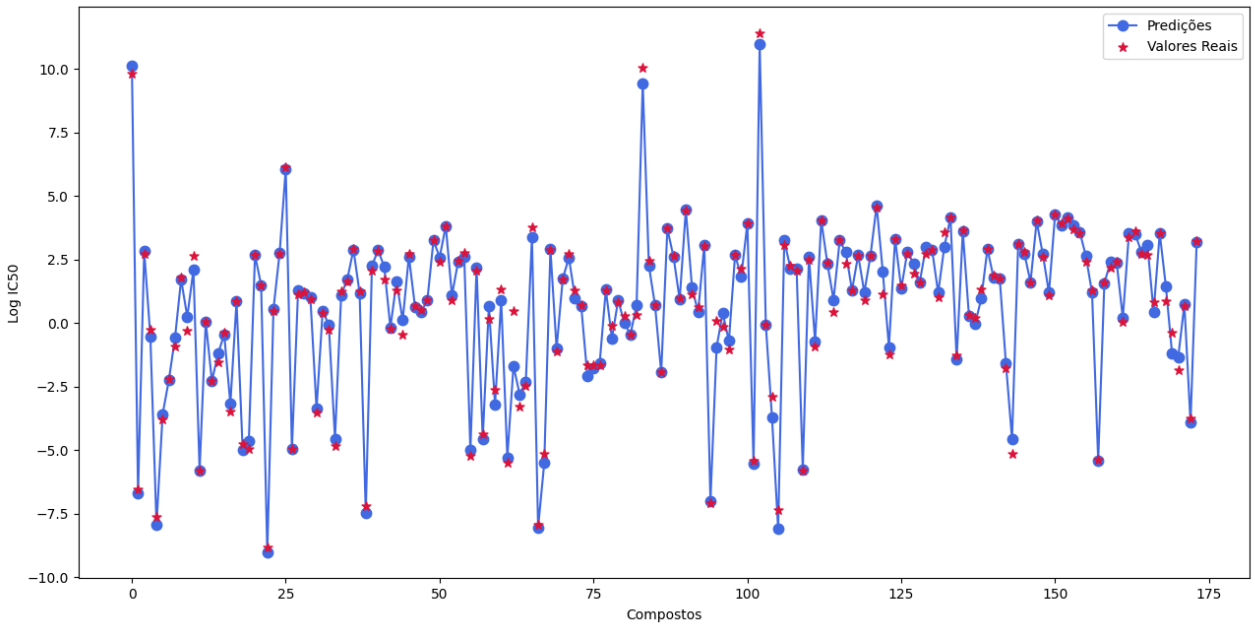


Figura 5.4: Médias dos resultados da predição do Modelo 1, representado pelas linhas azuis, comparadas à média do conjunto de dados de teste, as estrelas em vermelho. Através do gráfico é possível ter uma noção da distância do resultado para o esperado para cada composto.

5.1.4 Predição para fármacos com baixo IC_{50}

A partir de inspeção dos gráficos de resultados de predição, notou-se uma diferença menor entre os valores mais baixos apresentados. Considera-se os valores de IC_{50} mais indicados para tratamentos tendem a ser os mais baixos, significando que estes necessitam uma concentração menor para atingir seu efeito, diminuindo o nível de toxicidade ao paciente. A confirmação dessa observação pode indicar um resultado positivo ao melhorar a predição da rede para o grupo de compostos de maior interesse.

Para esse teste, foi proposto o cálculo do erro médio da predição utilizando apenas esse grupo de medicamentos. A ideia é comparar a eficiência do modelo em estimar o valor de IC_{50} com os medicamentos que apresentaram melhores resultados.

Foi então criado um subconjunto extraindo 25% dos medicamentos com menor média de IC_{50} nos dados do teste para medir o MSE da predição. Para este conjunto, o MSE estimado foi de 0,47, um resultado melhor em comparação com os resultados anteriores.

A Figura 5.5 apresenta uma comparação entre os valores preditos com os valores observados apenas para esse grupo de medicamentos. É possível observar um erro menor para esses fármacos, com alguns dos valores se ajustando quase completamente. Esse é um resultado importante ao considerar que estes tendem a ser os mais indicados para administração aos pacientes.

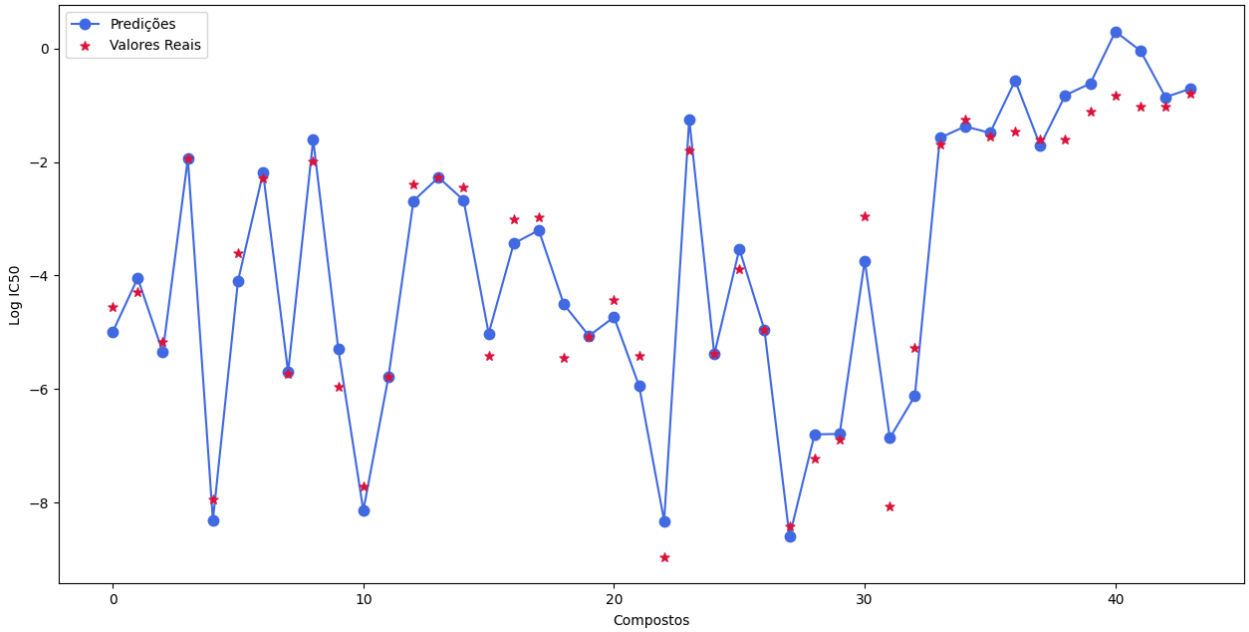


Figura 5.5: Gráfico de comparação da predição para o subconjunto de 25% dos fármacos com menores valores de IC_{50}. Para esse grupo de interesse as predições geraram um erro menor se comparado ao grupo completo de fármacos, uma vez que os maiores erros se concentravam nos fármacos com maiores IC_{50}.

5.1.5 Exploração de resultados dos fármacos estudados

Através dos gráficos apresentados na Figura 5.4, também pode-se destacar a identificação dos compostos que apresentam os maiores valores de IC_{50}. O conjunto de dados aponta que em média os cinco fármacos com maiores valores de IC_{50} são Phenformin, DMOG (Dimethyloxallyl Glycine), AICA Ribonucleotide, Mirin e UNC0638. Estes, correspondem aos picos observado no gráfico formado pela predição da rede e são fármacos que possuem como alvo processos envolvendo o metabolismo celular e a integridade do genoma.

Os altos valores para o IC_{50} podem indicar uma maior resistência ao fármaco, sendo necessário analisar o risco de efeito adverso aos pacientes devido a maior concentração necessária para o efeito sobre o alvo. A exploração dessas informações pode ser abordada em estudos posteriores de farmacologia e citotoxicidade.

Por outro lado, os compostos que apresentaram os menores valores de IC_{50} foram o Bortezomib, Docetaxel, Epothilone-b, Vinorelbine e Paclitaxel, respectivamente. A exemplo, o Docetaxel foi utilizado no estudo de Chiu et al. (2019) e foi apontada uma alta na sensibilidade de pacientes de câncer de esôfago para esse composto (25,3% dos casos). Esse fato foi relacionado a um aumento no taxa de mutação entre o grupo de pacientes sensíveis e aqueles mais resistentes a esse composto.

Quanto a informações sobre mutações, o conjunto de perfis genéticos apresentou o maior número de alterações nos genes TP53, TTN, MUC16, LRP1B e SYNE1. Estas alterações estão presentes em 87,50, 81,25, 45, 38,75 e 35% das linhagens celulares analisadas, respectivamente. Isso pode indicar quais fatores genéticos estão mais diretamente relacionados aos casos de câncer de esôfago e no impacto na resposta aos fármacos estudados. Além disso, é preciso considerar as variações entre os perfis de diferentes populações, especialmente para posterior aplicação sobre dados regionais.

O TP53, por exemplo, é frequentemente associado a diversos tipos de tumores, uma vez que está ligado diretamente à proteína p53, um supressor de tumores responsável por regular a divisão celular (Donehower et al., 2019). Alterações neste gene se refletem na instabilidade de cromossomos, sendo capazes de causar a deleção de genes responsáveis pela formação de outros supressores de tumores e cuja expressão elevada está associada ao aumento no ciclo de reprodução celular.

Destaca-se a possibilidade de estudos seguintes relacionando os genes com as mutações mais comuns, bem como a sua expressão, e como isso afeta a resposta apresentada pelos fármacos. Da mesma forma, a partir destes resultados podem ser exploradas abordagens que utilizem outras informações importantes sobre os tumores. No caso dos dados dos genes é possível investigar o uso de informações não só sobre a presença ou não de mutações, mas dos tipos de alterações presentes.

5.2 Implementação de modelo utilizando dados clínicos

Após a realização e avaliação dos primeiros experimentos, foi dado seguimento na implementação da rede. Seguindo a hipótese de uso das informações clínicas para melhoria da rede foi estudada a forma de inclusão desses dados, utilizando o modelo construído até o momento.

Para isso, foi coletado o conjunto de dados com informações clínicas do TCGA para 169 casos de câncer de esôfago. Após o pré-processamento e codificação das infor-

mações foi desenvolvido um terceiro autocodificador C. Seguindo os modelos criados anteriormente, essa rede foi criada para processar mais um conjunto de dados de entrada. Sua estrutura foi adaptada para conter menos camadas e unidades de processamento, uma vez que possui menos *features* que os conjuntos de mutação e expressão.

Da mesma forma, a rede final de predição foi adaptada para receber a nova entrada de dados. A camada de saída do *encoder* C, que passa a se ligar com a rede P, possui 32 unidades de processamento. Portanto, foi aumentado para 160 o número de neurônios na camada de entrada da rede P, que concatena o resultado dos 3 codificadores, M, E e agora C. A arquitetura do segundo modelo de rede criado pode ser visto na Figura 5.1 - C, onde as quatro estruturas que compõe a rede estão ilustradas.

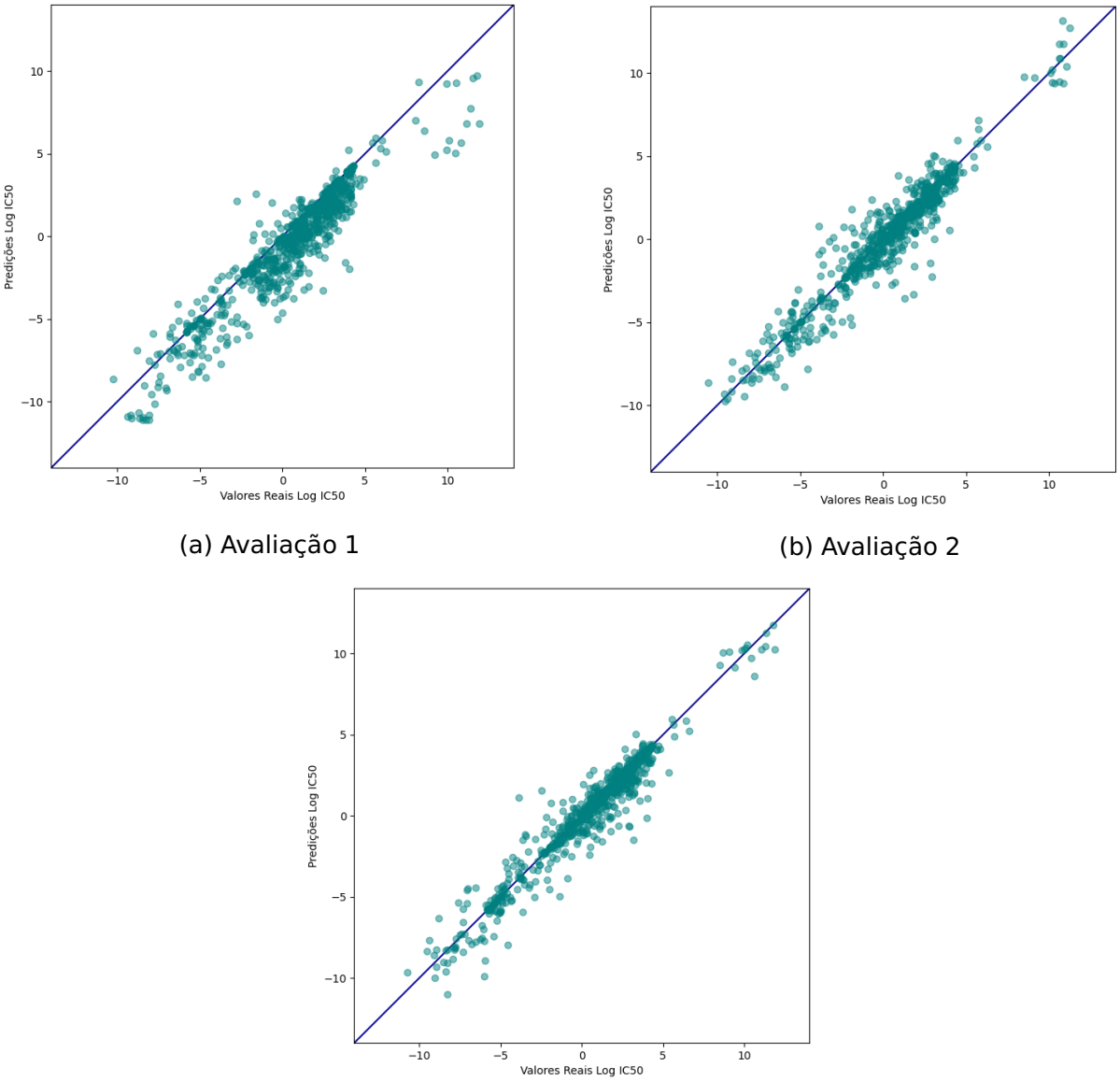
Para este modelo foram definidas três avaliações diferentes, para fins de comparação: na primeira avaliação a rede é treinada e testada com os mesmos dados do primeiro modelo, de expressão e mutação; na Avaliação 2 são usados parte dos dados clínicos do TCGA para atribuição às linhagens celulares do conjunto de teste, e feita sua predição; já na Avaliação 3, o conjunto de dados clínicos do TCGA que contém 169 entradas é divido entre 102 e 67 amostras, sendo a primeira parte usada para o treinamento do *autoencoder* C e as restantes pareadas com as amostras de linhagens celulares, assim permitindo o treinamento e teste do Modelo 2 com os dados clínicos.

Essa abordagem foi aplicada devido a ausência das informações clínicas das linhagens celulares nos bancos de dados. Para atribuição dos dados clínicos ao conjunto de linhagens foi utilizada a similaridade entre os valores de expressão entre os dados presentes no CCLE e no TCGA. Os resultados dos treinamentos e das diferentes avaliações desse modelo são descritos a seguir.

5.3 Resultado do treinamento

O *autoencoder* C foi treinado utilizando os dados clínicos do TCGA. Seu treinamento foi avaliado e obteve um valor de MSE de 0,13. Diferentemente das outras redes de codificação, este atingiu um baixo valor de erro executando todas as 100 épocas previstas.

Após treinamento as camadas de codificação foram extraídas, formando o *encoder* C. Junto com o treinamento dos *encoders* E e M para os casos, os resultados foram salvos e os codificadores gerados foram ligados a entrada da rede P. O Modelo 2 completo foi treinado por 100 épocas e foi feita a Avaliação 1. Para o treino foram usados os dados de mutação e expressão das linhagens celulares. Uma vez que esse dataset não possui dados clínicos atribuídos, o *input* desse conjunto foi nulo para esse experimento. A performance do modelo obteve valores de MSE de 0,74 para o treino e de 0,88 para o teste. Os valores de MAE computados foram de 0,54 e 0,47 para treino e teste, respectivamente.



(a) Avaliação 1

(b) Avaliação 2

(c) Avaliação 3

Figura 5.6: Representação de dispersão das predições utilizando o conjunto de teste com normalização min-max do Modelo 2, para as diferentes avaliações. É possível analisar pelos valores de IC_{50} que o Modelo 2 obteve uma qualidade de predição similar, porém com dispersão maior que o Modelo 1 para as avaliações com os dados clínicos ausentes.

Devido às limitações das informações usadas para o treinamento foi decidido atribuir os valores de dados clínicos para as linhagens celulares. Assim, foi definida a Avaliação 2, onde foi atribuído um conjunto de dados clínicos para as linhagens do conjunto de teste para analisar a predição considerando esses dados. Para isso isso foi aplicada validação cruzada sobre o modelo, pelo método k-fold, com *k* definido em 10. Assim a cada iteração, foram usados os conjunto de 10% dos dados clínicos de teste que não são utilizados no pré-treinamento do *autoencoder* C. Os dados rearranjados foram pareados com os dados de entrada para predição, avaliando esta para diferentes conjuntos de da-

dos. Isso foi feito para garantir que o resultado não seja dependente da seleção dos dados para atribuição.

O objetivo do experimento era avaliar o resultado de predição para estes, em comparação ao primeiro teste. Para cada interação os valores de MSE e MAE do teste foram registrados e ao final foi foram computados os valores gerais de 0,86 e 0,69. Ainda, avaliou-se que os resultados de cada interação não apresentaram grande variação, demonstrando que a qualidade da predição é consistente, independente do dataset. Isso também ilustra a combinação dos três conjuntos de dados e a capacidade de representação dos codificadores.

Já para a terceira avaliação, os dados clínicos foram utilizados tanto para treinamento quanto teste. Para isso, foi feita a partição do conjunto de dados clínicos em dois subconjuntos, de 102 e 67 entradas. O conjunto de 102 amostras foi utilizado para o treinamento do *autoencoder* C, enquanto o restante foi atribuído para o conjunto completo de dados de linhagens celulares, posteriormente divido em dados de treino e teste. Para essa avaliação o Modelo 2 obteve resultados de MSE de 0,70 e 0,72, e valores de MAE de 0,53 e 0,56. A diferença entre as avaliações pode ser vista através da Figura 5.6

Avalia-se que o modelo obteve resultados similares aos da primeira rede, ainda que seu erro seja maior para algumas avaliações ou métricas. A diferença para o primeiro modelo pode se dar pela quantidade de dados presente no dataset de dados clínicos, além de possuir informações mais heterogêneas. Ainda assim, o resultado pode ser considerado bom, uma vez que segue tendo um erro menor que o visto em trabalhos anteriores. Destaca-se também a necessidade de maiores trabalhos quanto a inclusão e atribuição dos dados clínicos, e de que forma estes podem contribuir aos modelos.

Observa-se uma distribuição de acordo com os valores esperados, porém com uma dispersão maior. Além disso, é possível identificar que há uma tendência a subestimar os valores de IC_{50} em comparação ao primeiro modelo. Esse resultado fica mais explícito ao analisar o gráfico de predições para cada fármaco. A Figura 5.7 apresenta a média dos valores preditos pela rede e os respectivos resultados esperados. Ainda que o erro médio seja maior, a rede foi capaz de predizer valores de acordo com o conjunto de teste. Analisando individualmente cada atributo, é possível visualizar que em geral o valor predito tende a ser menor que o valor esperado.

5.4 Comparação dos modelos

Os resultados obtidos foram analisados comparando os modelos criados e os trabalhos anteriores. Os valores para as métricas de avaliação MSE e MAE de treinamento e teste das diferentes redes e avaliações são descritos na Tabela 5.1. O trabalho de Chiu et al. (2019), descreve valores de MSE de 1,48 e 1,98 para treino e teste. Destaca-se a di-

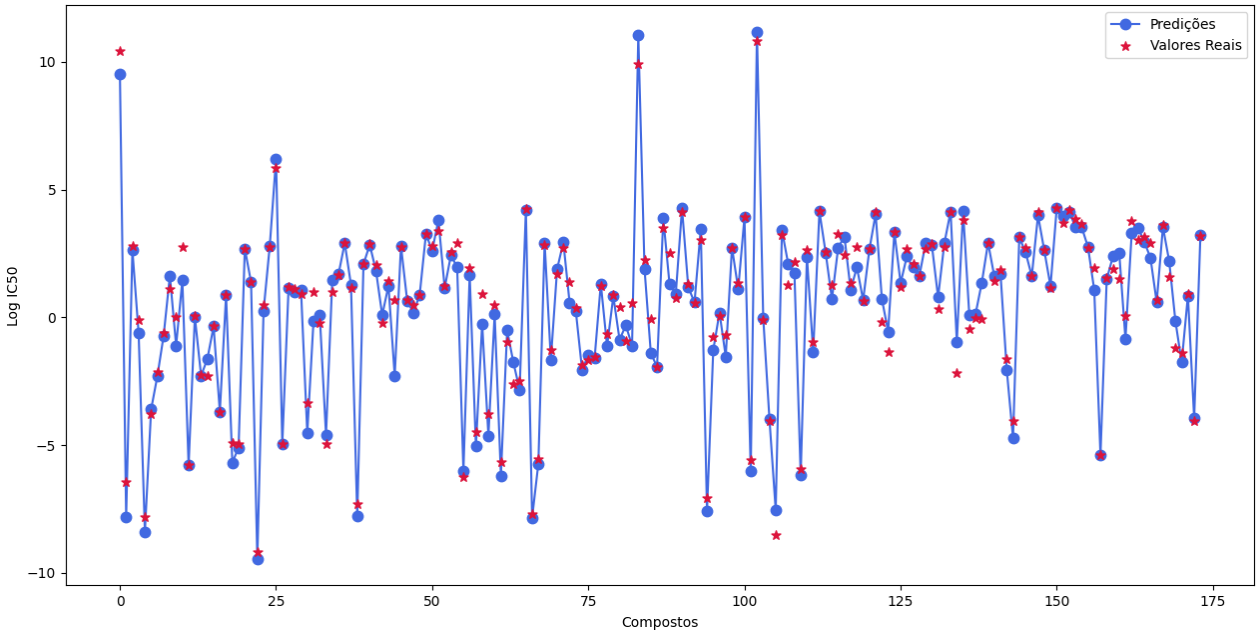


Figura 5.7: Médias dos valores de predição do Modelo 2, utilizando os dados de expressão, mutação e clínicos para treinamento e teste, comparadas à média do conjunto de dados de teste. Analisando os resultados individualmente é possível perceber que o modelo tende a subestimar os valores de IC_{50}.

ferença entre os experimentos, uma vez que o modelo do estudo usa o conjunto de dados com todos os tipos de câncer, tornando o uso dos valores apenas para referência. Ainda assim, a redução nos valores de MSE demonstram uma melhora no poder de predição da rede em relação ao estudo anterior.

Isso contribui para a hipótese inicial do trabalho que considera a possibilidade de melhora do aprendizado ao especializar o treinamento para casos do mesmo tipo de câncer. É preciso considerar que há uma diminuição significativa nos conjuntos de amostras processadas. Ainda assim, não só o volume de dados é menor como podem apresentar menor variação. Especialmente em comparação a um conjunto com 33 tipos de câncer onde, apesar de compartilhar características em comum, cada um possui suas especificidades em termos de expressão e mutações (Ramazzotti et al., 2018).

Os resultados do erro médio dos codificadores E e M refletem o *loss* computado durante o aprendizado. O valor de *loss* baixo atrelado ao treinamento interrompido antes das 100 épocas pode indicar um ajuste da rede sobre os dados. Esse fator reforça a necessidade da adição das regularizações utilizadas, sem que tenha comprometido os resultados.

Ao adicionar as camadas de codificação das informações clínicas observa-se um aumento no erro da rede de predição. Ainda assim, o modelo consegue predizer valores próximos ao esperado e com resultados melhores que no trabalho que utiliza informações de todos os tipos de câncer. Esse é considerado um resultado positivo, por estar próximo

Tabela 5.1: Comparação das métricas de avaliação para os diferentes modelos e experimentos. O Modelo 1, que utiliza dados de expressão e mutação, é o que apresenta menor MSE para as predições. O Modelo 2 é referente a arquitetura proposta para integração de dados clínicos, processados pelo *encoder* C. A avaliação 2 do Modelo 2 utiliza a mesma arquitetura e reporta os dados de erro registrados nos experimentos que atribuem os dados clínicos para as linhagens celulares do conjunto de teste, com uma pequena melhora nesse valor. Já a Avaliação 3 do Modelo 2 atribui os dados clínicos durante o treinamento e teste.

Modelo	**Dados Treino**	**Dados Teste**	**Treino**		**Teste**	
			MSE	**MAE**	**MSE**	**MAE**
Chiu et al. (2019)	Expressão + Mutação	Expressão + Mutação	1,48	-	1,98	-
Autoencoder E	Expressão	-	0,17	0,28	-	-
Autoencoder M	Mutação	-	0,16	0,16	-	-
Autoencoder C	Dados Clínicos	-	0,13	0,15	-	-
Modelo 1	Expressão + Mutação	Expressão + Mutação	0,69	0,53	0,74	0,33
Modelo 2 - Avaliação 1	Expressão + Mutação	Expressão + Mutação	0,74	0,54	0,88	0,47
Modelo 2 - Avaliação 2	Expressão + Mutação	Expressão + Mutação + Dados Clínicos	0,74	0,54	0,86	0,52
Modelo 2 - Avaliação 3	Expressão + Mutação + Dados Clínicos	Expressão + Mutação + Dados Clínicos	0,70	0,53	0,72	0,56

ao primeiro modelo, ainda assim é uma limitação a ser estudada em trabalhos futuros para melhora das predições. Isso pode ser obtido através de novas abordagens ou estratégias para integração dos dados clínicos.

Além disso, outro fator importante registrado durante o desenvolvimento é quanto a qualidade dos dados clínicos. Durante o pré-processamento destes, fica claro o problema da ausência de muitos dos dados do *dataset*. Assim, uma possível solução é a busca por novos meios de obtenção dos dados clínicos para maiores experimentos.

Tabela 5.2: Comparação dos conjuntos de dados utilizados para cada rede.

Modelo	**Origem**	**Amostras**	**Atributos**
Chiu et al. (2019)	TCGA + CCLE	9.059 + 622	18.281 + 256
Autoencoder E	TCGA	199	14.415
Autoencoder M	TCGA	199	13.864
Autoencoder C	TCGA	169	123
Rede P	CCLE	67	174

Para fins de comparação, a quantidade de dados envolvida em cada rede, assim como a fonte, está descrita na Tabela 5.2. A primeira entrada apresenta os dados usados no pré-treinamento dos autoencoders e no treinamento da rede descrita no trabalho de Chiu et al. (2019). A seguir estão as amostras dos conjuntos do TCGA, usados no treinamento dos autoencoders E, M e C, este último que apresenta uma redução no número de

amostras. Por fim, os conjuntos de dados do CCLE usados pela rede completa P, com a quantidade de amostras de expressão, mutação e dados clínicos testados.

A comparação entre os valores preditos pelas redes pode ser vista no gráfico de densidade apresentado na Figura 5.8, que contém a distribuição dos valores de IC_{50} dos *datasets* originais e das predições de cada modelo. Além disso, pode-se ver na Figura 5.8 que os valores preditos pelos modelos seguiram uma distribuição em acordo com a apresentada nos conjuntos de dados original e com os valores imputados. Apesar da diferença no erro médio, os valores preditos entre as duas redes possui uma distribuição similar.

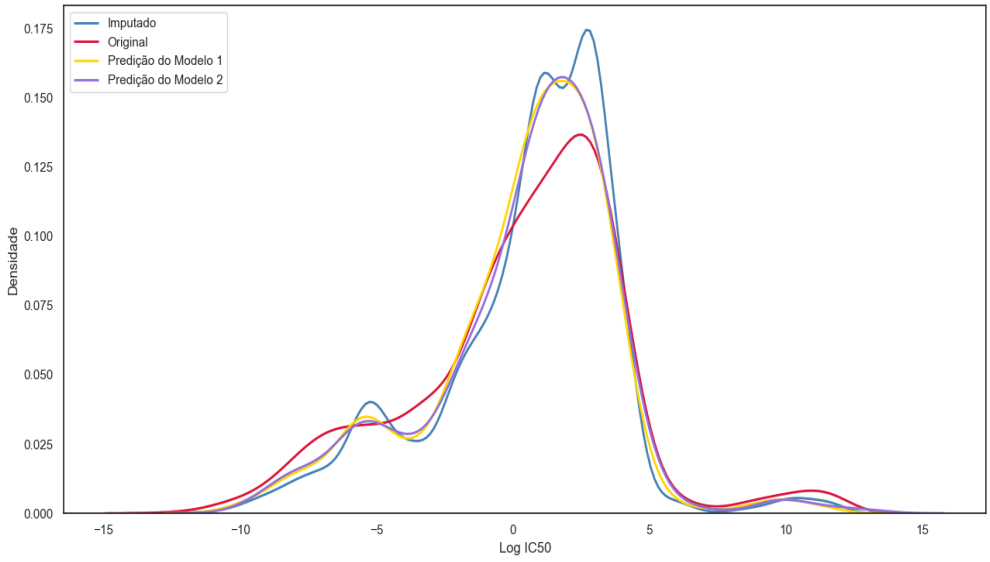


Figura 5.8: Distribuição dos valores de IC_{50} para as duas versões dos conjuntos de dados e as predições geradas pelos dois modelos. Em amarelo estão os valores para o Modelo 1, utilizando dados de expressão e mutação e em roxo os valores preditos pelo Modelo 2, com a inclusão dos dados clínicos. Observa-se que o Modelo 2 foi capaz de predizer valores próximos aos do Modelo 1.

6. CONCLUSÕES E TRABALHOS FUTUROS

Nesta dissertação foram exploradas as aplicações de técnicas de aprendizado profundo para a área da farmacogenômica, usando informações de expressão, mutação e dados clínicos para a predição de valores de IC_{50}. Os estudos realizados tiveram ênfase nos dados de tumores e fármacos utilizados no tratamento do câncer de esôfago. Assim, foram desenvolvidos dois modelos de redes neurais com o objetivo de estimar os melhores fármacos para cada perfil de paciente, devido a importância da área da farmacogenômica para o avanço de tratamentos mais adequados, especialmente para o câncer de esôfago, que possui destaque regional.

Primeiramente foi desenvolvida uma rede que utiliza os dados de expressão e de mutação para a tarefa de aprendizado. Os experimentos foram realizados com os dados de câncer de esôfago, considerando a hipótese inicial de melhoria nos resultados obtidos ao especializar o estudo para um único cenário de câncer. Estes serviram para avaliar os dados, bem como o pré-processamento necessário e identificar as adaptações do modelo de acordo com volume de dados.

Os treinamentos dos autocodificadores atingiram bons resultados em relação ao erro. Ainda assim foi observada a necessidade da utilização das técnicas de regularização para evitar o *overfitting* das redes e manter a generalização para novos dados. Isso assegurou a capacidade da tarefa extração de *features* e representação dos dados para o aprendizado, que se refletiu no resultado final da predição.

Avalia-se que a implementação do primeiro modelo teve resultados positivos e, a partir destes foram exploradas formas de aprimorar a rede. Busca-se considerar as informações clínicas devido a importância desses atributos e os possíveis fatores de risco identificados relacionados aos casos de câncer de esôfago. Além disso, outra motivação foi explorar esses dados uma vez que ainda são mais comuns de serem obtidos na prática clínica do que os de expressão e mutações.

Assim, dando seguimento no desenvolvimento, foi introduzida uma nova arquitetura com a integração dos dados clínicos. Para isso foi criado um novo codificador para representação das novas informações. Apesar dos resultados positivos, foi identificada uma das limitações da implementação, referente aos dados clínicos, especialmente sobre a quantidade de atributos sem informação. Os testes realizados indicaram que o modelo foi consistente quanto ao erro médio, mesmo ao considerar mais informações para predição. Além disso, os experimentos com a atribuição de dados clínicos às linhagens foi capaz de mostrar que os modelos obtêm uma boa capacidade de generalização ao ser submetido a novos conjuntos de dados.

No estudo foram avaliados 199 casos de câncer de esôfago e a quantificação de expressão para 14.415 genes, 13.864 mutações, coletados a partir do TCGA. O estudo

relacionou esses aos conjuntos de dados de expressão e mutação de 67 linhagens celulares, disponíveis na base de dados do CCLE. Para estas foram selecionadas informações de resposta de 174 fármacos utilizados no tratamento contra o câncer, disponíveis através do projeto GDSC.

Em ambos os experimentos as redes foram capazes de predizer valores de resposta consistentes com o conjunto de dados, confirmando a performance da predição. Os experimentos apresentaram resultados comparáveis aos trabalhos anteriores com diminuição no resultados de erro segundo as métricas utilizadas. As informações obtidas mostram-se promissoras para a continuidade e evolução do desenvolvimento dos modelos das redes neurais, testes com novos conjuntos de dados, além de abrir possibilidade para outras abordagens de aprendizado. O trabalho abordou um problema da área de farmacogenômica através da aplicação de técnicas de aprendizado de máquina sobre dados do câncer de esôfago.

6.1 Limitações

Apesar dos bons resultados obtidos, o trabalho possui desafios e limitações a serem exploradas. Entre elas, destaca-se:

Uma das primeiras dificuldades encontradas foi a reprodução dos experimentos do trabalho de Chiu et al. (2019). O estudo descreve a rede desenvolvida mas não há a disponibilização de código ou maiores detalhes sobre a versão dos bancos de dados utilizados.

Ao passo que a redução da análise para um tipo de câncer possa trazer benefícios, o número de amostras para uso do aprendizado consequentemente diminui. Um número maior das amostras poderia trazer uma melhor representação dos conjuntos para aprendizado. Isso também influenciou a utilização de dados clínicos, que possui um conjunto menor de amostras e com muitos dados ausentes.

Outra limitação é quanto a avaliação em comparação a outros trabalhos, pois apesar de encontrados trabalhos relacionados, o número de aplicações ainda é baixo. Assim, torna-se difícil a comparação de performance com outros estudos, explorando outros tipos de câncer, por exemplo.

Um dos objetivos iniciais seria a exploração dos dados clínicos coletados no Hospital São Lucas e a aplicação da rede sobre estes. Por motivos de tempo não foi obtido acesso aos dados até a finalização desse trabalho, fazendo com que essa tarefa não pudesse ser executada.

6.2 Perspectivas

Este trabalho demonstrou resultados positivos e melhorias e ainda pode explorar diferentes abordagens nas redes desenvolvidas e nos dados utilizados em trabalhos futuros.

Uma possibilidade de continuidade são as experimentações com a rede com apenas parte dos dados de entrada para predição. Dessa forma pode-se explorar se com os pesos aprendidos previamente o modelo é capaz de estimar o resultado de resposta apenas com os dados clínicos, por exemplo. Essa abordagem necessita mais estudos e ajustes da rede mas, além de ser uma aplicação importante na prática, poderia ajudar a esclarecer a importância de cada conjunto de dado para a predição. O mesmo pode ser investigado quanto a seleção de *features* e os tipos de informações utilizadas.

Visando superar uma das limitações encontradas, a busca por conjuntos de dados mais robustos pode aprimorar o resultado das predições. Isso pode ser feito buscando por outros projetos ou bases de dados biológicos e avaliando se estes são equivalentes aos utilizados neste estudo.

Além disso, estudos explorando os dados locais seguem como objetivos. Uma vez com acesso aos dados gerados pelo HSL, podem ser feitos experimentos adaptando a rede aos dados coletados e testando sua predição. Isso pode gerar estudos abordando o perfil de pacientes da população do Rio Grande do Sul.

Para fins de comparação, também podem ser aplicadas etapas de seleção de *features* para os dados biológicos. Assim, além da aplicação dos *autoencoders* pode-se reduzir a grande quantidade de atributos dos dados de expressão e mutação, desconsiderando aqueles que não agregam informações relevantes.

Ainda, a fim de avaliar os resultados obtidos, pode-se buscar novos trabalhos com preditores desenvolvidos na área. A exploração desses estudos pode ajudar nas comparações de resultados para outros tipos de câncer, outras aplicações das técnicas de aprendizado, como os *autoencoders*, ou trabalhos que utilizem dados clínicos. Seguindo o exemplo de outros trabalhos relacionados, uma possibilidade de evolução dos modelos seria explorar não só a predição dos IC_{50} mas também a interação entre os fármacos.

Ainda sobre a avaliação, são sugeridos mais experimentos utilizando os dados clínicos, e como incorporá-los no aprendizado. Pode-se utilizar uma parte destes ou combiná-los com *data sets* de outros bancos para obter resultados mais robustos sobre sua utilização para predição.

Do ponto de vista biológico, há a possibilidade de dar seguimento nos estudos explorando os resultados obtidos em função de entender o papel das alterações dos genes na resposta ao tratamento. Essa tarefa pode ser executada em conjunto com os colaboradores do projeto do LFA, investigando como as informações de expressão influenciam na

resistência ou sensibilidade dos pacientes. Além disso, é necessária uma etapa de avaliação quanto a aplicação sobre a população local, considerando o perfil de mutações em comparação aquele presente nos bancos de dados públicos.

REFERÊNCIAS BIBLIOGRÁFICAS

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C. e Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Recuperado de https://www.tensorflow.org/. Junho 2021.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT Press.

Arnal, M. J. D., Arenas, Á. F. e Arbeloa, Á. L. (Jul, 2015). Esophageal cancer: Risk factors, screening and endoscopic treatment in western and eastern countries. *World Journal of Gastroenterology*, vol. 21, pp. 7933.

Canzoneri, R., Lacunza, E. e Abba, M. C. (Jan, 2019). Genomics and bioinformatics as pillars of precision medicine in oncology. *Medicina (Buenos Aires)*, vol. 79, pp. 587–592.

Carvalho, A., Faceli, K., Lorena, A. e Gama, J. (Mar, 2011). Inteligência artificial–uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, vol. 2, pp. 45.

CCLE e GDSC (Nov, 2015). Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, vol. 528, pp. 84–87.

Chiu, Y.-C., Chen, H.-I. H., Gorthi, A., Mostavi, M., Zheng, S., Huang, Y. e Chen, Y. (Dez, 2020). Deep learning of pharmacogenomics resources: moving towards precision oncology. *Briefings in Bioinformatics*, vol. 21, pp. 2066–2083.

Chiu, Y.-C., Chen, H.-I. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J. e Chen, Y. (Jan, 2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, vol. 12, pp. 143–155.

Chollet, F. (2015). Keras. Recuperado de https://keras.io. Junho 2021.

Conesa, A. e Beck, S. (Out, 2019). Making multi-omics data accessible to researchers. *Scientific Data*, vol. 6, pp. 1–4.

De Barros, S., Ghisolfi, E., Luz, L., Barlem, G., Vidal, R., Wolff, F. e Grüber, A. (Jan, 2000). High temperature "mate"infusion drinking in a population at risk for squamous cell carcinoma of the esophagus. *Arquivos de Gastroenterologia*, vol. 37, pp. 25–30.

Donehower, L. A., Soussi, T., Korkut, A., Liu, Y., Schultz, A., Cardenas, M. e Lichtarge, O. (Ago, 2019). Integrated analysis of TP53 gene and pathway alterations in the cancer genome atlas. *Cell Reports*, vol. 28, pp. 1370–1384.

Eichelbaum, M., Ingelman-Sundberg, M. e Evans, W. E. (Fev, 2006). Pharmacogenomics and individualized drug therapy. *Annual Review of Medicine*, vol. 57, pp. 119–137.

Fagundes, R., Carli, D., Xaubet, R. e Cantarelli Jr, J. (Jun, 2016). Unchanging pattern of prevalence of esophageal cancer, overall and by histological subtype, in the endoscopy service of the main referral hospital in the central region of Rio Grande do Sul State, in Southern Brazil. *Diseases of the Esophagus*, vol. 29, pp. 603–606.

Fagundes, R. B., Abnet, C. C., Strickland, P. T., Kamangar, F., Roth, M. J., Taylor, P. R. e Dawsey, S. M. (Mai, 2006). Higher urine 1-hydroxy pyrene glucuronide (1-ohpg) is associated with tobacco smoke exposure and drinking mate in healthy subjects from rio grande do sul, brazil. *BMC cancer*, vol. 6, pp. 1–7.

Goldstein, D. B., Tate, S. K. e Sisodiya, S. M. (Jan, 2003). Pharmacogenetics goes genomic. *Nature Reviews Genetics*, vol. 4, pp. 937–947.

Goodfellow, I., Bengio, Y. e Courville, A. (2016). *Deep Learning*. MIT Press.

Gopisankar, M. G. (Abr, 2017). Cyp2d6 pharmacogenomics. *Egyptian Journal of Medical Human Genetics*, vol. 18, pp. 309–313.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A. e Staudt, L. M. (Set, 2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, vol. 375, pp. 1109–1112.

Guo, C., Xie, X., Li, J., Huang, L., Chen, S., Li, X. e Zhou, H. (Abr, 2019). Pharmacogenomics guidelines: Current status and future development. *Clinical and Experimental Pharmacology and Physiology*, vol. 46, pp. 689–693.

Hambali, M. A., Oladele, T. O. e Adewole, K. S. (Jun, 2020). Microarray cancer feature selection: review, challenges and research directions. *International Journal of Cognitive Computing in Engineering*, vol. 1, pp. 78–97.

He, K., Zhang, X., Ren, S. e Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034. IEEE.

Huang, F.-L. e Yu, S.-J. (Dez, 2018). Esophageal cancer: risk factors, genetic association, and treatment. *Asian Journal of Surgery*, vol. 41, pp. 210–215.

INCA (2021). Institutor nacional de câncer: Câncer de esôfago. Recuperado de https://www.inca.gov.br/tipos-de-cancer/cancer-de-esofago. Dezembro 2021.

Janiaud, P., Serghiou, S. e Ioannidis, J. P. (Jan, 2019). New clinical trial designs in the era of precision medicine: an overview of definitions, strengths, weaknesses, and current use in oncology. *Cancer Treatment Reviews*, vol. 73, pp. 20–30.

Jonge, P. J., van Blankenstein, M., Grady, W. M. e Kuipers, E. J. (Out, 2014). Barrett's oesophagus: epidemiology, cancer risk and implications for management. *Gut*, vol. 63, pp. 191–202.

Klein, M. E., Parvez, M. M. e Shin, J.-G. (Mar, 2017). Clinical implementation of pharmacogenomics for personalized precision medicine: barriers and solutions. *Journal of Pharmaceutical Sciences*, vol. 106, pp. 2368–2379.

Kuiava, V. A., Perin, A. T., Gurski, R. R., Madalosso, C. A. S., Hoppe, L. e Navarini, D. (Jan, 2018). Epidemiological profile of esophageal cancer mortality in rio grande do sul and its health regions. *Clinical & Biomedical Research*, vol. 38, pp. 213–217.

LeCun, Y., Bengio, Y. e Hinton, G. (Mai, 2015). Deep learning. *Nature*, vol. 521, pp. 436–444.

Li, M., Wang, Y., Zheng, R., Shi, X., Wu, F. e Wang, J. (Mai, 2019). Deepdsc: a deep learning method to predict drug sensitivity of cancer cell lines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, pp. 575–582.

Liang, Z., Huang, J. X., Zeng, X. e Zhang, G. (Ago, 2016). Dl-adr: a novel deep learning model for classifying genomic variants into adverse drug reactions. *BMC Medical Genomics*, vol. 9, pp. 195–204.

Libbrecht, M. W. e Noble, W. S. (Mai, 2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, vol. 16, pp. 321–332.

Lorena, A. C., Gama, J. e Faceli, K. (2000). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

Lubin, J. H., De Stefani, E., Abnet, C. C., Acosta, G., Boffetta, P., Victora, C. e Franceschi, S. (Out, 2014). Mate drinking and esophageal squamous cell carcinoma in south america: pooled results from two large multicenter case–control studies. *Cancer Epidemiology and Prevention Biomarkers*, vol. 23, pp. 107–116.

Luger, G. F. (2013). *Inteligência Artificial*. Pearson.

McCulloch, W. S. e Pitts, W. (Jan, 1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Mostavi, M., Chiu, Y.-C., Huang, Y. e Chen, Y. (Abr, 2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*, vol. 13, pp. 1–13.

Nagaraj, K., Sharvani, G. S. e Sridhar, A. (Jan, 2018). Emerging trend of big data analytics in bioinformatics: a literature review. *International Journal of Bioinformatics Research and Applications*, vol. 14, pp. 144–205.

NCI (2021). Understanding cancer: What is cancer? Recuperado de https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Dezembro 2021.

Nicholson, W. T., Formea, C. M., Matey, E. T., Wright, J. A., Giri, J. e Moyer, A. M. (2020). Considerations when applying pharmacogenomics to your practice. In: *Proceedings of the Mayo clinic*, pp. 218–230. Elsevier.

Norvig, P. e Russell, S. (2014). *Inteligência Artificial*. Elsevier Brasil.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. e Duchesnay, E. (Jan, 2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830.

Pevsner, J. (2015). *Bioinformatics and functional genomics*. John Wiley & Sons.

Puche, C. C., García, S., García, R., López, G. e Shahrour, F. (Mai, 2020). How can bioinformatics contribute to the routine application of personalized precision medicine? *Expert Review of Precision Medicine and Drug Development*, vol. 5, pp. 115–117.

Rafique, R., Islam, S. R. e Kazi, J. U. (Jul, 2021). Machine learning in the prediction of cancer therapy. *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4003–4017.

Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. e Sidow, A. (Out, 2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature Communications*, vol. 9, pp. 1–14.

Relling, M. V. e Evans, W. E. (Out, 2015). Pharmacogenomics in the clinic. *Nature*, vol. 526, pp. 343–350.

Sakellaropoulos, T., Vougas, K., Narang, S., Koinis, F., Kotsinas, A., Polyzos, A. e Kardala, E. (Dez, 2019). A deep learning framework for predicting response to therapy in cancer. *Cell Reports*, vol. 29, pp. 3367–3373.

Saunders, G., Baudis, M., Becker, R., Beltran, S., Béroud, C., Birney, E. e Drysdale, R. (Ago, 2019). Leveraging european infrastructures to access 1 million human genomes by 2022. *Nature Reviews Genetics*, vol. 20, pp. 693–701.

Short, M. W., Burgers, K. e Fry, V. (Jan, 2017). Esophageal cancer. *American Family Physician*, vol. 95, pp. 22–28.

Subramanian, I., Verma, S., Kumar, S., Jere, A. e Anamika, K. (Jan, 2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, vol. 14, pp. 1–24.

Tan, P.-N., Steinbach, M. e Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Vicente, A. E., Lumbreras, E., Hernández, J. M., Martín, M., Calles, A., Otín, C. L. e Taron, M. (Fev, 2016). Pharmacogenetics and pharmacogenomics as tools in cancer therapy. *Drug Metabolism and Personalized Therapy*, vol. 31, pp. 25–34.

W Caldwell, G., Yan, Z., Lang, W. e A Masucci, J. (Mai, 2012). The ic50 concept revisited. *Current Topics in Medicinal Chemistry*, vol. 12, pp. 1282–1290.

Wake, D. T., Ilbawi, N., Dunnenberger, H. M. e Hulick, P. J. (Dez, 2019). Pharmacogenomics: prescribing precisely. *Medical Clinics*, vol. 103, pp. 977–990.

Wang, Y., Yang, Y., Chen, S. e Wang, J. (Abr, 2021). Deepdrk: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Briefings in Bioinformatics*, vol. 22, pp. 1–11.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K. e Stuart, J. M. (Set, 2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, vol. 45, pp. 1113–1120.

WHO (2020). Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey. (Relatório técnico), World Health Organization.

WHO (2021a). World health organization: Cancer. Recuperado de https://www.who.int/news-room/fact-sheets/detail/cancer. Fevereiro 2022.

WHO (2021b). World health organization: Cancer overview. Recuperado de https://www.who.int/health-topics/cancer. Dezembro 2021.

Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S. e Thompson, I. R. (Dez, 2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, vol. 41, pp. 955–961.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br